BIOINFORMATICS ORIGINAL PAPER Vol. 27 no. 18 2011, pages 2494–2501
doi:10.1093/bioinformatics/btr426

Sequence analysis

Advance Access publication July 19, 2011

Identification of novel phosphorylation modification sites in human proteins that originated after the human–chimpanzee divergence

Dong Seon Kim and Yoonsoo Hahn*

School of Biological Sciences (BK21 Program) and Research Center for Biomolecules and Biosystems, Chung-Ang University, Seoul 156-756, Korea

Associate Editor: Alfonso Valencia

ABSTRACT

Motivation: Phosphorylation modifications of specific protein residues are involved in a wide range of biological processes such as modulation of intracellular signal networks. Here, we present the development and application of a bioinformatics procedure for systematic identification of human-specific phosphorylation sites in proteins that may have occurred after the human–chimpanzee divergence.

Results: We collected annotated human phosphorylation sites and compared each site to orthologous mammalian proteins across taxa including chimpanzee, orangutan, rhesus macaque, marmoset, mouse, dog, cow, elephant, opossum and platypus. We identified 37 human-specific gains of annotated phosphorylation sites in 35 proteins: 22 serines, 12 threonines and 3 tyrosines. The novel phosphorylation sites are situated in highly conserved segments of the protein. Proteins with novel phosphorylation sites are involved in crucial biological processes such as cell division (AURKB, CASC5, MKI67 and PDCD4) and chromatin remodeling (HIRA, HIRIP3, HIST1H1T, NAP1L4 and LRWD1). Modified phosphorylatable residues produce novel target sites for protein kinases such as cyclin-dependent kinases and casein kinases, possibly resulting in rewiring and fine-tuning of phosphorylation regulatory networks. The potential human-specific phosphorylation sites identified in this study are useful as candidates for functional analysis to identify novel phenotypes in humans.

Contact: hahny@cau.ac.kr

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on February 16, 2011; revised on June 30, 2011; accepted on July 13, 2011

1 INTRODUCTION

Humans have many distinct phenotypic traits compared with the other great apes, such as loss of body hair, upright posture, accelerated brain expansion and more complex cognitive abilities (Varki and Altheide, 2005). These traits must have originated as genetic modifications and undergone subsequent natural selection in the human lineage after humans diverged from their closest living relatives, the chimpanzees. The accelerated sequence substitution of proteins may be associated with the evolution of certain human-specific traits: for example, genes implicated in nervous system development were found to display significantly higher rates of protein evolution in the human lineage (Dorus *et al.*, 2004). The FOXP2 protein, which is known to be associated with speech and language in humans, acquired two amino acid substitutions specific to humans after the human–chimpanzee divergence (Enard *et al.*, 2002). The human FOXP2 protein has been reported to differentially regulate genes involved in central nervous system development compared with the chimpanzee ortholog (Konopka *et al.*, 2009). Interestingly, loss of proteins has also been implicated in the development of human-specific traits (Olson, 1999). For example, inactivation of the *MYH16* gene in humans may be related to the reduction of the masticatory muscle and expansion of the brain (Stedman *et al.*, 2004). Other proteins that became inactive only in humans include BASE, SERPINA13 and MOXD2 (Hahn and Lee, 2005, 2006; Hahn *et al.*, 2007).

Phosphorylation is a fundamental post-translational modification (PTM) of proteins that acts in many important biological processes, such as cellular signaling pathways through the modulation of protein function, stability, interaction and localization (Pawson and Scott, 2005). Several tens of thousands of phosphorylation modification sites have been characterized in human proteins using high-throughput proteomic analyses (Beausoleil *et al.*, 2004; Olsen *et al.*, 2006; Oppermann *et al.*, 2009; Wang *et al.*, 2008). Data regarding phosphorylation modification sites in human proteins are available via databases, such as UniProtKB/Swiss-Prot (UniProt Consortium, 2010), PHOSIDA (Gnad *et al.*, 2007) and dbPTM (Lee *et al.*, 2006).

It is likely that biologically important phosphorylation sites are evolutionarily conserved (Malik *et al.*, 2008). Tan *et al.* (2009a) identified 479 phosphorylation sites in 344 human proteins that appear to be positionally conserved between human and at least one species of fly, worm or yeast and proposed that these sites are involved in fundamental cellular processes. Nevertheless, the gain and loss of protein phosphorylation sites may rewire biological networks and be regarded as a motive force for adaptive evolution (Shou *et al.*, 2011; Vener, 1990). The gain or loss of phosphorylation sites in Cdk1 substrates in yeasts may have facilitated the evolution of kinase-signaling circuits (Holt *et al.*, 2009). Tan *et al.* (2009b) reported that tyrosine residues have been selectively lost through metazoan evolution to remove deleterious tyrosine phosphorylations as tyrosine kinases have been expanded. It is argued that evolution

*To whom correspondence should be addressed.

of phosphorylation regulatory network could play a prime role in generating phenotypic changes during organismal evolution (Moses and Landry, 2010). It has been proposed that the human-specific serine residue in the human FOXP2 protein may provide a novel phosphorylation site for protein kinase C (Enard *et al.*, 2002), although there is no supporting experimental evidence for this hypothesis.

In this study, we hypothesize that novel phosphorylation sites in the human proteins that originated after the human–chimpanzee divergence may have been involved in the evolution of human-specific traits through the strengthening and/or rewiring of existing protein interaction networks. To address this possibility, we developed a bioinformatics procedure to systematically identify novel human-specific phosphorylation sites. Using this procedure, we identified 37 novel phosphorylation sites that have arisen in human proteins. We also identified 31 phosphorylation sites that are conserved in humans and other mammals, but have been lost in the chimpanzee proteins.

2 METHODS

2.1 Datasets and bioinformatics tools

To identify annotated phosphorylation modification sites in human proteins, we downloaded the human UniProtKB/Swiss-Prot records from the UniProt database (http://www.uniprot.org/). The HomoloGene database of the National Center for Biotechnology Information (NCBI) (http://www.ncbi.nlm.nih.gov/homologene) was consulted for the initial collection of RefSeq accession numbers of orthologous mammalian protein sequences for each human protein, which were then retrieved from the NCBI RefSeq database (http://www.ncbi.nlm.nih.gov/RefSeq/). The data analyzed in this study were downloaded on January 4, 2011. Since the HomoloGene database includes only six mammalian species (human, chimpanzee, mouse, rat, dog and cow), we also accessed the mammalian genome assemblies available in the University of California Santa Cruz (UCSC) Genome Browser database (http://genome.ucsc.edu) to verify and predict more orthologous proteins in orangutan, rhesus macaque, marmoset and other non-primate mammals. The genome assemblies analyzed in this study include hg19 (human), panTro2 (chimpanzee), ponAbe2 (orangutan), rheMac2 (rhesus macaque), calJac3 (marmoset), mm9 (mouse), rn4 (rat), canFam2 (dog), felCat4 (cat), bosTau4 (cow), susScr2 (pig), equCab2 (horse), loxAfr3 (elephant), monDom5 (opossum) and ornAna1 (platypus).

