Dear Editor,

Thank you very much.

We are grateful for your comment and the ones raised by the three reviewers, all of which greatly enriched our perspective on the questions addressed in our manuscript.

I have received three reviews of your paper and they are all generally positive. I agree with the reviewers. Your manuscript would make a great contribution to MBE. I would like to invite you to submit a revised version of the manuscript where you address point-by-point all the comments raised by the reviewers. The suggestion of examining the correlating among parameters raised by one reviewer is important and I would expect to find such analyses in a revised version of the manuscript. I also think you could include a better description of similar previously applied models and analyses. For example, MBE 20: 1231-1239 contains a likelihood model for estimating varying N_e among lineages of a phylogeny. The difference between your method and the method in that paper seems to be mostly in the choice of a Bayesian versus frequentist inference framework. MBE 24: 228-235 also used similar models to estimate varying S for optimal codon usage on different lineages of a phylogeny together with varying Omega. And there might be other papers out there using similar approaches of modeling the probability of fixation to vary among different lineages of a phylogeny that I am forgetting. A literature review identifying differences and similarities with previous related approaches would improve the manuscript.

Testing and discussing the identifiability of N_e and μ is absolutely a good idea. As described in the revision destined to reviewer 3, we performed several analyses to examine the correlations between these two parameters. For each branch of the tree we draw a 2-D scatter plot for the joint posterior distribution of N_e and μ. We subsequently fit a linear regression and compute the coefficient of determination (r^2) for each branch of the tree. For a given MCMC, the distribution of r^2 across all branches is then represented as a violin plot. This analysis has been carried out for the simulated datasets (figure 7 in supplementary materials), the mammal datasets (figure 18 in supplementary materials) and the isopods datasets (figure 29 in supplementary materials). Altogether, the values of r^2 are low, below 0.01 on average for simulated datasets and below 0.1 on average for empirical datasets, providing quite some confidence that Ne and μ are indeed identifiable and are not strongly correlated.

We edited the introduction to introduce these previous attempts at capturing the variation in N_e among branches and to outline the differences between these previous approaches and our model:

A first attempt in this direction was proposed by Nielsen and Yang (2003), using a population-genetic argument to relate the distribution of dN/dS across sites with the underlying distribution of fitness effects. This first approach assumes that all non-synonymous mutations at a given site have the same selection coefficient. As a result of this assumption, there is a simple, one-to-one mapping between the dN/dS at a given site and the selection coefficient associated with all non-synonymous mutations at that site. In practice, different non-synonymous mutations are likely to have different fitness effects.. As a result, the substitution rate between each pair of codons can be predicted, as the product of the mutation rate and the fixation probability of the new codon, which is in turn dependent on the fitness of the initial and the final codons. Since the strength of selection is typically not homogeneous along the protein sequence, and depends on the local physicochemical requirements (Echave et al., 2016; Goldstein and Pollock, 2016, 2017), local changes in selective strength are usually taken into account by allowing for site-specific amino-acid fitness profiles. Site-specific amino-acid preferences are typically estimated either by penalized maximum likelihood (Tamuri and Goldstein, 2012; Tamuri et al., 2014), or in a Bayesian context, using an infinite mixture based on a Dirichlet process prior (Rodrigue et al., 2010; Rodrigue and Lartillot, 2014). This second approach is further considered below.

And latter in the introduction:
Conversely, since the mutation-selection formalism explicitly incorporates N_e as a parameter of the model, extending the model so as to let N_e vary across lineages is relatively straightforward. This idea was previously explored in the context of two mechanistic models, relying on the distribution of dN/dS across sites (Nielsen and Yang, 2003) or accounting for selection on codon usage (Nielsen et al., 2007). Doing this in the context of mutation-selection models with site-specific amino-acid preferences would provide provide an occasion to address several important questions: do we have enough signal in empirical sequence alignments, to estimate the evolutionary history of N_e along a phylogeny? Can we more generally revisit the question of the empirical correlations between N_e and ecological life-history traits (longevity, maturity, weight, size, ...), previously explored using classical dN/dS based models, but now in the context of this mechanistic framework?

