Comment on "Positive Selection of Tyrosine Loss in Metazoan Evolution"

Zhixi Su,[1] Wei Huang,[1] Xun Gu[1,2]*

Tan *et al.* (Reports, 25 September 2009, p. 1686) argued that loss of tyrosine residues from proteins in metazoans was driven by positive selection to remove potentially deleterious phosphorylation sites. We challenge this hypothesis, providing evidence that the high guanine-cytosine (GC) content of metazoan genomes was the primary driver in the loss of tyrosine residues.

Tan *et al.* (*1*) reported that the genome-wide frequency of the amino acid tyrosine (Y) is inversely associated with the number of cell types and the number of tyrosine kinases in budding yeast and 15 metazoan model organisms. To explain this observation, they argued that the evolutionary process of tyrosine loss must have been driven by positive selection that removed deleterious phosphorylation sites, an adaptive mechanism to allow an increase in the number of tyrosine kinases, which in turn facilitated an increase in the number of cell types in metazoans. We present strong evidence that the increased GC content in coding and flanking regions caused by directional mutational pressure or natural selection (*2*–*5*), as well as GC isochors in warm-blooded animals (*6*), is the main driver for the reduction in tyrosine content over metazoan evolution. This is simply because tyrosine is encoded by two AU(T)-rich codons (UAU and UAC) that are underrepresented in genomes having high GC content. Hence, a more plausible evolutionary scenario is that in metazoans 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.

As recognized by Tan *et al.* (*1*) and shown here in Fig. 1A, the inverse relationship between the number of cell types and tyrosine frequency collapses in the choanoflagellate *Monosiga brevicollis*, a unicellular species (with five to seven cell subtypes) (*7*) close to the metazoan lineage that contains a large number of tyrosine kinases (*8*). We note that at 54%, the GC content of the choanoflagellate genome (*9*) is much higher than that of the budding yeast (38%) and most metazoans (35% to 47%). Therefore, the low frequency of tyrosines in the choanoflagellate could be simply a consequence of its high GC content. To determine whether this is a general pattern, we conducted the following analysis on the genomes of budding yeast, metazoans, and choanoflagellate (*9*). We controlled for the GC isochores that exist in warm-blooded animals (*6*) by calculating the GC content in each species based only on the genomic regions that flank protein-coding genes (see the legend of Fig. 1B). It has been shown that the GC content of coding and contiguous noncoding sequences are highly correlated in vertebrate genomes (*10*). Impressively, Fig. 1B shows a high negative correlation between GC content and tyrosine frequency (Spearman's $R = -0.85$, $P < 1 \times 10^{-4}$; Pearson's $R = -0.81$, $P < 1 \times 10^{-4}$). We therefore conclude that GC content can explain most of the variation in tyrosine frequency among the genomes of budding yeast, metazoans, and choanoflagellate (see table S1).

The above regression analysis, which is also presented by Tan *et al.* (*1*), fails to correct for the topology of the phylogenetic tree that may inflate the significance level (*11*). Even after this potential bias is corrected (*11*, *12*), the negative re-

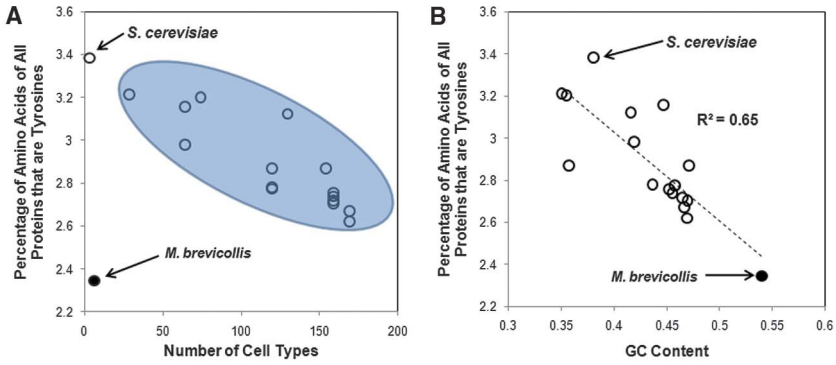

Fig. 1. Genomic GC-content bias and evolution of tyrosine kinase–related cell signaling. (**A**) Relationship between the frequency of tyrosine (Y) and the number of cell types in budding yeast (*Saccharomyces cerevisiae*) and 15 metazoans (*1*), plus a choanoflagellate (*M. brevicollis*). The number of cell types in each species was obtained from the literature (*1*, *7*). Tyrosine frequencies were calculated on all protein-coding genes; only the longest protein isoform was used. (**B**) Relationship between the frequency of tyrosine and the GC content. For each species, the GC content was calculated from upstream 2 kb and downstream 2 kb noncoding sequences surrounding all protein-coding genes. Different technical treatments, for example, analysis-based different cutoffs (such as 5 kb) in defining surrounding noncoding regions, or based on the GC content at the four-fold degenerate sites of genes (GC4) (Pearson's $R = -0.59$), GC content at third codon positions of codons (GC3) ($R = -0.62$), or GC content of all coding sequences ($R = -0.74$) (see SOM Text and table S1 for details).

Table 1. Correlations of amino acid frequency with GC content.

