MBE Advance Access published May 7, 2014

Signatures of Natural Selection on Mutations of Residues with Multiple Posttranslational Modifications

Vanessa E. Gray,[†,1] Li Liu,[†,1] Ronika Nirankari,[1] Peter V. Hornbeck,[2] and Sudhir Kumar[*,1,3,4]

[1]Center for Evolutionary Medicine and Informatics, Biodesign Institute, Arizona State University

[2]Cell Signaling Technology, Danvers, MA

[3]School of Life Sciences, Arizona State University

[4]Center for Genomic Medicine and Research, King Abdulaziz University, Jeddah, Saudi Arabia

[†]These authors contributed equally to this work.

***Corresponding author:** E-mail: s.kumar@asu.edu.

Associate editor: Helen Piontkivska

Abstract

Posttranslational modifications (PTMs) regulate molecular structures and functions of proteins by covalently binding to amino acids. Hundreds of thousands of PTMs have been reported for the human proteome, with multiple PTMs known to affect tens of thousands of lysine (K) residues. Our molecular evolutionary analyses show that K residues with multiple PTMs exhibit greater conservation than those with a single PTM, but the difference is rather small. In contrast, short-term evolutionary trends revealed in an analysis of human population variation exhibited a much larger difference. Lysine residues with three PTMs show 1.8-fold enrichment of Mendelian disease-associated variants when compared with K residues with two PTMs, with the latter showing 1.7-fold enrichment of these variants when compared with the K residues with one PTM. Rare polymorphisms in humans show a similar trend, which suggests much greater negative selection against mutations of K residues with multiple PTMs within population. Conversely, common polymorphisms are overabundant at unmodified K residues and at K residues with fewer PTMs. The observed difference between inter- and intraspecies patterns of purifying selection on residues with PTMs suggests extensive species-specific drifting of PTM positions. These results suggest that the functionality of a protein is likely conserved, without necessarily conserving the PTM positions over evolutionary time.

Key words: **posttranslational modification, proteomics, evolution.**

Posttranslational modifications (PTMs) modulate protein function for nearly every protein in eukaryotic proteomes (Mann and Jensen 2003; Seo and Lee 2004; Walsh et al. 2005). Although experimental studies have consistently attested the functional importance of PTMs (Gonzalez and Montminy 1989; Apweiler et al. 1999; Cohen 2000; Lemeer and Heck 2009; Nakajima et al. 2010), initial molecular evolutionary analyses found weak to no purifying selection acting on PTM residues (Gnad et al. 2007; Landry et al. 2009; Chen et al. 2010). By estimating the degree of purifying selection for each protein individually, we recently found that PTM residues are far more conserved than other residues (Gray and Kumar 2011). These results led us to predict that residues harboring multiple PTMs in humans will be under stronger natural selection and may be more highly conserved among species.

Many different types of amino acids harbor multiple PTMs, including lysine (K), which can be modified by ubiquitination, acetylation, methylation, and SUMOylation (Perlmann 1955; Freiman and Tjian 2003; Medzihradszky et al. 2004; Chen et al. 2007; Yang and Seto 2008). Each of these four PTM types of K is enzymatically regulated, thus only sites that attract overlapping modifying enzymes harbor multi-PTMs, such as protein Histone 3 at position 9 (H3K9) that is recognized by both acetylating and methylating enzymes (Kouzarides 2007). Each of the four PTM types covalently binds the same site of the K side-chain, so only one PTM modifies K at a given time. During ubiquitination and SUMOylation, enzymes attach small peptides called ubiquitin and Small Ubiquitin-related MOdifier (SUMO) to K within a target protein (Komander 2009; Creton and Jentsch 2010). These PTMs are associated with a myriad of molecular functions, such as degradation and localization (Komander 2009; Denuc and Marfany 2010). Acetylation and methylation are found anatomically throughout a cell, yet a multitude of studies focus on the nucleus to study the effects of histone modification on chromatin remodeling and gene expression (Strahl and Allis 2000; Li et al. 2007).