Reviewer: 1

Overall, I believe this work represents an important next-step in the development of HB-style models that explores how we might relax the constant N_e assumption. My comments are as follows:

1) On page 3 line 94, the authors suggest that dN/dS is not a parameter of their model (similarly there is no such parameter in equation 20 presenting the codon substitution component of the model). However, "omega" is a parameter of the model in the PhyloBayes MPI package. Can the authors please confirm whether omega is a model parameter or not? Looking now at the ThibaultLatrille/bayescode repository, it's not clear to me that "omega" has actually been taken out of the model in practice; e.g. comments in https://github.com/ThibaultLatrille/bayescode/blob/chronogram/src/lib/AAMutSelNeCodonMatrixBidimArray.hpp read: "It then a constructs a 2-dimensional-array of AAMutSelOmegaCodonSubMatrix ." I have not explored the code in depth, and of course a comment does not imply omega was definitely used, but it is certainly suggestive.

It is indeed true that "*omega*" has been implemented as a parameter of the model in our implementation, and that it can be fed as a parameter of the codon matrices "*AAMutSelOmegaCodonSubMatrix*". We actually leveraged this modelling approach combining mutation-selection fitness profiles and "*omega*" to detect specific sites inside a gene that show an excess of substitutions compared to the nearly-neutral expectation (https://doi.org/10.1093/molbev/msaa265).

However, in the model presented here, "*omega*" has been taken out of the model in practice, as it can be found at line 13-14 of "*AAMutSelNeCodonMatrixBidimArray.cpp*" when constructing the "*AAMutSelOmegaCodonSubMatrix*".

```
matrixbidimarray[i][j] = new AAMutSelOmegaCodonSubMatrix(
    codonstatespace, nucmatrix, fitnessarray->GetVal(j), 1.0, pop_size_array.at(i));
```

Where the 1.0 refers to "omega" parameter, set to a fixed value of 1.0 (multiplicative parameter).

As a side note, we are implementing a model where "*omega*" is also a branch-specific parameter (another component of the multivariate Brownian process), such as to test whether it correlates with N_e:
https://github.com/ThibaultLatrille/bayescode/tree/branchomega

2) For Figure 1, I recommend two minor changes to improve clarity (but note that overall, this is a great figure!):
a) For panel "A," are residue colors based on the same colors as tips? Communicating how the colors are related will be helpful. It is also challenging to distinguish red/orange (and I

do not have any color vision deficiency). If the colors are key for interpreting the figure, please ensure they can be universally distinguished.
b) It would be helpful to add the letter "i" in the top right text orange: "20 amino-acid fitnesses for each profile category (each site i of the alignment is assigned one of the K profile category)".

Thank you very much for these suggestions.

a) Residue colors are indeed misleading and the legend was missing. It was originally meant to discriminate synonymous (blue) and non-synonymous substitutions (red). But this information is not relevant and is just confusing, so all residues now have the same color.
b) Upon reflection on this comment, and for better clarity, we now display the site-specific amino-acid fitnesses as logo-plot (where the fitness for each amino-acid is stacked vertically for each site), instead of profile categories (which are then assigned to sites). The legend has been changed accordingly.

3) Page 5 line 130: "In the model introduced here, Ne and μ are allowed to vary between species (across branches)." The phrase "across branches" is somewhat ambiguous, rephasing this to "among branches" might be better to communicate that parameters do not vary *within* a particular branch.

Absolutely, we also edited a second occurrence page 15 line 409-410

4) The paragraph at the top of page 10 needs some significant copy editing. Are there 18 genes from 4 MSAs, or 18 MSAs? In addition, the methods behind this subsetting should be more clearly explained. How many subsets are analyzed?

The paragraph has been rephrased to:

We restricted our analyses to a random set 18 of orthologous genes, which are then concatenated into a single (multiple sequence alignment) MSA for analysis. To assess the reproducibility of our inference and check that the signal about variation in N_e is not driven by particular genes, we analysed in total 4 different concatenated MSA each containing 18 randomly sampled genes.

5) Jumping off the previous comment, in general through the manuscript, this subsetting aspect of the Methods was not clear to me. How do all the subsets relate to the full analysis of 77 placentals presented (and similarly for isopods)? Reading the mammals section of the SI, I see that 4 replicates of 18 randomly-chosen taxa were performed, but Table 7 shows four replicates for which there are more than 18 rows per replicate. This doesn't make sense to me with the current explanation of the methods.