The MUSCLE program (http://www.drive5.com/muscle/) was used to construct multiple sequence alignments (Edgar, 2004), which were then decorated using the BOXSHADE server (http://www.ch.embnet.org/software/BOX_form.html). Pairwise alignments between cDNAs and genomic sequences were performed to identify exons using the SIM4 program (http://globin.cse.psu.edu/dist/sim4/) (Florea *et al.*, 1998). We wrote some *ad hoc* PERL scripts to manipulate sequences, database records and software outputs.

2.2 Selection of candidates for gain or loss of phosphorylation sites during human and chimpanzee evolution

We downloaded human UniProtKB/Swiss-Prot records for phosphorylation modification site information. For each UniProt record, we examined the feature table (FT) lines showing post-translationally modified residues (key name 'MOD_RES'). When the lines contained one of the keywords 'Phosphoserine', 'Phosphothreonine' or 'Phosphotyrosine', the sequence record was collected. A total of 32 750 phosphorylation modification sites were retrieved for 6950 proteins.

To identify potential human-specific phosphorylation sites, we performed multiple sequence alignments for each human protein with orthologous mammalian proteins including chimpanzee proteins. We used the RefSeq accession numbers in UniProt records as keys to identify the corresponding HomoloGene datasets. The non-human RefSeq proteins for each HomoloGene dataset were extracted from the RefSeq protein database and combined with the UniProt human protein sequence to construct an orthologous protein sequence set. A total of 6425 orthologous protein sets were prepared. Multiple sequence alignments were produced using the MUSCLE program. For each aligned human phosphorylation site, when the orthologous chimpanzee sequence was not one of three phosphorylatable residues (serine, threonine or tyrosine), the position was identified as a potential human-specific phosphorylation site. A total of 90 annotated human phosphorylation sites were found to differ from the orthologous chimpanzee sequences.

2.3 Manual inspection of datasets

As a final step, we manually scrutinized the 90 candidate human phosphorylation sites to identify human-specific gains or chimpanzee-specific losses of phosphorylation sites. The mammalian ortholog sequences were retrieved from the HomoloGene database, the NCBI protein sequence database or predicted from genome assemblies available through the UCSC Genome Browser database. The sequencing qualities of the chimpanzee codons corresponding to the human phosphorylated amino acid residues were examined to ensure that any sequence differences between human and chimpanzee proteins were not due to sequencing errors.

When chimpanzee and orangutan sequences lacked a phosphorylatable residue at an aligned human phosphorylation site, we concluded that the human protein acquired the phosphorylation site after the human–chimpanzee divergence. When non-chimpanzee primates or other mammals exhibited the same amino acid sequence as the human phosphorylation site but the chimpanzee protein had a different sequence, we concluded that the chimpanzee protein had lost the phosphorylation site after the human–chimpanzee divergence. Of the 90 candidate sites, 37 were identified as unique to human proteins and 31 as lost in chimpanzee proteins. The remaining 22 cases were discarded due to ambiguous orthology relationship, incorrect prediction of the chimpanzee protein or low sequence quality of the orthologous chimpanzee codons.

2.4 Polymorphisms among modern humans and Neanderthal alleles

To identify possible polymorphisms in the codons for human-specific phosphorylation sites, we examined human polymorphism information available via the UCSC Genome Browser database. The polymorphism data were drawn from the simple nucleotide polymorphism database (dbSNP) build 132 and personal genomes including two anonymous Yoruban Africans (UCSC track name NA19240 and NA18507), three anonymous Europeans (NA12878, NA12891 and NA12892), J. Craig Venter (Venter), James Watson (Watson), one anonymous Han Chinese (YH) and one anonymous Korean (SJK). Detailed information on the SNP data was obtained from dbSNP (http://www.ncbi.nlm.nih.gov/SNP/).

We also analyzed the Neanderthal sequence data (Green *et al.*, 2010) available in the UCSC Genome Browser (the 'Neandertal Alleles in Human/Chimp Coding Non-synonymous Differences in Human Lineage' track) to examine Neanderthal alleles.

3 RESULTS

3.1 Novel phosphorylation sites in human proteins

We examined 32 750 annotated phosphorylation modification sites in 6950 human proteins to identify 37 human-specific phosphorylation

sites in 35 proteins that presumably originated after the human–chimpanzee divergence (Table 1). Multiple sequence alignments of phosphorylation sites and the surrounding regions of the representative proteins are shown in Figure 1. Alignments for all of the cases are shown in Supplementary Figure S1.

Two proteins, MKI67 and PDXDC1, each acquired two novel phosphorylation sites, while the remaining 33 proteins acquired one site each. The amino acid residues of the novel phosphorylation sites in human proteins include 22 serines, 12 threonines and 3 tyrosines. Among 37 annotated phosphorylation sites, 36 were experimentally identified either by large-scale analyses using mass spectrometry (35 sites) or by conventional molecular biology method (HIRA serine 687) (Hall *et al.*, 2001). Eight sites were detected in more than two independent experiments. None of these 37 sites has yet been assigned a biological function.

Interestingly, the regions surrounding human-specific phosphorylation sites are strongly conserved among mammalian orthologs, including those of the non-placental mammals opossum (marsupial) and platypus (monotreme). Representative proteins are ABCF1, ADNP, BAIAP2L1, HIRA, NAP1L4, NOP2, PDCD4, RUSC2, SNRPA, TARBP1, TP53BP1 and USP7 (Fig. 1 and Supplementary Fig. S1). The human ABCF1 protein, for example, has a phosphorylated threonine residue at position 108, where an alanine residue is found in all other mammals including opossum and platypus. Proteins that are highly conserved across mammalian taxa or at least among primates but that have undergone accelerated amino acid substitution during human evolution are proposed to be involved in the development of adaptive human phenotypes (Clark *et al.*, 2003). The proteins identified in this study exhibit not only human-specific amino acid changes but also novel phosphorylation sites, possibly resulting in dramatic effects on the functions of the proteins themselves, as well as those of their associated protein networks.