Using over 40,000 experimentally determined lysine PTMs sites, we tested the null hypothesis that K residues with single and multiple PTMs will show equal evolutionary conservation. This comparison was made feasible by the fact that over 3,500 K residues are known to be modified multiple times (fig. 1*A*) (Hornbeck et al. 2012). Evaluation of this hypothesis and the difference in the evolutionary conservations of singly and multiply modified sites will illuminate how the degree of protein modification affects the strength of purifying selection acting on a site.

Letter

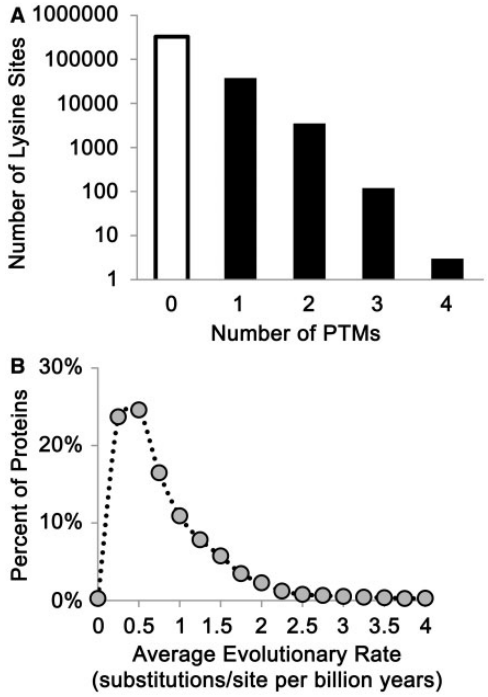

FIG. 1. Distribution of lysine (K) residues and protein evolutionary rates. (*A*) Logarithmic distribution of K residues with different numbers of PTMs. The vast majority of Ks are not modified. Among modified K residues, 96.6% (37,720) have a single PTM. (*B*) Distribution of evolutionary rates of proteins. The median (mean) evolutionary rate of proteins is 0.52 (0.73), which shows that a large number of proteins evolve slowly, but many evolve with faster rates as well.

Results

In total, we downloaded 41,681 PTMs from PhosphoSitePlus (http://www.phosphosite.org, last accessed January 13, 2014) (Hornbeck et al. 2012). These PTMs affect sites from 8,825 proteins, wherein we observe 29,056 ubiquitination sites, 14,764 acetylation sites, 1,292 methylation sites, and 549 SUMOylation sites. Of all PTM sites, 3,626 are modified by more than one PTM type, which we refer to as multi-PTM sites. The vast majority of multi-PTM sites is doubly modified (97%), with only three protein positions showing modification by four PTM types. For each K position in 8,825 proteins, absolute evolutionary rates were estimated by dividing the number of substitutions at that position in a 46-species protein alignment by the total elapsed time in the known phylogeny of 46-species (Kumar et al. 2009). Rate estimates were obtained for 325,982 positions with an unmodified K and 41,681 positions containing Ks with PTMs. The average rate of protein evolution calculated, based only on the absolute rates at K positions, shows that proteins evolve at vastly different rates (fig. 1*B*). Therefore, we compared modified (single- and multi-PTMs) and unmodified sites within each protein separately, following the procedure used in Gray and Kumar (2011).

We defined a measure of purifying selection, α, which is calculated by dividing the average evolutionary rate of modified K sites by that of unmodified K sites, for individual proteins (see Materials and Methods section). An α-value equal to 1.0 indicates that modified and unmodified sites evolve equivalently, whereas $\alpha < 1.0$ suggests that modified sites experience greater purifying selection than their unmodified counterparts in a single protein. For proteins with at least one K PTM, 4,992 (57%) of them experienced greater purifying selection on PTM sites compared with non-PTM sites. Considered separately, each K PTM type, that is, ubiquitination, acetylation, methylation, and SUMOylation, shows similar attributes of α. For these PTM types, we find 57%, 59%, 60%, and 62% of proteins with $\alpha < 1.0$, respectively (fig. 2*A*–*D*). In each case, >50% of proteins showed greater purifying selection acting on modified sites when compared with unmodified sites (proportion test: $P < 10^{-70}$). Therefore, the null hypothesis of no effect is rejected.

