Model misspecification confounds the estimation of rates, and exaggerates their time dependency

Brent C. Emerson[1,2*], Diego F. Alvarado-Serrano[3] and Michael J. Hickerson[3,4,5]

1. Island Ecology and Evolution Research Group, Instituto de Productos Naturales y Agrobiología (IPNA-CSIC), C/Astrofísico Francisco Sánchez 3, La Laguna, Tenerife, Canary Islands, 38206, Spain.

2. School of Biological Sciences, University of East Anglia, Norwich Research Park, Norwich NR4 7TJ, UK.

3. Biology Department, City College of New York, New York, NY, 10031, USA.

4. The Graduate Center, City University of New York, New York, NY, 10016, USA.

5. Division of Invertebrate Zoology, American Museum of Natural History, New York, NY 10024, USA.

* Contact author: bemerson@ipna.csic.es

Abstract

While welcoming the comment of Ho *et al.* (2015), we find little that undermines the strength of our criticism, and it would appear they have misunderstood our central argument. Here we respond with the purpose of reiterating that we are (i) generally critical of much of the evidence presented in support of the time dependent molecular rate (TDMR) hypothesis, and (ii) specifically critical of estimates of μ derived from tip-dated sequences that exaggerate the importance of purifying selection as an explanation for TDMR over extended timescales. In response to assertions put forward by Ho *et al.* (2015), we use panmictic coalescent simulations of temporal data to explore a fundamental assumption for tip-dated tree shape and associated mutation rate estimates, and the appropriateness and utility of the date-randomisation test. The results reveal problems for the joint estimation of tree topology, effective population size and μ with tip-dated sequences using BEAST. Given the simulations, BEAST consistently obtains incorrect topological tree structures that are consistent with the substantial overestimation of μ and under-estimation of effective population size. Data generated from lower effective population sizes were less likely to fail the date-randomisation test yet still resulted in substantially upwardly biased estimates of rates, bringing previous estimates of μ from temporally sampled DNA sequences into question. We find that our general criticisms of both the hypothesis of time-dependent molecular evolution, and Bayesian methods to estimate μ from temporally sampled DNA sequences, are further reinforced.

Introduction

In their opening paragraph, and then repeated within their comment, Ho *et al.* (2015) state that we (Emerson & Hickerson 2015) "claim that there is a lack of support for a time-dependent pattern in molecular rate estimates". This is not correct. What we argue for, both in our original paper and here, is that (i) there is a lack support for the inferred magnitude of TDMR patterns, and that (ii) explanations of purifying selection over extended timescales to reconcile differences between spontaneous μ and phylogenetic estimates of μ have been greatly exaggerated, largely because of issues with biased rate estimates derived from ancient DNA (aDNA) analyses. Neither in this response, nor in our original article, do we deny there to be evidence for time dependent patterns for molecular rate estimates. Nor do we deny that purifying selection will lead to lower values for spontaneous μ. What we argue for in our original article (Emerson & Hickerson 2015), but apparently misunderstood by Ho *et al.* (2015), is that the support for purifying selection underpinning these observed patterns is greatly overstated when most of the observed changes in estimates of μ can be explained as methodological artifacts. Purifying selection will lead to lower values for spontaneous μ. This is a truism that we have recognised previously (Emerson 2007). However, the assumption of Ho *et al.* (2015) that pattern is evidence for process exaggerates both the inferred extent of and timescale for rate reduction due to purifying selection. This is our central argument and cause for concern.

Evidence for pattern is not evidence of process

A substantial part of the comment of Ho *et al.* (2015) is devoted to presenting many examples of evidence for time-dependent rate estimates, although for nuclear data, Ho *et al.* (2015) acknowledge that there is no strong evidence for such a pattern. As stated above, we are not in denial of the many published estimates supporting the pattern for mtDNA, and as such our position is somewhat misrepresented by Ho *et al.* (2015). It is important to point out that, if a pattern can be explained by something other than the hypothesis (the hypothesis here being purifying selection), then the pattern itself cannot be used as evidence in support of the hypothesis. In this context, the examples presented by Ho *et al.* (2015) do not in themselves contradict the points raised in Emerson & Hickerson (2015), as these may be subject to the methodological issues raised in our original article. Indeed, some of the examples where we highlight methodological issues (e.g. *Caenorhabditis elegans*) are presented again by Ho *et al.* (2015) as supporting the hypothesis of time-dependent molecular evolution without further discussion of the concerns we raised. We focus the remainder of this response on specific points within the comment of Ho *et al.* (2015), where we feel they may have either failed to provide an adequate response, or misrepresented our work, when discussing the evidence for the hypothesis that purifying selection is the driver of TDMR estimates.

Adélie penguin data

In our original article (Emerson & Hickerson 2015) we pointed out that, in contradiction to the TDMR hypothesis (i.e. the hypothesis that molecular rate estimates decrease toward the past as a consequence of purifying selection) mean pedigree-based estimates of the mutation rate of mitochondrial DNA in Adélie penguins are lower than those inferred from aDNA. In response to this, Ho *et al.* (2015) make two points. They first suggest that the non-reporting of 95% credibility intervals may somehow limit the significance of our observation, and further claim there to be substantial overlap in the 95% credibility intervals between aDNA estimates and the pedigree estimate. They then state that we acknowledged that both the pedigree rate and aDNA rate estimates "greatly" exceed those inferred from fossil-calibrated analyses of birds. The first point is incorrect, and thus misrepresents our original work (Emerson & Hickerson 2015), as the 95% CI of one of the three published aDNA estimates of μ does not overlap with the pedigree-derived estimate of μ. The second point requires further context (see below) to understand the extent to which both pedigree and aDNA rates for Adélie penguins can be compared to a phylogenetic rate.

With regard to the first point, we stated in our original work (Emerson & Hickerson 2015) that the Adélie aDNA rate estimate of Ho *et al.* (2007a) is *significantly* higher than the pedigree rate. Thus, in contrary to the claim of Ho *et al.* (2015), there is no overlap among their 95% credibility intervals. We do not deny that the 95% credibility intervals of the aDNA rate estimates of Lambert *et al.* (2002) and Millar *et al.* (2008), which have a lower mean value than that of Ho *et al.* (2007a), overlap with the pedigree rate. However, this should not be seen as somehow undermining the discrepancy between these two aDNA rate estimates and the pedigree rate in a field (TDMR) where trends in mean values are frequently reported as support for the hypothesis.

