Response to reviewers

We thank the editor and the two reviewers for their constructive comments. We address each specific comment below. Original comments from editor/reviewers are shown in black, and our responses are shown in blue.

In addition, we would like to point out that, since submission, we have uncovered another improvement in LEISR relative to Rate4Site. Specifically, we have found out that Rate4Site only produces rate estimates for sites which are not gaps in the first observed sequence in the input sequence file, or a user-specified reference sequence. By contrast, LEISR infers rates for all alignment sites. We believe that our approach is clearly more desirable. We now highlight this issue in both the Abstract and Introduction when discussing LEISR improvements and/or Rate4Site drawbacks.

Comments from the Editor

1. L28-29: Due to the placement of the version in parentheses it reads at first as if HyPhy were only the (or a) leading molecular inference platform beginning with that version, which I think is not what you mean. Consider changing this to, e.g., "To this end, beginning with version X.Y.Z we introduce ...", or something else that removes that semantic dissonance.
   We have changed the wording accordingly to the following: To this end, we introduce a generalization of Rate4Site, which we term ``LEISR" (Likehood Estimation of Individual Site Rates). LEISR is available as part of a leading molecular evolution inference platform HyPhy starting with version 2.3.8.

2. L33: "In addition, LEISR is additionally MPI-enabled" -> "In addition" and "additionally" are redundant.
   We have removed the extra “additionally” (also noted by reviewer 1).

3. L40: "and it then for each site, it infers" -> either "and it then for each site infers", or "and for each site it then infers".
   Fixed.

4. L40: "a scaler parameter" -> Is "scaler" correct and not a typo? I'm used to "scaling parameter" or "scale parameter", though will defer to the authors whether "scaler" is the correct term here.

We agree with the reviewer than non-standard terminology is undesirable, and have switched to using “scaling parameter”.

5. L43: "HyPhy will write markdown-formatted status-indicators" -> please cite Markdown, even if only by URL. (Even if this seems sad, there remain many biologists unfamiliar with Markdown format.)
   This is a great point. We are not entirely certain what the most appropriate citation for Markdown is, but we have selected one based on the Wikipedia history of the language.

6. L45: "written to a JSON-formatted file" -> please cite JSON, even if only by URL. See comment about Markdown :-)
   Same as #5.

7. L48-49: "We note that a general description of HyPhy output JSON contents is available from http://www.hyphy.org."; While I don't doubt that the information is there somewhere, I couldn't find the page that would have that information. I suggest to give a more specific URL so that users don't have to go digging around.
   This is an excellent suggestion. To avoid potential reader confusion if the precise URL changes in the future, we have updated this sentence as follows: “We note that a general description of HyPhy output JSON contents is available from http://www.hyphy.org in the ``Resources'' tab.”

8. L64-70: I suggest to rephrase this text. There are repeated "Indeed"s, and it's a little confusing *why* users are encouraged to opt for no rate variation. It seems there are two reasons (leaking of rate variation information into branch lengths; layering LEISR as a non-parametric method over a parametric estimation method is inefficient), but these are a sentence removed and hence it's not quite clear for a reader whether that's the correct semantic parsing result.
   Following suggestions from Reviewer 1, we have broadly reorganized this paragraph. We have additionally changed the second “Indeed” to “In other words”. We believe the changes implemented have clarified the paragraph and our points about branch length optimization.

9. L101-106: In theory the Conclusion section should be about the key take-aways for a reader. I suggest considering to either retitle the section, or to move some of the information here to other parts. For example, could the request to cite Rate4Site to the acknowledgments? And the sentence about earlier versions

containing a more limited implementation of LEISR seems rather peripheral, perhaps that should be a footnote earlier where the 2.3.8 version is initially mentioned?
We have, as recommended, removed this section and relocated text as follows:

1. We have moved the first sentence about accessing LEISR in the HyPhy menu to the abstract.
2. We have moved the statement about 2.3.6 to as a footnote earlier as suggested.
3. We have moved the suggestion to cite Rate4Site to Acknowledgements.

10. The input tree isn't in the repo, and doesn't seem to be created as an initial step in the simulation script? There's commented out code to call ape's rtree() in R, but it's not clear whether that's something that can be done or must be done first. (I think reviewer #2 observed this too.)
In fact, the trees are in the repository contained in the input files "simulation_rates/rtree{25/50/100}.tre". We have expanded the repository's README to clarify.

