Reviewer comments are in black, and response to reviewers is in blue.

Comments from the Editor

I am pleased to tell you that it is potentially suitable for publication in GENETICS; however, both reviewers have substantial comments and concerns that need to be addressed in a revised manuscript…. It is of particular importance that you fully address concerns about the interpretation of your results and, as noted by both reviewers, to conjecture on why your results are as they are.

We thank the editor for these comments. In our revised manuscript, we have substantially expanded our investigation into the source of performance differences between one-rate and two-rate dN/dS inference approaches. We additionally include an expanded discussion of important caveats which bear on the proper interpretation of our findings.

Comments from Reviewer 1

In this manuscript, the authors assess the accuracy of maximum-likelihood (FEL), Bayesian (FUBAR) and counting-based (SLAC) methods to estimate per-site dN/dS values in the absence of positive selection. Accurate sitewise dN/dS estimates are important for assessing questions about the correlation of dN/dS with protein structural properties, such as solvent accessibility and contact number.

The authors address this question by simulating data under MutSel model, which combines a nucleotide substitution model with fitness values at each codon to produce a codon substitution model. The authors simulate alignments in which each codon site has a different set of codon fitnesses. They simulate under two conditions: one where all codons for an amino acid have equal frequency (“without codon bias”) and one where one codon has elevated frequency, and the remainder have identical frequency (“with codon bias”). They derive the “true” dN/dS value at each site by analytically taking long-term averages of “dN=Kn/Ln” and “dS=Ks/Ls”, as described in (their previous paper Spielman and Wilke 2015b).

The authors conclude that estimating both a synonymous rate (dS) and a non-synonymous rate (dN) for each site leads to worse estimates than estimating a single (dN/dS) rate. Simulating data with codon bias does not lead to an increase in the performance of the approaches that estimate dS separately. The suggested explanation is that attempting to estimate an extra parameter (dS) for each site leads to noisy dS estimates, so that the ratio of dN/dS is estimated badly. The authors also find that for a given tree length (Number of Branches * Branch Length), one should prefer longer branches as opposed to more sequences, in order to obtain accurate per-site estimates of dN/dS.

The paper has a couple of major problems that should be addressed before publication:

1. The authors write (pg 4) “In the presence of codon bias, both dN and dS varied at each site. As a consequence, there are two approaches for calculating the true site-wise dN/dS ratio: One can either calculate the ratio of each site's dN and dS values [True2], or one can take each site's dN value and divide by the average dS over the entire sequence [True1].”

This isn't correct. While the second quantity [True1] can be computed, it isn't a site-wise dN/dS ratio. It is instead a sitewise dN value, divided by an across-site dS average. In contrast, the first quantity is at least the ratio of the dN and dS values for a particular model employed at the given site.

This is important, because it means that figure 1C is misleading. Whereas in Figure 1A and 1B, the “2-rate” methods appear to approach the accuracy of the 1-rate methods as the number of sequences and branch length increase, in Figure 1C the accuracy of 2-rate methods remains substantially lower.

However, this is probably because in Figure 1C, the accuracy of the methods is determined by comparison to “True1”, which is not in fact a true sitewise property.
We agree that the True1 quantity is not a true site-wise property, as the reviewer recognizes. We had originally calculated the so-called “True1” and “True2” rates in order to have baseline comparison values whose properties reflected one-rate vs. two-rate inference models. We realize, however, that this consideration may lead to some confusion and ultimately may be misleading to readers. As such, we now focus exclusively on True2 (which is now simply referred to as “true”), and we have removed True1 comparisons from the manuscript.

2. The only measure of accuracy of the methods (e.g. Figure 1) is correlation. By showing only measures of correlation, the authors neglect the possibility of estimates that are inaccurate because of bias and scaling. Can the authors please include (perhaps in the supplement) an modified version of Figure 1 that assesses accuracy in terms of the bias and root-mean-squared error? This would offer some assurance that the estimate with higher correlations are indeed more accurate in the sense of having a smaller mean-squared error, or something?
We thank the reviewer for this excellent suggestion. We now additionally include RMSD measures (as well as measures for the variance in residuals, as per reviewer 2’s request) to compare true and inferred dN/dS ratios, both in the main text and supplementary materials. Note that a supplementary figure for estimator bias was already in the original manuscript, and it remains in this revised version.

3. The authors description of models without codon bias is incorrect (pg 2, materials and methods). The authors seem to assume that codon bias occurs when different codons for the same amino acid have a different equilibrium frequency. In fact, codon bias under the MutSel model occurs when different codons have different *fitnesses*. See for example the fMutSel0 model in Yang and Nielsen (2008). The authors decision to force all codons to have the same equilibrium frequency when there is no codon bias actually creates codon bias, unless the frequencies of all nucleotide frequencies are 0.25.
We respectfully disagree with the reviewer’s statement about MutSel models. In our original simulations, nucleotide frequencies were indeed all set to 0.25, and hence different equilibrium frequencies directly imply different fitness values and vice versa. As described in-depth in Sella and Hirsh 2005 (http://www.pnas.org/content/102/27/9541), frequencies are mathematically equivalent to fitness values when mutation rates are symmetric, which they indeed were under our HKY mutation model. In this circumstance, the frequencies of two codons will be the same if and only if the fitnesses of the two codons are the same. Therefore, the implementation of codon bias was indeed mathematically consistent and correct. Note that, as described in the next response, we have now included additional simulations with unequal nucleotide frequencies. Here, the fitness values from the original HKY simulations were used, and frequencies (for dN/dS calculation) were computed numerically directly from the nucleotide frequencies and codon fitness values, as described in the manuscript.

4. Inspection of the author's simulation script 'simulate_alignments.py' suggests that the author's are implicitly assuming that the HKY mutation process does, in fact, have equal nucleotide frequencies. Please explicitly mention this and make an argument for using equal nucleotide frequencies.
In retrospect, we feel (perhaps as the reviewer does as well) that employing an HKY model with equal frequencies was maybe overly simplistic. As such, we have performed additional simulations using an HKY model with unequal nucleotide frequencies. The fitnesses for these simulations were the same as for the HKY simulations with equal nucleotide frequencies. Both sets of simulations displayed the same general trends regarding one-rate vs. two-rate inference performance. We now primarily discuss simulations with unequal nucleotide frequencies in the main text, and we have placed results for the original simulations with equal frequencies in the supplementary materials.