With regard to the second point, we recognize that the mean values for all aDNA rate estimates and the pedigree-derived rate estimate of μ are higher than the bird phylogenetic divergence rate of 0.208 mutations/site/Myr presented by Shields and Wilson (1987) that has been used in previous comparisons (e.g. Lambert *et al.* 2002; Millar *et al.* 2008). However, there are several features of this phylogenetic rate estimate that limit its use for comparative purposes.

Firstly, it is not a general bird rate estimate, it is an estimate derived from the analysis of 5 species of geese. A difference between a phylogenetically derived mutation rate for geese, and aDNA or pedigree-derived rates for penguins may equally be explainable by fundamental differences between these very different, phylogenetically distant taxonomic groups. Secondly, the phylogenetic rate is probably underestimated, as recognised by Shields and Wilson (1987), due to the difficulty of estimating genetic divergences from restriction fragment analysis.

Comparing pedigree-derived rate estimates with phylogenetic rate estimates

We have previously pointed out, using *Caenorhabditis elegans* as an example, that a mutation accumulation line or pedigree-derived estimate of μ for a given taxa can only be considered high if it exceeds a taxonomically relevant phylogenetic rate (Emerson & Hickerson 2015). We provide an additional example of this problem above, with the inappropriate comparison of Adélie penguin pedigree and aDNA-derived estimates of μ with a phylogenetic estimate of μ derived from geese. Rather than providing suitable comparisons within their reply, Ho *et al.* (2015) continue to cite the spontaneous mutation rate for *C. elegans* (Denver *et al.* 2004), as well as *Drosophila melanagoster* (Keightley et al. 2014) and *Heliconius melpomene* (Keightley *et al.* 2015), as being higher than "corresponding phylogenetic estimates". There are no phylogenetic estimates within the response of Ho *et al.* (2015), nor within the original articles, with the exception of Keightley *et al.* (2015), who note that applying their spontaneous mutation rate to estimate the age of the *Heliconius* suggests that the fossil-calibrated age for the genus is approximately correct. The spontaneous rate is however higher than the fossil rate, and as pointed out by Keightley *et al.* (2015), further work is needed to reconcile the two estimates. But the difference itself is not evidence for the TDMR hypothesis when alternative equally plausible explanations exist. For example, a difference could arise because (1) the data sets being compared are very different (whole genome vs a non-random set of protein coding genes), or (2) only secondary calibration points were used for the phylogeny (i.e. there are no fossil Heliconiini). But let's assume the difference is real. What does it tell us? It tells us that purifying selection results in the underestimation of spontaneous μ when using a phylogenetic calibration. What it does not tell us is the timescale over which this occurs, and thus such data is uninformative about the timescale for the TDMR hypothesis.

Estimates of μ from temporally sampled DNA, and their lack of validation

Ho *et al.* (2015) take issue with our claim that, while many studies have produced estimates of μ from aDNA, none have provided validation of their estimates independently of the Bayesian implementation within BEAST (Drummond *et al.* 2012) from which they were derived (Emerson & Hickerson 2015). To support that we are “demonstrably wrong”, they cite two tests to evaluate the information content of time-structured data. However, these either have not provided, or do not provide, independent estimates of μ. The first of these, the regression of tree height against sampling time of Fitch *et al.* (1991) can, with some caveats, be used to estimate μ but has not, to our knowledge, ever been used to validate a Bayesian estimate of μ. The second test cited by Ho *et al.* (2015), that of Ramsden *et al.* (2009), which has been further developed by Duchêne *et al.* (2015), is not independent. It is a test of information content, where the Bayesian estimate of μ is compared to the distribution of μ estimated when dates are randomised across the tree. Thus, our original assertion still stands - Bayesian estimates of μ have yet to be independently validated.

Measurably evolving populations, date-randomisation and μ

Ho *et al.* (2015) provide a summary of the date-randomisation test, first presented by Ramsden *et al.* (2009) to test for sufficient signal within temporally sampled DNA data sets to estimate μ and divergence dates. It is important to consider what the 95% credibility interval of the date-randomised rate estimate represents. Ho *et al.* (2015) correctly point out that the two data sets presented in the schematic trees in Fig. 2 of Emerson & Hickerson (2015) would yield positive and misleading estimates of μ. We agree with this, but we do not agree with their conclusion that both data sets do not represent “measurably evolving populations”. On the contrary, both data sets do represent measurably evolving populations. The definition of genetic change in populations used by Ho *et al.* (2015) and elsewhere (e.g. Drummond *et al.* 2003; Ewing *et al.* 2004) is of mutation between sampling time points. However, it has been long understood that genetic change in populations involves changes in allele frequencies under the dynamic between mutation, selection and drift (Hartl & Clark 2007), and it is important to clarify that the mutation rate μ is the rate of mutation along any branch of a sampled gene genealogy, rather than being the rate of new mutations within a population or rate of mutational turnover between sampling time points. For example, due to the coalescent process, the vast majority of mutations between two temporally different samples can often occur at times older than either of the samples. As recognised by Ho *et al.* (2015), the sampling scenarios in panels C and D of Fig. 2 (Emerson & Hickerson 2015) will yield non-zero estimates of μ. Ho *et al.* (2015) also suggest that both data sets would fail the date-randomisation test of Ramsden *et al.* (2009). We agree that they probably would fail (although that can only be assessed by direct analysis). However, from this point we disagree with Ho *et al.* (2015), and the accepted interpretation of the date-randomisation test - that if the empirical estimate exceeds the 95% confidence intervals from the randomised distribution, then the empirical value is a reliable estimate of μ.