11. The repository should be permanently archived as a snapshot prior to publication, and the archive URL be mentioned in the paper. This can be quite easily and cost-free accomplished through Zenodo (https://zenodo.org) and Figshare (https://figshare.com) through their respective Github integrations, which will subsequently link the archived snapshot to the live repository on Github. Both also issue DOIs for such deposits. See the guide here (which uses Zenodo): https://guides.github.com/activities/citable-code/. Other possibilities exist of course too.
Upon acceptance, we will archive as suggested using the linked github/Zenodo guide. We will indicate the final DOI in the repository README upon acceptance, as well.

12. Consider easy to accomplish ways to state dependencies more formally. For example, for Python the requirements.txt file does allow to give source code repository URLs, including for Git, instead of PyPy packages https://pip.readthedocs.io/en/stable/reference/pip_install/#vcs-support. I realize that because there are dependencies both for the Python and for the R part, covering both ends equally well may be too challenging. That said, steps that can be taken in this direction and that are relatively easy to take should be considered.

We have added a “requirements.txt” file for managing Python dependencies. There does not seem to be an equivalent file type for CRAN, so instead we added the script “install_R_requirements.R” which will install all necessary packages. The README has additionally been updated to explain how these scripts can be run.

13. - The Python script has several try-catch blocks that in essence make failures completely silent, as not even the error is printed. Perhaps that's justified and failure for any of these blocks is entirely inconsequential, but even if so, that's not obvious from the code, and should hence be stated in comments.
We have updated the Python script to print statements indicating to the users when failures occur.

Comments from Reviewer 1 (Ashley Teufel)

1. Line 28: Is HyPhy really “the leading molecular evolution inference platform”? It may be, but this seems like a pretty bold statement considering that you don’t give a reference. Consider changing the sentence to “a leading molecular evolution inference platform”.
We agree that this sentence was an overreach, and we have changed the article “the” to “a” as suggested.

2. Line 31: This sentence makes it seem like LEISR can only be used on non-coding data.
We have changed the sentence to read “..,thereby providing a flexible and fast platform for rate inference **that may complement codon-level rate inference**”

3. Line 33: Change to “In addition, LEISR is MPI-enabled”
We have removed the extra “additionally”.

4. Line 50: I am not suggesting you implement this in the current version, but in future versions, it would be nice if the normalized estimates were reported as well.
We appreciate this suggestion and will keep it in mind for future releases. At this time, we prefer to only report non-normalized data so as to not confuse/mislead users who are unsure whether to employ normalized or non-normalized data. This way, users can choose the best procedure which suits their needs instead of relying on our chosen normalization scheme.

5. Line 59: The first sentence of this paragraph is very long and hard to unpack. Consider breaking it into two sentences.
   We have revised this sentence as follows (note that the revision includes a bit more detail about differences between approaches):
   “Rate4Site offers two statistical frameworks for rate inference:, maximum-likelihood (ML) (Pupko et al., 2002) and empirical Bayes (Mayrose et al., 2004). Their ML framework is a “fixed effects” approach where a separate parameter is inferred for each site’s rate. Their empirical Bayes framework, by contrast, employs a “random effects” approach where site rates are drawn from a prior gamma distribution. The LEISR implementation is analogous to the rate4site ML approach.”

   We also note that, for organizational purposes, we moved this paragraph up in the manuscript to directly follow the paragraph introducing LEISR’s algorithm.

6. Line 74: A 100 sites seems pretty small. I get that it was probably chosen to keep the run-time of simulations and rate inferences short and I have no qualms with using shorter sequences. However, adding a sentence that points out that the validity of this analysis is not dependent on protein size may help to strengthen the manuscript.
   We have added the following sentence to this section:
   “Although our simulations consisted relatively short gene sequences of only 100 sites, LEISR will show similar accuracy for longer gene sequences due to its fixed effects approach.”

7. Line 81: The word “during” is repeated
   Fixed.

8. Line 89: It's confusing that this paragraph starts out talking about Figure 1, then switches to Figure 2, and then goes back to talking about Figure 1. Consider moving the discussion that starts on line 94 to after the sentence that ends on line 90, so that Figure 1 is discussed then Figure 2 is discussed.
   We have reorganized these sentence as suggested.

9. Line 90: I assume the data shown in Figure 2 is the raw data and not normalized, but maybe it should be stated.

In fact this data have been normalized by the mean, which is stated in line 88: “Finally, for each alignment inference, we normalized rate estimates by dividing all rates by the mean site rate estimate, as described earlier.”

