

Testing whether Metazoan Tyrosine Loss Was Driven by Selection against Promiscuous Phosphorylation

Siddharth Pandya,¹ Travis J. Struck,¹ Brian K. Mannakee,^{1,2} Mary Paniscus,^{1,3} and Ryan N. Gutenkunst^{*1}

¹Department of Molecular and Cellular Biology, University of Arizona

²Division of Epidemiology and Biostatistics, Mel and Enid Zuckerman College of Public Health, University of Arizona

³Graduate Interdisciplinary Program in Genetics, University of Arizona

*Corresponding author: E-mail: rgutenk@email.arizona.edu.

Associate editor: Gregory Wray

Abstract

Protein tyrosine phosphorylation is a key regulatory modification in metazoans, and the corresponding kinase enzymes have diversified dramatically. This diversification is correlated with a genome-wide reduction in protein tyrosine content, and it was recently suggested that this reduction was driven by selection to avoid promiscuous phosphorylation that might be deleterious. We tested three predictions of this intriguing hypothesis. 1) Selection should be stronger on residues that are more likely to be phosphorylated due to local solvent accessibility or structural disorder. 2) Selection should be stronger on proteins that are more likely to be promiscuously phosphorylated because they are abundant. We tested these predictions by comparing distributions of tyrosine within and among human and yeast orthologous proteins. 3) Selection should be stronger against mutations that create tyrosine versus remove tyrosine. We tested this prediction using human population genomic variation data. We found that all three predicted effects are modest for tyrosine when compared with the other amino acids, suggesting that selection against deleterious phosphorylation was not dominant in driving metazoan tyrosine loss.

Key words: phosphorylation, tyrosine, promiscuous, solvent accessibility, structural disorder, expression level, allele frequency.

Introduction

Protein phosphorylation is a key posttranslational modification that can reversibly regulate protein activity, binding, or localization, and the evolution of phospho-regulation is receiving increasing attention (Moses and Landry 2010; Tan 2011). Phosphorylation is common; the PhosphoSitePlus database (Hornbeck et al. 2012) reports more than 150,000 phosphorylation sites among more than 17,000 human proteins. Because the kinases that carry out phosphorylation may be promiscuous (Ubersax and Ferrell 2007), many of these phosphorylations may be nonfunctional (Lienhard 2008). Notably, Landry et al. (2009) estimated, based on evolutionary conservation, that up to 65% of phosphorylation sites may be nonfunctional. (Although some nonconserved sites may function by regulating bulk electrostatics [Tan et al. 2010].)

Tyrosine phosphorylation, in particular, is key to cell–cell communication in metazoans (multicellular animals) (Hunter 2009). The basic repertoire of tyrosine kinases existed before the divergence of metazoans from filastereans (Clarke et al. 2013; Ringrose et al. 2013), but tyrosine kinases have diversified dramatically in the metazoan lineage (Shiu and Li 2004; Lim and Pawson 2010; Ringrose et al. 2013). Moreover, there is a strong correlation between the number of tyrosine kinases encoded in a species' genome and the number of cell types in that species.

Tan et al. (2009) recently noted a strong negative correlation in metazoans between the number of genomically

encoded tyrosine kinases and the overall abundance of tyrosine within proteins (supplementary fig. S1, Supplementary Material online). Tyrosine is metabolically costly, but Tan et al. (2009) argued that is unlikely to be driving the correlation, because tryptophan and phenylalanine are similarly costly but show, respectively, strongly positive and weakly negative correlations (supplementary fig. S1, Supplementary Material online). Tan et al. (2009) also noted that the single-celled choanoflagellate *Monosiga brevicollis*, which possesses an unusually large number of tyrosine kinases, uses an unusually small amount of tyrosine in its proteome, consistent with the metazoan correlation. Lastly, they found that the relative deficit of tyrosine in human versus *Saccharomyces cerevisiae* yeast proteins was larger in proteins not observed to be tyrosine-phosphorylated in humans. Together, these observations led Tan et al. (2009) to propose that the correlation between metazoan tyrosine kinase count and tyrosine usage was driven by selection to avoid deleterious tyrosine phosphorylation. Promiscuous tyrosine phosphorylation could be deleterious by interfering with protein function or stability. For example, the fission yeast *Schizosaccharomyces pombe* possesses no native tyrosine kinases, and expressing the c-Src tyrosine kinase in *Sc. pombe* is lethal (Superti-Furga and Fumagalli 1993).

In a technical comment on Tan et al. (2009), Su et al. (2011) argued that a more parsimonious hypothesis was that the loss of tyrosine in metazoans was driven by an increase in genomic guanine-cytosine (GC) content, because

tyrosine is encoded by codons that are GC-poor. Consistent with this argument, Su et al. (2011) noted that other amino acids encoded by GC-rich or -poor codons are also enriched or depleted, respectively, in metazoan genomes. (Su et al. [2011] also pointed out that correlation analyses between species characters must account for confounding phylogenetic relationships [Felsenstein 1985].) In response, Tan et al. (2011) noted that the most relevant measures of GC content focus on coding regions and that tyrosine content correlates more strongly with number of tyrosine kinases than with those measures of GC content, suggesting that GC content does not fully explain the decline in tyrosine content. They also noted that phenylalanine, which is also encoded by a GC-poor codon and differs from tyrosine only by the phosphorylatable hydroxyl group, does not show the preferential loss in nontyrosine-phosphorylated proteins that tyrosine shows. Moreover, phenylalanine and tryptophan residues in yeast are less likely to be substituted for tyrosine in humans if a protein is not observed to be tyrosine phosphorylated than if it is. Lastly, among tyrosine residues that are conserved between human and yeast, Tan et al. (2011) used their NetPhorest algorithm (Miller et al. 2008) to show that human tyrosine residues are less likely to reside in known targeting motifs of tyrosine kinases, suggesting that flanking bases may also be evolving to reduce phosphorylation. More recently, Kutchko and Siltberg-Liberles (2013) noted that the expansion of metazoan tyrosine kinases was paralleled by an expansion in metabolic pathways that consume tyrosine to synthesize neurotransmitters and hormones. Diversion of tyrosine from protein synthesis to those metabolic pathways may thus also contribute to the metazoan loss of proteomic tyrosine.

