Comparing linked versus unlinked character models for species-tree inference

Kerry Cobb[1] and Jamie R. Oaks*[1]

[1]Department of Biological Sciences & Museum of Natural History, Auburn University, 101 Rouse Life Sciences Building, Auburn, Alabama 36849

March 24, 2023

*Corresponding author: joaks@auburn.edu

1 Introduction

Current model-based methods of species tree inference require biologists to make difficult decisions about their genomic data. They must decide whether to assume (1) sites in their alignments are each inherited independently ("unlinked"), or (2) groups of sites are inherited together ("linked"). If assuming the former, they must then decide whether to analyze all of their data or only putatively unlinked variable sites. Our goal in this chapter is to use simulated data to help guide these choices by comparing the robustness of different approaches to errors that are likely common in high-throughput genetic datasets.

Reduced-representation genomic data sets acquired from high-throughput instruments are becoming commonplace in phylogenetics (Leaché and Oaks, 2017), and usually comprise hundreds to thousands of loci from 50 to several thousand nucleotides long. Full likelihood approaches for inferring species trees from such datasets can be classified into two groups based on how they model the evolution of orthologous DNA sites along gene trees within the species tree—those that assume (1) each site evolved along its own gene tree (i.e., each site is "unlinked") (Bryant et al., 2012; Maio et al., 2015), or (2) contiguous, linked sites evolved along a shared gene tree (Liu and Pearl, 2007; Heled and Drummond, 2010; Ogilvie et al., 2017; Yang, 2015). We will refer to these as unlinked and linked-character models, respectively. For both models, the gene tree of each locus (whether each locus is a single site or a segment of linked sites) is assumed to be independent of the gene trees of all other loci, conditional on the species tree. Methods using linked character models become computationally expensive as the number of loci grows large, due to the estimation or numerical integration of all of the gene trees (Yang, 2015; Ogilvie et al., 2017). Unlinked-character models on the other hand are more tractable for a large number of loci, because estimating individual gene trees is avoided by analytically integrating over all possible gene trees (Bryant et al., 2012; Maio et al., 2015). Whereas unlinked-character models can accommodate a larger number of loci than linked-character models, most genetic data sets comprise linked sites and unlinked-character models are unable to utilize the aggregate information about ancestry contained in such linked sites.

Investigators are thus faced with decisions about how best to use their data to infer a species tree. Should they use a linked-character method that assumes the sites within each locus evolved along a shared gene tree? Ideally, the answer would be "yes," however this is not always computationally feasible and the model could be violated by intralocus recombination. Alternatively, should investigators remove all but one single-nucleotide polymorphism (SNP) from each locus and use an unlinked-character model? Or, perhaps they should apply the unlinked-character method to all of their sites, even if this violates the assumption that each site evolved along an independent gene tree. Important considerations in such decisions include the sources of error and bias that result from reduced-representation protocols, high-throughput sequencing technologies, and the processing of these data.

Most reduced-representation sequencing workflows employ amplification of DNA using polymerase chain reaction (PCR) which can introduce mutational error at a rate of up to 1.5×10^{-5} substitutions per base (Potapov and Ong, 2017). Furthermore, current high-throughput sequencing technologies have non-negligible rates of error. For example, Illumina sequencing platforms have been shown to have error rates as high as 0.25% per base (Pfeiffer et al., 2018). In hope of removing such errors, it is common for biologists to filter out variants

that are not found above some minimum frequency threshold (Rochette et al., 2019; Linck and Battey, 2019). The effect of this filtering will be more pronounced in data sets with low or highly variable coverage. Also, to avoid aligning paralogous sequences, it is common to remove loci that exceed an upper threshold on the number of variable sites (Harvey et al., 2015). These processing steps can introduce errors and acquisition biases, which have been shown to affect estimates derived from the assembled alignments (Harvey et al., 2015; Huang and Knowles, 2016; Linck and Battey, 2019). Given these issues are likely common in high-throughput genomic data, downstream decisions about what methods to use and what data to include in analyses should consider how sensitive the results might be to errors and biases introduced during data collection and processing.

Our goal is to determine whether linked and unlinked character models differ in their robustness to errors in reduced-representation genomic data, and whether it is better to use all sites or only SNPs for unlinked character methods. Linked-character models can leverage shared information among linked sites about each underlying gene tree. Thus, these models might be able to correctly infer the general shape and depth of a gene tree, even if the haplotypes at some of the tips have errors. Unlinked character models have very little information about each gene tree, and rely on the frequency of allele counts across many characters to inform the model about the relative probabilities of all possible gene trees. Given this reliance on accurate allele count frequencies, we predict that unlinked character models will be more sensitive to errors and acquisition biases in genomic data. To test this prediction that linked character models are more robust to the types of errors contained in reduced-representation data, we simulated data sets with varying degrees of errors related to miscalling rare alleles and heterozygous sites. Our results support this prediction, but also show that with only two species, the region of parameter space where there are differences between linked and unlinked character models is quite limited. Further work is needed to determine whether this difference in robustness between linked and unlinked character models will increase for larger species trees.