Amino acid	Genetic codons	Spearman's *R*†
	Encoded by GC-rich codons	
Proline (P)	CCU,CCC,CCA,CCG	0.78***
Alanine (A)	GCU,GCC,GCA,GCG	0.69**
Glycine (G)	GGU,GGC,GGA,GGG	0.73***
Tryptophan (W)	UGG	0.60*
	Encoded by AU-rich codons	
Phenylalanine (F)	UUU,UUC	−0.70**
Tyrosine (Y)	UAU,UAC	−0.85***
Asparagine (N)	AAU,AAC	−0.82***
Lysine (K)	AAA,AAG	−0.52*
Isoleucine (I)	AUU,AUC,AUA	−0.83***
Methionine (M)	AUG	−0.69**

†Conventional regression analysis: *$P < 0.05$ **$P < 0.01$ ***$P < 0.001$. After correcting the effect of phylogenetic topology (see SOM), all remain statistically significant ($P < 0.05$) except for lysine and tryptophan, although the trend is the same, probably due to the small sample size used in this study.

[1]MOE Key Laboratory of Contemporary Anthropology and Center for Evolutionary Biology, School of Life Sciences, Institutes of Biomedical Sciences, Fudan University, Shanghai 200433, China. [2]Department of Genetics, Developmental and Cell Biology, Iowa State University, Ames, IA 50011, USA.

*To whom correspondence should be addressed. E-mail: xgu@iastate.edu

lationship between the GC content and tyrosine frequency remains significant [$P < 1 \times 10^{-3}$; see Supporting Online Material (SOM) text]. Moreover, the GC-content hypothesis predicts a similar trend in other amino acids encoded by AU(T)-rich codons, such as phenylalanine (F), asparagine (N), lysine (K), isoleucine (I), and methionine (M), and an opposite trend for those amino acids encoded by GC-rich codons, such as proline (P), alanine (A), glycine (G), and tryptophan (W). Table 1 shows these two patterns as predicted, which is consistent with what is observed in bacteria (*12*). For most amino acids encoded by GC-intermediate codons (*12*), the effects of GC content are weak. These pervasive correlations suggest that it is not necessary to assume a unique adaptive mechanism to explain the variation in tyrosine frequencies among these organisms.

Our analysis provides insight into the evolution of tyrosine kinases in metazoans and choanoflagellates. Subsequent to the time these lineages split more than ~1 billion years ago, the GC content increased independently in both lineages. These increases in GC content would be expected to have resulted in losses of tyrosine sites. As suggested by Tan *et al*. (*1*), removing deleterious tyrosine phosphorylation sites may have facilitated the expansion of tyrosine kinases by gene duplications, which occurred independently in metazoans and choanoflagellates. In any case, the inverse relationship between the number of tyrosine kinases and tyrosine frequency holds in both lineages (*1*). Probably driven by natural selection, the metazoan ancestor used the increasing numbers of tyrosine kinases to enhance the evolutionary capability toward the stage of multicellularity. In choanoflagellates, the exact role of these species-specific tyrosine kinases during the process of adaptation at the organismal level remains unknown.

In conclusion, Tan *et al*. (*1*) suggested that expansion of tyrosine kinases drove the conversion of tyrosine to other amino acids. The difficulty with this argument, however, is that the frequencies of amino acids unrelated to phosphorylation are also highly correlated with the number of protein kinases, or cell types. In contrast, our hypothesis easily explains this observation. The observed changes (reductions and increases) in amino acid frequencies are simply secondary correlations due to changes in GC content (Table 1). Although tyrosine loss was clearly a by-product of GC content variation, it remains unknown exactly how it participated in the emergence in metazoans of potentially adaptive tyrosine signaling networks (*13*).

References and Notes

1. C. S. H. Tan *et al*., *Science* **325**, 1686 (2009).
2. N. Sueoka, *Proc. Natl. Acad. Sci. U.S.A.* **85**, 2653 (1988).
3. G. Bernardi, G. Bernardi, *J. Mol. Evol.* **24**, 1 (1986).
4. M. Lynch, *Proc. Natl. Acad. Sci. U.S.A.* **104** (suppl. 1), 8597 (2007).
5. L. Duret, N. Galtier, *Annu. Rev. Genomics Hum. Genet.* **10**, 285 (2009).
6. G. Bernardi *et al*., *Science* **228**, 953 (1985).
7. S. B. Hedges, J. E. Blair, M. L. Venturi, J. L. Shoe, *BMC Evol. Biol.* **4**, 2 (2004).
8. G. Manning, S. L. Young, W. T. Miller, Y. Zhai, *Proc. Natl. Acad. Sci. U.S.A.* **105**, 9674 (2008).
9. N. King *et al*., *Nature* **451**, 783 (2008).
10. M. Costantini, G. Bernardi, *Gene* **410**, 241 (2008).
11. J. Felsenstein, *Am. Nat.* **125**, 1 (1985).
12. X. Gu, D. Hewett-Emmett, W. H. Li, *Genetica* **102-103**, 383 (1998).
13. X. Gu, *J. Exp. Zool. Mol. Dev. Evol.* **312B**, 75 (2009).

Acknowledgments: This work was supported by the Shanghai Rising-Star Program (09QA1400200) to Z.S. We are grateful to P. Schnable and two anonymous referees for their insightful comments and suggestions that have improved the manuscript substantially. We also thank Tan *et al*. for pointing out erroneously computed GC4 correlations with tyrosine frequency and other discrepancies in earlier versions of this manuscript, which have now been corrected.

Supporting Online Material

www.sciencemag.org/cgi/content/full/332/6032/917-a/DC1
SOM Text
Figs. S1 to S3
Table S1
References

21 January 2010; accepted 11 April 2011
10.1126/science.1187374