10. You claim that LEISR can use nucleotide data beyond the capacity of Rate4Site, but all of the analysis appears to be done on proteins. Is Rate4Site’s ability to deal with nucleotide data so poor that a comparison between Rate4Site and LEISR cannot be done? Would inference using nucleotide data be expected to work as well as analysis of protein data? Consider sneaking a sentence in somewhere that addresses if any differences are expected when analyzing nucleotide data.
In fact, the Rate4Site command line version only provides the JC69 model for nucleotide rate inference, whereas we extend to also include HKY85 and GTR. We have observed (not shown in paper) that inferences on nucleotide datasets using JC69 in Rate4Site and LEISR show the same level of agreement as with protein data ($r^2 = 0.994$ for an alignment of 10 taxa and 100 sites). We now state in the paper that the same robust relationship seen with protein rates is observed with nucleotide data:
“We note that nucleotide rate inferences under the JC69 model (the only currently available nucleotide model in the Rate4Site command line version) show similarly strong agreement between LEISR and Rate4Site.”

Response to Reviewer #2 (Jesse Bloom)

The paper is clearly written and the background is adequate. The authors provide clear links to the actual analysis code on GitHub used for this paper.

At the beginning of paragraph 2, the authors refer to `HyPhy` as "the leading molecular evolution inference platform." There are also other widely used packages, so it might be better to change "the leading" to "a leading."
We have changed this accordingly, also noted by Reviewer 1.

1. In Approach, the authors never clearly state if LEISR estimates the tree topology or not. My impression is that it does not. If this is so, it should be made clear that the tree is provided along with the alignment.
You are correct that LEISR does not estimate topology; it only optimizes branch lengths. We now state this explicitly when describing LEISR’s algorithm:

“As input, LEISR requires a phylogeny and multiple sequence alignment, and its algorithm proceeds in two steps. It first obtains estimates of alignment-wide branch lengths (considering the input topology as fixed)...”

2. For the test simulations, the authors never explain the topology / branch lengths of the tree. This seems likely to strongly affect the results in the figures (presumably the results are much better with trees with long independent branches). So some information should be given about the tree used.
We agree and have now provided this information:
“Tree lengths (sum of branch lengths) for each tree, respectively, were 13.85, 27.32, and 52.83, and all trees had a mean branch length of ~0.27.”

3. In the introduction they note that `Rate4Site` does not handle *large alignment* or *short sequences* particularly well. It is implied that `LEISR` address both of these issues but no evidence is presented to backup this claim. This could be addressed by simply giving a table of the respective runtimes of the programs, and providing an examples where `LEISR` converges well and `Rate4Site` does not.
We agree that this would be an ideal analysis which we attempted to implement. However, we found that runtimes are largely comparable between LEISR and Rate4Site with the exception of Rate4Site runs which error out virtually immediately due to underflow issues. We in fact had difficulty running Rate4Site on the simulations presented, which we indicated in the manuscript with the following sentence:
“Note that the default number of rate categories for this step in Rate4Site is 16, but Rate4Site failed with errors for all 100-taxa simulations.”

We now augment this explanation with the additional sentence:
“For those runs which completed, we observed comparable run times between LEISR and Rate4Site.”

4. Although not obligatory, it might be nice if Figure 2 also compared the inferred rates to the *true* values used for the simulation. In the sense that this paper just describes a re-implementation of Rate4Site, this is not required. However, it could be useful to have some sense of how accurate the methods are as well.
The original Rate4Site paper in fact performed this analysis, and given the comparability between our implementation and theirs, we believe our rates are accurate. That said, we have some forthcoming projects which examine this

issue more in depth, so we prefer to keep the scope of this paper just to the software implementation.

5. It would be nice to compare `Rate4Site` and `LEISR` on an empirical data to show that they give similar results on real as well as simulated data.
We agree that this would be helpful. We there have run both Rate4Site and LEISR on an established alignment of HRH1 sequences and present results in Figure 3. Results demonstrate that platforms are fully comparable for both simulated and empirical data.

6. The paper takes a bit of time to explain why you should normalize by the median, not the mean, but in the results they normalize by the mean. Why not normalize by the median?
For this revision, we have decided to tone down our explanation of median vs. mean normalization. This makes our choice to normalize by the mean more consistent. We change text to the following:
“That said, in certain circumstances, empirical rate distributions may be overdispersed and zero-inflated. In such cases, we suggest to normalize by the median rather than the mean, should normalization be desired.”