The sub-sampling procedure was undoubtedly not clearly stated. More specifically, for each replicate, we randomly choose 18 genes which are then concatenated. If a gene is not present for one the 77 taxa, gaps are added to the MSA for this specie. We verified that for each species the concatenated MSA does not contains only gaps (meaning that all 18 genes are absent in this species). Once the analysis is done, we obtain an estimation of N_e for each extant species, which is shown in table 7. This procedure is repeated on 4 different replicates. Hence we obtain for each species 4 different estimations of N_e.

6) Jumping off again...The authors state only that the subsetting was performed "for computational reasons." Much more information is needed here to explain why this procedure was necessary and what the specific "computational reasons" are. Ie what resources were needed and what was the runtime for analyzing "only" 18 genes? Is this also an MPI implementation? This will also help readers know if the presented method is feasible for their data. The overall limitations of the method associated with computational constraints should be discussed. Indeed, the only caveat about computational resources in the manuscript is page 13 line 364, "In addition, the empirical fitting of the model requires more computing resources." A lot more is needed here.

The program has not been implemented in MPI so far. And we agree that feasibility should be discussed, we added this paragraph to the Discussion:

Moreover, providing a computationally more efficient implementation of the model would be important for broader application. Currently, running the program on an MSA of 18 mammals genes (77 extant species, and on the order of 15000 nucleotide sites) for 4000 iterations of the chain (1000 are left as burn-in) takes approximately 2-4 weeks of computations, which is quite long although still accessible for reasonably small datasets. Increasing the computational efficiency could be achieved by several means: first, parallelizing the program would be relatively easy in the present case, in particular, by dispatching genes over multiple cores. Second, a large fraction of the computing time is spent in updating the fitness profiles, and thus, fixing them to empirical values or using pre-estimated profiles under a constant N_e would lead to a substantial acceleration.

7) Suggestions to improve readability of Figure 2:
a) Please move panel labels A-F to the top-left outer corners of each panel.
b) Please add information to the caption indicating what the simulation for C/F is.
This information is provided for A/D and B/E, but not C/F.

a) Top-left corner is absolutely more appropriate for the labels than bottom-right, and also more consistent with the other figures.
b) Accounting for site epistasis in the context of selection for protein stability (C/F)

8) In general, the methods and parameterizations for the simulations need a lot more detail, even considering the SI. Eg, the authors state that fitnesses were extracted from Bloom's deep mutational scanning data, but they do not describe how specifically these values were incorporated into simulations. Another example, this sentence, "Mutations are drawn randomly based on mutation rates" leaves much to be desired. Please take some time to seriously clarify the precise approach to each simulation, including the size of simulations (I had a hard time finding the trees used for simulation, the codon lengths,

etc), all initial parameterizations and how those values were ascertained and specifically applied to each simulation, how specifically parameters are setup to vary across phylogenies, etc.

The fitness vectors used in this study are extracted from Bloom (2017), which were experimentally determined by deep mutational scanning for 498 codon sites of the nucleoprotein in Influenza Virus strains (as human host). Although the Influenza Virus is phylogenetically far away from vertebrates, the experimentally determined profiles in these strains capture the structural constrains exerted on proteins. For each codon site z of our simulation, we assign randomly one the 498 fitness profile (sampling with replacement) experimentally determined, which altogether determines the selection coefficient for any non-synonymous mutation.
We edited the supplementary materials to include the fitness profiles displayed as logo plots.

In our simulations, the tree is composed of 77 species (available in newick format in supplementary), the tree root is 150 million years old, the initial mutation rate is 10^{-8} per site per generation and the initial generation time is 10 years. The simulator starts from an initial sequence at equilibrium, composed of 15,000 codon sites.

Mutations are drawn based on a user-defined nucleotide matrix, where our simulations used a symmetric time-reversible mutation matrix.

We edited the supplementary materials to include a more detailed description of the simulations, in particular, by providing the parameters and configurations used to produce alignments.

The manuscript has been edited to clarify these points.

9) The estimated N_e's appear to be relative to some initial population size for the simulations; 10^-1 is not itself a reasonable N_e value. Can the authors please add this further context? I see this information in the Fig 3 caption, "N_e values are relative to the root, which is arbitrarily set to one," but this information needs to be more emphasized in the simulation methods and/or explanation of the model itself. For example, when reading Figure 2, one will not yet have read this information about root being fixed at N_e=1. The only formal statement I could find about this is buried on page 13 line 363 in the Discussion, so it must be introduced earlier.

