**Response to Reviewers**

We would like to thank the reviewers Nicolas Lartillot and Sergei Kosakovsky Pond for their constructive comments. A detailed, point-by-point response to individual comments follows below.

**Associate Editor**

I also enjoyed reading the manuscript. I agree with the reviewers' comments in general. Please make an effort to address the reviewers' comments in your revised version.

We are pleased to see that you enjoyed reading the manuscript. We have made every effort to address the reviewer comments.

1. When you write (pg 4) that "with notable exception 33, I think you miss many studies. Please make an effort to better search the literature on this point.

We now write “but see” instead of “with notable exception”, and we have added the following references:

1. Schoniger M, von Haeseler A. 1995. Performance of the maximum likelihood, neighbor joining, and maximum parsimony methods when sequence sites are not independent. Syst Biol 44:533 – 547.
2. Minin V, Abdo Z, Joyce P, Sullivan J. 2003. Performance-based selection of likelihood models for phylogeny estimation. Syst Biol 52:674 – 683.
3. Yap VB, Lindsay H, Easteal S, Huttley G (2010) Estimates of the effect of natural selection on protein-coding content. Mol Biol Evol 27:726 – 734.
4. Rubinstein N, Faigenboim-Doron A, Mayrose I, Pupko T. 2011. Evolutionary models accounting for layers of selection in protein-coding genes and their impact on the inference of positive selection. Mol Biol Evol 28:3297 – 3308.

If the editor knows of other specific references, we are happy to include them.

**Reviewer 1, Nicolas Lartillot**

I really enjoyed reading this insightful and well-written manuscript. The analyses conducted in this work are timely: mutation-selection models have now been around for a while, and we still lack a clear vision of the exact relationship between them and more classical dN/dS codon models. In this respect, the present manuscript clearly represents a very useful contribution. In addition, it is so well written, with everything so clearly explained, that there is not much to add or to comment. I would have only relatively minor points to discuss, concerning minor aspects of the arguments developed in the manuscript (thus, never questioning the results themselves).

Thanks a lot for these kind comments.

1. Strictly speaking, I don’t really agree with the statement, made several times across the manuscript, that mutation-selection models cannot accomodate positive, diversifying selection…. On the computational side, site-independent fluctuations of fitness coefficients could be implemented using Markov-modulated models, very much like those explored by Simon Whelan (2008). So, perhaps the Authors could be slightly less definitive on this question.

In a similar spirit, I don’t agree with the idea that the use of mut-sel models is ‘only justified under conditions of strictly purifying selection’. In fact, even if you don’t explicitly model fluctuating selection as just suggested in point 1 above — and thus, even if you use versions of mut-sel models that are inherently unable to express diversifying selection — you can still use them as null-models, such that rejecting them in specific ways (e.g. observing too many non-syn substitutions compared to what would be expected under the null) could be interpreted as indicating the presence of positive selection.

Thanks for pointing this out. We have softened our language in this part and now discuss more precisely the utility of mut-sel models (specifically regarding the HB framework vs other frameworks which allow fluctuating landscapes) in the context of positive selection. We also have added a sentence mentioning Markov-modulated models.

1. Concerning the relationship between dN/dS and scaled selection coefficients: if I am correct, using MG-parameterization, under unbiased mutation and assuming no selection on synonymous mutations, dN/dS is just the average relative fixation probability of non-synonymous mutations (relative to neutral)… I think you can get this relation by simplifying equations 8 and 9 and then by putting them together. So, this is only a suggestion, but the Authors could point out this fact, before deriving the more complicated (and less easily interpretable) relation between dN/dS and selection coefficients in the general case.

Thanks for pointing this out. It turns out that dN/dS is indeed the average relative non-synonymous fixation probability, but under the more simplified condition of uniform mutation rates (not unbiased). We now point out this fact, although we do so after the formal dN/dS derivation.

1. I must say I do not quite understand the point of such mathematical manipulations. What is so special about the GY-style framework, or more generally about the GTR parameterization? I mean, even if MG-style models were not expressible in GTR form, what would that change?

There are multiple reasons why we found the GTR parameterization useful: First, it provides a 3-line proof that the F1x4 codon frequencies are the steady-state frequencies of the MG-style model. Without this parameterization, obtaining the steady-state frequencies would require a lengthy calculation. Second, if MG-style models were not expressible in GTR form, then they could not be described by a time-reversible Markov process. Thus, none of the useful mathematical properties that we know are true for time-reversible Markov processes, such as detailed balance, would be guaranteed for MG models. While one could still use such models and fit them to sequence data, one would have to be more careful in carrying out these fits, and hence it is important to know whether time reversibility holds or not. Third, we found the GTR parameterization helpful for practical reasons. We had previously written a simulator that could simulate sequences using a GY-style model. With the GTR parameterization of the MG model, it was trivial to convert our simulator to simulate MG-style models. Note that these simulations were **not**, however, used in the present paper.

1. Perhaps the Authors could also mention that, in the limit of no selection on non-syn substitutions, mutation-selection models tend to a MG-like model with omega=1 — and not a GY-like parameterization. Which you may see as an argument suggesting that MG parameterizations have a better mechanistic justification.

Thanks for pointing this out. We have added a brief discussion to this effect in the section “Biased dN/dS estimates under asymmetric mutation models.”

1. Concerning the discrepancy between model selection and dN/dS accuracy: perhaps it is just that AIC is not what you want to this end? If I remember correctly, AIC is known to be good primarily for prediction, not for inference. So, what about other criteria (BIC?, or others?). “

We now provide BIC values (Table S3) as well. We find that our overall conclusions remain unchanged.