In four cases (HIRIP3, LRWD1, PAG1 and PDXDC1 serine 737), human phosphorylation sites produce the same amino acid residues in non-primate mammals but not in other primates. The human LRWD1 protein, for example, has a phosphorylated serine at position 259, where other primates have glycines (Supplementary Fig. S1, No. 14). In contrast, some non-primate mammals including mouse, cat, horse, pig and cow exhibit a serine residue at this position. Mouse and primates belong to Euarchontoglires and cat, horse, pig and cow to Laurasiatheria. Therefore, in these cases, it is probable that phosphorylation sites were acquired in ancestral proteins before divergence of Euarchontoglires and Laurasiatheria but were lost in primate ancestors, again becoming phosphorylatable during human evolution.

3.2 Notable cases of human-specific gains of phosphorylation sites

3.2.1 ABCF1 The human ABCF1 (also known as ABC50) protein, which is a member of the superfamily of ATP-binding cassette (ABC) transporters, has a human-specific phosphorylated threonine residue at position 108 (Fig. 1A). The ABCF1 protein was reported to interact with eukaryotic initiation factor eIF2 and promote translation initiation in mammalian cells by binding to ribosomes (Paytubi *et al.*, 2008, 2009). Serine 109 and serine 140 of the ABCF1 protein were found to be phosphorylated by casein kinase 2 (CK2), and mutations of these sites resulted in marked decreases in the association of eIF2 with ribosomes (Paytubi *et al.*, 2008). Currently, the human ABCF1 protein has nine annotated phosphorylation sites. It is possible that modification of the human-specific phosphorylation site threonine 108 can positively or negatively affect the function of the adjacent CK2 target serine 109.

3.2.2 AURKB and CASC5 The human AURKB and CASC5 proteins acquired novel phosphorylation sites serine 7 and serine 1076, respectively (Supplementary Fig. S1, Nos 3 and 7). The human AURKB protein or aurora kinase B is a serine/threonine kinase that functions in the attachment of the mitotic spindle to the centromere during cell division for accurate chromosome segregation (Lampson *et al.*, 2004). AURKB interacts with many proteins including INCENP, BIRC5 (also known as Survivin) and NINL (also known as Nlp) and modulates their functions via phosphorylation (Honda *et al.*, 2003; Yan *et al.*, 2010).

CASC5 (also known as KNL1) has been reported to oppose AURKB activity by promoting dephosphorylation of AURKB substrates (Liu *et al.*, 2010). Interestingly, CASC5 itself is a substrate for AURKB and is down-regulated by phosphorylation (Welburn *et al.*, 2010). Both the AURKB and CASC5 proteins exhibit many phosphorylation targets, with 13 and 21 annotated sites, respectively. The human-specific phosphorylation of the AURKB serine 7 and CASC5 serine 1076 might result in fine adjustments in protein interaction networks and in the human cell division process.

3.2.3 HIRA, HIRIP3 and NAP1L4 The HIRA protein is an evolutionarily conserved histone chaperone that preferentially places the variant histone H3.3 in nucleosomes and that plays an important role in the formation of senescence-associated heterochromatin foci (Zhang *et al.*, 2005). The HIRIP3 protein is also a histone-binding protein (Lorain *et al.*, 1998). The human HIRA and HIRIP3 proteins are heavily phosphorylated, with 14 and 35 annotated sites, respectively. Interestingly, the human HIRA and HIRIP3 proteins acquired novel phosphorylation targets serine 687 and serine 300, respectively (Fig. 1B and Supplementary Fig. S1, No. 10, respectively). HIRA and HIRIP3 proteins have been reported to directly interact with each other (Lorain *et al.*, 1998). The human-specific HIRA serine 687 is located in the H2B-binding region (Lorain *et al.*, 1998) and reported to be phosphorylated by the cyclin-dependent kinase 2 (CDK2) (Hall *et al.*, 2001).

NAP1L4 is also a histone chaperone that belongs to the nucleosome assembly protein (NAP) family and can interact with both core and linker histones (Okuwaki *et al.*, 2010; Rodriguez *et al.*, 1997). The human NAP1L4 (Fig. 1C) has 13 phosphorylation sites with one human-specific site, serine 139. Modifications of human HIRA, HIRIP3 and NAP1L4 proteins might be implicated in chromatin-remodeling evolution.

3.2.4 LWRD1 The human LRWD1 [also known as origin recognition complex-associated (ORCA)] protein mediates the origin recognition complex (ORC) in chromatin which is critical for the initiation of pre-replication complex assembly in G1 and chromatin organization in post-G1 cells (Shen *et al.*, 2010). The human LRWD1 has seven annotated phosphorylation sites, of which the serine 259 site is human-specific. Other primates have a glycine residue at this position (Supplementary Fig. S1, No 14).

Table 1. Gains of phosphorylation modification sites in the human proteins that were identified in this study