Regardless of whether a dataset passes the randomization test or not, estimates of μ from temporally sampled data using BEAST may be overestimated because of other population genetic (drift and the coalescent) and sampling processes, as well as phylogenetic constraints that BEAST imposes on temporally sampled data (Box 1). Citing Duchêne *et al.* (2015), Ho *et al.* (2015) point out that data sets that fail the date-randomisation test tend to yield overestimates of μ, which could be taken to suggest that data sets that pass the test provide meaningful approximations of μ. This is not the case. A careful examination of Duchêne *et al.* (2015) reveals that data sets can pass the test *and* yield significant overestimates of μ, where the the 95% confidence interval of the estimate does not include μ. In fact, the parameter space within which both the estimation of μ is correct, and the test is passed, is limited (Fig. 1 of Duchêne *et al.* 2015). The take home point is that passing the date-randomisation test is not validation for an estimation of μ using the BEAST temporally sampled model. To more fully explore this dynamic, we have conducted coalescent simulations of temporally sampled data, matching parameters commonly associated with ancient mtDNA data, and show that BEAST can systematically overestimate μ given temporally sampled data due to incorrect topological estimates that arise from constraining tip dates (Box 1).

TDMR for some genomes, and not for others?

Ho *et al.* (2015) suggest that there is scant evidence for an observed TDMR pattern in nuclear genomes. It will be interesting to see what is learned from new genomic data as it emerges, although it is worth pointing out that much of this observed discrepancy between nDNA and mtDNA evaporates if the studies using tip-dating methods with ancient mitochondrial DNA are confirmed to be the non-trivial overestimates as suggested from our simulation-based exploration. Furthermore, their assertion that “unfortunately, there remains considerable uncertainty about nuclear mutation rates in humans”, is vague and misleading, as the various papers show strong evidence that there is genetic variation for the mutation rate and that paternal age can drive differences in mutation rates (e.g. Scally & Durbin 2012; Thomas & Hahn 2014). It also seems somewhat incongruous for Ho *et al.* (2015) to criticise us for reporting short-term estimates of μ for nuclear data, while they themselves report such data when they believe it to support their argument (e.g. Denver *et al.* 2004; Keightley *et al.* 2014; Keightley *et al.* 2015, but see comments above).

Bison data and the Bayesian estimation of μ from temporally sampled DNA

Ho *et al.* (2015) cast doubt on two aspects of our reanalysis of the *Bison bison* data first published by Shapiro *et al.* (2004) and reanalysed by Ho *et al.* (2015). Their concerns regarding the impact of fixing effective populations size are vague and misleading, as they seem to suggest that there are "other parameters" in the cataclysmic demographic model that might somehow explain our results. As we have made all our input files publicly available, it is not clear why Ho *et al.* (2015) do not quantitatively assess their concern. A reanalysis exploring their parameters of concern would suffice. We therefore see nothing in the argument of Ho *et al.* (2015) regarding the fixing of modern effective population size for *B. bison*, that explains our results.

With regard to their other doubt, Ho *et al.* (2015) state that fixing the root age of the analysis explains our result because "removing the sequences from older samples to reduce the sampling window preferentially removes older branches in the gene tree". In doing so, Ho *et al.* (2015) assume a correlation between DNA sequence sampling time, and the coalescence time of the sampled sequence, which is in stark contrast to expectations under the standard Kingman coalescent for a single panmictic population without size change or subdivision (Tajima 1983). When we examined this assumption of Ho *et al.* (2015) it was apparent that, when compared to an unconstrained tree of the *B. bison* data, constraining the tree with tip dates positively contributes to such a correlation. The maximum clade credibility tree for the *B. bison* data with tip date constraints is topologically very different from the unconstrained tree, with DNA sequences of older age branching more basally within the tip date-constrained tree (Appendix S1, Supporting Information). As an explanation for this, we can only conclude that enforcing tip dates as a constraint contributes to the overestimation of μ, due to additional mutation change in the tree required to accommodate topological difference. We further explore these issues using coalescent simulations of temporally sampled data under a single panmictic population and find that indeed BEAST tends to incorrectly misestimate the gene genealogies as well as consistently overestimate μ given the sample size and temporal distribution of tips of the *B. bison* data (Box 1). Our analyses (Box 1) call into question all previous estimates of μ from tip-dated sequences using BEAST.

Ho *et al.* (2015) seem to be dismissive of their *B. bison* data, suggesting it to be small by current measures. It is in fact among the biggest data sets that have been analysed to date, providing an apparently compelling example of significance with respect to the date-randomisation test (Ho *et al.* 2011). Their argument that bigger data sets for a greater variety of genes will yield more decisive results will only be realised if the concerns we raise both here and in Emerson & Hickerson (2015) are taken on board. There are clear and identifiable problems with the estimation of μ from temporally sampled sequences, and not all these problems will necessarily be solved with more data.

Conclusions

After responding to the comment Ho *et al.* (2015), we find that our general criticisms of both (i) the hypothesis of time-dependent molecular evolution, and (ii) methods to estimate μ from temporally sampled DNA sequences, are further reinforced. As we have previously pointed out (Emerson & Hickerson 2015), much of the perceived support for the time-dependent molecular evolution hypothesis comes from overestimates of μ that are derived from phylogenetic analyses of temporally calibrated aDNA using the Bayesian program BEAST. Such estimates of μ have been argued to be evidence against calibration error as a sufficient explanation for patterns of TDMR (Ho *et al,* 2011). In this article we clearly identify a positive bias in the estimation of μ from tip-dated gene trees with BEAST that appears to be associated with the interaction between effective population size and enforcing the age of DNA sequences when reconstructing the topologies of the gene genealogies. Together with previously raised concerns (Debruyne & Poinar 2009; Emerson 2007; Emerson & Hickerson 2015; Navascués & Emerson 2009; Ramakrishnan & Hadly 2009) it is now clear that published estimates of μ using aDNA data should be considered unreliable, particularly if it cannot be shown that analyses underpinning the estimates did not result in topological differences between tip-date constrained and unconstrained trees. As we have pointed out, much of the remaining evidence for patterns of TDMR estimates can be explained without resorting to selection, suggesting no more than a limited temporal contribution of purifying selection to reconcile differences between spontaneous μ and phylogenetic estimates of μ.