The frequency distributions of α were not normally distributed, thus we used median as a summary statistic. The median α for all proteins was 0.88, indicating that evolutionary selective pressures eliminate 12% more mutations from modified sites than unmodified K sites. The use of medians supported these conclusions, because they fell in the range of 0.73–0.88 (see fig. 2 legend for values). Thus, up to 27% more mutations are eliminated from protein positions affected by PTMs than unmodified sites. Furthermore, each PTM type has unique evolutionary properties, which is reflected in the differences among the respective statistical distributions (fig. 2*A*–*D*). Additionally, these four PTM types show patterns similar to those reported for N-linked glycosylation and phosphorylations of serine, threonine, and tyrosine residues (Gray and Kumar 2011).

Of 8,825 proteins, 1,493 (17%) contain sites with two or more modification types. The α-value was <1.0 for 63% of these proteins, which indicates higher conservation (6% difference) when compared with singly modified sites (56%, fig. 2*E*). This difference was statistically significant in a Kolmogorov–Smirnov test for the equality of distributions of α between single- and multi-PTMs ($P < 10^{-16}$). In particular, over 28% of the proteins with multi-PTMs show $\alpha = 0$, which more than doubles the frequency for proteins containing only singly modified sites (12%, χ^2 test: $P < 10^{-30}$). Overall, there are small differences between the evolutionary conservation of single- and multi-PTM K positions, which could be taken to suggest that the K residues involved in multi-PTMs are not much more important than those with single PTMs. An alternative is that K PTMs are species-specific, which would not leave an imprint for detection in multispecies analyses. To evaluate these alternatives, we examined the relative frequencies of known disease variants of K residues with one, two, and three PTMs in humans. An exploration of the Human Gene Mutation Database (HGMD) (Cooper et al. 2006) produced a large number of K residues ($n = 868$) to be associated with disease phenotypes; 747 mapped to unmodified K, 104 mapped to K with one PTM, 16 with two PTMs, and only one with three PTMs.

For K sites with one PTM, the rate of HGMD variant occurrence is 0.28 per 100 residues, which is 20% larger than that for unmodified K sites (0.23 per 100 residues, χ^2 test: $P = 0.09$). However, the rate of HGMD variant occurrence increases significantly to 0.46 per 100 residues for Ks with two PTMs