This information was indeed buried in the Discussion and the Methods (subsections *Branch dependent traits* and *Codon substitution rates)* while it should unquestionably have been introduced earlier in the manuscript.
We added this information in the section *New approaches:*
Of note, since N_e and f are confounded parameters (equation 3), the effective population size at the root is set to 1 for identifiability of the fitness profiles. As a result, all values of *Ne*

along the phylogeny are relative to that of the root, with a value of $N_e > 1$ reflecting an increase in N_e along the branches (respectively a decrease for $N_e < 1$) compared to the N_e at the root.

10) Figure 3 is rather hard to read. At a minimum, I recommend:
a) Increasing the size of the scale bars (but please also add the note to the caption that branch lengths neutral subs/unit time in mutation-selection models)
b) Ensuring distinct color schemes between OTUs and branches; The orange/yellow/purple colors in the taxa match the legend scales which is confusing.
c) Is there any reasoning to the specific color scheme of clades, or are colors just alternating to indicate clades more generally? Please make sure that that allalternating colors are distinguishable for individuals with color vision deficiencies.
d) My guess is that all the mammal icons came from http://phylopic.org/ ; please make sure to attribute these images to them appropriately.

a) The scale is time (the legend was missing), and is set to 1.0 from root to leaves (the tree is a chronogram). Fossil records (at least one) can allow to scale time in absolute (millions years).
b & c) Color schemes for clades were indeed misleading and non-informative, it has been changed to a grey/white color alternance between clades, which is more consistent and parsimonious.
d) Indeed pictograms are from http://phylopic.org, which is now edited in the manuscript.

11) Page 9 line 211, "assumes a mostly nearly neutral regime." This is true, as the authors already pointed out based prior literature, ONLY if the regime is at equilibrium. Is the size of the phylogenies examine here sufficient to achieve an equilibrium process? This would require high branch lengths. One way to check is to see if observed AA frequencies in simulations roughly match simulation parameters. If they are very different, then the simulation does not achieve stationarity and we cannot assume an observed nearly-neutral process (see Jones 2019 which the authors cite).

It is in fact a good question: whether a model with a fixed fitness landscape but changes in N_e is still to be considered nearly-neutral. However, please note that, in the present case, for the analyses under the mutation-selection model, we do not assume to process to be at equilibrium since N_e is changing along the phylogeny, and we don't assume this either in our simulations. More globally, we don't need to assume a fast return to equilibrium upon changes in N_e for our reconstructions under the model to be valid.

On the other hand, our point here was just that most genes that are detected under positive selection by site models (i.e. with dN/dS>1 for some sites) are probably experiencing strongly fluctuating fitness landscapes, at least locally along the sequence (Rodrigue and Lartillot, 2016). Conversely, we don't think that the kind of variation in N_e inferred or simulated in our experiments is susceptible to create such strong positive selection patterns over the whole tree, so, it seems relatively safe to exclude those genes, as just not fitting the assumptions of the model.

We thus edited the sentence to be more accurate:
Of note, the mutation-selection model considered here assumes that the fitness profiles do not change with time. In contrast, some genes might experience fluctuating fitness landscapes through time. Such fluctuations are in fact one main cause of ongoing adaptation (Mustonen and Lässig, 2009; Rodrigue and Lartillot, 2016). For that reason, genes for which positive selection was detected using a site codon model were excluded from the analysis.

12) What are the correlations in Table 1/described in text starting page 9 line 222? Are these spearman? In addition, in this paragraph on page 10, the presentation of ranges is really confusing - I initially interpreted something like \rho = [-0.84, -0.83] as a credible interval of some kind. After looking at Table 1, I see this is a range of different correlations performed. It might help to just remove these ranges from the text and just refer readers to the table.

The correlation coefficients are given by the covariance matrix of the multivariate Brownian process. As such they are already corrected for phylogenetic inertia.
Subsection 5.7 (Correlation between traits) of the section Materials and Methods, and more specifically equation 23, relates correlation coefficient to the covariance matrix and how they are computed. They are averaged over the posterior distribution, and statistical support is assessed based on the posterior probability of having a positive (or negative) value for the coefficient.
The representation of the different values as a range was undoubtedly very confusing, and was removed as a consequence.