Acknowledgements

We thank Sebastián Duchêne for providing access to simulation files from Duchêne *et al.* (2015).

References

Anderson CN, Ramakrishnan U, Chan YL, Hadly EA (2005) Serial SimCoal: a population genetics model for data from multiple populations and points in time. *Bioinformatics*, **21**, 1733-1734.

Debruyne R, Poinar HN (2009) Time dependency of molecular rates in ancient DNA data sets, a sampling artifact? *Systematic Biology*, **58**, 348-359.

Denver DR, Morris K, Lynch M, Thomas WK (2004) High mutation rate and predominance of insertions in the *Caenorhabditis elegans* nuclear genome. *Nature*, **430**, 679-682.

Drummond AJ, Pybus OG, Rambaut A, Forsberg R, Rodrigo AG (2003) Measurably evolving populations. *Trends in Ecology and Evolution*, **18**, 481-488.

Drummond AJ, Suchard MA, Xie D, Rambaut A (2012) Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Molecular Biology and Evolution*, **29**, 1969-1973.

Duchêne S, Duchêne D, Holmes EC, Ho SYW (2015) The performance of the date-randomisation test in phylogenetic analyses of time-structured virus data. *Molecular Biology and Evolution*, **32**, 1895-1906.

Emerson BC (2007) Alarm Bells for the molecular clock? No support for Ho et al.'s model of time-dependent molecular rate estimates. *Systematic Biology*, **56**, 337-345.

Emerson BC, Hickerson MJ (2015) Lack of support for the time-dependent molecular evolution hypothesis. *Molecular Ecology*, **24**, 702-709.

Ewing G, Nicholls G, Rodrigo A (2004) Using temporally spaced sequences to simultaneously estimate migration rates, mutation rate and population sizes in measurably evolving populations. *Genetics*, **168**, 2407-2420.

Fitch WM, Leiter JME, Li XQ, Palese P (1991) Positive Darwinian evolution in human influenza-a viruses. *Proceedings of the National Academy of Sciences of the United States of America*, **88**, 4270-4274.

Hartl DL, Clark AG (2007) *Principles of population genetics* Sinauer Associates, Sunderland, MA.

Ho SYW, Duchêne S, Molak M, Shapiro B (2015) Time-dependent estimates of molecular rates: Evidence and causes. *Molecular Ecology*, **this issue**.

Ho SYW, Kolokotronis S-O, Allaby RG (2007a) Elevated substitution rates estimated from ancient DNA sequences. *Biology Letters*, **3**, 702-705.

Ho SYW, Lanfear R, Phillips MJ*, et al.* (2011) Bayesian estimation of substitution rates from ancient DNA sequences with low information content. *Systematic Biology*, **60**, 366-374.

Ho SYW, Shapiro B, Phillips MJ, Cooper A, Drummond A (2007b) Evidence for time dependency of molecular rate estimates. *Systematic Biology*, **56**, 515-522.

Keightley PD, Ness RW, Halligan DL, Haddrill PR (2014) Estimation of the spontaneous mutation rate per nucleotide site in a *Drosophila melanogaster* full-sib family. *Genetics*, **196**, 313-320.

Keightley PD, Pinharanda A, Ness RW*, et al.* (2015) Estimation of the spontaneous mutation rate in *Heliconious melpomene*. *Molecular Biology and Evolution*, **32**, 239-243.

Kuhner MK, Yamato J, Felsenstein, J (1995) Estimating effective population size and mutation rate from sequence data using Metropolis-Hastings sampling. *Genetics*, **140**, 1421-1430.

Lambert DM, Ritchie PA, Millar CD*, et al.* (2002) Rates of evolution in ancient DNA from Adelie penguins. *Science*, **295**, 2270-2273.

Millar CD, Dodd A, Anderson JM*, et al.* (2008) Mutation and evolutionary rates in Adélie penguins from the Antarctic. *Plos Genetics*, **4**, e1000209.

Navascués M, Emerson BC (2009) Elevated substitution rate estimates from ancient DNA: model violation and bias of Bayesian methods. *Molecular Ecology*, **18**, 4390-4397.

Ramakrishnan U, Hadly EA (2009) Do complex population histories drive higher estimates of substitution rate in phylogenetic reconstructions? *Molecular Ecology*, **18**, 4341-4343.

Ramsden C, Holmes EC, Charleston MA (2009) Hantavirus evolution in relation to its rodent and insectivore hosts: No evidence for codivergence. *Molecular Biology and Evolution*, **26**, 143-153.

Scally A, Durbin R (2012) Revising the human mutation rate: implications for understanding human evolution. *Nature Reviews Genetics*, **13**, 745-753.

Shapiro B, Drummond AJ, Rambaut A*, et al.* (2004) Rise and fall of the beringian steppe bison. *Science*, **306**, 1561-1565.

Shields GF, Wilson AC (1987) Calibration of mitochondrial DNA evolution in geese. *Journal of Molecular Evolution*, **24**, 212-217.

Steel MA, Penny P (1993) Distributions of tree comparison metrics - some new results. *Systematic Biology*, **42**, 126-141.

Tajima F (1983) Evolutionary relationship of DNA sequences in finite populations. *Genetics*, **105**, 437-460.

Thomas GWC, Hahn MW (2014) The human mutation rate is increasing, even as it slows. *Molecular Biology and Evolution*, **31**, 253-257.

Author contributions

B.C.E., D.A.S. and M.J.H contributed equally to the preparation of this manuscript. All simulations were conducted by D.A.S.

Data Accessibility

All scripts for the simulations conducted within this manuscript and an example BEAST input file are available from https://diegofalvarado-s@bitbucket.org/diegofalvarado-s/tdmra_simulations.git. Bison DNA sequences and their sampling dates can be found within the online supporting information associated with Emerson & Hickerson (2015), doi: 10.1111/mec.13070.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Appendix S1 Maximum clade credibility trees for the *Bison bison* data of Ho *et al.* (2007a) with and without age constraints enforced for the tips.