While specific model-selection criteria may outperform others in certain cases, we believe that the point we are making is more general: If data were generated under mechanistic process A and the models we fit all describe variants of mechanistic process B, then by definition none of the models can truly fit the data. For this reason, which model will end up as the best fitting model is not clear *a priori*, nor how much bias this model will have in the key parameters of interest. It is rather likely, though, that the model will end up being over-parameterized, to correct for the mechanistic lack of fit.

The above reasoning suggests that more stringent model-fit criteria, i.e., criteria that more strictly penalize additional nuisance parameters, will often perform better.

1. Still concerning GY versus MG — and also concerning the fact that AIC does not select the most accurate model: Nicolas Rodrigue did some work on this question (Rodrigue et al, 2008). To be fair, this work did not really have anything to say about the accuracy of estimated dN/dS (simply because, at the time, we did not have any way to define what we would mean by ‘true’ dN/dS, which is precisely what the Authors of the present manuscript provide us with). Also, that old article explored but very few (only 3!) empirical datasets, thus allowing for only limited conclusions. But still, the results then obtained suggested that Bayes factors would in fact select the correct model — which again raises the question of the relevance of AIC in the present context.

Thanks for pointing this out. We now cite this work, and we also mention that further work along these lines is needed.

1. ‘since their introduction in the 1990s’: this sentence seems to suggest that Markov models of sequence evolution were introduced in the 1990s, which is of course wrong. I guess that what the Authors means is that the Markov \*codon\* models were introduced in the 1990s.

Yes, this is what we mean. We have made the suggested change.

1. S = 2 Ne s. Or, S = 4 Ne s in diploids ?

We now mention that the numerical factor would be 4 in the case of diploids.

1. ‘Elucidating the relationship between these competing modeling frameworks…’: these \*complementary\* modeling frameworks?  
   We have made this edit.
2. notation is a bit confusing: s\_i for the source, conflicts with s indicating the selection coefficient.

Good point, we’ve replaced *si* by *oi* and now write “origin” instead of “source”.

1. speaking of all of those reasons why dN/dS>1 may not just be due to diversifying selection, perhaps also mention GC-biased gene conversion (Ratnakumar et al, 2010).

We now mention this in the “Conclusions” section, and we have added the Ratnakumar reference.

**Reviewer 2**

I enjoyed reading the manuscript (not only because it recommends our software package and models). It is concise (a rarity), well-written, and convincing. Given the popularity and general confusion surrounding the myriads of models, tools, and interpretations of codon models, this paper is timely and should be read by anyone who develops and seriously uses such models.

Thanks a lot for these kind comments.

1. The AIC ranking is counterintuitive at first, but upon further reflection it makes sense because F61 (with its 60 parameters) will produce the most accurate estimate of codon frequencies (the ML estimates actually, under the multinomial model), which are at the driving force of "goodness-of-fit" for MutSel models (which define nearly everything in terms of frequencies). When you discuss AIC, remember that in the model with the lowest AIC minimizes the Kullbach-Leibler (KL) distance between the set of all models and the (unknown) true model; there never is an assumption that the correct model will be recovered.

Thanks for pointing this out. We have expanded our discussion on these issues in the (new) section “Model with best fit is not the model that yields the most accurate parameter estimates” in response to this comment as well as the remarks by Nicholas Lartillot regarding alternative model-fit criteria.

1. I read the caption to Table 1 as implying that k = 3 for all models. This is not correct for AIC calculation, because you need to include ALL parameters estimated from the data (not just those estimated by ML). This won't change the conclusions (Delta AICs are quite large), but it will change the numbers a little.

You are correct. We have recalculated the AIC values and updated Table 1 accordingly. As you suspected, we have indeed found no substantive change.

1. When describing MG and GY style models, you don't indicate which codon frequency estimator was used for MG models. Please elaborate. Did you estimate MG nucleotide frequency parameters using ML?

We used F1x4 frequencies for MG1 and F3x4 for MG3 as the state frequencies of the model. In both cases, we used empirical nucleotide frequencies, not ML estimates. We have updated the methods sections to include this information.

1. Because you use r^2 in Figure 1, it may be instructive to indicate linear regression lines (I presume p-values for association are highly significant in both cases).

We have added regression lines to the figure as well as p-values (which are highly significant) to the text.

1. The statement, "In addition, because MutSel models are based explicitly on population  genetics theory, these simulated alignments are likely far more similar  to real sequence data than are alignments simulated under a dN/dS-based  model, and therefore provide excellent benchmarking data.” is too strong.  MutSel models make DIFFERENT and not necessarily better assumptions, e.g. that fitness values are fixed over time and sites (definitely not the case in most proteins), etc. The point about benchmarking models with data that don't conform to underlying model assumptions is valid, however.

Thanks for pointing this out. We have toned down our wording.

1. Is the fact that you can scale the rates in MG models to get a canonical q\_ij \* pi\_j parameterization (Appendix 2) used anywhere in the text? (I didn't find anything other than a statement that you can do it).

This parameterization justifies the F1x4 frequencies for the MG1 model. We now state this explicitly. See also our response to comment 3 of Nicolas Lartillot.

1. A few pedantic points from a former mathematician In Appendix A, the F (x,y) function (15), is undefined for x = y as written. l’Hˆopital’s rule simply defines the limit of the function as |x-y| -> 0, (not the value for x = y), and the function can be EXTENDED to have the value of 0 when x = y. On the same page please use "Taylor series expansion" (not Taylor expansion).  In equation (12), you could combine the first two cases as max (1, F\_j/F\_i) \* m\_ij

We have made the requested changes.