13) Figure 4 comments
a) Please add a-c labels instead of left/middle/right to assist readers in interpreting the figure.
b) I recommend moving the pvalues and r^2 values (general comment – interesting to see some frequentist statistics pop up in this paper! not a problem, just interesting) outside of the panels to more easily view the boxplots.
c) Again, are these N_e relative to an arbitrary root value of N_e=1?
d) What exactly is the data being analyzed for this figure, as it relates to the isopod data subsetting? Authors state that 4 replicates of 12 randomly-chosed taxa were analyzed (again, please add much more explanation and clarity to this procedure). How many taxa are actually in each boxplot?
e) The boxplot lines themselves are about the same size as the plot's background grid. The authors might wish to increase the width of the boxplot outlines to clearly distinguish.
f) Three analyses were performed on the same isopod dataset, so there should be an error correction to the p-values (though they are so slow this won't change any results).

a) Absolutely, it is now more consistent with the others figures.

b) The figure is indeed more coherent by moving the p-values and r^2 values outside the boxes.

c) Yes, absolutely.

d) The paragraph describing this subsetting has been updated:

To assess the reproducibility of our inference, we analysed in total 6 different concatenated MSA each containing 12 randomly sampled genes. The 6 different concatenated MSA showed similar trends in the change of μ and N_e between pairs of replicates (see supplementary). A statistical analysis performed on the pooled estimation of N_e across the 6 different concatenated MSA exhibits a statistically significant reduction in N_e for underground or depigmented species, or for species with visual impairment.

e) The boxplot have been filled (in blue) to clearly outline them against the background.

f) Absolutely

14) Page 12 line 300 sentence is very confusing: "However, if the trends are in right direction, the magnitude of the changes inferred across the phylogeny is surprisingly narrow and does not match independent empirical estimates of the variation in those clades." What is "the right direction?" Is that an EXPECTED direction? Which direction? Which specific trends? Etc.

The paragraph has been edited to:

However, although the changes in N_e are in the expected direction (negative correlation with body size, weight and maturity), the magnitude of the changes inferred across the phylogeny is surprisingly narrow (at most a factor 9.2 in the extremal case in mammals). This range does not match independent empirical estimates of the variation in mammals, where synonymous diversity varies by a factor at least 10 between species (Galtier, 2016). In animals, the synonymous diversity roughly spans two orders of magnitude, whereas N_e varies considerably more across species, by a factor of 10^3 (Galtier and Rousselle, 2020). For instance, effective population sizes estimated based on population genomic data are of the order of 10 000 in humans (Li and Durbin, 2011), and 100 000 in mice (Geraldes et al., 2008). Thus, clearly, our approach underestimates1 the true variation. Different mechanisms not accounted for by the model could explain this result.

15) When discussing factors that could influence the model results and/or future model developments in Discussion, it might be worth mentioning the site-invariant mutation rate. While this method varies μ across branches, in theory this parameter could be varied across sites as well.

The section *Discussion* was updated to:

This approach can be further extended in other directions.

First, the mutation rate (μ) is considered site-invariant, an assumption which could be relaxed by introducing site-specific mutation rate to account for variation in mutation rate along the sequence.

Second, currently, our model also assumes no selection on codon usage...

16) In equation 20, please use the phrase "nearest-neighbors" instead of "neighbors" when referring to codons with >1 substitution.

Nearest-neighbors is indeed more appropriate.

17) A good deal of the references are not correctly formatted and/or missing information such as journal, year, pages, etc. Please have a look at references.

We are ashamed of having left so many formatting issues. These are corrected in the revised manuscript.

18) Grammar and spelling comments:
a) Page 8 line 204, "taxnonomic" → taxonomic
b) Page 11 line 273: "This results contrast" → These results contrast (or, this result contrasts?)
c) Page 20 line 527, "mutant" → mutations (unless I have missed the meaning here, in which case please revise the sentence)
d) Fig 4 legend: "All three qualitative trait" → traits
e) Please remove the comma page 2 line 25, "Since the realization,".
f) Fig 2 caption line 29 should be "the population startS", and please remove the comma before ", and generation time of 10 years."
g) I suspect the last sentence in Table 1's caption should say "generation time" instead of "generation rate."
h) Page 10 line 262, "species that DID not" (not dit)

Really sorry for these grammar and spelling errors.
g) generation rate meant to describe the inverse of generation time, but this is confusing, hence the sentence has been rephrased to :
Moreover if the mutation rate per generation is considered constant in first approximation, the mutation rate per unit of time is thus negatively correlated to generation time. As a result, the mutation rate per unit of time is positively correlated to N_e (since generation time and N_e are negatively correlated).