Box 1: Tree shape and the overestimation of μ from tip-dated sequences

Constraining the tip dates within a phylogeny is expected to change branch lengths, but it might be less clear why topological relationships inferred from identical patterns of sequence variation should change. As can be seen in Appendix S1 (Supporting Information), the maximum clade credibility tree for *Bison bison* (Ho *et al.* 2007b; Shapiro *et al.* 2004) with tip date constraints is topologically very different from an unconstrained tree, with changes involving DNA sequences of older age branching more basally within the tip date-constrained tree, as would be expected if the panmictic effective population size was small. In some cases these rearrangements do not appear to increase the inferred amount of mutational change within the tree, as the change in gene tree topology does not disrupt patterns of shared derived variation, yet in other cases patterns of shared derived variation within the unconstrained tree are disrupted, increasing homoplasy and thus inferring additional mutational change within the tip-dated tree. One obvious outcome of an increase in the inferred number of mutational changes in a tip-date constrained tree is that the estimation of μ will also increase.

To explore this behavior, we followed a simulation procedure similar to that of Duchêne *et al.* (2015) — the main difference being the use of an explicit coalescent simulator, BayesSSC (Anderson *et al.* 2005) instead of BEAST (Drummond *et al.* 2012) to generate the input tree topologies given known effective population sizes (N) and mutation rates (μ), and a tip date distribution similar to the *B. bison* data (pipeline is available at https://bitbucket.org/diegofalvarado/tdmra_simulations). We have found that trees inferred by BEAST for tip-dated sequences tend to enforce an age-based coalescent pattern on the posterior distribution of gene trees. This pattern would be expected given small effective population sizes, despite true N being 483,827 and 1,451,481 individuals in the simulation models that generated the simulated datasets. One likely culprit is how the the compound demographic parameter ($\theta = 4N\mu$ where N is the effective population size and μ is the per site per generation per genealogical lineage mutation rate) is decoupled into joint estimates of N and μ in BEAST. Under a standard panmictic coalescent model, it is only possible to estimate the compound parameter θ rather than its components (N and μ) unless one of the two parameters are known or assumed (Kuhner et al. 1995). In contrast, the tip-dated panmictic coalescent model employed in BEAST allows decoupling the posterior estimates of θ into N and μ using the temporal-mutational information provided from the age-inforced tips of the posterior distribution of gene genealogies. As true N becomes larger, the tip-dated constraints result in inferred gene tree topologies that increasingly depart from the true gene tree topologies (Fig. I). This increasing level of phylogenetic inferential error corresponding with increasing levels of false homoplasy, which in turn corresponds with overestimates of μ and underestimates of *N*. In other words, underestimates of *N* result in older samples coalescing more basally than younger samples in the inferred topologies, and the consequences of this dynamic appear to be more severe when the true *N* was larger (Fig. I). As true *N* is larger, the magnitude of *N* underestimation and μ overestimation becomes more severe with inferred gene tree topologies becoming more age-constrained from the true topologies (Fig. II).

Of note is that under these simulations such overestimates of μ did not typically pass the date-randomisation test, yet this was less the case under the smaller true *N* (Fig. II). Under a coalescent model with small sized populations, one would expect genealogical coancestry between samples of similar age (i.e., age-based coalescence), and as expected, the simulations reveal that the probability of this is inversely related to population size (Figure III). At the same time, the randomization of tip ages has a stronger impact on rate estimates when disrupting patterns of age-based coalescence in the original tree, and hence, the date-randomization test is more likely passed when the true gene genealogy has a tighter age-coalescent time association (such as under relatively small effective population sizes; Figure III). Accordingly, as can be seen in Fig. IV, the association of coalescence time with sample age is much stronger for the bison data when compared to patterns obtained when simulating under a panmictic coalescent population model. Such a pattern is expected for population structure and/or small *N*. We suggest that even though the bison data was likely generated under scenarios that differed from what we explored in our simulations, the systematic overestimates of μ and underestimates of *N* are likely to still be at play with these estimates being biased by the consequences of large effective population sizes, population subdivision and/or local colonisation/extinction. Clearly this is in need of further evaluation with simulations that capture the demographic complexity and the patterns of tip-dates and coalescent times that are observed in real data.

Given that topological inconsistencies in BEAST appear to be associated with biasing estimates of both the number and age of DNA mutations together with overestimates of μ and underestimates of *N*, we make the following two suggestions. Firstly it would seem relevant to report the agreement between the topologies of tip-date constrained and unconstrained trees when reporting estimates of μ. Secondly, we suggest that while previous approaches using coalescent simulation have been useful to demonstrate that, under some conditions, BEAST can successfully estimate μ from tip-dated sequences of virus sequences (e.g. Duchêne *et al.* 2015), the complex conditions underlying temporally sampled ancient DNA with respect to sample sizes, effective population sizes, generation times, and subdivision need to be more fully examined to understand when estimates of μ from BEAST may be positively biased. Our simulations show that estimates of μ from such data can be systematically upwardly biased, and as such a more thorough exploration of the impacts of sample characteristics, historical demographics and analysis settings is needed to better understand the underlying causes of the methodological artifacts we have revealed. Our simulations also suggest that all previous estimates of μ from temporally sampled DNA sequence data using BEAST need a thorough reexamination before they can be accepted.

