**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. We have performed some of the additional simulations, described in responses below. For those which we did not perform, we justify.. There is now relatively little emphasis placed on the difference between inference methods. Instead, we focus primarily on the effect of alignment filtering.

**Comments by Reviewer 1**

So, 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.

We thank you very much and we have done this.

**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.

This is probably a good point, and we are going to check out the priors for this to determine where the increase in TPR is coming from for both PAML and FUBAR. We add this to the discussion section.

However, note that the same map from inferred alignment to true alignment was used for each alignment, including all of its filtered derivatives. We emphasize this point in the manuscript.

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.

We largely agree with this point. To account for this, we have conducted additional simulations which feature a relatively higher proporation of neutrally evolving sites (see next response).

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.

Again, we agree with this point. We have conduced an additional simulation set according to evolutionary parameters inferred from GP41. This distribution does feature a mean closer to one, as is discussed in the revised manuscript.

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, fixing dS at 1 and allowing 100 grid points along dN to account for the reduced grid dimensionality.

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 all above information in Table 1.

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.

We have added some areas to the discussion section to point this possibility out. Moreover, rather than recommending that filtering not be used, we suggest instead that filtering is most beneficial, although minimally, in larger data sets.

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?

Our random effect in fact has a sample size of four (unfiltered alignment, GuidanceP filtered alignment, BMweightsP filtered alignment, and PDweightsP filtered alignment). Therefore, we do not feel that this scenario suffers from overfitting. [Moreover, we note that when the random effect is excluded, the magnitude effects are still the same – it’s just significance that changes.]