The Tan et al. (2009) promiscuous phosphorylation hypothesis remains controversial. Changes in GC content do not account for all the correlation between number of tyrosine kinases and tyrosine content (Su et al. 2011; Tan et al. 2011). Definitive evidence is, however, lacking as to whether the remaining correlation is driven by selection against promiscuous phosphorylation. We thus sought to independently test several predictions of the promiscuous phosphorylation hypothesis. 1) Selection and thus tyrosine loss should be stronger for residues with greater propensity to be phosphorylated, based on accessibility (Ubersax and Ferrell 2007) or structural disorder (Collins et al. 2008). 2) Selection and thus tyrosine loss should be stronger for proteins that are more abundant and thus more prone to promiscuous phosphorylation (Levy et al. 2012). 3) In species with many tyrosine kinases, selection should disfavor mutations that create tyrosine compared with mutations that remove tyrosine, causing a difference between their allele frequency distributions (Wright 1938). We focused on these predictions, in particular, because they are all relatively insensitive to the changes in genomic GC content that have occurred during metazoan evolution; predictions 1 and 2 because they focus not on the total amount of tyrosine but rather on its distribution within the proteome; prediction 3 because it compares alleles within a single species.

Results and Discussion

In testing our first two predictions, we followed Tan et al. (2009) by focusing on comparisons between humans and the budding yeast *Sa. cerevisiae*, because humans possess a large number of tyrosine kinases (89), and yeast are the phylogenetically closest eukaryotes that possess no conventional tyrosine kinases. To minimize potential biases due to changes in protein function, we restricted our analyses to proteins that are orthologous between human and yeast. The lack of tyrosine kinases in yeast may be an evolutionarily derived state (Shiu and Li 2004; Suga et al. 2012). To control for phenomena associated with the yeast-specific loss of tyrosine kinases, we ran similar analyses comparing humans and the fruit fly *Drosophila melanogaster*, which possesses the smallest number of tyrosine kinases (29) among the metazoan model organisms considered by Tan et al. (2009). In all cases, comparisons with fruit fly yield similar conclusions to comparisons with yeast.

Phosphorylation Propensity

Our first prediction was that tyrosine residues that are more likely to be promiscuously phosphorylated will be more strongly selected against and thus preferentially lost. We assessed promiscuous phosphorylation propensity via solvent accessibility and structural disorder. We used these broad-scale predictors rather than a site-specific phosphorylation predictor (Trost and Kusalik 2011), because even state-of-the-art site-specific predictors have modest sensitivity and specificity, particularly when applied to species other than that on which they were trained (Dou et al. 2014). Moreover, we expect that promiscuously phosphorylated sites will be in relatively weak motifs that will be particularly challenging for site-specific predictors. We thus focused on the general biophysical preference for kinases to target accessible and unstructured sites.

We initially planned to assess solvent accessibility using Protein Data Bank structures (PDB; Bernstein and Koetzle 1977). For the metazoan species considered by Tan et al. (2009), however, the available PDB sequences do not follow the genomic trend of decreasing tyrosine content with increasing number of tyrosine kinases (supplementary fig. S1A, Supplementary Material online; Pearson correlation ≈ -0.19 , $P \approx 0.57$), presumably due to biases in the proteins for which structures have been solved. We thus turned to computational sequence-based predictions of solvent accessibility, using SPINE X, which is based on a multistep neural network (Faraggi et al. 2011). We validated this approach using the PhosphoSitePlus database (Hornbeck et al. 2012), showing that human tyrosine phosphorylation propensity does indeed increase strongly with predicted solvent accessibility (fig. 1A).

The deleterious phosphorylation hypothesis predicts that tyrosine content should decline most dramatically for the residues most likely to be phosphorylated, in this case, those with high solvent accessibility. Figure 1B shows the distributions of solvent accessibility for tyrosine residues in human and yeast, among orthologous proteins. The human

