Response to Comment on “Positive Selection of Tyrosine Loss in Metazoan Evolution”

Chris Soon Heng Tan,1,2,3 Erwin M. Schoof,4,5,* Pau Creixell,4,5,* Adrian Pasculescu,1 Wendell A. Lim,6 Tony Pawson,1,2 Gary D. Bader,1,2,3 Rune Linding1,4,†

Su et al. claim guanine-cytosine (GC) content variation can largely explain the observed tyrosine frequency variation, independent of adaptive evolution of cell-signaling complexity. We found that GC content variation, in the absence of selection for amino acid changes, can only maximally account for 38% of the observed tyrosine frequency variation. We also uncovered other mechanisms acting to reduce tyrosine phosphorylation that further support our previous proposal.

The expression of Src tyrosine kinase in the unicellular yeast Schizosaccharomyces pombe and Saccharomyces cerevisiae, whose genomes encode no tyrosine kinase, is toxic due to deleterious tyrosine phosphorylation (1, 2). This raises the question of how metazoans have evolved to accommodate the numerous tyrosine kinases encoded in their genomes (e.g., 90 tyrosine kinases in humans). We proposed that an observed tyrosine depletion (OTD) in metazoan proteins—in addition to tyrosine phosphatase and spatiotemporal/contextual regulation of tyrosine kinases—helps reduce deleterious tyrosine phosphorylation (3, 4). We further proposed that this accounts, at least in part, for the observed negative correlation of genomic tyrosine frequency (FreqTyr) with cell type number and tyrosine kinase number (NumTyrKin) across different species (3). The mutational forces changing GC content, briefly mentioned in our original paper and highlighted by Su et al. (5), might have facilitated the OTD. However, we disagree with Su et al. that “biased nucleotide substitutions (A/T → G/C) removed spurious tyrosine phosphorylation sites, a random genomic dynamics independent of the adaptive evolution of cell-signaling complexity” and that GC content variation can explain most of the observed variation in FreqTyr.

Genetic mutations generate phenotypic variation for selection forces to act upon. Hence, the observed high GC content in metazoans does not nullify our proposal. The concern is whether the OTD has been passively driven, in the absence of selection for amino acid changes, by (i) the mutational forces behind high GC content and (ii) the selection for their effects at the nucleotide level (e.g., transcription and translation) that we collectively termed GC directional force (GC force). We computed a GC-content measure minimally influenced by amino acid selection: GC content at the third position of all four-fold degenerate codons (GC4). We correlated FreqTyr with GC4 content to quantify how much OTD could be passively driven by GC force on protein-coding regions that can directly affect amino acid changes, in contrast to Su et al. who focus on GC content in noncoding regions. There are eight amino acids encoded by four-fold degenerate codons. The GC content at the third position for each of them (GC3) and the computed GC4 showed strong correlation with each other (see Fig. 1A). Thus, we conclude that GC4 content is a robust readout of global GC directional forces minimally influenced by amino acid selection. We also computed GC content at all codon positions (GC123) for comparison.

We found that variation in GC4, GC123, and NumTyrKin can individually account for up to 37.6, 56.2, and 73.4% of variation in FreqTyr, respectively (R2) (Fig. 1B). Adopting the approach by Su et al. to correct for phylogenetic relationship of species analyzed (gradual Brow- nian model), the statistical significance for the negative correlation of FreqTyr with GC4, GC123, and NumTyrKin is $P = 8.5 \times 10^{-3}$, $1.6 \times 10^{-3}$, and $1.8 \times 10^{-3}$ (subroutine test in R software), respectively (6). The GC4 content of a subset of species does correlate negatively with FreqTyr (Fig. 1B, yellow ellipse), but the trend is reversed in several other lineages (Fig. 1B, blue ellipse). In contrast, FreqTyr covaries more consistently with NumTyrKin across different lineages (Fig. 1B). Performing an orthogonal analysis using a generalized additive linear model (7, 8), which models the effects of multiple parameters on a variable simultaneously, we also observed that FreqTyr correlates better with NumTyrKin than with GC4 (Table 1). These results invalidate Su et al.’s claim that GC content variation alone can explain most of the observed variation in FreqTyr. As GC directional forces on coding regions directly influence amino acid frequency in the absence of selection forces, we question why Su et al. chose to emphasize the correlation of FreqTyr with GC content variation in noncoding regions while not commenting on the results of GC4 variation in coding regions. As we are studying protein evolution, it is more relevant to focus on protein-coding DNA regions.