2 Methods

2.1 Simulations of error-free data sets

For our simulations, we assumed a simple two-tipped species tree with one ancestral population with a constant effective size of N_e^R that diverged at time τ into two descendent populations (terminal branches) with constant effective sizes of N_e^{D1} and N_e^{D2} (Fig. 1). For two diploid individuals sampled from each of the terminal populations (4 sampled gene copies per population), we simulated 100,000 orthologous biallelic characters under a finite-sites, continuous-time Markov chain (CTMC) model of evolution. We simulated 100 data sets comprised of loci of four different lengths—1000, 500, 250, and 1 characters. We assume each locus is effectively unlinked and has no intralocus recombination; i.e., each locus evolved along a single gene tree that is independent of the other loci, conditional on the species tree. We chose this simple species tree model for our simulations to help ensure any differences in estimation accuracy or precision were due to differences in the underlying linked and unlinked character models, and *not* due to differences in numerical algorithms for searching species and

gene tree space. Furthermore, we simulated biallelic characters, because unlinked-character multi-species coalescent models (Bryant et al., 2012; Oaks, 2019) that are most comparable to linked-character models (Heled and Drummond, 2010; Ogilvie et al., 2017) are limited to characters with (at most) two states.

We simulated the two-tipped species trees under a pure-birth process (Yule, 1925) with a birth rate of 10 using the Python package DendroPy (Version 4.40, Commit eb69003; Sukumaran and Holder, 2010). This is equivalent to the time of divergence between the two species being Exponentially distributed with a mean of 0.05 substitutions per site. We drew population sizes for each branch of the species tree from a Gamma distribution with a shape of 5.0 and mean of 0.002. We simulated 100, 200, 400, and 100,000 gene trees for data sets with loci of length 1000, 500, 250, and 1, respectively, using the contained coalescent implemented in DendroPy. We simulated linked biallelic character alignments using Seq-Gen (Version 1.3.4) (Rambaut and Grass, 1997) with a GTR model with base frequencies of A and C equal to 0 and base frequencies of G and T equal to 0.5. The transition rate for all base changes was 0, except for the rate between G and T which was 1.0.

2.2 Introducing Site-pattern Errors

From each simulated dataset containing linked characters described above, we created four datasets by introducing two types of errors at two levels of frequency. The first type of error we introduced was changing singleton character patterns (i.e., characters for which one gene copy was different from the other seven gene copies) to invariant patterns by changing the singleton character state to match the other gene copies. We introduced this change to all singleton site patterns with a probability of 0.2 and 0.4 to create two datasets from each simulated dataset. The second type of error we introduced was missing heterozygous gene copies. To do this, we randomly paired gene copies from within each species to create two diploid genotypes for each locus, and with a probability of 0.2 or 0.4 we randomly replaced one allele of each genotype with the other. For the unlinked character dataset comprised of a single site per locus, we only simulated singleton character pattern error at a probability of 0.4.

2.3 Assessing Sensitivity to Errors

For each simulated data set with loci of 250, 500, and 1000 characters, we approximated the posterior distribution of the divergence time (τ) and effective population sizes (N_e^R , N_e^{D1} , and N_e^{D2}) under an unlinked-character model using ecoevolity (Version 0.3.2, Commit a7e9bf2; Oaks, 2019) and a linked-character model using the StarBEAST2 package (Version 0.15.1; Ogilvie et al., 2017) in BEAST2 (Version 2.5.2; Bouckaert et al., 2014). For both methods, we specified a CTMC model of character evolution and prior distributions that matched the model and distributions from which the data were generated. The prior on the effective size of the root population in the original implementation of ecoevolity was parameterized to be relative to the mean effective size of the descendant populations. We added an option to ecoevolity to compile a version where the prior is specified as the absolute effective size of the root population, which matches the model in StarBEAST2 and the model we used to generate the data. The linkage of sites within loci of our simulated data violates

the unlinked-character model of ecoevolity (Bryant et al., 2012; Oaks, 2019). Therefore, we also analyzed each data set with ecoevolity after selecting, at most, one variable character from each locus; loci without variable sites were excluded.