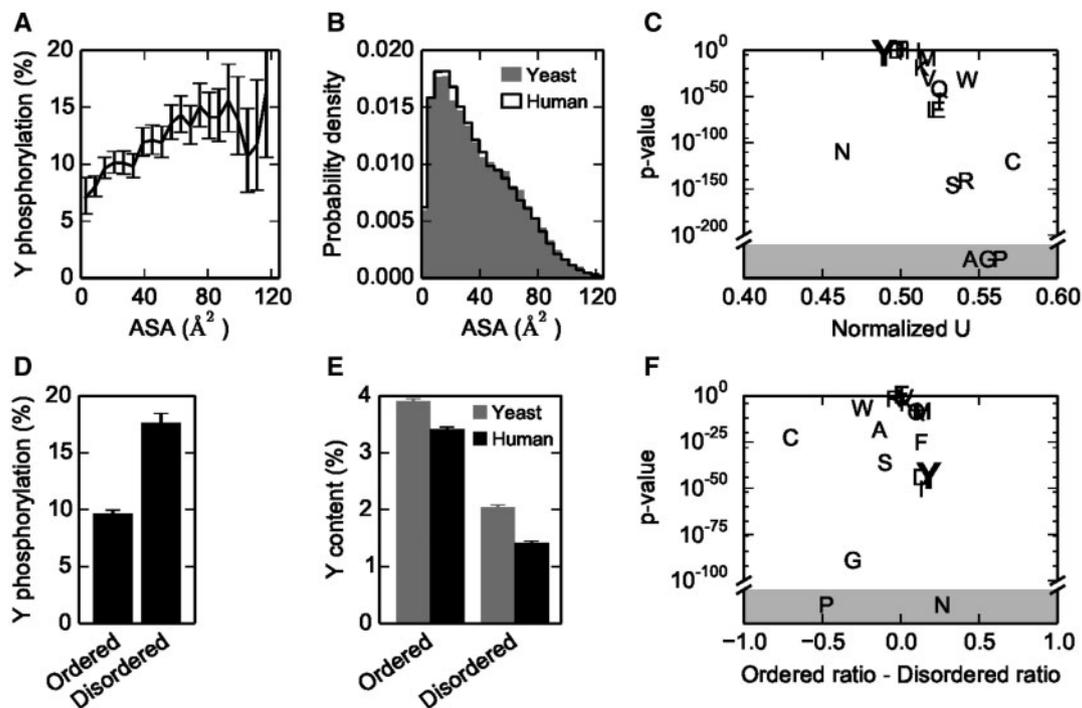


Fig. 1. Effects of phosphorylation propensity. (A) Observed fraction of tyrosine residues phosphorylated versus predicted absolute solvent accessibility (ASA) in human proteins. (B) Distributions of tyrosine ASA in budding yeast and human proteins. See [supplementary figure S3, Supplementary Material online](#), for other amino acids. (C) Normalized Mann–Whitney U statistics and corresponding P values comparing ASA distributions between yeast and human. Shaded region is $P < 10^{-200}$. See [supplementary table S1, Supplementary Material online](#), for numerical values. (D) Observed fraction of tyrosine residues phosphorylated versus structural disorder in human proteins. (E) Tyrosine content versus disorder in yeast and human proteins. See [supplementary figure S10, Supplementary Material online](#), for other amino acids. (F) Difference between ratios of human-to-yeast amino acid content for ordered and disordered residues and corresponding P values. Shaded region is $P < 10^{-100}$. See [supplementary table S1, Supplementary Material online](#), for numerical values. Throughout figure, data are from longest one-to-one orthologous proteins, and error bars denote two-standard deviation uncertainties. See [supplementary figure S4, Supplementary Material online](#), for similar comparisons between human and fruit fly and [supplementary figure S6, Supplementary Material online](#), for comparisons using homology models to PDB structures to estimate solvent accessibility.

distribution is only slightly shifted toward lower accessibility (Mann–Whitney $U/[n_H n_Y] = 0.489$), but the difference is statistically significant ($n_H = 36,756$, $n_Y = 39,006$, $P \approx 2.7 \times 10^{-7}$ two-tailed). To assess the biological significance of this result, we compared tyrosine with the other amino acids. Among all 20 amino acids, tyrosine has the 16th-largest difference in accessibility distributions between human and yeast (fig. 1C and [supplementary table S1, Supplementary Material online](#)). Similar results were found when we compared humans with fruit flies ([supplementary fig. S4 and table S2, Supplementary Material online](#)) and when we used homology models to PDB structures to estimate solvent accessibility ([supplementary fig. S6 and table S2, Supplementary Material online](#)). Consistent with the slight shift in the overall absolute solvent accessibility (ASA) distribution, the human ortholog has a lower median tyrosine solvent accessibility in 45% of pairwise comparisons between orthologous human and yeast proteins ([supplementary fig. S8A, Supplementary Material online](#)). This is statistically significantly different from 50% (binomial test, $n = 2,167$, $P \approx 2.8 \times 10^{-6}$ two-tailed), but among all 20 amino acids, tyrosine has only the 13th most statistically significant difference in this analysis ([supplementary fig. S8B and table S2, Supplementary Material online](#)).

We next considered the effect of structural disorder, using SPINE-D to predict whether residues are ordered or disordered within human and yeast orthologous proteins (Zhang et al. 2012). As expected (Iakoucheva et al. 2004; Collins et al. 2008), we found in humans that residues in predicted disordered regions are more likely to be phosphorylated (fig. 1D). If tyrosine were selected against irrespective of phosphorylation propensity, then an equal proportion of tyrosine should have been lost in human ordered and disordered regions. Conversely, the deleterious phosphorylation hypothesis predicts that the proportional tyrosine deficit in human relative to yeast should be larger in disordered regions. This is particularly true because a larger fraction of residues are disordered in human proteins (39.3%) versus in yeast proteins (34.3%), which in the absence of amino acid composition change would tend to increase promiscuous phosphorylation. The human tyrosine content deficit is indeed larger in disordered regions (fig. 1E); the tyrosine content of yeast ordered residues is $3.91 \pm 0.2\%$ versus $3.41 \pm 0.2\%$ for human ordered residues, and the tyrosine content of yeast disordered residues is $2.04 \pm 0.2\%$ versus $1.41 \pm 0.2\%$ in human disordered residues. The proportional human-to-yeast tyrosine content in ordered regions is thus $3.41/3.91 \approx 0.872$ versus $1.41/2.08 \approx 0.691$ in disordered regions.

The difference in ratios of 0.181 ± 0.013 is statically different from zero, based on a normal approximation ($P \approx 8 \times 10^{-46}$, two-tailed). Among all 20 amino acids, tyrosine has the sixth-largest magnitude of difference in content ratios between ordered and disordered residues, which is the fifth-most statistically significant (fig 1F; supplementary table S1, Supplementary Material online). We found similar results comparing human against fruit fly (supplementary fig. S4 and table S2, Supplementary Material online).

The larger relative difference in tyrosine content in disordered versus ordered regions (fig. 1E) is consistent with the deleterious phosphorylation hypothesis, but it is also consistent with the overall faster rate of evolution in disordered regions (Brown et al. 2002). Moreover, the tyrosine content difference is not exceptional when compared against the other amino acids (fig. 1F), suggesting that any effect caused by selection against deleterious phosphorylation is modest compared with the forces that shape the evolution of the other amino acids. Although statistically significant, the difference in tyrosine solvent accessibility distributions we found was slight (fig. 1B). In particular, tyrosine had one of the smallest differences in accessibility distributions among the 20 amino acids (fig. 1C). The overall increase in evolutionary rate with solvent accessibility (Franzosa and Xia 2009) should enhance any effect of selection against highly exposed tyrosine. The small effect we observe thus suggests that highly exposed tyrosine is not under particularly strong negative selection against promiscuous phosphorylation. Another possible explanation for the lack of stronger selection against exposed or disordered tyrosine is that promiscuous phosphorylation may itself be less deleterious when it occurs on such tyrosine. However, by regulating bulk electrostatics (Serber and Ferrell 2007; Strickfaden et al. 2007) or binding (Pawson et al. 2001), phosphorylation in disordered regions can be just as functional as in ordered regions (e.g., Holt et al. 2009), and thus promiscuous phosphorylation in disordered regions could be just as deleterious as in ordered regions. Our results suggest that selection against promiscuous phosphorylation is similarly weak in both ordered and disordered regions.