Although we previously acknowledged that other mechanisms may have contributed to the OTD, we supported our proposal with empirical data that human proteins with no detectable phosphotyrosines (non-pTyr proteins) have lost considerably more tyrosines than known tyrosine-phosphorylated proteins (pTyr proteins), as compared with their S. cerevisiae orthologs (3). Here, we extended this analysis to S. pombe and reached the same conclusion (Fig. 2A). Yeasts are compared because they are the known eukaryotes phylogenetically closest to human that lack a dedicated phosphotyrosine signaling system, and we implicitly assume both species are informative of the ancestral tyrosine frequency. Our conclusion is supported by the observation that phenylalanine is not lost preferentially in either protein group (Fig. 2A). This is crucial because phenylalanine and tyrosine are structurally identical except for a phosphorylatable hydroxyl group on tyrosine, and are each encoded by two AA-rich codons (Phe: TTT and TTC; Tyr: TAT and TAC), and thus should be subjected to a similar degree of GC force. Hence, the observed preferential loss of tyrosine is due to the presence of its hydroxyl group and strongly supports the argument that signaling fidelity is a driving force behind OTD.

During evolution, phenylalanine and tyrosine are commonly substituted for each other because of their similar physicochemical properties and encoding codons [see BLOSUM matrices, for example (9)]. We investigated their substitution pattern from yeast to human. As expected, tyrosines in yeast are most frequently substituted with phenylalanine (~35%) in human, whereas phenylalanine and tryptophan in yeast are frequently substituted with tyrosine in human (Fig. 2B). However, we observe that the substitution of phenylalanine and tryptophan in yeast by tyro- 

1. Samuel Lunenfeld Research Institute, Mount Sinai Hospital, Toronto, Canada. 2. Department of Molecular Genetics, University of Toronto, Toronto, Canada. 3. The Donnelly Centre, University of Toronto, Toronto, Canada. 4. The Institute of Cancer Research (ICR), London, UK. 5. Cellar Signal Integration Group (CSI6), Center for Biological Sequence Analysis (CBS), Department of Systems Biology, Technical University of Denmark (DTU), DK-2800 Lyngby, Denmark. 6. Howard Hughes Medical Institute and Department of Cellular and Molecular Pharmacology, University of California, San Francisco, California, USA.

*These authors contributed equally to this work.
†To whom correspondence should be addressed. Email: linding@cbs.dtu.dk
Next, for the set of tyrosines that are conserved between human and \textit{S. cerevisiae} as identified from sequence alignments, we applied the NetPhorest algorithm (10) to investigate their propensity to be phosphorylated by human tyrosine kinases that are influenced by residues flanking the tyrosines on primary sequences. In general, we found that tyrosines from human are less phosphorylatable by human tyrosine kinases than the corresponding tyrosines from \textit{S. cerevisiae} (Fig. 2C) ($P = 5.34 \times 10^{-5}$, Wilcoxon Test). This suggests that there are selection forces favoring mutations flanking tyrosines that reduce tyrosine phosphorylation.

In summary, we contend that our analysis invalidates Su et al.'s claims. Furthermore, we uncovered multiple plausible mechanisms acting to reduce tyrosine phosphorylation. Hence, we suggest that using GC content variation as an Occam's Razor, without considering the physicochemical properties of amino acids in relation to the study of biological systems, can compromise our understanding of similar phenomena. Su et al. concur with us that low \textit{FreqTyr} permitted the expansion of tyrosine kinases through reducing deleterious phosphotyrosines. Reciprocally, an increased \textit{NumTyrKin}, other than promoting