Reviewer: 2

This is a very nice contribution adding a variable population size to the inference of mutation-selection models. The paper is very well written, very clear, and the results are reasonable. There are some limitations, most notably with respect to the range of Ne values inferred, but these limitations are clearly stated and convincing reasons for these limitations are provided.

I am happy with the paper as is and don't have any specific comments that need to be addressed.

Thank you very much, we edited the manuscript to reflects the suggestions of the editor, reviewer 1 and 3.

Reviewer: 3

The authors extended their Bayesian MCMC implementations of a mutation-selection model of codon substitution to allow the population size and mutation rate to vary among branches of the species phylogeny, and use the model to analyse several datasets to estimate branch-specific population sizes. While the model is able to suggest variable population sizes the range seems far too narrow, suggesting that some important components are still missing or not right in the model. Overall, the analyses have not offered much novel biological insight in the data analysis. However, the models are interesting and have potential so it is to be hoped that future explorations and improvements may lead to greater utility. Overall the paper is clearly written and reads very well. I support publication of the paper.

I have only a few minor comments. I am afraid that I don't have good ideas for identifying the problems that the authors face, for example, why the estimated N showed such narrow ranges.

1) I wonder whether it is useful to include a discussion of identifiability or information content of such models. It seems that some parameters are strongly correlated or only weakly identifiable, which may make it hard to fit the model to realistically sized datasets. Could you for a few selected branches make a 2-D scatter plot for N and u, for example, to see whether they two are strongly correlated.

Testing and discussing the identifiability is absolutely a good idea.
To test whether N_e and μ are identifiable, for each branch of the tree we draw a 2-D scatter plot for N_e and μ, where each point is a step in the MCMC procedure. We subsequently fit a linear regression and compute the coefficient of determination (r^2) for each branch of the tree. For a given MCMC, the distribution of r^2 across all branches is then represented as a violin plot. This analysis has been carry out for the simulated datasets (figure 7 in supplementary materials), the mammal datasets (figure 18 in supplementary materials) and the isopods datasets (figure 29 in supplementary materials).
Altogether, the distribution of r^2 is relatively low, below 0.01 on average for simulated datasets and below 0.1 on average for empirical datasets, providing confidence that N_e and μ are indeed identifiable and not strongly correlated.

2) I initially thought that joint analysis of thousands of genes may be useful for teasing apart the multiple parameters, because N and life history characters have genome-wide effects while selection is expected to be gene-specific. However on second thought, this may not be useful either since the model already assumes site-specific amino acid profiles so that the among-gene variation is already accommodated in the model by the among-site variation. I suppose selection on codon usage may show gene-wide effects

(for example, some genes have much high GC3 than others), but I am not sure whether features like that can be built into the model, and whether they may help to tease apart the multiple parameters.

It is theoretically possible to build gene-wide codon usage effect by introducing gene-wide fitness parameters for synonymous codons (as in Nielsen & Yang, 2007) in addition to site-specific amino-acid fitness profiles, as we currently do. This feature would still fit into the mutation-selection framework. However, practically, the current implementation working on a concatenated alignment is not adapted to accommodate such a change. As a perspective, under a new MPI implementation where each gene would have its own process, it would be definitely be worth implementing and exploring the effect of codon usage.

3) p.1 “unreasonable hypothesis”, perhaps change to “unrealistic assumption”?

Absolutely

4) p.2 “Since the realization, by Zuckerkandl and Pauling (1965) that genetic”. Delete the comma.

Absolutely

5) p.2 “Codon models have been used to empirically measure such changes in the efficacy of purifying selection along phylogenies, either by allowing for different dN/dS values in different parts of the tree (Dutheil et al., 2012), or by estimating dN/dS independently for every branch of the tree (Popadin et al., 2007).”
Should you cite Yang (1998), Yang and Nielsen (1998) for branch models of dN/dS variation, or are those papers too old to be relevant?