We analyzed the data sets simulated with 1-character per locus (i.e., unlinked data) with ecoevolity. Our goal with these analyses was to verify that the generative model of our simulation pipeline matched the underlying model of ecoevolity, and to confirm that any behavior of the method with the other simulated data sets was not being caused by the linkage violation.

For ecoevolity, we ran four independent Markov chain Monte Carlo (MCMC) analyses with 75,000 steps and a sample frequency of 50 steps. For StarBEAST2, we ran two independent MCMC analyses with 20 million steps and a sample frequency of 5000 steps. To assess convergence and mixing of the ecoevolity and StarBEAST2 MCMC chains, we computed the effective sample size (ESS; Gong and Flegal, 2016) and potential scale reduction factor (PSRF; the square root of Equation 1.1 in Brooks and Gelman, 1998) from the samples of each parameter, and considered an ESS value greater than 200 and PSRF less than 1.2 (Brooks and Gelman, 1998) to indicate adequate convergence and mixing of the chains. Based on preliminary analyses of simulated data sets without errors, we chose to discard the first 501 and 201 samples from the MCMC chains of ecoevolity and StarBEAST2, leaving 4000 and 7600 posterior samples for each data set, respectively.

2.4 Project repository

The full history of this project has been version-controlled and is available at https://github.com/kerrycobb/align-error-sp-tree-sim, and includes all of the data and scripts necessary to produce our results.

3 Results

3.1 Behavior of linked (StarBEAST2) versus unlinked (ecoevolity) character models

The divergence times estimated by the linked-character method, StarBEAST2, were very accurate and precise for all alignment lengths and types and degrees errors, despite poor MCMC mixing (i.e., low ESS values) for shorter loci (Figs. 2–4). For data sets without error, the unlinked-character method, ecoevolity, estimated divergence times with similar accuracy and precision as StarBEAST2 when all characters are analyzed (Figs. 2–4). However when alignments contained errors, ecoevolity underestimated very recent divergence times with increasing severity as the frequency of errors increased (Figs. 2–4); estimates of older divergence times were unaffected.

The biased underestimation of divergence times by ecoevolity in the face of errors was coupled with overestimation of the ancestral effective population sizes (Figs. 5–7). When analyzing the alignments without errors, ecoevolity essentially returned the prior distribution on the effective size of the ancestral population (Figs. 5–7). Despite poor MCMC mixing, StarBEAST2 consistently estimated the effective size of the ancestral population better

than ecoevolity and was unaffected by errors in the data (Figs. 5–7), and the precision of StarBEAST2's estimates of N_e^R increased with locus length.

Estimates of the effective size of the descendant populations are largely similar between StarBEAST2 and ecoevolity; both methods underestimate the descendant population sizes when the data sets contain errors, and this downward bias is generally worse for ecoevolity (Figs. 8–10). The degree of underestimation increases with the rate of errors in the data sets for both StarBEAST2 and ecoevolity, and the results were largely consistent across different locus lengths. (Figs. 8–10).

When we apply ecoevolity to data sets simulated with unlinked characters (i.e., data sets simulated with 1-character per locus), we see the same patterns of biased parameter estimates in response to errors (Fig. 11) as we did with the linked loci (Figs. 2–4). These results rule out the possibility that the greater sensitivity of ecoevolity to the errors we simulated is due to violation of the method's assumption that all characters are unlinked.

3.2 Analyzing all sites versus SNPs with ecoevolity

The unlinked character model implemented in ecoevolity assumes that ortholgous nucleotide sites evolve independently along separate gene trees. The data however, were simulated under a model assuming that contiguous linked sites evolve along a shared gene tree. It would thus be a violation of the ecoevolity model to include all sites in the analysis. However, avoiding this violation by removing all but one variable site per locus drastically reduces the amount of data. When analyzing the simulated data sets without errors, the precision and accuracy of parameter estimates by ecoevolity was much greater when all sites of the alignment were used relative to when a single SNP per locus was used despite violating the model (Figs. 2–10). This was generally true across the different lengths of loci, however, the coverage of credible intervals is lower with longer loci. Analyzing only SNPs does make ecoevolity more robust to the errors we introduced. However, this robustness is due to the lack of information in the SNP data leading to wide credible intervals, and in the case of population size parameters, the marginal posteriors essentially match the prior distribution (Figs. 8–10).

3.3 Coverage of credible intervals

The 95% credible intervals for divergence times and effective population sizes estimated from alignments without error in StarBEAST2 had the expected coverage frequency in that the true value was within approximately 95% of the estimated credible intervals. This was also true for ecoevolity when analyzing data sets simulated with unlinked characters (i.e., no linked sites). This coverage behavior is expected, and helps to confirm confirm that our simulation pipeline generated data under the same model used for inference by StarBEAST2 and ecoevolity. As seen previously (Oaks, 2019), analyzing longer linked loci causes the coverage of ecoevolity to be lower, due to the violation of the model's assumption that the sites are unlinked.