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{fig1.png}
\caption{(A) GC content at the third codon position (GC3) for each set of four-fold degenerate codons in all coding sequences in different species. The four-fold degenerate codon set of each amino acid are GCN (alanine), CGN (arginine), GGN (glycine), CTN (leucine), CCN (proline), TCN (serine), ACN (threonine), and GTN (valine). The species are sorted with decreasing evolutionary distance from human. The GC3 content for these eight amino acids are highly correlated with each other (average Pearson's correlation = 0.85, SD = 0.11). GC3 content for all sets of four-fold degenerate codons are summed to derive the GC4 content for each species. The GC4 content is strongly correlated with GC3 content of these eight amino acids across the species analyzed (average Pearson's correlation = 0.96, SD = 0.036). (B) Correlation of genomic tyrosine frequency (\textit{FreqTyr}) with GC4 content and inferred number of tyrosine kinases (\textit{NumTyrKin}) in species analyzed. \textit{NumTyrKin} are inferred as previously described (3) except for \textit{Monosiga brevicollis}, which is based on (14). The low \textit{FreqTyr} and high \textit{NumTyrKin} observed in \textit{M. brevicollis} is consistent with our proposed evolutionary model (3). The species analyzed are budding yeast (\textit{S. cerevisiae}), choanoflagellate (\textit{M. brevicollis}), worm (\textit{Caenorhabditis elegans}), sea squirt (\textit{Ciona intestinalis}), fly (\textit{Drosophila melanogaster}), mosquito (\textit{Anopheles gambiae}), zebrafish (\textit{Danio rerio}), tetraodon pufferfish (\textit{Tetraodon nigroviridis}), Japanese pufferfish (\textit{Takifugu rubripes}), frog (\textit{Xenopus tropicalis}), chicken (\textit{Gallus gallus}), dog (\textit{Canis familiaris}), cow (\textit{Bos taurus}), mouse (\textit{Mus musculus}), rat (\textit{Rattus norvegicus}), chimpanzee (\textit{Pan troglodytes}), and human (\textit{Homo sapiens}). Looking at the closest pairs of species in five lineages (\textit{R.nor/M.mus}, \textit{P.tro/H.sap}, \textit{C.fam/B.tau}, \textit{T.rub/T.nig}, and \textit{D.meli/A.gam}), a negative relationship between \textit{FreqTyr} and \textit{NumTyrKin} is observed for all five species pairs. Contradicting Su et al.'s claim, a positive correlation between \textit{FreqTyr} and GC4 is observed for three species pairs (\textit{C.fam/B.tau}, \textit{T.rub/T.nig}, and \textit{D.meli/A.gam}).}
\end{figure}
Fig. 2. (A) Frequency of tyrosine, phenylalanine, and tryptophan in human tyrosine-phosphorylated (pTyr) proteins, human non-tyrosine-phosphorylated (non-pTyr) proteins, and their protein orthologs in budding yeast (S. cerevisiae) and fission yeast (S. pombe).

Tyrosine, but not phenylalanine and tryptophan, is preferentially depleted in non-pTyr proteins over pTyr proteins. Classification of pTyr and non-pTyr human proteins is based on empirical data [as previously described (3)], which are available in (15). Human protein orthologs in budding yeast and fission yeast are inferred using Inparanoid as described in (3), and only proteins with an inferred one-to-one orthologous relationship are used. Phenylalanine and tyrosine are structurally identical except for the hydroxyl group on tyrosine that renders the residue phosphorylatable. Tryptophan is also included for comparison because it is physicochemically similar to phenylalanine and tyrosine. Statistical significance indicated is for differences in the distribution between pTyr and Non-pTyr protein sets for observed differences in amino acid frequency of human-yeast orthologous protein pairs, computed with Mann-Whitney test (one-tailed) as previously described (3) using R (16).

(B) Observed substitution of phenylalanine and tryptophan by tyrosine between human and yeasts. Pairwise sequence alignments between orthologous protein pairs are performed using MAFFT with default parameters (17). We then computed the frequency of mutated phenylalanine and tryptophan being substituted with tyrosine from yeast to human and vice versa. To reduce error due to faulty alignments, residues with fewer than five identical aligned flanking residues out of 10 positions (five on each side) are excluded. We observed that the substitution rate of phenylalanine by tyrosine from yeast to human for pTyr proteins is similar to the rate for human to yeast. However, the rate is significantly lower for non-pTyr from yeast to human. Statistical significance of observed differences is computed with Fisher’s exact test (one-tailed) using R.