Protein Abundance

The deleterious phosphorylation hypothesis predicts that the deficit in human versus yeast tyrosine content should be larger in more abundant proteins, because they should suffer more promiscuous phosphorylation (Levy et al. 2012). The hypothesis (Kutchko and Siltberg-Liberles 2013) that metabolic limitation is driving the tyrosine deficit also predicts a larger deficit in more abundant proteins, because they consume more of the metabolic budget than less abundant proteins. Defining an evolutionarily relevant effective protein abundance is challenging in multicellular organisms with complex development, so we used the Codon Adaptation Index (CAI; Sharp and Li 1987) of the yeast ortholog as a proxy for protein abundance. The CAI measures the degree to which a protein is encoded using preferred codons, and it is a commonly used proxy for protein abundance in yeast

(Ghaemmaghami et al. 2003), particularly in evolutionary analyses (e.g., Drummond et al. 2005; Wall et al. 2005).

For tyrosine, we found a weak, but statistically significant, positive Pearson correlation between difference in tyrosine content between human and yeast orthologs and yeast CAI (fig. 2; $r \approx 0.071$, $P \approx 8 \times 10^{-4}$ two-tailed). This indicates that the human tyrosine deficit is smaller in more highly expressed proteins. Among all amino acids, tyrosine shows the 12th-strongest dependence of differences in amino acid content on yeast CAI (fig. 2B and supplementary table S1, Supplementary Material online). Moreover, although GC content is higher in human than in yeast across the genome, the difference is larger for highly expressed genes (supplementary fig. S16, Supplementary Material online; Lercher et al. 2003; Sémon et al. 2005), which would enhance any removal of tyrosine by selection. Because yeast CAI is an imperfect proxy for metazoan protein abundance, we carried out similar analyses comparing human and fruit fly using fruit fly protein abundance (supplementary fig. S12, Supplementary Material online; Wang et al. 2012) and comparing mouse and yeast using mouse gene expression (supplementary fig. S14, Supplementary Material online; Su et al. 2004; Drummond and Wilke. 2008), in both cases averaging across tissues and developmental stages. Both these analyses yielded similar results to our yeast analysis.

We found smaller differences in tyrosine content for more abundant proteins, contrary to the prediction of the deleterious phosphorylation and metabolic limitation hypotheses. This result is, however, consistent with the well-known reduction in evolutionary rates of proteins with higher expression (Pál et al. 2001). The fact that tyrosine is unexceptional in our protein abundance analysis thus suggests that selection against tyrosine due to deleterious phosphorylation or metabolic limitation is modest compared with the more general selection against protein misfolding (Drummond et al. 2005) and misinteraction (Yang et al. 2012) that is thought to drive the overall correlation between evolutionary rate and expression level.

Allele Frequencies

The deleterious phosphorylation hypothesis predicts that, in species with many tyrosine kinases, mutations that create tyrosine should, on average, be selectively disfavored compared with mutations that remove tyrosine. In other words, the distribution of fitness effects (DFE) of mutations that create tyrosine should be shifted toward more negative values, relative to the DFE of mutations that remove tyrosine (Eyre-Walker and Keightley 2007). Patterns of genetic polymorphism contain a great deal of information about the DFE (Eyre-Walker et al. 2006; Boyko et al. 2008). A shift in the frequency spectrum of segregating polymorphisms toward lower frequencies is indicative of moderate negative selection, which lowers the probability of alleles reaching high frequency. A lack of polymorphism is indicative of strong negative selection, which purges mutations from the population before they rise to even modest frequency. We used high-coverage exome sequencing data from 1,092 human

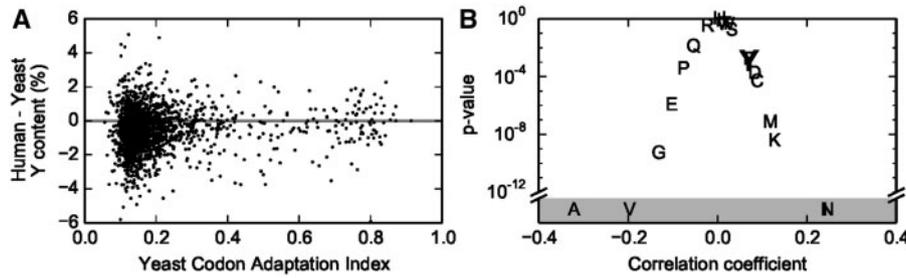


Fig. 2. Effects of protein abundance. (A) For one-to-one human-yeast protein orthologs, difference in tyrosine content versus yeast CAI. See [supplementary figure S11, Supplementary Material](#) online, for other amino acids. (B) Pearson correlation between difference in amino acid content and yeast CAI and corresponding *P* values. Amino acids with $P < 10^{-12}$ are plotted in the shaded region. See [supplementary table S1, Supplementary Material](#) online, for numerical values, [supplementary figure S12, Supplementary Material](#) online, for human–fruit fly comparison using protein abundances, and [supplementary figure S14, Supplementary Material](#) online, for mouse–yeast comparison using gene expression.