3.4 MCMC convergence and mixing

Most sets of StarBEAST2 and ecoevolity MCMC chains yielded samples of parameters with a PSRF less than 1.2, indicative of convergence. However, we do see poor mixing (ESS < 200) of the StarBEAST2 chains as the length of loci decreases (Figs. 2–10; yellow indicates ESS < 200, red indicates PSRF > 1.2, green indicates both) We only see evidence of poor mixing and convergence for ecoevolity when applied to data sets with errors. This is in contrast to StarBEAST2, for which the frequency and degree of poor MCMC behavior is largely unaffected by the type or frequency of errors. The proportion of simulation replicates where StarBEAST2 had an ESS of the ancestral population size less than 200 was high across all analyses (Figs. 5–7). For the descendant population size, StarBEAST2 had better ESS values across all analyses, with the exception of rare estimates of essentially zero when analyzing 250 bp loci (Figs. 8–10).

4 Discussion

Phylogeneticists seeking to infer species trees from large, multi-locus data sets are faced with difficult decisions regarding assumptions about linkage across sites and, if assuming all sites are unlinked, what data to include in their analysis. With the caveat that we only explored trees with two species, the results of our simulations provide some guidance for these decisions. As we predicted, the linked-character method we tested, StarBEAST2, was more robust to the sequencing errors we simulated than the unlinked character method, ecoevolity. However, even with only two species in our simulations, the current computational limitations of linked-character models was apparent from the poor sampling efficiency of the MCMC chains, especially with shorter loci. For data sets with more species and many short loci, linked character models are theoretically appealing, but current implementations may not be computationally feasible. The unlinked character method, ecoevolity, was more sensitive to sequence errors, but was still quite robust to realistic levels of errors and is more computationally feasible thanks to the analytical integration over gene trees.

Overall, for data sets with relatively long loci, as is common with sequence-capture approaches, it might be worth trying a linked-character method. If computationally practical, you stand to benefit from the aggregate information about each gene tree contained in the linked sites of each locus. However, if your loci are shorter, as in restriction-site-associated DNA (RAD) markers, you are likely better off applying an unlinked-character model to all of your data, even though this violates an assumption of the model. Below we discuss why performance differs between methods, locus lengths, and degree of error in the data, and what this means for the analyses of empirical data.

4.1 Robustness to character-pattern errors

As predicted, the linked-character model of StarBEAST2 was more robust to erroneous character patterns in the alignments than the unlinked-character model of ecoevolity. This is most evident in the estimates of divergence times, for which the two methods perform very similarly when there are no errors in the data (Row 1 of Figs. 2–4). When errors are introduced, the divergence time estimates of StarBEAST2 are unaffected, but ecoevolity

underestimates recent divergence times as both singleton and heterozygosity errors become more frequent (Rows 2–5 of Figs. 2–4). However, ecoevolity divergence-time estimates are only biased at very recent divergence times, and the effect disappears when the time of divergence is larger than about $8N_e\mu$.

These patterns make sense given that both types of errors we simulated reduce variation *within* each species. Thus, it is not too surprising that the unlinked-character model in ecoevolity struggles when there is shared variation between the two populations (i.e., most gene trees have more than two lineages that coalesce in the ancestral population). The erroneous character patterns mislead both models that the effective size of the descendant branches is smaller than they really are (Figs. 8–10). To explain the shared variation between the species (i.e. deep coalescences) when underestimating the descendant population sizes, the unlinked-character model of ecoevolity simultaneously reduces the divergence time and increases the effective size of the ancestral population. Despite also being misled about the size of the descendant populations (Figs. 8–10), the linked-character model of StarBEAST2 seems to benefit from more information about the general shape of each gene tree across the linked sites and can still maintain an accurate estimate of the divergence time (Figs. 2–4)and ancestral population size (Figs. 5–7).

This downward biased variation within each species becomes less of a problem for the unlinked-character model as the divergence time gets larger, likely because the average gene tree only has a single lineage from each species that coalesces in the ancestral population. As the coalesced lineage within each species leading back to the ancestral population becomes a large proportion of the overall length of the average gene tree, the proportion of characters that either show fixed differences between the species or are invariant likely provides enough information to the unlinked character model about the time of divergence to overcome the downward biased estimates of the descendant population sizes.