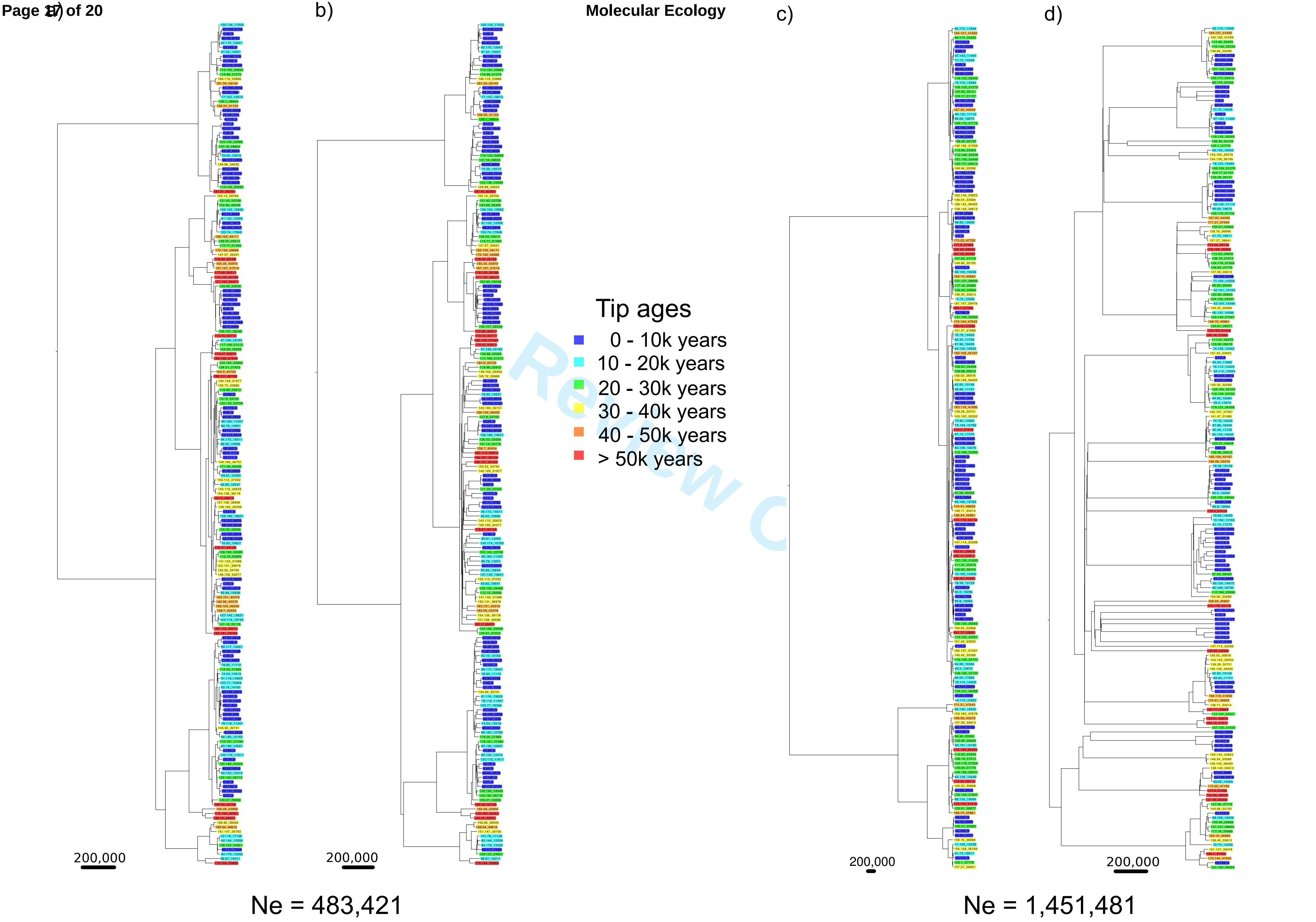
Figure I. Comparison of simulated and recovered tree topology for tip-dated sequence data using BEAST (Drummond *et al.* 2012). Note that the tree topology inferred by BEAST (b and d) is markedly different from the tree used to simulate the sequences (a and c) that serve as input to BEAST. This problem is accentuated under comparatively larger population sizes (Robinson-Foulds distance between a and b = 250, between c and d = 260; weighted-path difference (Steel & Penny 1993) between a and b = 0.77, between c and d = 20.09). Tips are coloured based on age to highlight the tendency for age-based coalescent events (i.e. tendency of younger samples to cluster as ingroups to older samples) in BEAST-estimated trees.

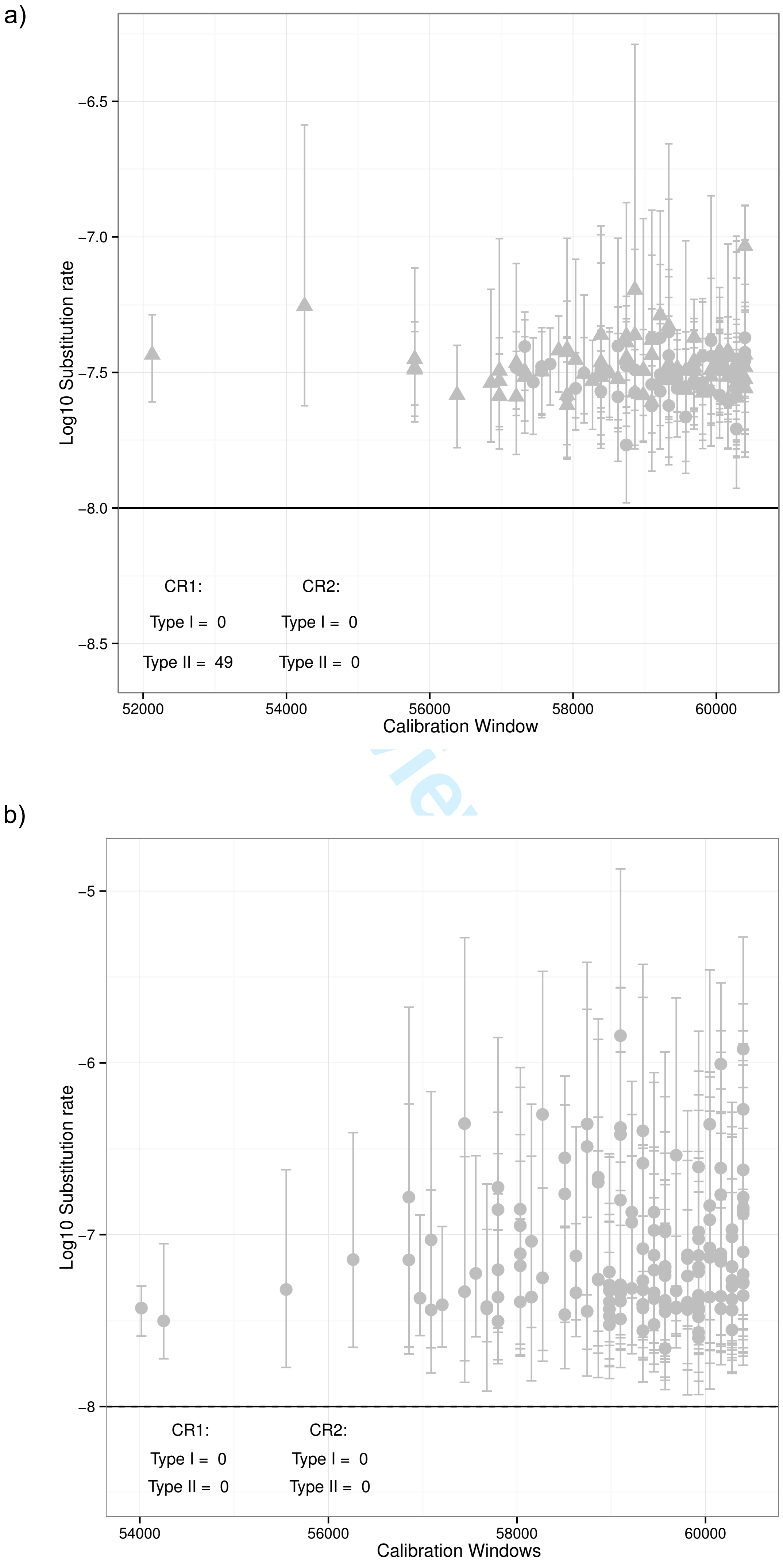
Figure II. Estimates of the substitution rate (in log 10 scale) against the width of the calibration window under two different populations sizes: (a) $N = 483{,}427$; (b) $N = 1{,}451{,}481$. The solid horizontal line represent the true simulated rate (mean=1e-8, sd=5%). Symbols represent the mean rate estimate for each simulation, with the error bars showing the 95% credible intervals. We conducted 10 randomizations for the date-randomization test for all data sets. Circles denote rate estimates that failed the test according to both criteria CR1 and CR2 (Duchêne *et al.* 2015), whereas triangles denote those that failed according to CR2 only. Numbers of type I and type II errors are shown for each rate treatment.