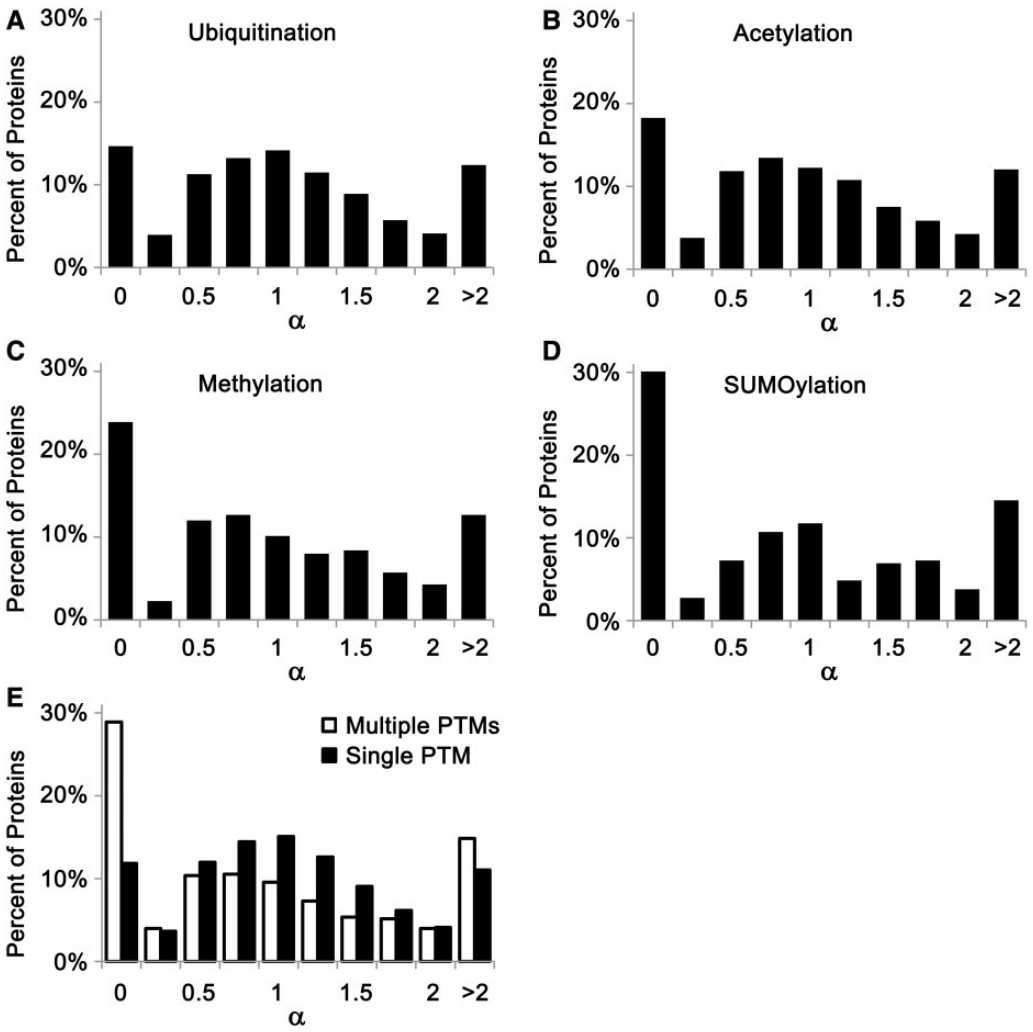

FIG. 2. Distribution of protein-specific estimates of selection on PTM positions (α). The distributions for ubiquitination (*A*) and acetylation (*B*) are similar to each other with the median α-value of 0.86 and 0.80, respectively. The distributions of methylation (*C*) and SUMOylation (*D*) are similar and both have the median α-value of 0.73. There is a clear distinction between α-value distributions for single and multiple PTM protein positions (*E*).

(χ^2 test: $P < 0.01$) and to 0.84 per 100 residues for Ks with three PTMs (fig. 3*A*). That is, positions with greater numbers of PTMs are more likely to be associated with disease phenotypes and the magnitudes of differences are much larger than those observed in multispecies evolutionary analyses (see fig. 2).

These patterns establish that the preservation of K residues harboring multi-PTMs is extremely important within a species and that this importance only leaves a modest footprint on long-term evolutionary trends of conservation. To test this interpretation of results from the analysis of HGMD variation, we examined population polymorphisms at modified and unmodified K sites using the 1000 Genomes Project data (Consortium 2010). If multi-PTM K residues experience stronger purifying selection, then we expect the population variants at those residues to occur at lower allele frequencies (Subramanian and Kumar 2006). Indeed, we observed significantly different patterns of allele frequency distributions (fig. 3*B*). Consistent with predictions, the fraction of rare variants (allele frequency <1%) at polymorphic sites increases steadily from unmodified K residues to one, two, and three PTMs K residues (88.8%, 91.1%, 93.8%, and 100%, respectively). Conversely, the fraction of common variants (allele frequency >5%) dropped from 5.6% at positions harboring unmodified K residues to 0% at positions harboring K residues with three PTMs. Therefore, the analysis of population data confirmed our findings of strong within-species constraints on multi-PTM positions.

