

**d_N/d_S does not show positive selection drives separation of polar-tropical SAR11 populations**

Large-scale sequencing of environmental microorganisms (including genomes, metagenomes, and metatranscriptomes) now allows researchers to apply sequence-based evolutionary genetics approaches to microbial ecology questions, including studying how microbes adapt to a diversity of natural environments. There is, however, a danger for misapplication and misinterpretation if the caveats and limitations of these evolutionary methods are not fully appreciated. Here, we highlight an example of such problems in an evolutionary analysis of potential positive selection within strains of SAR11 marine bacteria (Brown et al, 2012). We feel that these issues deserve wider recognition, particularly as evolutionary approaches become more widespread among environmental microbiologists.

In the vast majority of genes, the ratio ($\omega$) of non-synonymous substitutions per non-synonymous site ($d_N$) to synonymous substitutions per synonymous site ($d_S$) is $< 1$, which is evidence of purifying selection, acting to remove deleterious mutations (most of which are non-synonymous). If $\omega$ is $> 1$, then this may be indicative of positive selection favoring amino-acid replacements (Hughes and Nei, 1988, Hughes, 1999). If $\omega$ is not statistically different from 1, then there is evidence for neither purifying nor positive selection. Positive selection is not the only factor that can cause $\omega$ to be $> 1$; since $\omega$ is a ratio, an unusually high $\omega$ value may indicate either an unusually high $d_N$ or an unusually low $d_S$. For example, purifying selection on synonymous mutations can lead to a low $d_S$ and thus high $\omega$ (Pamrley and Hurst, 2007). Approaches for calculating $d_S$ and $d_N$ can be classified into two categories, based on whether the underlying nucleotide evolutionary model takes into account the base/codon frequency bias (Supplementary Table S1; Zhang et al, 2006). For instance, the Yang and Nielsen method (YN00) considers the base/codon frequency bias (Yang and Nielsen, 2000), whereas the Nei and Gojobori method (NG86) does not (Nei and Gojobori, 1986). Taking base/codon frequency into account can be important when nucleotide content biases are sufficiently strong to reduce the observed number of synonymous substitutions.

In a recent publication, Brown et al (2012) calculated $\omega$ using the NG86 method over full gene alignments as well as over 60-codon sliding windows for orthologs shared by two SAR11 genomes, IMCC9063 from polar waters and HIMB114 from tropical waters. They identified 116 genes that had at least one window showing $\omega$ $> 1$, and claimed that these genes are under positive selection (hereafter, 116 PS genes). They conducted a similar analysis with two other polar-tropical pairs of genomes, ACE_P3.2 (polar) versus HIMB114 (96 PS genes), and ACE_P1a.1 (polar) versus HTCC7211 (tropical) (50 PS genes). However, there are problems with the interpretation of $\omega$ in these analyses. Here, we present in-depth analyses of $\omega$ calculation for orthologous genes shared by IMCC9063 and HIMB114. We did not include the other two genome pairs of Brown et al (2012) because the ACE_P3.2 and ACE_P1a.1 sequences were derived from metagenomes, which could be subject to assembly errors. However, similar average nucleotide identity (ANI) between all pairs (74%, 70%, and 76%) suggests that conclusions from the IMCC9063/HIMB114 analysis will also be relevant to these pairs.

The 1083 orthologous genes shared by IMCC9063 and HIMB114 (Supplementary Table S2) were aligned using the same approach as in Brown et al (2012). The median of G + C content at the third codon position (%GC3) is 0.20 among all orthologous genes and 0.19 among the 116 PS genes, suggesting a strong bias in base/codon usage. Using PAML software (Yang, 1997), the NG86 method (which was used by Brown et al and does not consider base/codon frequency bias) and YN00 method (which considers base/codon frequency bias) gave similar estimates of $d_S$ (median = 0.29 and 0.27, respectively), but very different estimates of $d_N$ (median = 1.47 and 3.64, respectively) (Supplementary Table S3). We hypothesized that the failure to account for base/codon frequency bias in genomes with evidence for strong bias causes the NG86 approach to underestimate $d_S$ in these data sets.

To test this hypothesis, we employed several other methods in common use for calculating $\omega$ (implemented in the KaKs_Calculator software of Zhang et al, 2006). Methods that assume equal base frequency (hereafter, ‘equal frequency methods’; these include NG86, LWL85, LPB93, MLWL04, and MLPB04) consistently produced greater estimates of $\omega$ and lower estimates of $d_S$ than those which take into account the base/codon frequency bias (hereafter, ‘frequency bias methods’; these include YN00, GY94, MYN06, MS06, and MA06; these 10 methods are all referred in Zhang et al, 2006), but comparable results for $d_N$. For simplicity, only comparison of MS06 with NG86 was shown (Figure 1), but the results of other frequency bias methods were similar. The best-fit model for $\omega$ calculation among the 14 candidate DNA sequence evolutionary models (Supplementary Table S1; Zhang et al, 2006) was always one of the latter models (Supplementary Table S4) using corrected Akaike information criterion (AICc).