From the ecoevolity results, we also see that when faced with heterozygosity errors, accuracy decreases as locus length increases. In contrast, accuracy of ecoevolity is not affected by locus length when analyzing data sets with singleton erros. This pattern makes sense in light of how we generated these errors. We introduced singleton errors per-site and heterozygosity errors per-locus. Thus, the same per-locus rate of heterzygosity errors affects many more sites of a dataset with 1000bp loci compared to dataset with 250bp loci.

Unsurprisingly, the MCMC sampling performance of StarBEAST2 declines with decreasing locus length. There is less information in the shorter loci about ancestry, and thus more posterior uncertainty about the gene trees. This forces StarBEAST2 to traverse a much broader distribution of gene trees during MCMC sampling, which is difficult due to the constraints imposed by the species tree. This decline in MCMC performance in StarBEAST2 does not appear to correlate with poor parameter estimates and the distribution of estimates is generally as good or better than those from ecoevolity. However, this might be due to fact that there is no uncertainty in the species tree in any of our analyses, because there are only two species. As the number of species increases, it seems likely that the MCMC performance will further decline and start to affect parameter and topology estimates.

4.2 Relevance to empirical data sets

It is reassuring to see the effect of sequence errors on the unlinked-character model is limited to a small region of parameter space, and is only severe when the frequency of errors in the data is large. Our simulated error rate of 40% is likely higher than the rate that these types of errors occur during most sample preparation, high-throughput sequencing, and bioinformatic processing. However, empirical alignments likely contain a mix of different sources of errors and biases from various steps in the data collection process. Also, real data are not be generated under a known model with no prior misspecification. Violations of the model might make these methods of species-tree inference more sensitive to lower rates of error.

The degree to which a dataset will be affected by errors from missing heterozygote haplotypes and missing singletons will be highly dependent on the method used to reduce representation of the genome, depth of sequencing coverage (i.e., the number of overlapping sequence reads at a locus), and how the data are processed. To filter out sequencing errors, most pipelines for processing sequence reads set a minimum coverage threshold for variants or a minimum minor allele frequency. This can result in the miscalling or removal of true variation, especially if coverage is low due to random chance or biases in PCR amplification and sequencing. Processing the data in this way can result in biased estimates of parameters that are sensitive to the frequencies of rare alleles (Huang and Knowles, 2016; Linck and Battey, 2019). If the thresholds for such processing steps are stringent, it could introduce levels of error greater than our simulations.

4.3 Recommendations for using unlinked-character models

When erroneous character patterns cause ecoevolity to underestimate the divergence time it also inflates the effective population size of the ancestral population. We are seeing values of $N_e^R\mu$ consistent with an average sequence divergence between individuals *within* the ancestral population of 3%, which is almost an order of magnitude larger than our prior mean expectation (0.4%). Thus, looking for unrealistically large population sizes estimated for internal branches of the phylogeny might provide an indication that the unlinked-character model is not explaining the data well. However, there is little information in the data about the effective population sizes along ancestral branches, so the parameter that might indicate a problem is going to have very large credible intervals. Nonetheless, many of the posterior estimates of the ancestral population size from our data sets simulated with character-pattern errors are well beyond the prior distribution.

Whether using linked or unlinked-character models with empirical high-throughput data sets, it is good practice to perform analyses on different versions of the aligned data that are assembled under different coverage thresholds for variants or alleles. Variation of estimates derived from different assemblies of the data might indicate that the model is sensitive to the errors or acquisition biases in the alignments. This is especially true for data where sequence coverage is low for samples and/or loci. Given our findings, it might be helpful to compare the estimates of the effective population sizes along internal branches of the tree. Seeing unrealistically large estimates for some assemblies of the data might indicate that the model is being biased by errors or acquisition biases present in the character patterns.

Consistent with what has been shown in previous work (Oaks, 2019; Oaks et al., 2019), ecoevolity performed better when all sites were utilized despite violating the assumption that all sites are unlinked. This suggests that investigators might obtain better estimates by analyzing all their data under unlinked-character models, rather than discarding much of it to avoid violating an assumption of the model. Given that the model of unlinked characters implemented in ecoevolity does not use information about linkage among sites (Bryant et al., 2012; Oaks, 2019), it is not surprising that this model violation does not introduce a bias. Linkage among sites does not change the gene trees and site patterns that are expected under the model, but it does reduce the variance of the those patterns due to them evolving along fewer gene trees. As a result, the accuracy of the parameter estimates is not affected by the linkage among sites within loci, but the credible intervals become too narrow as the length of loci increase (Oaks, 2019; Oaks et al., 2019). However, it remains to be seen whether the robustness of the model's accuracy to linked sites holds true for larger species trees.