In summary, we find that K residues with multi-PTMs exhibit higher evolutionary conservation than those with a single PTM. However, we observed significant differences between the strengths of purifying selection inferred using the long-term multispecies evolutionary analysis and using the short-term evolutionary analysis of human disease and polymorphism variation. The latter shows that the negative selection on multi-PTM K sites within humans is much greater for such positions. One reason for long-term evolutionary history to not capture the importance of K modifications is that the physical site of modification may drift within a protein. For example, Beltrao et al. (2012) describe a site in Rdi1p, where a

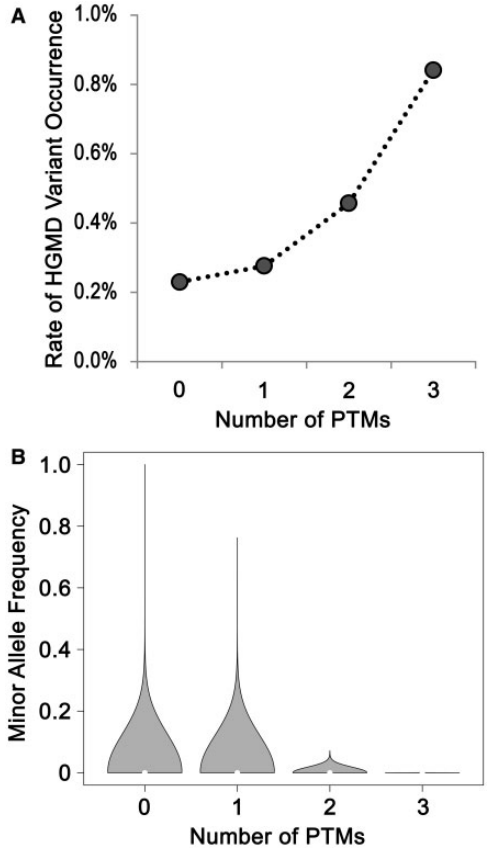

FIG. 3. Analysis of disease-associated variants and population polymorphisms at K sites with and without PTMs. (*A*) The frequency of K residues per 100 positions with zero, one, two, and three PTMs that contain a disease-associated variant. (*B*) Violin plots showing the allele frequency distributions at polymorphic sites that harbor K residues with zero, one, two, and three PTMs.

serine at position 40 is phosphorylated in *Saccharomyces cerevisiae* and *Candida albicans*, but the human ortholog is missing this phosphorylation. Instead, another modification on a tyrosine residue at position 20 is found, which has been shown to interact with the same interaction partner as the phosphoserine sites of *S. cerevisiae* and *C. albicans*. Although it is not known if such compensatory evolutionary patterns of PTM are common, our results from multispecies analyses predict that to be the case, because long-term evolutionary rate estimates do not show strong and persistent selective pressures on PTM positions across species. Our results suggest that knowledge of human disease variation and polymorphisms will ameliorate this problem, as collective knowledge rapidly grows as a consequence of exome sequencing from disease populations.

Materials and Methods

The PTM data set was downloaded from PhosphoSitePlus (http://www.phosphosite.org, last accessed January 13, 2014) (Hornbeck et al. 2012), which is currently the largest repository of experimentally verified human K PTMs. Many K sites have been reported to harbor more than one PTM. Only proteins with at least one of the following PTM types were included: Ubiquitination, acetylation, methylation, and SUMOylation. The UCSC Genome Browser (Dreszer et al. 2012) was used to map each PTM site onto a 46-species alignment using an NCBI RefSeq identifier (Pruitt et al. 2012). The PTM data sets and UCSC data sets were cross-referenced based on chromosomal coordinates with duplicates removed. All PTM sites that could not be mapped to a RefSeq accession were excluded from this study.

To obtain the protein-specific measure, α, we divide the average absolute evolutionary rate (r) of the modified amino acid sites by r of modified K sites for each protein, individually. For example, for sites harboring two modified K sites, we calculate α_2 by dividing the average evolutionary rates of these two sites by that of unmodified K sites (r_u) within a single protein ($\alpha_2 = r_2/r_u$). An α-value <1.0 indicates that modified sites are more conserved than unmodified sites, and consequently a larger fraction of mutations at modified sites are eliminated from a population via purifying selection than unmodified sites. Absolute evolutionary rates are estimated by dividing the number of amino acid substitutions at a single protein site among 46 species by the elapsed evolutionary time since species' divergence, as done in Gray and Kumar (2011).