Definitively a wrongdoing of not citing these seminal papers, sentence rephrased to:
Codon models allowing for variation in dN/dS across branches (Yang and Nielsen, 1998; Yang, 1998; Dutheil et al., 2012) have been used to empirically measure such changes in the efficacy of purifying selection along phylogenies.

6) Should “log-Brownian” be “exponential Brownian”. Log-normal is a misnomer as it should have been called exp-normal, but I have not seen “log-Brownian” used before. Another term is geometric Brownian motion.

Agreed with the terminology of a geometric Brownian process.

p.6 figure 2 legend, 1e^-8 should be changed to 10^-8, as the way you write is looks like exp(-8). There are a few other occurrences of this in the ms.

This notation was undoubtedly confusing. We found three occurrences in the caption of figures 2 and 3 that have been edited.

7) p.6 figure 2. What do we expect to see in those plots if the theory and program implementation are correct? The program is complex, so debugging (for example telling apart errors and bugs from mcmc mixing problems) looks hard to me. Could you comment briefly how the program is validated.

The plots shown in this figure (panels A and D) show a rather close match between inferred and true branch lengths, as well as between inferred and true N_e. In itself, it is a good, although perhaps qualitative, argument in favor of the soundness of our implementation. Also, please note that The implementation used for inference (https://github.com/ThibaultLatrille/bayescode/) is completely separated from the implementation of the simulator (https://github.com/ThibaultLatrille/SimuEvol). Hence, if one (or both) implementation were wrong, it would be unlikely that the inference (BayesCode) could uncover from the alignment the same branch-specific N_e simulated by the other program generating the alignment (SimuEvol).

More quantitatively, on panel D for which simulations and inferences are modelled equivalently, point estimates for N_e for inference are close to the simulated value, with a precision=$|N_e^{inferred}$-$N_e^{simulated}|/N_e^{simulated}$ of 81%. Secondly, the errors on the point estimates for N_e are small compared to the variation in N_e across branches, with a z-score=$|N_e^{inferred}$-$N_e^{simulated}|/\sigma(N_e^{simulated})$ of 0.21. Thirdly, 91% of the credible intervals shown on the figure cover the true value.

On panel E, for which simulations are under a Wright-Fisher model accounting for small population size effects, point estimates for N_e for inference are close to the simulated value, with a precision=$|N_e^{inferred}$-$N_e^{simulated}|/N_e^{simulated}$ of 79%. Secondly, the errors on the point estimates for N_e are small compared to the variation in N_e across branches, with a z-score=$|N_e^{inferred}$-$N_e^{simulated}|/\sigma(N_e^{simulated})$ of 0.25. Thirdly, 90% of the credible intervals shown on the figure cover the true value.

On panel F, for which simulations are under a model of selection for protein stability accounting for site epistasis, point estimates for N_e for inference are further to the simulated value, with a precision=$|N_e^{inferred}$-$N_e^{simulated}|/N_e^{simulated}$ of 47%. Secondly, the errors on the point estimates for N_e are large compared to the variation in N_e across branches, with a z-score=$|N_e^{inferred}$-$N_e^{simulated}|/\sigma(N_e^{simulated})$ of 0.94. Finally, only 21% of the credible intervals shown on the figure cover the true value.

8) p.9 “we restricted our analyses to small concatenates”, “The different concatenate showed similar trends”. concatenate is apparently a verb. Please rephrase.

The paragraph was rephrased to:
We restricted our analyses to a random set 18 of orthologous genes, which are then concatenated into a single multiple sequence alignment (MSA) for analysis. To assess the reproducibility of our inference and check that the signal about variation in N_e is not driven by a particular sampled gene set, we analysed in total 4 different concatenated MSA each containing 18 randomly sampled genes.

Also in the caption of figure 3, we added:
Inferred phylogenetic history of N_e (left) and μ (right) across placental mammals, based on an analysis of a concatenation of 18 CDS randomly chosen among single-copy orthologs putatively under an exclusively purifying selection regime (see methods).

9) p.9 “The estimated covariance matrix (table 1) gives...” This paragraphs has quite a few typos and grammatical errors. Please fix.