No	Gene	UniProt	Position	Amino acid[a]	Human RefSeq	Refs.[b]	Title
1	*ABCF1*	ABCF1_HUMAN	108	T ? A A A A	NM_001025091	Beausoleil *et al*., 2004; Kim *et al*., 2005; Olsen *et al.*, 2006; Imami *et al.*, 2008; Dephoure *et al.*, 2008; Gauci *et al.*, 2009; Mayya *et al.*, 2009	ATP-binding cassette, sub-family F
2	*ADNP*	ADNP_HUMAN	921	S S L L L L	NM_015339	Gauci *et al.*, 2009	Activity-dependent neuroprotector homeobox
3	*AURKB*	AURKB_HUMAN	7	S ? A A A A	NM_004217	Nousiainen *et al*., 2006; Wang *et al.*, 2008; Daub *et al.*, 2008	Aurora kinase B
4	*BAIAP2L1*	BI2L1_HUMAN	416	T T A A A A	NM_018842	Gauci *et al.*, 2009	BAI1-associated protein 2-like 1
5	*C1orf131*	CA131_HUMAN	163	T ? I I I I	NM_152379	Matsuoka *et al*., 2007	Chromosome 1 open reading frame 131
6	*C14orf50*	CN050_HUMAN	392	Y Y H H H H	NM_172365	Molina *et al*., 2007	Chromosome 14 open reading frame 50
7	*CASC5*	CASC5_HUMAN	1076	S S N N N N	NM_170589	Nousiainen *et al*., 2006; Wang *et al.*, 2008; Daub *et al.*, 2008; Dephoure *et al.*, 2008; Mayya *et al.*, 2009	Cancer susceptibility candidate 5
8	*FASN*	FAS_HUMAN	2198	S S N N N S	NM_004104	Dephoure *et al.*, 2008	Fatty acid synthase
9	*HIRA*	HIRA_HUMAN	687	S S G G G G	NM_003325	Hall *et al*., 2001	HIR histone cell–cycle regulation defective homolog A (*S.cerevisiae*)
10	*HIRIP3*	HIRP3_HUMAN	300	S S G G G G	NM_003609	Molina *et al*., 2007	HIRA interacting protein 3
11	*HIST1H1T*	H1T_HUMAN	160	T S A A A V	NM_005323	Olsen *et al.*, 2006	Histone cluster 1, H1t
12	*KRT8*	K2C8_HUMAN	438	S ? G G G G	NM_002273	Wang *et al.*, 2008	Keratin 8
13	*LDHA*	LDHA_HUMAN	10	Y ? H H H H	NM_005566	Mayya *et al.*, 2009	Lactate dehydrogenase A
14	*LRWD1*	LRWD1_HUMAN	259	S S G G G G	NM_152892	Cantin *et al*., 2008; Dephoure *et al.*, 2008; Gauci *et al.*, 2009; Mayya *et al.*, 2009	Leucine-rich repeats and WD repeat domain containing 1
15	*MKI67*	KI67_HUMAN	1256	S S P P P P	NM_002417	Dephoure *et al.*, 2008	Antigen identified by monoclonal antibody Ki-67
16	*MKI67*	KI67_HUMAN	3042	S ? P P P P	NM_002417	Wang *et al.*, 2008	Antigen identified by monoclonal antibody Ki-67
17	*NAP1L4*	NP1L4_HUMAN	139	S S N N N N	NM_005969	Molina *et al*., 2007	NAP 1-like 4
18	*NOP2*	NOP2_HUMAN	140	T T M M M M	NM_001033714	Olsen *et al.*, 2006	NOP2 nucleolar protein homolog (yeast)
19	*PAG1*	PAG1_HUMAN	380	S ? G G G G	NM_018440		Phosphoprotein associated with glycosphingolipid microdomains 1
20	*PDCD4*	PDCD4_HUMAN	80	S S N N N N	NM_014456	Dephoure *et al.*, 2008	Programmed cell death 4 (neoplastic transformation inhibitor)
21	*PDXDC1*	PDXD1_HUMAN	737	S ? G G G G	NM_015027	Olsen *et al.*, 2006	Pyridoxal-dependent decarboxylase domain containing 1
22	*PDXDC1*	PDXD1_HUMAN	779	S S P P P P	NM_015027	Olsen *et al.*, 2006; Matsuoka *et al*., 2007	Pyridoxal-dependent decarboxylase domain containing 1
23	*RCSD1*	CPZIP_HUMAN	284	S ? P A A A	NM_052862	Mayya *et al.*, 2009	RCSD domain containing 1
24	*RIF1*	RIF1_HUMAN	1148	S S A A A T	NM_018151	Gauci *et al.*, 2009	RAP1 interacting factor homolog (yeast)
25	*RREB1*	RREB1_HUMAN	1315	T ? K E K K	NM_001003698	Dephoure *et al.*, 2008	ras responsive element binding protein 1
26	*RRP8*	RRP8_HUMAN	223	S ? P P P P	NM_015324	Beausoleil *et al*., 2006; Dephoure *et al.*, 2008	Ribosomal RNA processing 8, methyltransferase, homolog (yeast)
27	*RUSC2*	RUSC2_HUMAN	1368	S ? P P P P	NM_001135999	Dephoure *et al.*, 2008	RUN and SH3 domain containing 2
28	*SNRPA*	SNRPA_HUMAN	131	T ? A A A A	NM_004596	Dephoure *et al.*, 2008	snRNP polypeptide A
29	*SVIL*	SVIL_HUMAN	270	S S P P S P	NM_021738	Zahedi *et al*., 2008; Cantin *et al*., 2008	Supervillin
30	*TARBP1*	TARB1_HUMAN	1442	S S P P P P	NM_005646	Daub *et al.*, 2008	TAR (HIV-1) RNA binding protein 1
31	*TCOF1*	TCOF_HUMAN	581	T T A A A A	NM_001135243	Dephoure *et al.*, 2008	Treacher Collins-Franceschetti syndrome 1
32	*TERF1*	TERF1_HUMAN	434	S S C C C C	NM_017489	Dephoure *et al.*, 2008	Telomeric repeat binding factor (NIMA-interacting) 1
33	*TJP2*	ZO2_HUMAN	1131	T ? P L L L	NM_004817	Molina *et al*., 2007, Dephoure *et al.*, 2008	Tight junction protein 2 (zona occludens 2)
34	TP53BP1	TP53B_HUMAN	394	T ? M M M M	NM_005657	Gauci *et al.*, 2009	Tumor protein p53 binding protein 1
35	USP7	UBP7_HUMAN	54	T ? N N N N	NM_003470	Mayya *et al.*, 2009	Ubiquitin-specific peptidase 7 (herpes virus-associated)
36	USP36	UBP36_HUMAN	874	Y Y H H H H	NM_025090	Olsen *et al.*, 2006	Ubiquitin-specific peptidase 36
37	ZNF638	ZN638_HUMAN	809	T ? A A A A	NM_014497	Oppermann *et al*., 2009	Zinc finger protein 638

[a]Amino acid sequences of primate species: human, Neanderthal, chimpanzee, orangutan, rhesus macaque and marmoset. ?, unknown.
[b]References to experimental evidence of phosphorylation.