The full set of available disease-associated mutations was downloaded from the HGMD (http://www.hgmd.org/, last accessed May 12, 2013) (Cooper et al. 2006). To identify K sites with known disease-associated mutations, we cross-referenced the HGMD data set with the PTM data set based on chromosomal coordinates. In total, 868 sites were shared between the two data sets; 121 of these shared sites affect PTM sites and the remainder affects unmodified K sites. To measure the propensity for modified sites to harbor disease-associated mutations, we calculated the proportion of modified sites with disease-associated mutations in all modified K sites. This ratio represents the rate of occurrence of disease variants and is calculated by dividing the number of modified disease-associated sites by the total number of modified sites. These calculations were completed for PTM sites with one, two, and three modifications. To determine the rate of occurrence of disease variants for unmodified sites, we divided the number of unmodified disease-associated sites by the total number of unmodified sites.

To determine the tendency for PTM sites to coincide with polymorphic sites in the human population, we implemented an approach similar to the aforementioned methods for human disease-associated mutations. A set of human polymorphic sites was obtained from the 1000 Genomes Project (http://www.1000genomes.org/, phase 1 release v3, last accessed March 10, 2013) (Consortium 2010). After identifying protein sites common to the data set from 1000 Genomes Project and our modified and unmodified lysine data sets, we measured the abundance of modified sites at polymorphic sites. In total, 592 and 4,749 polymorphisms were found at modified and unmodified K sites, respectively. Variants with population frequency $<1\%$ were defined as rare variants; and variants with population frequency $>5\%$ were defined as common variants. The rate of rare or common variant occurrence was calculated by taking the ratio of the

number of variants to the total number of polymorphic sites. For example, to obtain the rate of occurrence of rare variants for singly modified sites, we divided the number of single PTM sites with rare variants by the number of polymorphic sites with single PTM.

Acknowledgments

The authors thank Ms. Carol Williams for edits and Maxwell Sanderford for assistance in data preparation and analysis execution. This research was supported by the National Institutes of Health grants HG002096-12 and LM0110834-03 to S.K., a training grant GM071798 to Brenda Hogue, and the National Science Foundation graduate research fellowship DGE-1256082 to V.E.G.

References

Apweiler R, Hermjakob H, Sharon N. 1999. On the frequency of protein glycosylation, as deduced from analysis of the swiss-prot database. *Biochim Biophys Acta.* 1473:4–8.

Beltrao P, Albanèse V, Kenner LR, Swaney DL, Burlingame A, Villén J, Lim WA, Fraser JS, Frydman J, Krogan NJ. 2012. Systematic functional prioritization of protein posttranslational modifications. *Cell* 150: 413–425.

Chen S, Chen F, Li W. 2010. Phosphorylated and non-phosphorylated serine and threonine residues evolve at different rates in mammals. *Mol Biol Evol.* 11:2548–2554.

Chen Y, Sprung R, Tang Y, Ball H, Sangras B, Kim SC, Falck JR, Peng J, Gu W, Zhao Y. 2007. Lysine propionylation and butyrylation are novel post-translational modifications in histones. *Mol Cell Proteomics.* 6: 812–819.

Cohen P. 2000. The regulation of protein function by multisite phosphorylation—a 25 year update. *Trends Biochem Sci.* 25:596–601.

Consortium GP. 2010. A map of human genome variation from population-scale sequencing. *Nature* 467:1061–1073.

Cooper DN, Stenson PD, Chuzhanova NA. 2006. The human gene mutation database (HGMD) and its exploitation in the study of mutational mechanisms. Chapter 1. *Curr Protoc Bioinformatics.* Unit 1.13.

Creton S, Jentsch S. 2010. Snapshot: the sumo system. *Cell* 143: 848–848.e1.

Denuc A, Marfany G. 2010. Sumo and ubiquitin paths converge. *Biochem Soc Trans.* 38:34–39.