4.4 Other complexities of empirical data in need of exploration

Our goal was to compare the theoretical performance of linked and unlinked character models, not their current software implementations. Accordingly, to minimize differences in performance that are due to differences in algorithms for exploring the space of gene and species trees, we restricted our simulations to two species model and a small number of individuals. Nonetheless, exploring how character-pattern errors and biases affect the inference of larger species trees would be informative. The species tree topology is usually a parameter of great interest to biologists, so it would be interesting to know whether the linked model continues to be more robust to errors than the unlinked model as the number of species increases. We saw the MCMC performance of StarBEAST2 decline concomitantly with locus length in our simulations due to greater uncertainty in gene trees. Given that data sets frequently contain loci shorter than 250 bp, it is important to know whether good sampling of the posterior of linked-character models becomes prohibitive for larger trees. Also, ecoevolity greatly overestimated the effective size of the ancestral population in the face of high rates of errors in the data. Exploring larger trees will also determine whether this behavior is limited to the root population or is a potential problem for all internal branches of the specie tree.

Exploring other types of errors and biases would also be informative. To generate alignments of orthologous loci from high-throughput data, sequences are matched to a similar portion of a reference sequence or clustered together based on similarity. To avoid aligning paralogous sequences it is necessary to establish a minimum level of similarity for establishing orthology between sequences. This can lead to an acquisition bias due to the exclusion of more variable loci or alleles from the alignment (Huang and Knowles, 2016). Furthermore, when a reference sequence is used, this data filtering will not be random with respect to the species, but rather there will be a bias towards filtering loci and alleles with greater sequence divergence from the reference. Simulations exploring the affect of these types of data acquisition biases would complement the errors we explored here.

In our analyses, there was no model misspecification other than the introduced errors (except for the linked sites violating the unlinked-character model). With empirical data, there are likely many model violations, and our prior distributions will never match the

distributions that generated the data. Introducing other model violations and misspecified prior distributions would thus help to better understand how species-tree models behave on real data sets. Of particular concern is whether misspecified priors will amplify the effect of character-pattern errors or biases.

We found that character-pattern errors that remove variation from within species can cause unlinked-character models to underestimate divergence times and overestimate ancestral population sizes in order to explain shared variation among species. This raises the question of whether we can explicitly model and correct for these types of data collection errors in order to avoid biased parameter estimates. An approach that could integrate over uncertainty in the frequency of these types of missing-allele errors would be particularly appealing.

5 Acknowledgments

This work was supported by the National Science Foundation (grant number DEB 1656004 to JRO). Most of the computational work for this project was performed on the Auburn University Hopper Cluster. This work is contribution number 938 of the Auburn University Museum of Natural History.

References

Bouckaert, R., J. Heled, D. Kühnert, T. Vaughan, C.-H. Wu, D. Xie, M. A. Suchard, A. Rambaut, and A. J. Drummond. 2014. BEAST 2: A Software Platform for Bayesian Evolutionary Analysis. PLoS Computational Biology 10:e1003537.

Brooks, S. P. and A. Gelman. 1998. General methods for monitoring convergence of iterative simulations. Journal of Computational and Graphical Statistics 7:434–455.

Bryant, D., R. Bouckaert, J. Felsenstein, N. A. Rosenberg, and A. RoyChoudhury. 2012. Inferring Species Trees Directly from Biallelic Genetic Markers: Bypassing Gene Trees in a Full Coalescent Analysis. Molecular Biology and Evolution 29:1917–1932.

Gong, L. and J. M. Flegal. 2016. A practical sequential stopping rule for high-dimensional Markov chain Monte Carlo. Journal of Computational and Graphical Statistics 25:684–700.

Harvey, M. G., C. D. Judy, G. F. Seeholzer, J. M. Maley, G. R. Graves, and R. T. Brumfield. 2015. Similarity thresholds used in DNA sequence assembly from short reads can reduce the comparability of population histories across species. PeerJ 3:e895.

Heled, J. and A. J. Drummond. 2010. Bayesian Inference of Species Trees from Multilocus Data. Molecular Biology and Evolution 27:570–580.

Huang, H. and L. L. Knowles. 2016. Unforeseen consequences of excluding missing data from next-generation sequences: Simulation study of rad sequences. Systematic biology 65:357–365.

Hunter, J. D. 2007. Matplotlib: A 2D graphics environment. Computing In Science & Engineering 9:90–95.

Leaché, A. D. and J. R. Oaks. 2017. The Utility of Single Nucleotide Polymorphism (SNP) Data in Phylogenetics. Annual Review of Ecology, Evolution, and Systematics 48:69–84.