The paragraph was edited:
The estimated covariance matrix (table 1) gives a global synthetic picture of the patterns of covariation between the mutation rate per unit of time μ, the effective population size N_e and the three LHTs. First, μ covaries negatively with body mass, age at sexual maturity and longevity (table 1). These correlations,which were previously reported (Lartillot and Delsuc, 2012; Nabholz et al., 2013) probably reflect generation time effects (Lanfear et al., 2010; Gao et al., 2016). Similarly, and more interestingly in the present context, N_e covaries negatively with LHTs (table 1). This is consistent with the expectation that small-sized and short-lived species tend to be characterized by larger effective population sizes (Romiguier et al., 2014). Of note, these results mirror previous findings, based on classical codon models, showing that dN/dS tends to be positively correlated with LHTs (Lartillot and Delsuc, 2012; Nabholz et al., 2013; Figuet et al., 2017). This positive correlation between dN/dS and LHTs was also recovered on the present dataset, using a classical dN/dS based codon model (supplementary materials). Interestingly, the correlation between dN/dS and LHTs is weaker than the correlation between our inferred N_e and LHTs, as expected if the variation in dN/dS indirectly (and imperfectly) reflects the underlying variation in $N_{e.}$ Finally, N_e and μ are positively correlated in their variation ($\rho = 0.44$), which might simply reflect the fact that both covary negatively with LHTs. The partial-correlation coefficients (see supplementary) between N_e and LHTs are not significantly different from 0. However, this might simply be due to the very strong correlation between the three LHTs considered here, such that controlling for any one of them removes most of the signal contributed by the empirically available variation between species.

10) p.11 “mutation-selection models of the HB family ”. What is the HB family?

Halpern and Bruno, but it is true that the acronym has not been introduced, hence the sentence has been rephrased to:
Thus far, the development of mutation-selection models of the Halpern and Bruno (1998) family (Rodrigue et al.,2010; Tamuri and Goldstein, 2012) has mostly focused...

p.12 “However, if the trends are in right direction”, perhaps change “if” to “although”.

Absolutely

11) p.13 “(allowing for at most 50 distinct profile categories).” 50 categories sound to me like a lot, if they are really distinct.

Firstly, for a concatenated alignment of 18 different genes, we expect the codon sites to fall under a wide variety of profiles. But, as you are suggesting, 50 profile categories sounds a lot, and you are right since experiments made by a collaborator (not published yet) suggest that over 40 profile categories the results are equivalent in most cases (no gain by adding new profiles). However, we wished to a bit conservative such as to care for “unusual” cases, and also to dampen the effect of a potentially changing N_e along the sequence, since this effect would be captured by the “extra” categories. The downfall is the extra computational time required.

12) p.13 “the signal about the intensity of drift comes from the relative rate of non-synonymous substitutions”. Do you mean “synonymous substitutions”.

Under the assumption that synonymous mutations are neutral, the synonymous substitution rate (Q) is independent on the intensity of drift (Q=μ). Only the non-synonymous substitutions are carrying signal on the strength of drift, with a reduced rate of non-synonymous substitutions.

The sentence was rephrased to:
the signal about the intensity of drift comes from the rate of non-synonymous substitutions relative to that of synonymous substitutions.

13) p.15 “Dir(1/6, 6)”. Is this usually written as Dir(1,1,1,1,1,1). For example Dir(1,1) is U(0,1). The same comment about Dir(1/4, 4).

Agreed, edited in the manuscript and in figure 5.

14) p.17 “For a given node, all the nodes pointing toward him (upstream) are its dependencies which determines its distribution.” Rewrite the sentence. him should be it.

The sentence was rephrased to:
All the nodes pointing toward a given node (upstream) are its dependencies which determines its distribution.

15) p.19 “detailed substitution history H for all sites along the tree.” Is this actually more expensive than calculating P(t) and summing over ancestral states at the internal nodes if the branch is long with many substitutions?

Obtaining the detailed substitution history (H) is indeed time-consuming, and it is more costly than calculating P(t) and summing over ancestral states at the internal nodes. However, without H, calculating P(t) and summing over ancestral states at the internal nodes would be necessary for every move of any parameter of the model. Since the model is parameter intensive, it is much less costly to first compute *H*, and then move the parameters based on this detailed substitution history (which is much more efficient). Every point in the chain of the MCMC starts by computing *H* based on the current parameters of the model.

16) The References section has formatting issues, with missing page numbers, inconsistent capitalization of titles, etc. These need to be fixed.

We are ashamed of having left so many formatting issues. These are corrected in the revised manuscript.