(C) Lower phosphorylation propensity of tyrosines in human compared to S. cerevisiae. Tyrosines conserved between human and S. cerevisiae, as identified by pairwise sequence alignments, are tested for their phosphorylation propensity by human tyrosine kinases using the NetPhorest algorithm (10). Proteins with experimentally observed phospho-tyrosines (pTyr) are tested separately from proteins that have no experimentally observed phospho-tyrosines (non-pTyr). The median probabilities are labeled for each data set, and the indicated P values were calculated using the Wilcoxon test.
tyrosine-removing mutations, could also constrain new tyrosine appearance to favor freq_{tyr} change unidirectionally. We found it interesting, based on Su et al.’s result, that freq_{tyr} negatively correlates better with GC content in flanking non-coding regions than with GC4 content and wonder whether CpG dinucleotide site and transcription initiation are attributing factors. If increased GC content does associate with increased gene and protein expression (e.g., abundance and multi-tissue expression) reported in studies (11–13), we speculate that it helps to remove more tyrosines to counter increased encounters of cellular proteins with tyrosine kinases.

References and Notes
5. Z. Su, W. Huang, X. Gu, Science 332, 917 (2011); www.sciencemag.org/cgi/content/full/332/6032/917-a.
6. To address Su et al.’s criticism directly, we applied the gradual Brownian model as described in the supporting online material (SOM) of Su et al., using the species tree inferred by them. The contrast program in PHYLIP software suite computed \( R = -0.66 \) and \( R = -0.83 \) for the correlation of \( \text{freq}_{tyr} \) with \( \text{GC4} \) and \( \text{Num}_{\text{tyr Kin}} \) respectively. Even without this sophisticated analysis, we already observed that \( \text{freq}_{tyr} \) shows a stronger covariance with \( \text{Num}_{\text{tyr Kin}} \) than with \( \text{GC4} \) among pairs of closely related species in a phylogenetic-aware manner (see Fig. 1B). The computed contrast values are in the SOM for this paper.

Acknowledgments: We thank B. Liu for reference assistance, reviewers for their critical comments, and all our colleagues for their feedback on the initial finding that has greatly improved the present work. R.L. and this work are supported by a Human Frontier Science Program (HFSP) Career Development Award.

Supporting Online Material
www.sciencemag.org/cgi/content/full/332/6032/917-b/DC1
SOM Text
References
19 March 2010; accepted 26 April 2011
10.1126/science.1188535

Table 1. Statistical linear models for correlation between tyrosine frequency \( (\text{freq}_{\text{tyr}}) \) and one or more of the variables \( \text{GC4} \) content \( (\text{GC4}) \) and tyrosine kinase number \( (\text{Num}_{\text{tyr Kin}}) \). Results are produced with R (16). AIC, Akaike information criterion (lower values indicate a better model). \( \Pr \), probability. The last column is a visual indicator of the significance of the estimated parameters: ***) \( P < 0.001; **P < 0.01; *P < 0.05 \).

| Statistical linear model | AIC  | \( R^2 \) | \( R^2 \) adjusted | Variables       | Estimate | Std. Error | t   | \( \Pr (>|t|) \) | P   |
|--------------------------|------|----------|---------------------|-----------------|----------|------------|-----|----------------|-----|
| \( \text{freq}_{\text{tyr}} \sim \text{GC4} \) | 0.80 | 0.363    | 0.32                | (Intercept)     | 3.63     | 0.265      | 13.71| \( 6.83 \times 10^{-10} \) | *** |
|                           |      |          |                     | \( \text{GC4} \) | -0.0147  | 0.00505    | -2.92| \( 1.05 \times 10^{-2} \) | *   |
|                           |      |          |                     | (Intercept)     | 3.37     | 0.0840     | 40.11| <2 \times 10^{-16}  | *** |
|                           |      |          |                     | \( \text{Num}_{\text{tyr Kin}} \) | -0.0075  | 0.00105    | -6.45| \( 1.09 \times 10^{-5} \) | *** |
| \( \text{freq}_{\text{tyr}} \sim \text{GC4} + \text{Num}_{\text{tyr Kin}} \) | -14.13| 0.735    | 0.72                | (Intercept)     | 3.71     | 0.143      | 26.05| \( 2.92 \times 10^{-13} \) | *** |
|                           |      |          |                     | \( \text{GC4} \) | -0.00808 | 0.00291    | -2.78| \( 1.49 \times 10^{-2} \) | *   |
|                           |      |          |                     | \( \text{Num}_{\text{tyr Kin}} \) | -0.00579  | 0.00094    | -6.18| \( 2.39 \times 10^{-5} \) | *** |