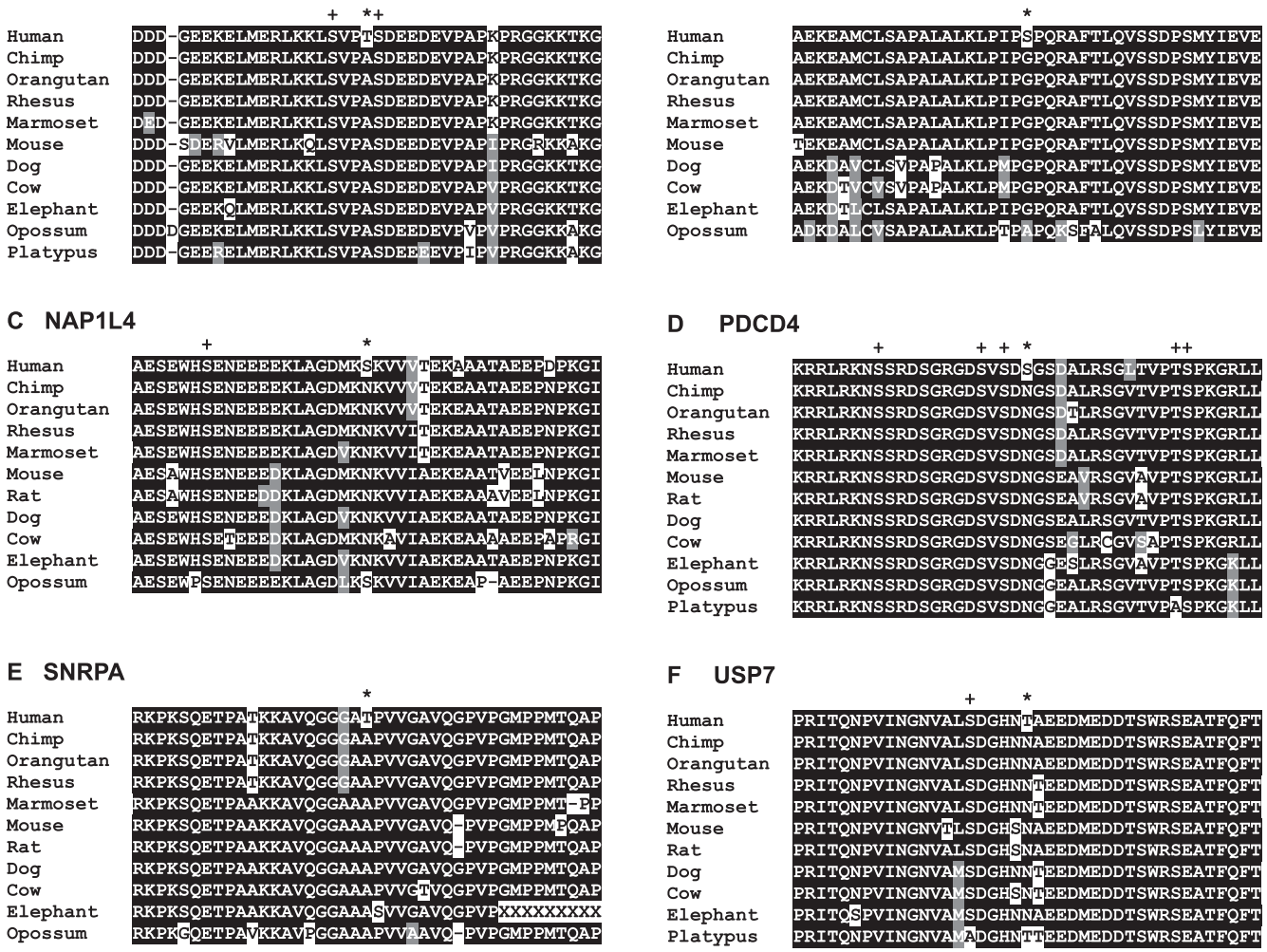

Fig. 1. Multiple sequence alignments of representative cases of human-specific gains of phosphorylation sites. The human-specific phosphorylation sites (*), shared sites (+) and the surrounding regions (20 aligned positions on both sides) for each case are presented. Residues that were identical in ≥50% of the aligned proteins are highlighted with black backgrounds. Opossum NAP1L4 and platypus USP7 exhibit the same phosphorylatable amino acid residue as does the respective human protein, probably reflecting independent gains.

3.2.5 MKI67 and PDCD4 The human MKI67 (also known as Ki-67) is a nuclear protein that is associated with and may be necessary for cellular proliferation (Gerdes *et al.*, 1991). Overexpression of MKI67 is associated with many cancers, including neuroendocrine carcinomas and colorectal cancers (Erler *et al.*, 2011; Ma *et al.*, 2010). The human MKI67 protein, which has 124 phosphorylation sites, acquired two novel phosphorylation sites, serine 1256 and serine 3042 (Supplementary Fig. S1, Nos 15 and 16, respectively). In contrast, the PDCD4 protein is a tumor suppressor (Göke *et al.*, 2004; Yang *et al.*, 2006). The human PDCD4 protein acquired a novel phosphorylation site, serine 80, resulting in a total of nine phosphorylation sites. All other mammals that have been studied, including opossum and platypus, have an asparagine residue at this site (Fig. 1D). The mammalian PDCD4 proteins show strong sequence conservation, indicating that they play important roles in mammalian biology.

3.2.6 SNRPA and TARBP1 The SNRPA protein (also known as U1A) is associated with stem/loop 2 of the U1 small nuclear RNA (snRNA) and forms the spliceosomal U1 small nuclear ribonucleoprotein (snRNP), which is required for mRNA splicing (Law *et al.*, 2006). The SNRPA proteins are highly conserved in mammals, and the human SNRPA has a single novel phosphorylation site, threonine 131 (Fig. 1E). TARBP1 is also an RNA binding protein that was initially identified as a binding protein of the human immunodeficiency virus-1 (HIV-1) trans-activation-responsive (TAR) RNA (Sheline *et al.*, 1991). TARBP1 binds directly to DICER1 and loads small interfering RNA into the RNA-induced silencing complex (RISC) (MacRae *et al.*, 2008). The human TARBP1 protein also acquired a single novel phosphorylation site, serine 1442. These modifications may alter RNA biology in humans.

3.2.7 USP7 and USP36 The USP7 and USP36 proteins are cysteine proteases that function as deubiquitinating enzymes (de Bie *et al.*, 2010; Endo *et al.*, 2009). Protein ubiquitination is one of the fundamental regulatory PTMs that control intracellular protein signal networks (Komander, 2009). The human USP7 and USP36 proteins acquired human-specific phosphorylation sites, threonine 54 and tyrosine 874, respectively (Fig. 1F and Supplementary Fig. S1, No 36, respectively). It is notable that the threonine 54 of the USP7 protein is embedded in a strongly conserved region. The newly acquired phosphorylatable residues of these deubiquitinating enzymes may provide novel regulatory phosphorylation targets in human cells.

3.3 Polymorphisms among modern humans and Neanderthal alleles

Human polymorphism data drawn from the UCSC Genome Browser database were examined to identify possible amino acid-altering polymorphisms in codons encoding phosphorylated residues. We found non-synonymous SNPs at the phosphorylation sites in C1orf131 and NOP2 proteins.

The first base of the codon for the C1orf131 protein position 163 shows an A/G polymorphism (dbSNP accession number rs115635619). The allelic codons are ACA and GCA, encoding threonine and alanine, respectively. Interestingly, the corresponding chimpanzee codon is ATA, encoding isoleucine which is common in other primates, indicating that the both ACA (threonine) and GCA (alanine) are derived alleles. The frequency data available in dbSNP show that the ACA (threonine) allele frequency is 98.3% in Africans, indicating that the phosphorylatable allele is prevalent.

