Response to Reviewers

Comments by the Associate Editor

Both reviewers have good points, but I agree with the first reviewer that this manuscript is best published as a Letter. I will encourage the Senior Editor to invite you to submit a revised version of the manuscript in this form where you answer all the issues raised by the reviewers. The additional simulations required by the second reviewer could possibly be omitted if you instead include a discussion of the limitations of the simulation set-up. You could perhaps tone down the emphasis on the difference between the methods.

We have rewritten the manuscript according to MBE Letter requirements. Moreover, we did conduct additional simulations, which are described below, as requested by Reviewer #2. There is now relatively little emphasis placed on the difference between inference methods FUBAR and PAML. Instead, we have re-focused our manuscript to emphasize the effects of alignment filtering.

In addition, edits that we implemented in response to reviewer comments have produced some new results not present in the first manuscript submission. Due to these new results, as well as the suggestion that we re-format our manuscript as an MBE Letter, our revised manuscript has changed substantially from the original and contains some additional analyses (e.g. a more thorough treatment of alignment false positive rates). However, the overall conclusions have remained largely unchanged.

Comments by Reviewer 1

I recommend that the authors revise the manuscript so that it meets the criteria for the "Letter" category. Revision would largely involve reducing its length.

Again, as suggested, we have revised our manuscript to adhere to MBE criteria for Letter publications.

Comments by Reviewer 2

Does filtering the alignment change the inferred distribution of dN/dS to make a dramatic difference (I can buy that if a sufficiently large proportion of sites is filtered)? For example, what happens to the MLE of the omega > 1 class in M8 when comparing filtered and unfiltered data? My intuition is that this estimate will be generally LOWER for filtered alignments, allowing the detection of sites simulated with omega closer to 1 more reliably. The same can be extended to FUBAR, by examining how much weight is assigned to each dN/dS value. Another explanation for the apparent boost in power is that the mapping between an aligned site and the corresponding simulated site is influenced by the filtering.

We have examined the different inferred dN/dS distributions for both FUBAR and PAML between filtered and unfiltered alignments. We were unable to recover any clear trend relating either the PAML dN/dS >1 MLE or the cumulative FUBAR grid weights for dN/dS > 1 to circumstances in which power increased.

Additionally, the same map relating sites from an inferred alignment to the true alignment was used for a given unfiltered alignment and all of its filtered derivatives, so this is also not an explanation for any observed increases in gpower observed. We now clarify this point in our Supplementary Methods section.

Thus, if FPR is a quantity of interest, it is essential to include a non-trivial proportion of sites simulated at or near neutrality, otherwise a false positive would only occur if a site simulated under reasonably strong constraint (if the log-normal is peaked), is misclassified as POSITIVELY selected… It would be instructive to consider a different selective profile, something like an antiviral factor (APOBEC3G or TRIM5alpha, see the corresponding papers from Harmit Malik's group), or the well studies sperm lysin (or a self-incompatibility locus). The distribution of dN/dS would quite different (mean closer to one), and a more of a challenge to classify.

We largely agree with these points regarding the potential issues with our simulation setup of choice. Therefore, we have conducted additional simulations according to the dN/dS selective profile of HIV-1 env subunit GP41, as recommended by Reviewer 2. This new selective distribution does feature a mean dN/dS closer to 1 (in fact, it is 0.89), which we discuss in the revised manuscript:

> We simulated protein-coding sequences using Indelible (Fletcher and Yang 2009) according to two selective profiles: H1N1 influenza hemagluttinin (HA), which featured a mean dN/dS = 0.37, and HIV-1 envelope protein subunit GP41, which featured a mean dN/dS = 0.89. We used these selective profiles because, while both genes contain positively selected regions (Bush et al. 1999; Frost et al. 2001; Bandawe et al. 2008; Meyer and Wilke 2012), most sites in HA are either under strong purifying or positive selection, whereas relatively more sites in GP41 have dN/dS values near 1, making positive-selection inference more challenging.

Since INdelible does not support dS variation, it is probably easiest to "balance" the comparison, by running a version of FUBAR which does assumes a constant dS but puts a more dense grid on dN (at the same computational cost).

We have re-calculated all FUBAR results with a 1D grid that fixes dS at 1. Moreover, as recommended by Reviewer 2, we specified 100 dN grid points to account for the reduced dimensionality resulting from removed dS variation. We explain this methodological approach in Supplementary Materials:

> For FUBAR inference, we used mostly default parameters, except when specifying grid dimensionality. As neither Indelible nor PAML simulates sequences without dS variation, we specified that FUBAR only consider dN variation, in order to make results from FUBAR and PAML fully comparable. We additionally specified 100 grid points to account for the reduced grid dimensionality caused by ignoring dS variation.

I don't have a good sense of how many alignment sites were being filtered out on average; are we talking about only a few per simulation? Perhaps this information can be included in Table 1, e.g. for each filtering strategy and sequence count, add a column with median and IQR or percent of sites filtered per replicate. In Table 1, it would be useful to see the TPR under the _correct_ alignment.

We have now included information regarding the percent, and absolute number, of filtered sites in Table S3. We additionally have included a column in Table 1 reporting the mean true alignment TPR.

Larger trees can lead to "difficult alignments" even if mean root path length is reasonably low, e.g. simply because it is harder to infer a good guide tree from pairwise sequence distances or other "crude" metrics for many sequences. I wonder if this is a possible area where filtering may be beneficial? For example, in Table 1 FUBAR was starting to show improvement for 60 and 150 sequences when using filtering.

After recalculating FUBAR results with dS fixed to 1, we found that, when analyzed with FUBAR, mean TPR either increased or decreased for 60-sequence simulation sets. Howevever, filtering universally increased mean TPR (although minimally) for the 158-sequence simulation sets. We made the following changes to emphasize that larger alignments might expect more benefits from filtering than smaller alignments, as follows.
In the subsection *Guidance-based filters have a minimal effect on positive-selection inference,* we state,

> However, we emphasize that, for both the HA and GP41 simulation sets of 158 taxa, all filters significantly reduced FPR and increased TPR, although all effect magnitudes were minimal.

Additionally, in the subsection *Discussion and Conclusions*, we state,

> Thus, we did recover a slight trend suggesting that MSA filtering should be reserved for larger MSAs, which universally featured both a TPR increase and a FPR decrease, on average.

Unless I am mistaken, including the simulation count as a random effect in the mixed effects model relies on the sample size of two to estimate the random effect regression coefficients; does this lead to model overfitting, because there effectively is a separate model parameter per simulation?

We are afraid the reviewer is mistaken here. The lme4 package models random effects as normally distributed random variables with mean 0, thus only a single variance needs to be estimated for each random effect. For the fixed effects, we need to estimate as many parameters as we have filtering algorithms (7, including unfiltered, in the revised ms, versus 2 in our the original draft, where we analyzed normalization methods and phylogenetic correction methods separately). Therefore, we estimate a total of 8 parameters (1 random effect, 7 fixed effects) from 700 data points (7 algorithms times 100 replicates). There is no risk of overfitting. In the previous version of the manuscript, we estimated 3 parameters (1 random effect, 2 fixed effects) from 200 data points (2 algorithms times 100 replicates), which also posed no risk of overfitting.