Linck, E. and C. J. Battey. 2019. Minor allele frequency thresholds strongly affect population structure inference with genomic data sets. Molecular Ecology Resources 19:639–647.

Liu, L. and D. K. Pearl. 2007. Species Trees from Gene Trees: Reconstructing Bayesian Posterior Distributions of a Species Phylogeny Using Estimated Gene Tree Distributions. Systematic Biology 56:504–514.

Maio, N. D., D. Schrempf, and C. Kosiol. 2015. PoMo: An Allele Frequency-Based Approach for Species Tree Estimation. Systematic Biology 64:14.

Oaks, J. R. 2019. Full Bayesian Comparative Phylogeography from Genomic Data. Systematic Biology 68:371–395.

Oaks, J. R., C. D. Siler, and R. M. Brown. 2019. The comparative biogeography of philippine geckos challenges predictions from a paradigm of climate-driven vicariant diversification across an island archipelago. Evolution 73:1151–1167.

Ogilvie, H. A., R. R. Bouckaert, and A. J. Drummond. 2017. StarBEAST2 Brings Faster Species Tree Inference and Accurate Estimates of Substitution Rates. Molecular Biology and Evolution 34:2101–2114.

Pfeiffer, F., C. Gröber, M. Blank, K. Händler, M. Beyer, J. L. Schultze, and G. Mayer. 2018. Systematic evaluation of error rates and causes in short samples in next-generation sequencing. Scientific Reports 8.

Potapov, V. and J. L. Ong. 2017. Examining Sources of Error in PCR by Single-Molecule Sequencing. PLOS ONE 12:e0169774.

Rambaut, A. and N. C. Grass. 1997. Seq-Gen: An application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. Bioinformatics 13:235–238.

Rochette, N. C., A. G. Rivera-Colón, and J. M. Catchen. 2019. Stacks 2: Analytical methods for paired-end sequencing improve RADseq-based population genomics. Molecular Ecology 28:4737–4754.

Sukumaran, J. and M. T. Holder. 2010. DendroPy: A Python library for phylogenetic computing. Bioinformatics 26:1569–1571.

Yang, Z. 2015. The BPP program for species tree estimation and species delimitation. Current Zoology 61:854–865.

Yule, G. U. 1925. A mathematical theory of evolution, based on the conclusions of Dr. J. C. Willis, F.R.S. Philosphical Transactions of the Royal Society B 213:21–87.

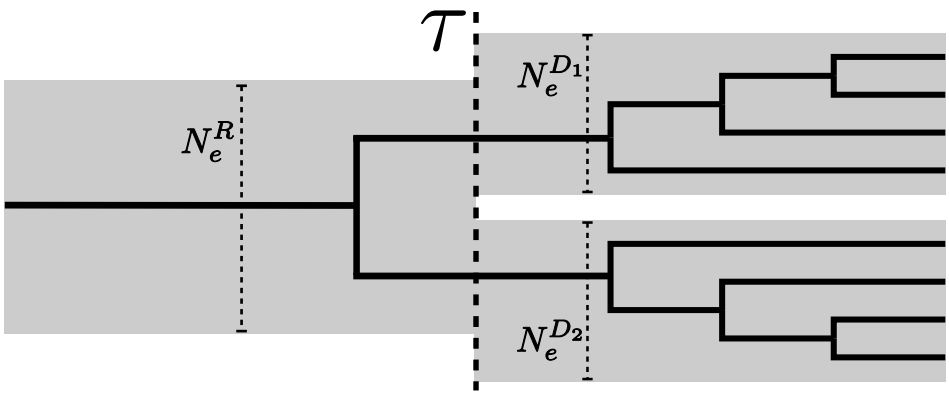

Figure 1. An illsustration of the species-tree model we used to simulate data. N_e^R , N_e^{D1} , and N_e^{D2} represent the constant effective population sizes of the root, and each of the two terminal populations. τ reperesents the instantaneous separation of the ancestral population into two descendant populations. One hypothetical gene tree is shown to illustrate the gene trees simulated under a contained coalescent process for 4 haploid gene copies sampled from each of the terminal branches of the species tree.