The second base of the codon for the NOP2 protein position 140 exhibits a C/T polymorphism (rs35556146). The codon variants are ACG and ATG, encoding threonine (derived) and methionine (ancestral), respectively. The derived phosphorylatable allele is prevalent in modern humans: 92.5% in Africans and 100% in both Europeans and Asians.

In a modern human-Neanderthal genome comparison, we have found that 20 derived sites are shared with Neanderthal DNA, indicating that the acquisition of these phosphorylatable residues occurred before the divergence of modern humans and Neanderthals. No information is available for the rest of the cases.

3.4 Losses of phosphorylation sites in chimpanzee proteins

When annotated human phosphorylation sites are shared with other primates and mammals, but different non-phosphorylatable residues are found in chimpanzees compared with those in the other animals, we conclude that the phosphorylation site was lost in chimpanzees after the human–chimpanzee divergence. By applying this condition, we identified 31 chimpanzee proteins with losses of annotated phosphorylation sites (Supplementary Table S1). Multiple sequence alignments surrounding the mutated sites are shown in Supplementary Figure S2. In most cases, the annotated phosphorylation sites are conserved across mammals except in the chimpanzee, implying that phosphorylation at the corresponding position was acquired early in mammalian evolution. For example, the human ASPM protein residue number 1103 is serine, which is conserved across mammals including opossum and platypus (Supplementary Fig. S2, No. 2). However, the chimpanzee protein has a cysteine residue at the corresponding position. The original phosphorylated amino acid sequences lost in chimpanzee proteins include 20 serines, 10 threonines and 1 tyrosine.

4 DISCUSSION

We identified 37 cases of novel phosphorylation site gains in 35 human proteins that likely arose after the human–chimpanzee divergence. Human proteins with novel phosphorylation modification sites are involved in various biological processes: for example, AURKB, CASC5, MKI67, PDCD4 in the cell cycle; HIRA, HIRIP3, HIST1H1T, NAP1L4 and LRWD1 are involved in chromatin remodeling; SNRPA and TARBP1 are involved in RNA biology; and USP7 and USP36 are involved in protein deubiquitination.

The number of novel phosphorylation sites in human proteins and the number of lost sites in chimpanzee proteins are very similar, 37 versus 31, indicating that gain and loss events occurred at comparable rates in both taxa. The distributions of changed residues are also similar. The amino acid residues of the 37 novel human phosphorylation sites include 22 serines, 12 threonines and 3 tyrosines. In chimpanzee proteins, 20 serines, 10 threonines and 1 tyrosine have been mutated to non-phosphorylatable residues. The conversion to or from serine is more frequent than other conversions. This difference may be explained in terms of the numbers of codons in the genetic code. There are six codons for serine, four for threonine and two for tyrosine. If a mutation is completely random, it is likely that the serine codon will appear more often than will threonine or tyrosine.

The distribution of the lost phosphorylatable residues may simply be due to the frequencies of the phosphorylated residues in the proteome. The human protein dataset analyzed in this study contains 23 977 phosphoserines, 5610 phosphothreonines and 3163 phosphotyrosines. Because phosphoserine is more common than phosphothreonine or phosphotyrosine, it is expected that phosphoserine is lost more often at similar mutation rates. However, we cannot rule out the possibility that there could be evolutionary constraints regarding the gain or loss of phosphorylation sites. It is possible that the relatively small number of gain or loss of phosphotyrosine sites reflects the fact that phosphotyrosine sites are involved in more biologically important processes than are phosphothreonine or phosphoserine sites. It has been proposed that, as tyrosine kinases have been expanded during metazoan evolution, tyrosine residues have been selectively removed to minimize possibly deleterious tyrosine phosphorylations (Tan *et al.*, 2009b).

The casein kinase 1 (CK1) family of protein kinases functions as signal transduction pathway regulators in eukaryotic cells and recognizes the motif S-X-X-S/T* (Gnad *et al.*, 2007). The three novel phosphorylation sites of ABCF1, MKI67 (serine 1256) and TARBP1 match the CK1 motif. CK2 has been implicated in several cellular processes including cell-cycle control and DNA repair, and it recognizes the motif S/T*-X-X-E, where E is aspartic acid (Gnad *et al.*, 2007). There are six human-specific phosphorylation sites (ABCF1, ADNP, PDXDC1 serine 737 and serine 779, RUSC2 and USP7) that match the CK2 motif. We assume that the advent of new target sites for the various kinases in these proteins may result in evolution of protein regulatory networks. There are four CK1 sites

(C17orf56, EIF4G1, NEK1 and PLEKHG3) and three CK2 sites (BSN, CENPF and PRR11) that are lost in chimpanzee proteins.

The CDKs recognize the consensus sequence S/T*-P-X-K/R, where S/T* is the phosphorylated serine or threonine, P is proline, X is any amino acid, K is lysine and R is arginine (Moses *et al.*, 2007). They also weakly recognize the sequence S/T*-P. Among 37 novel human-specific phosphorylation sites, two cases (BAIAP2L1 and HIRA) match the strong consensus and 12 cases (BAIAP2L1, CASC5, HIRA, HIST1H1T, LRWD1, RCSD1, RRP8, RUSC2, SNRPA, SVIL and TJP2) match the weak consensus. Similarly, out of 31 sites lost in chimpanzee, two (GBF1 and MELK) and 10 cases (C17orf56, CCNO, COPB2, EIF4G1, EPN1, MEK1, RAD52, TP53BP1, WDR62 and WRAP53) match the strong and weak consensus sequences, respectively.

The similar rate of phosphorylation site gain in humans and loss in chimpanzees is consistent with the previous report that CDK consensus sites in clusters turn over rapidly (Moses *et al.*, 2007). When we examined surrounding regions of the phosphorylation sites, either gained in humans or lost in chimpanzees, we found at least one additional phosphorylation site within a 20-residue distance in 26 human-specific gains and in 21 chimpanzee-specific losses. And hence, in about two-thirds of the cases, gain or loss of phosphorylation site occurred in clusters. Nevertheless, gain of novel phosphorylation sites may drive regulatory evolution as demonstrated that acquisition of CDK consensus sites in a replication protein Mcm3 is associated with the evolution of CDK-mediated shuttling of mini-chromosome maintenance complex in *Saccharomyces cerevisiae* (Moses *et al.*, 2007).