Thus, equal frequency methods do not provide reliable estimates of $d_S$ and $\omega$ for base/codon frequency biased genomes, such as IMCC9063 and HIMB114. Further, because
synonymous sites are saturated with substitutions when $d_S > 1.0$ (Ochman et al., 1999), the finding that the $d_S$ of most orthologous genes was $\geq 3$ using the frequency bias methods (Figure 1B) suggests that even these methods cannot accurately estimate $\omega$ at the level of genetic divergence of these genome pairs.

Because stochastic errors can be substantial in a short sequence, the calculation of $\omega$ over 60-codon sliding windows by Brown et al. (2012) was particularly problematic. Using this approach, $d_S$ is significantly correlated with %GC3 in 76% of all genes and 70% of the 116 PS genes (Supplementary Table S3). Among the significantly correlated genes, 92% are positively correlated (Supplementary Table S3). We identified 107 genes, which contain at least one window of 60 codons showing $\omega > 1$ using the NG86 method employed by Brown et al. (2012), including 92 that were found in the 116 PS genes. For all 2951 windows with $\omega > 1$ (Supplementary Table S5), the Spearman rank correlation between $d_S$ and %GC3 was 0.506 ($P<0.001$). The phenomenon of the correlation between $d_S$ and %GC3 suggests that the small sliding window analysis is not appropriate; G + C content is a significant determinant of the calculated $d_S$ and thus affects $\omega$ for these windows. This is true whatever factors (including stochastic error, mutational bias, or natural selection) cause variation in G + C content.

A final concern in the approach of Brown et al. (2012) is that the statistical significance of cases in which $\omega > 1$ was not determined. A Fisher’s exact test (implemented in the KaKs_Calculator) showed that only one orthologous pair (HIMB114_0556 and SAR11G3_00373) contained at least one 60-codon window showing $\omega > 1$ with statistical significance (Supplementary Table S6). Here, $\omega$ was estimated by NG86 using KaKs_Calculator, which is slightly different from that estimated by NG86 using PAML. This ortholog pair contains five such windows, all of which have extremely low %GC3, ranging from 0.08 to 0.11, and low $d_S$, compared with other windows in the same gene, yet $d_S$ is only slightly greater than the median, suggesting that the low G + C content reduced the number of observed synonymous substitutions in these windows, resulting in $\omega > 1$. In addition, this orthologous gene encodes a putative porphobilinogen deaminase; there is no evidence showing that it is involved in cold adaptation.

In conclusion, the finding of Brown et al. (2012) that positive selection drives biogeographic separation of polar and tropical phytopotypes of SAR11 populations is based on an evolutionary model that does not account for the low G + C/high codon bias of the genomes being analyzed. Our analysis of evolutionary statistics using optimized evolutionary models does not support the contention that 116 genes shared by polar and tropical SAR11 populations have diverged as a result of positive selection.

Supplementary information
Supplementary information is available at the Molecular Systems Biology website (www.nature.com/msb).

Acknowledgements
We thank MA Moran for discussion, K Konstantinidis for providing the script for computing ANI, Z Zhang for advice on using KaKs_Calculator, and C English with graphics. This research was funded by grants from the Gordon and Betty Moore Foundation and the National Science Foundation (MCB-0702125).

Haiwei Luo1 and Austin L Hughes2
1Department of Marine Sciences, University of Georgia, Athens, GA, USA and
2Department of Biological Sciences, University of South Carolina, Columbia, SC, USA

Conflict of interest
The authors declare that they have no conflict of interest.

References
Brown MV, Lauro FM, DeMaere MZ, Muir L, Wilkins D, Thomas T, Riddle MJ, Fuhrman JA, Andrews-Pfannkoch C, Hoffman JM, McQuaid JB, Allen A, Rintoul SR, Cavicchioli R (2012) Global biogeography of SAR11 marine bacteria. Mol Syst Biol 8: 595
Hughes AL, Nei M (1988) Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. Nature 335: 167–170
Hughes AL (1999) Adaptive Evolution of Genes and Genomes Oxford University Press, New York, USA
Nei M, Gojobori T (1986) Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol 3: 418–426
Ochman H, Elwyn S, Moran NA (1999) Calibrating bacterial evolution. Proc Natl Acad Sci USA 96: 12638–12643
Parmley J, Hurst L (2007) How common are introgene windows with Ka > Ks owing to purifying selection on synonymous mutations? J Mol Evol 64: 646–655

Yang Z (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. Bioinformatics 13: 555–556

Yang Z, Nielsen R (2000) Estimating synonymous and non-synonymous substitution rates under realistic evolutionary models. Mol Biol Evol 17: 32–43

Zhang Z, Li J, Zhao XQ, Wang J, Wong GK, Yu J (2006) Ka/Ks_Calculator: calculating Ka and Ks through model selection and model averaging. Genomics, Proteomics Bioinformatics 4: 259–263