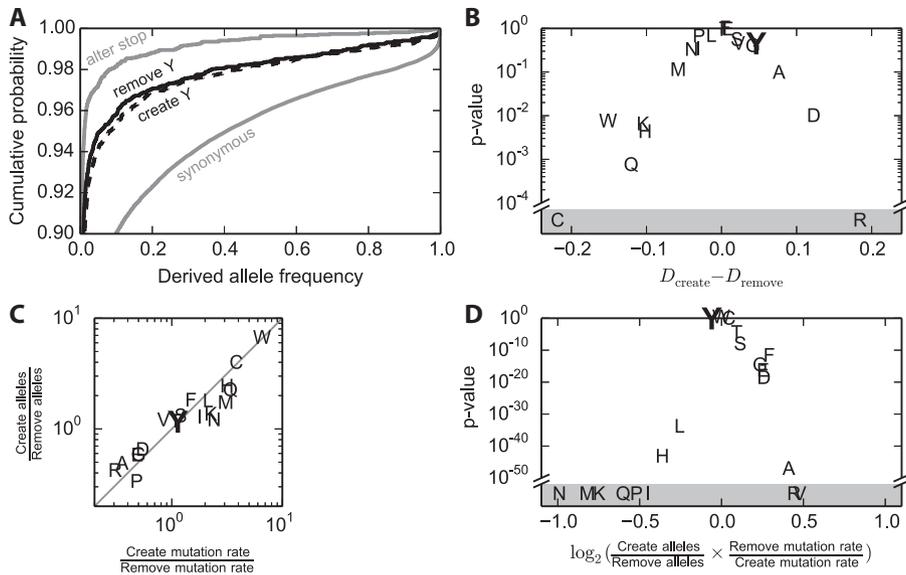


Fig. 3. Allele frequencies and counts. (A) Cumulative distributions of minor allele frequencies for mutations that alter stop codons (typically deleterious; gray), remove tyrosine (solid black), create tyrosine (dotted black), or do not change the coded amino acid (approximately neutral; gray). For confidence intervals and other amino acids, see [supplementary figure S17, Supplementary Material](#) online. (B) Statistical significance versus difference in Tajima’s *D* for alleles that create or remove amino acids. Shaded region is $P < 10^{-4}$. (C) Relative numbers of observed alleles that create or remove particular amino acids versus expected relative mutation rates based on a context-dependent mutation rate model. (D) Deviations between allele count and mutation ratios for amino acid creation or removal and corresponding *P* values.

individuals generated by phase 1 of the 1000 Genomes Project (1000 Genomes Project Consortium 2012) to assess both regimes of negative selection for mutations that create or remove tyrosine.

Figure 3A shows the frequency spectra of alleles that create or remove tyrosine. For comparison, also shown are the spectra for synonymous alleles, which do not change the protein and are expected to be relatively selectively neutral, and for stop codon-altering alleles, which truncate or extend proteins and are expected to particularly deleterious. We quantified the frequency spectra differences between alleles that create or remove tyrosine using Tajima’s *D* (Tajima 1989), a widely used measure of how skewed a frequency spectrum is relative to the standard neutral model. More negative values of *D* indicate a greater shift toward rare alleles and thus stronger negative selection. Tyrosine has the 10th largest magnitude of

difference in *D* ($D_{\text{create}} - D_{\text{remove}} \approx 0.046$), but that difference is not statistically significant ($P \approx 0.38$, two-tailed permutation test). Cysteine, glutamine, and arginine do, however, show significant differences ($P < 0.05/20$). As a sharper test of the promiscuous phosphorylation hypothesis, we specifically compared alleles that remove versus create tyrosine in disordered regions, again finding no statistically significant difference ([supplementary fig. S18, Supplementary Material](#) online; $D_{\text{create}} - D_{\text{remove}} \approx -0.051$, $P \approx 0.63$, two-tailed permutation test).

Our analysis of the frequency spectrum suggests that alleles that create versus remove tyrosine have similar distributions of moderately deleterious effects. To test for differences among strongly deleterious alleles, we asked whether the observed ratio of numbers of segregating alleles that create versus remove tyrosine matches that expected from the

influx of new mutations. To calculate expected relative rates of mutation (supplementary table S3, Supplementary Material online), we used a context-dependent mutation rate model based on the flanking two bases (Hwang and Green 2004). Such local context does not explain all variation in mutation rates (Hodgkinson and Eyre-Walker 2011), but we expect that larger-scale sources of rate variation average out in our analysis, because sites where mutations could create or remove tyrosine are scattered roughly equally throughout the genome. The ratio of numbers of segregating alleles that create versus remove tyrosine is statistically indistinguishable from the ratio expected under this mutation model (fig. 3A and B; $P \approx 0.10$ two-tailed). In contrast, 17 other amino acids show a statistically significant difference after Bonferroni correction (fig. 3D and supplementary table S1, Supplementary Material online; $P < 0.05/20$), although some differences could be caused by biases in the mutation rate model.

We found only small and statistically nonsignificant differences between the relative frequencies of and numbers of alleles that create versus remove tyrosine in humans. This is consistent with relative neutrality between these classes of mutations, suggesting that any selection against deleterious tyrosine phosphorylation is modest in humans.

Other Amino Acids

Because in our genomic analyses even small effects can be statistically significant, throughout we have assessed biological significance by comparing tyrosine against the other amino acids. Our results thus touch on hypotheses regarding other amino acids.

Phenylalanine is similar biochemically to tyrosine, differing only in that it lacks a phosphorylatable hydroxyl group, and Tan et al. (2009, 2011) used differences with phenylalanine to argue for phosphorylation-driven effects on tyrosine. In our analyses of residue solvent accessibility, residue disorder, protein abundance, and allele frequencies, phenylalanine behaves similarly to tyrosine, with deviations from the null model in the same direction, although less statistically significant (supplementary table S1, Supplementary Material online). The one exception is that phenylalanine shows an excess of creation versus removal alleles compared with the predicted relative mutation rates, although this excess is not exceptional when compared with the other amino acids (fig. 3D).

Cysteine is enriched in multicellular eukaryotes (supplementary fig. S2, Supplementary Material online; Miseta and Csutora 2000), and we found that this enrichment is particularly strong in regions of high solvent accessibility (fig. 1C and supplementary fig. S3, Supplementary Material online) and structural disorder (fig. 1F and supplementary fig. S10, Supplementary Material online). This is consistent with the hypothesis that this enrichment is driven by the acquisition of disulfide bonds to promote protein stability (Wong et al. 2010).

In contrast, asparagine is particularly depleted in regions of high solvent exposure (fig. 1C and supplementary fig. S3, Supplementary Material online) or structural disorder

(fig. 1F and supplementary fig. S10, Supplementary Material online). Additionally, the ratio of segregating alleles that create versus remove asparagine is markedly lower than expected from the predicted mutation rates (fig. 3C and D). These observations are consistent with the hypothesis that mammalian genomes are depleted in asparagine repeats (Kreil and Kreil 2000) due to selection against excessive glycosylation (Karlin et al. 2002). Notably, segregating alleles that remove asparagine in a repeat (defined as being flanked by another asparagine) have higher frequencies, on average, than alleles that create or remove isolated asparagine (supplementary fig. S19, Supplementary Material online; $D \approx 0.166$; $P \approx 0.047$, two-tailed), consistent with selection to remove asparagine repeats.