The UniProt database annotation records include phosphorylation sites of non-human proteins. In a parallel study, we surveyed sites to identify human-specific loss of ancestral phosphorylation sites. We retrieved data for 21 374 phosphorylation sites (14 328 serines, 3714 threonines and 3332 tyrosines) in 5618 non-human proteins. However, we were unable to identify any instances of human-specific loss, probably because most of the annotated phosphorylation sites in non-human proteins available in the UniProt database were not determined in independent experiments, but were deduced based on sequence similarities to human protein phosphorylation data. Use of high-throughput analysis data for non-human mammals, such as recently reported phosphorylation data for mouse tissues (Huttlin *et al.*, 2010), may enable us to collect human-specific loss of phosphorylation sites in the future.

We also encountered limitations of the HomoloGene and RefSeq databases. We found that many of the predicted chimpanzee protein sequences in the RefSeq database are not complete or are not accurate, possibly due to the draft quality of the current chimpanzee assembly (Taudien *et al*., 2006). If high-quality chimpanzee protein sequences are eventually obtained, we may be able to identify more instances of gain and loss of phosphorylation sites and other PTM sites.

The procedure presented in this study may be applied to other PTM sites such as *N*-glycosylation, *O*-glycosylation, acetylation, methylation, SUMOylation, ubiquitination, glycosylphosphatidylinositol anchoring and proteolytic cleavage sites. We are currently researching these modifications. Recent advancements in large-scale analyses of the human proteomes have greatly increased the amount of PTM data that is available (Gnad *et al.*, 2007; Lee *et al.*, 2006). Through the analysis of these large datasets, our knowledge of the evolution of PTM sites and associated protein networks will be greatly expanded.

In conclusion, we identified 37 novel phosphorylation sites in human proteins that arose after the human–chimpanzee divergence. We propose that functional study of these human-specific phosphorylation candidates may explain the molecular mechanism of some human-specific phenotypes.

Funding: The National Research Foundation of Korea (2009-0071595); the Next-Generation BioGreen 21 Program (SSAC2011-PJ008220), Rural Development Administration, Republic of Korea.

Conflict of Interest: none declared.

REFERENCES

Beausoleil,S.A. *et al.* (2004) Large-scale characterization of HeLa cell nuclear phosphoproteins. *Proc. Natl Acad. Sci. USA*, **101**, 12130–12135.

Beausoleil,S.A. *et al*. (2006) A probability-based approach for high-throughput protein phosphorylation analysis and site localization. *Nat. Biotechnol*., **24**, 1285–1292.

Cantin,G.T. *et al*. (2008) Combining protein-based IMAC, peptide-based IMAC, and MudPIT for efficient phosphoproteomic analysis. *J. Proteome Res*., **7**, 1346–1351.

Clark,A.G. *et al.* (2003) Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios, *Science*, **302**, 1960–1963.

Daub,H. *et al*. (2008) Kinase-selective enrichment enables quantitative phosphoproteomics of the kinome across the cell cycle. *Mol. Cell*, **31**, 438–448.

de Bie,P. *et al.* (2010) Regulation of the Polycomb protein RING1B ubiquitination by USP7. *Biochem. Biophys. Res. Commun.*, **400**, 389–395.

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

Dorus,S. *et al.* (2004) Accelerated evolution of nervous system genes in the origin of Homo sapiens, *Cell*, **119**, 1027–1040.

Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.

Enard,W. *et al.* (2002) Molecular evolution of FOXP2, a gene involved in speech and language. *Nature*, **418**, 869–872.

Endo,A. *et al.* (2009) Nucleophosmin/B23 regulates ubiquitin dynamics in nucleoli by recruiting deubiquitylating enzyme USP36. *J. Biol. Chem.*, **284**, 27918–27923.

Erler,B.S. *et al.* (2011) CD117, Ki-67, and p53 predict survival in neuroendocrine carcinomas, but not within the subgroup of small cell lung carcinoma. *Tumour Biol.*, **32**, 107–111.

Florea,L. *et al.* (1998) A computer program for aligning a cDNA sequence with a genomic DNA sequence. *Genome Res.*, **8**, 967–974.

Gauci,S. *et al*. (2009) Lys-N and trypsin cover complementary parts of the phosphoproteome in a refined SCX-based approach. *Anal. Chem*., **81**, 4493–4501.

Gerdes,J. *et al.* (1991) Immunobiochemical and molecular biologic characterization of the cell proliferation-associated nuclear antigen that is defined by monoclonal antibody Ki-67. *Am. J. Pathol.*, **138**, 867–873.

Göke,R. *et al.* (2004) Programmed cell death protein 4 (PDCD4) acts as a tumor suppressor in neuroendocrine tumor cells. *Ann. NY Acad. Sci.*, **1014**, 220–221.

Gnad,F. *et al.* (2007) PHOSIDA (phosphorylation site database): management, structural and evolutionary investigation, and prediction of phosphosites. *Genome Biol.*, **8**, R250.

Green,R.E. *et al.* (2010) A draft sequence of the Neandertal genome. *Science*, **328**, 710–722.

Hahn,Y. and Lee, B. (2005) Identification of nine human-specific frameshift mutations by comparative analysis of the human and the chimpanzee genome sequences. *Bioinformatics*, **21**, i186–i194.

Hahn,Y. and Lee, B. (2006) Human-specific nonsense mutations identified by genome sequence comparisons. *Hum. Genet.*, **119**, 169–178.

Hahn,Y. *et al.* (2007) Inactivation of MOXD2 and S100A15A by exon deletion during human evolution. *Mol. Biol. Evol.*, **24**, 2203–2212.

Hall,C. *et al.* (2001) HIRA, the human homologue of yeast Hir1p and Hir2p, is a novel cyclin-cdk2 substrate whose expression blocks S-phase progression. *Mol. Cell. Biol.*, **21**, 1854–1865.

Holt,L.J. *et al.* (2009) Global analysis of Cdk1 substrate phosphorylation sites provides insights into evolution. *Science*, **325**, 1682–1686.

Honda,R. *et al.* (2003) Exploring the functional interactions between Aurora B, INCENP, and survivin in mitosis. *Mol. Biol. Cell*, **14**, 3325–3341.

Huttlin,E.L. *et al.* (2010) A tissue-specific atlas of mouse protein phosphorylation and expression. *Cell*, **143**, 1174–1189.

Imami,K. *et al.* (2008) Automated phosphoproteome analysis for cultured cancer cells by two-dimensional nanoLC-MS using a calcined titania/C18 biphasic column. *Anal. Sci*., **24**, 161–166.

Kim,J.E. *et al*. (2005) Global phosphoproteome of HT-29 human colon adenocarcinoma cells. *J. Proteome Res*., **4**, 1339–1346

Komander,D. (2009) The emerging complexity of protein ubiquitination. *Biochem. Soc. Trans.*, **37**, 937–953.

