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.

5. The authors should describe their simulation methodology better.
The authors should describe how the mutation model and the codon fitnesses are determined for each column. The fact that the authors are using a dubious approach to codon bias is partially hidden because they do not supply the MutSel rate matrix and describe how the parameters are set. For example, it is currently unclear what the nucleotide frequencies are for the HKY model that appears to be used to model the neutral nucleotide mutation process. Since these nucleotide frequencies affect the rate matrix, please specify how they are set, and how they are related to codon frequencies and to codon fitnesses.
We thank the reviewer for identifying an area of our manuscript needing clarification. We have substantially expanded our explanation of simulation strategies, especially given the inclusion of the new simulations. We also include a precise explanation for how codon frequencies are derived from the combination of codon fitness and nucleotide frequency values.

6. The authors do not make a case in the introduction that the MutSel model is the right model to use when assessing the benefits of models that allow site-to-site variation in synonyous rate. They simply write in the discussion “We did not, however, examine how one- and two-rate inference models compare when dS variation is driven by mutational rather than selective processes.” However, if real data has variation in synonymous rate across sites, then ignoring this process when simulating would be misleading.
We now include a justification in the Introduction for our use of MutSel models in the context of this study.

Specifically, our use of the MutSel models is a novel strategy for asking how dN/dS models perform when they are misspecified, that is when the simulation model does not match the inference model. Given that any real-world analysis will naturally be misspecified (real sequences do not evolve according to a dN/dS model), our MutSel simulation approach allows for unique insight into how dN/dS models behave on data generated according to a different mechanism. In this regard, at this time, the MutSel is in fact the only possible non-dN/dS model one could use for simulation and still have a true dN/dS to which inferences may be compared. Such a scenario might be possible with other non-MutSel models, although currently there is no other framework for which a dN/dS has been defined. Therefore, while the reviewer may not find our use of MutSel models ideal, there is no alternative, and we feel that this study provides a valuable contribution to understanding how dN/dS-inference models behave on data simulated under a distinct mechanism. We additionally note that at least three other research groups have used MutSel simulation to benchmark either phylogenetic or evolutionary rate inference, and hence there is an existing (albeit small, at this stage) precedent for our simulation strategy.

7. The title implies that the authors are comparing models, when they actually seem to be comparing inference methods. It may be true, for example, that for data simulated under the MutSel model, that the 2-rate models really do not capture the variation in synonymous rates well. The authors make the point in (I think) their previous paper that, if only 1 codon per amino acid is allowed, then dS is always 0, and dN/dS is infinity. That is a point of concern about the 2-rate model as such. However, the authors do not focus on that kind of concern in this manuscript. Instead, this manuscript focuses only on methods that attempt to infer evolutionary parameters of a single site. In such an endeavor, it might be wise to avoid a model with more parameters even if it is more realistic. This is not a critique of the model per se, but only of inference under the model. The confusion of these two concerns (model adequacy versus estimation accuracy) made this paper hard to follow.
We fully agree with the reviewer’s comment that adding more parameters (e.g. dS) to a site-specific estimation is unwise. However, this has not been a prevailing view (nor one which has undergone rigorous

testing) in the molecular evolution community. We therefore believe that this paper makes a valuable contribution which examines and indeed demonstrates the validity of the reviewer's own argument.

However, we respectfully disagree with the reviewer's assessment that we are comparing methods rather than models, although the reviewer does highlight a place in the manuscript which could use clarification. In the manuscript, we actually compare both models and methods. Specifically, we consider the comparison of 1-rate and 2-rate parameterizations as comparing models (i.e. rate matrix), but SLAC vs. FEL vs. FUBAR as comparing methods. Indeed, these three methods all utilize the same rate matrices, but utilize different algorithms for estimating model parameters. We have now added a sentence into the *Results* subsection "Approach" to clarify precisely how we use the words "model" (rate matrix) vs. "method" (algorithmic implementation used to estimate model parameters). Furthermore, we now use the either the word "framework" or "approach" to broadly describe one-rate vs. two-rate approaches when "model" or "method" is either imprecise or overly specific.