Conclusions

We tested several predictions of the hypothesis that selection against deleterious phosphorylation drove tyrosine loss in metazoans. Within proteins, we found that human and yeast have very similar distributions of tyrosine solvent accessibility (fig. 1B), consistent with evolution that is neutral with respect to accessibility and thus promiscuous phosphorylation propensity. Also within proteins, we found that the tyrosine deficit in humans relative to yeast was not exceptionally larger for residues more likely to be phosphorylated on the basis of structural disorder (fig. 1E). Among proteins, we found that the tyrosine deficit was not larger for proteins that are more likely to be phosphorylated because they are abundant (fig. 2). Among mutations, we found in humans that alleles that create versus remove tyrosine have similar frequency spectra and that the numbers of such alleles match the mutation influx (fig. 3), which is evidence for neutrality between creation and removal of tyrosine. All these results contradict the predictions of the promiscuous phosphorylation hypothesis and are consistent with a model in which tyrosine loss is neutral with respect to phosphorylation.

Our results cast doubt on the hypothesis that metazoan tyrosine loss was an adaptation to prevent promiscuous phosphorylation, but it remains unclear whether any other selective forces contributed substantially to the loss. A definitive answer will likely demand a quantitative model of how nonadaptive forces cause amino acid content to drift over evolutionary time. Such a model would need to build on recent efforts to model the roles that selection and GC-biased gene conversion play in the evolution of genomic GC content (e.g., Capra et al. 2013). The nonadaptive evolutionary forces of mutation and genetic drift can create complex and seemingly adaptive phenotypes (Lynch 2007), particularly in organisms with relatively small effective population sizes, such as many metazoans (Lynch and Conery 2003). The loss of tyrosine may simply be one such nonadaptive phenotype.

Materials and Methods

Protein sequence data for *Homo sapiens* and *Sa. cerevisiae* were obtained from Ensembl release 72 (Flicek et al. 2014), using the pep.all files that contain the “super-set of all transcripts resulting from Ensembl known, novel and

pseudogene predictions.” One-to-one orthologs between human and yeast were identified as the longest proteins in each of 2,180 orthologous gene sets annotated in Ensembl release 72. Numbers of genomic tyrosine kinases were obtained on January 7, 2014 from the SMART database version 7.0 (Letunic et al. 2012), by searching for TyrKc domains in “Genomic” mode.

Phosphorylation Propensity

Data on experimentally observed human tyrosine phosphorylation were obtained from the PhosphoSitePlus database (Hornbeck et al. 2012) on July 3, 2013. We used Ensembl BioMart (Kinsella et al. 2011) to map Ensembl protein identifiers to the UniProt accessions by which PhosphoSitePlus indexes its data, succeeding for 1,885 of our 2,180 human orthologs. Among those proteins, there were 4,180 phosphorylated tyrosine residues in the PhosphoSitePlus database, distributed among 1,324 proteins. We analyzed the 4,132 of these phosphorylated tyrosine residues that we could map to our Ensembl sequences, based on the position and flanking context reported in the PhosphoSitePlus database. Residue solvent accessibility was predicted using SPINE X 2.0 (Faraggi et al. 2011) and structural disorder predicted using SPINE-D, for all human-yeast one-to-one orthologs, with the exception of 18 for which disorder could not be predicted due to \cup or \times codons in the human sequence.

Uncertainties on phosphorylation and amino acid-content percentages in figure 1A, D, and E were estimated based on a binomial distribution as $\sqrt{P(100 - P)/N}$, where P is percentage of interest and N is the total number of residues considered. P values in figure 1F are based on a normal approximation, with standard deviations propagated through the ratios of human to yeast amino acid contents for ordered and disordered residues and then through the difference in these ratios.

Protein Abundance

CAI was calculated in BioPython (Cock et al. 2009), using the adaptation index from Sharp and Li (1987). Six ortholog pairs that corresponded to yeast mitochondrial genes were excluded from the analysis.

Allele Frequencies

Allele frequency data were obtained from phase 1 of the 1000 Genomes Project (1000 Genomes Project Consortium 2012). We used only autosomal single-nucleotide polymorphisms (SNP) characterized by high-coverage exome sequencing for which all 1,092 individuals were successfully called. We inferred ancestral states from the four-way EPO (Enredo, Pecan, Ortheus) alignments between human, chimp, orangutan, and rhesus macaque provided by the project, and we only used SNPs for which all three outgroups agreed on the ancestral state. The effects of those SNPs on the protein were inferred from the project’s annotations based on GENCODE Release 7. If there were multiple annotations indicating different effects on different transcripts, we discarded that SNP from our analysis.

To calculate relative total rates of mutation to and from different amino acids in the human genome, we applied the tri-nucleotide mutation rate matrix inferred for the human lineage by Hwang and Green (2004). For each possible codon and flanking bases (“padded codon”), we applied the nine possible mutations to the codon itself, looking up the rate for that mutation in the Hwang and Green (2004) matrix and tracking what amino acids that mutation would create or remove. We also counted the total occurrences of each padded codon in the longest isoforms of 22,665 human protein-coding genes, using data from Ensembl release 72. The total mutation rate for creating or removing a given amino acid is then the sum over all relevant padded codon mutations of the mutation rate times the total number of occurrences. Supplementary table S3, Supplementary Material online, reports these total rates, normalized by the largest total rate.

To assess the statistical significance of differences in Tajima’s D (fig. 3B), we ran 10,000 permutations of alleles between the classes create and remove for each amino acid. Standard deviations of allele count ratios (fig. 3C) and confidence intervals on allele frequency distributions in Supplementary figure S17, Supplementary Material online, were calculated by bootstrapping 1,000 times in 1 Mb chunks over the 1000 Genomes data. Standard deviations of mutation rate ratios (fig. 3D) were calculated by bootstrapping 1,000 times over longest transcripts in our human genomic data. P values comparing ratios of creation and removal alleles and mutations (fig. 3D) were calculated by propagating standard deviations and approximating the distribution of ratios of ratios by a normal distribution.

Supplementary Material

Supplementary materials and methods, tables S1–S3, and figures S1–S19 are available at *Molecular Biology and Evolution* online (<http://www.mbe.oxfordjournals.org/>).