Konopka,G. *et al.* (2009) Human-specific transcriptional regulation of CNS development genes by FOXP2. *Nature*, **462**, 213–217.

Lampson,M.A. *et al.* (2004) Correcting improper chromosome-spindle attachments during cell division. *Nat. Cell Biol.*, **6**, 232–237.

Law,M.J. *et al.* (2006) The role of RNA structure in the interaction of U1A protein with U1 hairpin II RNA. *RNA*, **12**, 1168–1178.

Lee,T.Y. *et al.* (2006) dbPTM: an information repository of protein post-translational modification. *Nucleic Acids Res.*, **34**, D622–D627.

Liu,D. *et al.* (2010) Regulated targeting of protein phosphatase 1 to the outer kinetochore by KNL1 opposes Aurora B kinase. *J. Cell Biol.*, **188**, 809–820.

Lorain,S. *et al.* (1998) Core histones and HIRIP3, a novel histone-binding protein, directly interact with WD repeat protein HIRA. *Mol. Cell. Biol.*, **18**, 5546–5556.

Ma,Y.L. *et al.* (2010) Immunohistochemical analysis revealed CD34 and Ki67 protein expression as significant prognostic factors in colorectal cancer. *Med. Oncol.*, **27**, 304–309.

MacRae,I.J. *et al.* (2008) In vitro reconstitution of the human RISC-loading complex. *Proc. Natl Acad. Sci. USA*, **105**, 512–517.

Malik,R. *et al.* (2008) Comparative conservation analysis of the human mitotic phosphoproteome. *Bioinformatics*, **24**, 1426–1432.

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

Mayya,V. *et al.* (2009) Quantitative phosphoproteomic analysis of T cell receptor signaling reveals system-wide modulation of protein-protein interactions. *Sci. Signal*., **2**, ra46.

Molina,H. *et al*. (2007) Global proteomic profiling of phosphopeptides using electron transfer dissociation tandem mass spectrometry. *Proc. Natl Acad. Sci. USA*, **104**, 2199–2204.

Moses,A.M. and Landry,C.R. (2010) Moving from transcriptional to phospho-evolution: generalizing regulatory evolution? *Trends Genet*., **26**, 462–467.

Moses,A.M. *et al.* (2007) Regulatory evolution in proteins by turnover and lineage-specific changes of cyclin-dependent kinase consensus sites, *Proc. Natl Acad. Sci. USA*, **104**, 17713–17718.

Nousiainen,M. *et al.* (2006) Phosphoproteome analysis of the human mitotic spindle. *Proc. Natl Acad. Sci. USA*, **103**, 5391–5396.

Okuwaki,M. *et al.* (2010) Functional characterization of human nucleosome assembly protein 1-like proteins as histone chaperones. *Genes Cells*, **15**, 13–27.

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

Olson,M.V. (1999) When less is more: gene loss as an engine of evolutionary change. *Am. J. Hum. Genet.*, **64**, 18–23.

Oppermann,F.S. *et al.* (2009) Large-scale proteomics analysis of the human kinome. *Mol. Cell. Proteomics*, **8**, 1751–1764.

Pawson,T. and Scott,J.D. (2005) Protein phosphorylation in signaling–50 years and counting. *Trends Biochem. Sci.*, **30**, 286–290.

Paytubi,S. *et al.* (2008) The N-terminal region of ABC50 interacts with eukaryotic initiation factor eIF2 and is a target for regulatory phosphorylation by CK2. *Biochem. J.*, **409**, 223–231.

Paytubi,S. *et al.* (2009) ABC50 promotes translation initiation in mammalian cells. *J. Biol. Chem.*, **284**, 24061–24073.

Rodriguez,P. *et al.* (1997) Functional characterization of human nucleosome assembly protein-2 (NAP1L4) suggests a role as a histone chaperone. *Genomics*, **44**, 253–265.

Sheline,C.T. *et al.* (1991) Two distinct nuclear transcription factors recognize loop and bulge residues of the HIV-1 TAR RNA hairpin. *Genes Dev.*, **5**, 2508–2520.

Shen,Z. *et al.* (2010) A WD-repeat protein stabilizes ORC binding to chromatin. *Mol. Cell*, **40**, 99–111.

Shou,C. *et al.* (2011) Measuring the evolutionary rewiring of biological networks. *PLoS Comput. Biol.*, **7**, e1001050.

Stedman,H.H. *et al.* (2004) Myosin gene mutation correlates with anatomical changes in the human lineage. *Nature*, **428**, 415–418.

Tan,C.S. *et al.* (2009a) Comparative analysis reveals conserved protein phosphorylation networks implicated in multiple diseases. *Sci. Signal.*, **2**, ra39.

Tan,C.S. *et al.* (2009b) Positive selection of tyrosine loss in metazoan evolution. *Science*, **325**, 1686–1688.

Taudien,S. *et al.* (2006) Should the draft chimpanzee sequence be finished? *Trends Genet.*, **22**, 122–125.

UniProt Consortium (2010) The Universal Protein Resource (UniProt) in 2010. *Nucleic Acids Res.*, **38**, D142–D148.

Varki,A. and Altheide,T.K. (2005) Comparing the human and chimpanzee genomes: searching for needles in a haystack. *Genome Res.*, **15**, 1746–1758.

Vener,A.V. (1990) Protein phosphorylation: a motive force for adaptive evolution. *Biosystems*, **24**, 53–59.

Wang,B. *et al.* (2008) Evaluation of the low-specificity protease elastase for large-scale phosphoproteome analysis. *Anal. Chem.*, **80**, 9526–9533.

Welburn,J.P. *et al.* (2010) Aurora B phosphorylates spatially distinct targets to differentially regulate the kinetochore-microtubule interface. *Mol. Cell*, **38**, 383–392.

Yan,J. *et al.* (2010) Aurora B interaction of centrosomal Nlp regulates cytokinesis. *J. Biol. Chem.*, **285**, 40230–40239.

Yang,H.S. *et al.* (2006) Tumorigenesis suppressor Pdcd4 down-regulates mitogen-activated protein kinase kinase kinase kinase 1 expression to suppress colon carcinoma cell invasion. *Mol. Cell. Biol.*, **26**, 1297–1306.

Zahedi,R.P. *et al*. (2008) Phosphoproteome of resting human platelets. *J. Proteome Res*., **7**, 526–534.

Zhang,R. *et al.* (2005) Formation of MacroH2A-containing senescence-associated heterochromatin foci and senescence driven by ASF1a and HIRA. *Dev. Cell*, **8**, 19–30.