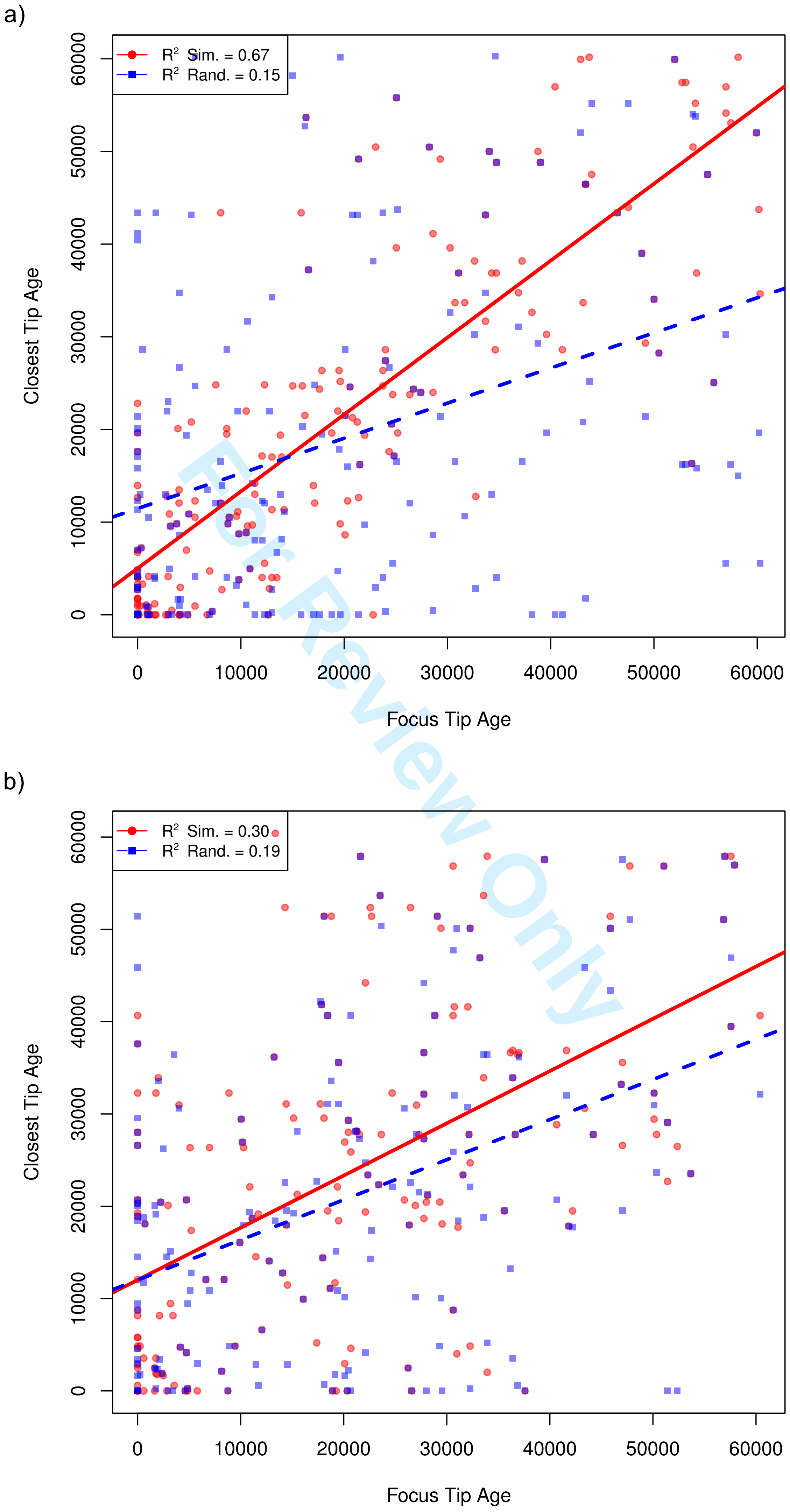
Figure III. Clustering of tip ages in BEAST-obtained trees based on simulated samples for (a) $N = 483{,}427$ and (b) $N = 1{,}451{,}481$. The ages of pairs of closest related tips is depicted, with original values represented in red, and date-randomised values represented in blue. Note how the difference between date-randomised and original data is smaller when the effective population size is comparatively larger, making it less likely to pass the test proposed by Duchene *et al.* (2015).