Dreszer TR, Karolchik D, Zweig AS, Hinrichs AS, Raney BJ, Kuhn RM, Meyer LR, Wong M, Sloan CA, Rosenbloom KR, et al. 2012. The ucsc genome browser database: extensions and updates 2011. *Nucleic Acids Res.* 40:D918–D923.

Freiman RN, Tjian R. 2003. Regulating the regulators: lysine modifications make their mark. *Cell* 112:11–17.

Gnad F, Ren S, Cox J, Olsen J, Macek B, Oroshi M, Mann M. 2007. Phosida (phosphorylation site database): management, structural and evolutionary investigation, and prediction of phosphosites. *Genome Biol.* 8:R250.

Gonzalez GA, Montminy MR. 1989. Cyclic amp stimulates somatostatin gene transcription by phosphorylation of creb at serine 133. *Cell* 59: 675–680.

Gray VE, Kumar S. 2011. Rampant purifying selection conserves positions with posttranslational modifications in human proteins. *Mol Biol Evol.* 28:1565–1568.

Hornbeck PV, Kornhauser JM, Tkachev S, Zhang B, Skrzypek E, Murray B, Latham V, Sullivan M. 2012. Phosphositeplus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse. *Nucleic Acids Res.* 40:D261–D270.

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

Kouzarides T. 2007. Chromatin modifications and their function. *Cell* 128:693–705.

Kumar S, Suleski MP, Markov GJ, Lawrence S, Marco A, Filipski AJ. 2009. Positional conservation and amino acids shape the correct diagnosis and population frequencies of benign and damaging personal amino acid mutations. *Genome Res.* 19: 1562–1569.

Landry C, Levy E, Michnick S. 2009. Weak functional constraints on phosphoproteomes. *Trends Genet.* 25:193–197.

Lemeer S, Heck A. 2009. The phosphoproteomics data explosion. *Curr Opin Chem Biol.* 13:414–420.

Li B, Carey M, Workman JL. 2007. The role of chromatin during transcription. *Cell* 128:707–719.

Mann M, Jensen O. 2003. Proteomic analysis of post-translational modifications. *Nat Biotechnol.* 21:255–261.

Medzihradszky KF, Darula Z, Perlson E, Fainzilber M, Chalkley RJ, Ball H, Greenbaum D, Bogyo M, Tyson DR, Bradshaw RA, et al. 2004. O-sulfonation of serine and threonine: mass spectrometric detection and characterization of a new posttranslational modification in diverse proteins throughout the eukaryotes. *Mol Cell Proteomics.* 3:429–440.

Nakajima M, Koga T, Sakai H, Yamanaka H, Fujiwara R, Yokoi T. 2010. N-glycosylation plays a role in protein folding of human ugt1a9. *Biochem Pharmacol.* 79:1165–1172.

Perlmann GE. 1955. The nature of phosphorus linkages in phosphoproteins. *Adv Protein Chem.* 10:1–30.

Pruitt KD, Tatusova T, Brown GR, Maglott DR. 2012. Ncbi reference sequences (refseq): current status, new features and genome annotation policy. *Nucleic Acids Res.* 40:D130–D135.

Seo J, Lee K. 2004. Post-translational modifications and their biological functions: proteomic analysis and systematic approaches. *J Biochem Mol Biol.* 37:35–44.

Strahl BD, Allis CD. 2000. The language of covalent histone modifications. *Nature* 403:41–45.

Subramanian S, Kumar S. 2006. Evolutionary anatomies of positions and types of disease-associated and neutral amino acid mutations in the human genome. *BMC Genomics* 7:306.

Walsh CT, Garneau-Tsodikova S, Gatto GJ. 2005. Protein posttranslational modifications: the chemistry of proteome diversifications. *Angew Chem Int Ed Engl.* 44:7342–7372.

Yang XJ, Seto E. 2008. Lysine acetylation: codified crosstalk with other posttranslational modifications. *Mol Cell.* 31:449–461.