Acknowledgments

This work was supported by the National Science Foundation (grant number DEB-1146074) and by the National Institutes of Health (grant number P30-GM092391). B.K.M. was also supported by a scholarship from the Phoenix chapter of the ARCS Foundation. The authors thank “Gnana” S. Gnanakaran for helpful discussions, and Ping-Hsun “Benson” Hsieh for help with preliminary analyses of allele frequencies.

References

- 1000 Genomes Project Consortium. 2012. An integrated map of genetic variation from 1,092 human genomes. *Nature* 491(7422):56–65.
- Bernstein F, Koetzle T. 1977. The Protein Data Bank: a computer-based archival file for macromolecular structures. *Eur J Biochem* 80(2):319–324.
- Boyko AR, Williamson Shindap AR, Degenhardt JD, Hernandez RD, Lohmueller KE, Adams MD, Schmidt S, Sninsky JJ, Sunyaev SR, White TJ, et al. 2008. Assessing the evolutionary impact of amino acid mutations in the human genome. *PLoS Genet* 4(5):e1000083.
- Brown CJ, Takayama S, Campen AM, Vise P, Marshall TW, Oldfield CJ, Williams CJ, Dunker AK. 2002. Evolutionary rate heterogeneity in proteins with long disordered regions. *J Mol Evol* 55(1):104–110.

- Capra JA, Hubisz MJ, Kostka D, Pollard KS, Siepel A. 2013. A model-based analysis of GC-biased gene conversion in the human and chimpanzee genomes. *PLoS Genet.* 9(8):e1003684.
- Clarke M, Lohan AJ, Liu B, Lagkouvardos I, Roy S, Zafar N, Bertelli C, Schilde C, Kianianmomeni A, Bürglin TR, et al. 2013. Genome of *Acanthamoeba castellanii* highlights extensive lateral gene transfer and early evolution of tyrosine kinase signaling. *Genome Biol.* 14(2):R11.
- Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B, et al. 2009. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* 25(11):1422–1423.
- Collins MO, Yu L, Campuzano I, Grant SGN, Choudhary JS. 2008. Phosphoproteomic analysis of the mouse brain cytosol reveals a predominance of protein phosphorylation in regions of intrinsic sequence disorder. *Mol Cell Proteomics.* 7(7):1331–1348.
- Dou Y, Yao B, Zhang C. 2014. PhosphoSVM: prediction of phosphorylation sites by integrating various protein sequence attributes with a support vector machine. *Amino Acids* 46(6):1459–1469.
- Drummond DA, Bloom JD, Adami C, Wilke CO, Arnold FH. 2005. Why highly expressed proteins evolve slowly. *Proc Natl Acad Sci U S A.* 102(40):14338–14343.
- Drummond DA, Wilke CO. 2008. Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell* 134(2):341–352.
- Eyre-Walker A, Keightley PD. 2007. The distribution of fitness effects of new mutations. *Nat Rev Genet.* 8(8):610–618.
- Eyre-Walker A, Woolfit M, Phelps T. 2006. The distribution of fitness effects of new deleterious amino acid mutations in humans. *Genetics* 173(2):891–900.
- Faraggi E, Zhang T, Yang Y, Kurgan L, Zhou Y. 2011. SPINE X: improving protein secondary structure prediction by multistep learning coupled with prediction of solvent accessible surface area and backbone torsion angles. *J Comput Chem.* 33(3):259–267.
- Felsenstein J. 1985. Phylogenies and the comparative method. *Am Nat.* 125(1):1–15.
- Flicek P, Amode MR, Barrell D, Beal K, Billis K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fitzgerald S, et al. 2014. Ensembl 2014. *Nucleic Acids Res.* 42(Database issue):D749–D755.
- Franzosa EA, Xia Y. 2009. Structural determinants of protein evolution are context-sensitive at the residue level. *Mol Biol Evol.* 26(10):2387–2395.
- Ghaemmaghami S, Huh W-K, Bower K, Howson RW, Belle A, Dephore N, O'Shea EK, Weissman JS. 2003. Global analysis of protein expression in yeast. *Nature* 425(6959):737–741.
- Hodgkinson A, Eyre-Walker A. 2011. Variation in the mutation rate across mammalian genomes. *Nat Rev Genet.* 12(11):756–766.
- Holt LJ, Tuch BB, Villén J, Johnson AD, Gygi SP, Morgan DO. 2009. Global analysis of Cdk1 substrate phosphorylation sites provides insights into evolution. *Science* 325(5948):1682–1686.
- 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(Database issue):D261–D270.
- Hunter T. 2009. Tyrosine phosphorylation: thirty years and counting. *Curr Opin Cell Biol.* 21(2):140–146.
- Hwang DG, Green P. 2004. Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. *Proc Natl Acad Sci U S A.* 101(39):13994–14001.
- Iakoucheva LM, Radivojac P, Brown CJ, O'Connor TR, Sikes JG, Obradovic Z, Dunker AK. 2004. The importance of intrinsic disorder for protein phosphorylation. *Nucleic Acids Res.* 32(3):1037–1049.
- Karlin S, Brocchieri L, Bergman A, Mrzcek J, Gentles AJ. 2002. Amino acid runs in eukaryotic proteomes and disease associations. *Proc Natl Acad Sci U S A.* 99(1):333–338.
- Kinsella RJ, Kähäri A, Haider S, Zamora J, Proctor G, Spudich G, Almeida-King J, Staines D, Derwent P, Kerhornou A, et al. 2011. Ensembl BioMart: a hub for data retrieval across taxonomic space. *Database* 2011:bar030.
- Kreil D, Kreil G. 2000. Asparagine repeats are rare in mammalian proteins. *Trends Biochem Sci.* 25(6):270–271.
- Kutchko KM, Siltberg-Liberles J. 2013. Metazoan innovation: from aromatic amino acids to extracellular signaling. *Amino Acids* 45(2):359–367.
- Landry CR, Levy ED, Michnick SW. 2009. Weak functional constraints on phosphoproteomes. *Trends Genet.* 25(5):193–197.
- Lercher MJ, Urrutia AO, Pavlíček A, Hurst LD. 2003. A unification of mosaic structures in the human genome. *Hum Mol Genet.* 12(19):2411–2415.
- Letunic I, Doerks T, Bork P. 2012. SMART 7: recent updates to the protein domain annotation resource. *Nucleic Acids Res.* 40(Database issue):D302–D305.
- Levy ED, Michnick SW, Landry CR. 2012. Protein abundance is key to distinguish promiscuous from functional phosphorylation based on evolutionary information. *Philos Trans R Soc Lond B Biol Sci.* 367(1602):2594–2606.
- Lienhard GE. 2008. Non-functional phosphorylations? *Trends Biochem Sci.* 33(8):351–352.
- Lim WA, Pawson T. 2010. Phosphotyrosine signaling: evolving a new cellular communication system. *Cell* 142(5):661–667.
- Lynch M. 2007. The frailty of adaptive hypotheses for the origins of organismal complexity. *Proc Natl Acad Sci U S A.* 104(Suppl 1), 8597–8604.
- Lynch M, Conery JS. 2003. The origins of genome complexity. *Science* 302(5649):1401–1444.
- Miller ML, Jensen LJ, Diella F, Jørgensen C, Tinti M, Li L, Hsiung M, Parker SA, Bordeaux J, Sicheritz-Ponten T, et al. 2008. Linear motif atlas for phosphorylation-dependent signaling. *Sci Signal.* 1(35):ra2.
- Miseta A, Csutora P. 2000. Relationship between the occurrence of cysteine in proteins and the complexity of organisms. *Mol Biol Evol.* 17(8):1232–1239.
- Moses AM, Landry CR. 2010. Moving from transcriptional to phospho-evolution: generalizing regulatory evolution? *Trends Genet.* 26(11):462–467.
- Pál C, Papp B, Hurst LD. 2001. Highly expressed genes in yeast evolve slowly. *Genetics* 158:927–931.
- Pawson T, Gish GD, Nash P. 2001. SH2 domains, interaction modules and cellular wiring. *Trends Cell Biol.* 11(12):504–511.
- Ringrose JH, van den Toorn HWP, Eitel M, Post H, Neerinx P, Schierwater B, Altelar AFM, Heck AJR. 2013. Deep proteome profiling of *Trichoplax adhaerens* reveals remarkable features at the origin of metazoan multicellularity. *Nat Commun.* 4(1):1408.
- Sémon M, Mouchiroud D, Duret L. 2005. Relationship between gene expression and GC-content in mammals: statistical significance and biological relevance. *Hum Mol Genet.* 14(3):421–427.
- Serber Z, Ferrell JE. 2007. Tuning bulk electrostatics to regulate protein function. *Cell* 128(3):441–444.
- Sharp PM, Li WH. 1987. The codon adaptation index—a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res.* 15(3):1281–1295.
- Shiu S-H, Li W-H. 2004. Origins, lineage-specific expansions, and multiple losses of tyrosine kinases in eukaryotes. *Mol Biol Evol.* 21(5):828–840.
- Strickfaden SC, Winters MJ, Ben-Ari G, Lamson RE, Tyers M, Pryciak PM. 2007. A mechanism for cell-cycle regulation of MAP kinase signaling in a yeast differentiation pathway. *Cell* 128(3):519–531.
- Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, Block D, Zhang J, Soden R, Hayakawa M, Kreiman G, et al. 2004. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc Natl Acad Sci U S A.* 101(16):6062–6067.
- Su Z, Huang W, Gu X. 2011. Comment on “Positive selection of tyrosine loss in metazoan evolution”. *Science* 332(6032):917.
- Suga H, Dacre M, de Mendoza A, Shalchian-Tabrizi K, Manning G, Ruiz-Trillo IN. 2012. Genomic survey of premetazoans shows deep