8. Finally, if the reason that the 2-rate models do not perform as well as the 1-rate models is because of difficulty in estimating dS, then why does FUBAR2 have the same problems as FEL2? My understanding of FUBAR2 is that it is able to pool information across sites, and thus should be less noisy that FEL2 in its estimates of the dS rate.
This is an excellent question. FUBAR is indeed expected to be less noisy than FEL, although the manifestation of this property will depend largely on the properties of the dataset. If, for example, the dataset is small or divergence is low, then FUBAR will show lower levels of noise in its estimates. With that said, FUBAR will likely be a more biased estimator at these low levels of information, precisely because it pools information across sites and there may not be enough site-specific signal for unbiased inference. Conversely, FEL will be less noisy and more accurate than FUBAR under asymptotic conditions (approaching infinite time and/or number of taxa).

Note that our RMSD analysis now demonstrates that FUBAR is actually less noisy than FEL and SLAC, so the reviewer may find this result more in line with expectations for FUBAR's algorithm.

Comments from Reviewer 2 (Sergei Pond)
This is a straightforward but interesting simulation-based paper which compares three different methods for inferring site-specific dN/dS estimates. As one of the developers of all these methods, I agree that we never really (but see point 2 below) evaluated how well they might do at inferring dN/dS ratios at individual sites. In fact, we tended to argue that such point estimates are inherently unreliable for realistic-sized data, and instead emphasized site-specific _tests_ of individual hypotheses (like is dN>dS?). For other applications, like those of how to explain variation in dN/dS through mechanistic parameters (RSA, etc), the quality of estimates is more important. Spielman et al investigate the performance of SLAC, FEL, and FUBAR on _point estimates_ of site-specific dN/dS on data simulated under Mutation Selection models. I have no objections to how the study was done, and most of the conclusions.
We thank the reviewer for these kind comments.

I do however, have a serious objection to how the title of the paper and the discussion frame the finding. I have learned in my relatively few years as an active academic, that one has to be careful not to overgeneralize, and this is what I think happens here. A random reader will, for example, take the title to mean that single rate model ALWAYS outperform two-rate models, and NOT that this is CONDITIONED on simulating the data under the MutSel model, and INFERRING the parameters under a different model. This

is NOT the desired outcome, because the literature is already quite confusing for the random practitioner (take the infamous Yokoyama et al rhodopsin paper which berated the entire CLASS of methods based on ONE example). Here are some specific issues I would like the authors to address before I can recommend the paper for publication.
We fully agree, and we have expanded the Abstract, Introduction, and Discussion to highlight that our results are conditioned on a misspecified model, and we offer guidelines for how to interpret (and not interpret) our results.

0). Based on general statistical principles, asymptotic biases in MLEs will arise when the model is misspecified. Clearly, this is the case here. Variation in dS (as applied in FUBAR, SLAC, and FEL) does NOT model selection for a specific synonymous codon in the MutSel framework (codon bias). In fact, this is not the intuition I had for the dS variation in the 2005 paper at all; the idea was to simply accommodate varied level of variability across sites (i.e. lower dS for conserved sites, possibly higher dS for unconstrained sites; the primary driver was to remove false positives for dN>dS due to variable sites, where gene-wide dS is too low). This should be reflected in the title and discussion. I freely grant the authors license to do the analysis as they chose to do it, but I would posit that with a PROPER parameterization of variation in dS, SOME two-rate models WILL do better in the MutSel variation framework than one-rate models. This contradicts the title of the paper.
We broadly agree with the reviewer's suggestions. We have changed the title of the paper to a less divisive one: "A comparison of one-rate and two-rate inference frameworks for site-specific *dN/dS* estimation," to avoid overgeneralization. In addition, we explain (as noted in the previous response comment) how issues with model misspecification influence our results and their interpretation.

(a). Check to see if there is any correlation between dS inferred in the two-rate models and the \gamma parameter used by the authors to simulate codon usage bias. I would expect there to be some inverse correlation (higher \gamma, lower dS). This analysis would indicate how badly the model is misspecified. If (as I suspect) is is misspecified BADLY, then the conclusions need to be modified accordingly.
We have checked this relationship, and indeed there is a strong negative relationship. However, we contend that the extent of model misspecification is quite similar for dN and dS, not just for dS alone. The MG model will assume a single rate for all nonsynonymous and all synonymous changes. For a codon bias MutSel model, *both* of these MG rate parameters would be misspecified - all nonsynonymous changes would experience a unique rate, and similarly the fixation rate between synonymous mutations would differ. Therefore, we believe that the model misspecification is more generalized and not specific to dS.