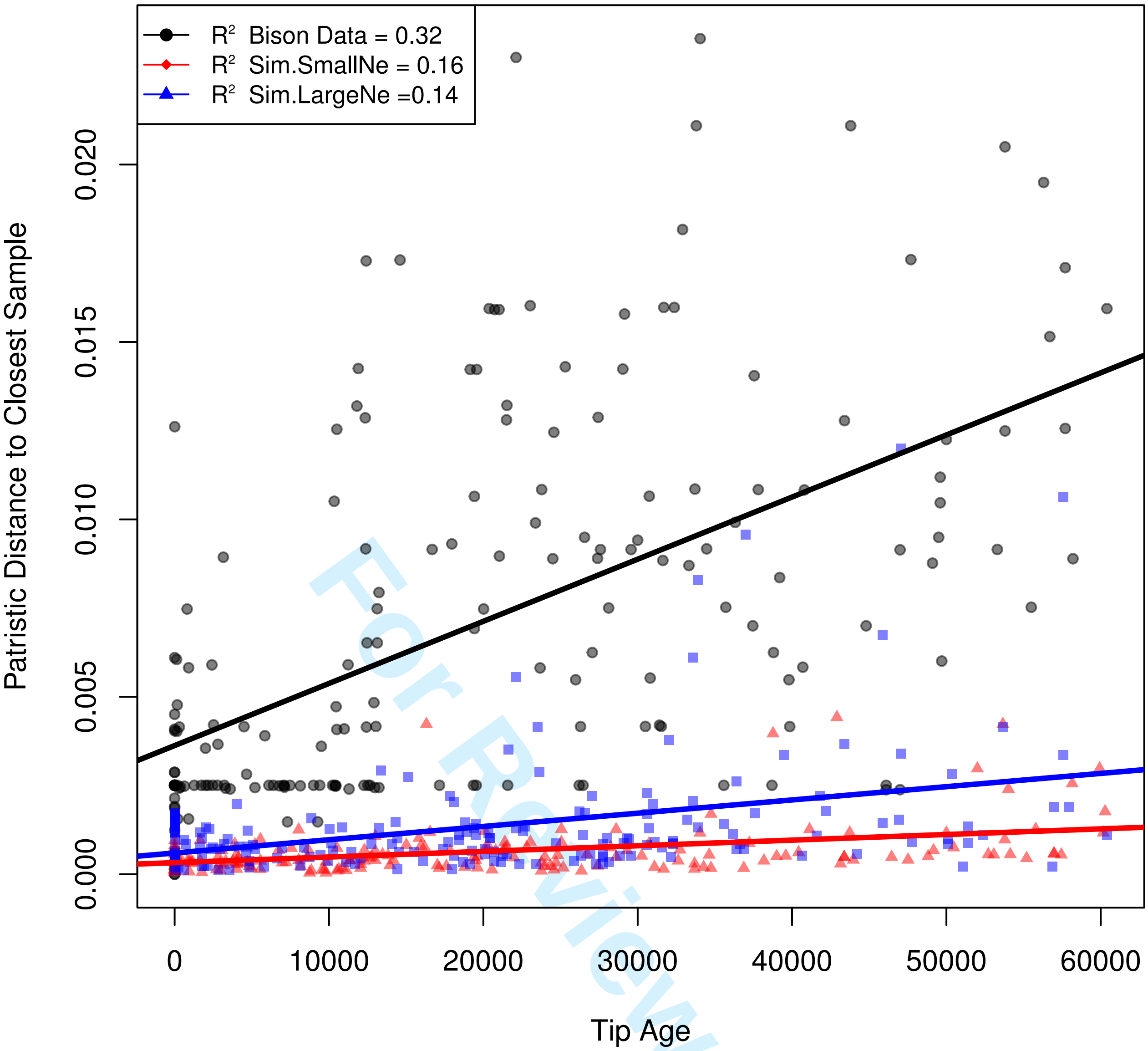
Figure IV. Association between tip-age and relative coalescence time. Patristic distance is used as an indicator for the time of coalescence of each sample in the tree. Note the empirical bison dataset (black) (Ho *et al.* 2007b) shows a much tighter association than any of the simulated datasets (small $N = 483{,}427$ in blue, large $N = 1{,}451{,}481$ in red) indicating a strong tendency for samples to

coalesce together based on their age in this dataset. Such a pattern is expected under small effective population sizes and/or population structure.


a)
b)
c)
d)
Tip ages
🟦 0 - 10k years
🟦 10 - 20k years
🟩 20 - 30k years
🟨 30 - 40k years
🟧 40 - 50k years
🟥 > 50k years
200,000
200,000
200,000
200,000
Ne = 483,421
Ne = 1,451,481


a)
−6.5
−7.0
−7.5
−8.0
−8.5
Log10 Substitution rate
CR1:
Type I = 0
Type II = 49
CR2:
Type I = 0
Type II = 0
52000
54000
56000
58000
60000
Calibration Window
b)
−5
−6
−7
−8
Log10 Substitution rate
CR1:
Type I = 0
Type II = 0
CR2:
Type I = 0
Type II = 0
54000
56000
58000
60000
Calibration Windows


a)
R² Sim. = 0.67
R² Rand. = 0.15
Closest Tip Age
0
10000
20000
30000
40000
50000
60000
Focus Tip Age
0
10000
20000
30000
40000
50000
60000
b)
R² Sim. = 0.30
R² Rand. = 0.19
Closest Tip Age
0
10000
20000
30000
40000
50000
60000
Focus Tip Age
0
10000
20000
30000
40000
50000
60000


R² Bison Data = 0.32
R² Sim.SmallNe = 0.16
R² Sim.LargeNe =0.14
Patristic Distance to Closest Sample
0.000
0.005
0.010
0.015
0.020
0
10000
20000
30000
40000
50000
60000
Tip Age