- conservation of cytoplasmic tyrosine kinases and multiple radiations of receptor tyrosine kinases. *Sci Signal*. 5(222):ra35.
- Superti-Furga G, Fumagalli S. 1993. Csk inhibition of c-Src activity requires both the SH2 and SH3 domains of Src. *EMBO J*. 12(7):2625–2634.
- Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123(11):585–595.
- Tan CSH. 2011. Sequence, structure, and network evolution of protein phosphorylation. *Sci Signal*. 4(182):mr6.
- Tan CSH, Jørgensen C, Linding R. 2010. Roles of “junk phosphorylation” in modulating biomolecular association of phosphorylated proteins? *Cell Cycle* 9(7):1276–1280.
- Tan CSH, Pasculescu A, Lim WA, Pawson T, Bader GD, Linding R. 2009. Positive selection of tyrosine loss in metazoan evolution. *Science* 325(5948):1686–1688.
- Tan CSH, Schoof EM, Creixell P, Pasculescu A, Lim WA, Pawson T, Bader GD, Linding R. 2011. Response to comment on “Positive selection of tyrosine loss in metazoan evolution”. *Science* 332(6032):917.
- Trost B, Kusalik A. 2011. Computational prediction of eukaryotic phosphorylation sites. *Bioinformatics* 27(21):2927–2935.
- Ubersax JA, Ferrell JE. 2007. Mechanisms of specificity in protein phosphorylation. *Nat Rev Mol Cell Biol*. 8(7):530–541.
- Wall DP, Hirsh AE, Fraser HB, Kumm J, Giaever G, Eisen MB, Feldman MW. 2005. Functional genomic analysis of the rates of protein evolution. *Proc Natl Acad Sci U S A*. 102(15):5483–5488.
- Wang M, Weiss M, Simonovic M, Haertinger C, Schrimpf SP, Hengartner MO, von Mering C. 2012. PaxDb, a database of protein abundance averages across all three domains of life. *Mol Cell Proteomics*. 11(8):492–500.
- Wong JWH, Ho SYW, Hogg PJ. 2010. Disulfide bond acquisition through eukaryotic protein evolution. *Mol Biol Evol*. 28(1):327–334.
- Wright S. 1938. The distribution of gene frequencies under irreversible mutation. *Proc Natl Acad Sci U S A*. 24(2):253–259.
- Yang J-R, Liao B-Y, Zhuang S-M, Zhang J. 2012. Protein misinteraction avoidance causes highly expressed proteins to evolve slowly. *Proc Natl Acad Sci U S A*. 109(14):E831–E840.
- Zhang T, Faraggi E, Xue B, Dunker AK, Uversky VN, Zhou Y. 2012. SPINE-D: accurate prediction of short and long disordered regions by a single neural-network based method. *J Biomol Struct Dyn*. 29(4):799–813.