(b). Under the data generated under the correct model (direct variation in rates), I would expect SLAC < FUBAR < FEL. This is because SLAC is simply a biased method which applies a poor man's binomial mis-approximation, FUBAR uses data-set wide smoothing (and a grid), and FEL directly estimates the rates. The fact that SLAC does better than ML-based methods, tells me that the models are misspecified quite badly.
We now include a sentence in the Discussion indicating that SLAC's performance may be a result of the model misspecification.

(c). Asymptotic simulations (2048 taxa with 0.64 branch lengths) are nice, but they can never be attained in practice. If you have >100 expected substitutions per site, the sequences will never look homologous, could not be aligned, and could not be analyzed. Please consider using real large scale trees to estimate what the methods would do. Super-saturated alignments are perfect for modeling equilibrium dynamics (because the evolutionary process does not spend any substantial part of the

tree away from this state).
We have performed additional simulations along five empirical phylogenies of varying size to complement our balanced tree analyses, and we have accordingly added a new section to the paper. We identified broadly the same results from these simulations, although for most trees, the correlations were indeed all $0.4 < r < 0.6$.

(d). I would argue that MOST dN/dS applications are in cases when there is NO evolutionary equilibrium (e.g. in the context of viral evolution, fitness values change over time). This needs to be addressed in the text.
Agreed. We now address this point in the Discussion.

(e). I would also argue that MOST dN/dS applications care about dN>dS (not modelled in the MutSel) framework.
Again, agreed. We now emphasize this point more clearly.

1). It seems that for realistic levels of divergence, ALL methods do relatively poorly (e.g. r ~ 0.5). What happens to the _variance_ in error estimates? Does this behave consistently as sample size increases?
We have examined the variance in residuals, and indeed it behaves consistently as sample size increased. We include a figure with this trend in Figure S4.

2). If you take a look at figure 1 from our "Not so different paper", you can see a very specific case when two-rate models do _qualitatively_ better than single-rate models -- simulate some data with constant (dS), throw in a proportion of neutrally evolving sites with high dS, and watch single-rate models cram this dS variation into the only thing they CAN vary (dN), and happily call them all selected. Can the authors convince me how in this context a single rate model can be better at inferring dN/dS?
The context the reviewer describes is one, in biological terms, where a specific region or regions of a gene have an elevated mutation rate. In this context, we agree that the approach described would be required. We now point out this specific scenario in the Discussion, and we suggest to readers that this approach may be preferred if they suspect, based on experimental information or biological intuition, mutational hotspots within the gene.

3). I think one other missing part of the puzzle is to explain WHY in this context single rate dN/dS models do better (as does an ugly and ad hoc SLAC approximation). My explanation is that they are more robust to the specific type of model misspecification; but I would like to be proven wrong.
We have investigated this issue more thoroughly, by examining correlation and RMSD between FEL- and FUBAR-inferred dN and dS estimates each with the true values. We find that dN is much more accurately estimated than is dS (Figures 3-4). We found that, in fact, synonymous changes occur with a much lower frequency than do nonsynonymous changes, by a ratio of ~2.5:1 (obtained by counting changes using simulated ancestral sequences). We therefore suggest that dN/dS models do much better because of the inherent relative difficulty in estimating a precise dS, whereas estimating dN is a comparatively “easy” task. Thus, we find that the issue boils down to a straight-forward small sample problem, where there is more information to estimate dN than there is to estimate dS on a site-wise basis.

4). Clearly adding codon bias kills the accuracy of ALL methods. How would the authors propose to modify a two-rate parameterization to deal with it?

We would generally recommend one-rate parameterizations over two-rate parameterizations. davIlf one wishes, however, to perform dS inference at individual sites, we would advocate the use of methods such as FRESCo (we now mention this method in the Discussion). This method fits both a one-rate and a two-rate model at each site (the original paper considered regions rather than sites, but the same principle applies), and then performs hypothesis testing to identify whether that site/region shows evidence for excess synonymous selection. When the test is significant, then one might consider the site-wise dS for further analysis.