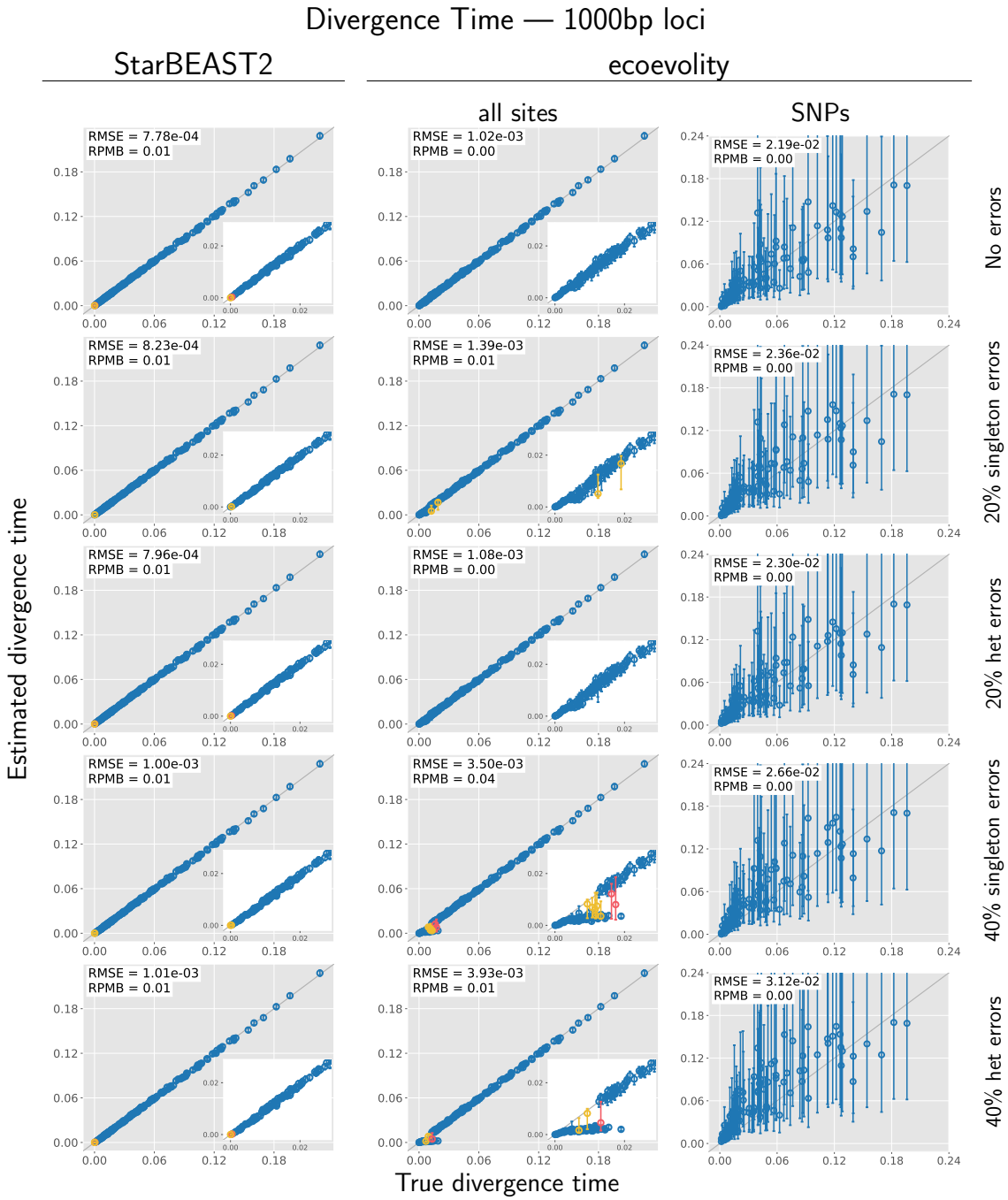

Figure 2. **Accuracy and precision of divergence-time estimates (in units of expected substitutions per site) with 1000 base pair loci.** The left column shows estimates from StarBEAST2, and the center and right column shows estimates from ecoevolity using all sites and (at most) one SNP per locus. The top row shows estimates from 200 data sets simulated without character-pattern errors. Rows labelled 20% and 40% singleton errors show estimates from the same alignments after singleton site patterns were changed to invariant sites with probabilities 0.2 and 0.4, respectively. Rows labelled 20% and 40% het errors show estimates from the same (error-free) alignments after we randomly paired gene copies within each species into two diploid genotypes, and for each genotype we randomly replaced one allele with the other with probability 0.2 and 0.4, respectively. Each plotted circle and associated error bars represent the posterior mean and 95% credible interval. Circles and error bars are colored yellow if the effective sample size (ESS) of the estimate was less than 200, red if the potential scale reduction factor (PSRF) was greater than 1.2, and green if both conditions were true. The root mean square error (RMSE) and rate of poor MCMC behavior (RPMB) is given for each plot, the latter of which is the proportion of estimates with ESS < 200 and/or PSRF > 1.2. We generated the plots using matplotlib Version 3.1.1 (Hunter, 2007).

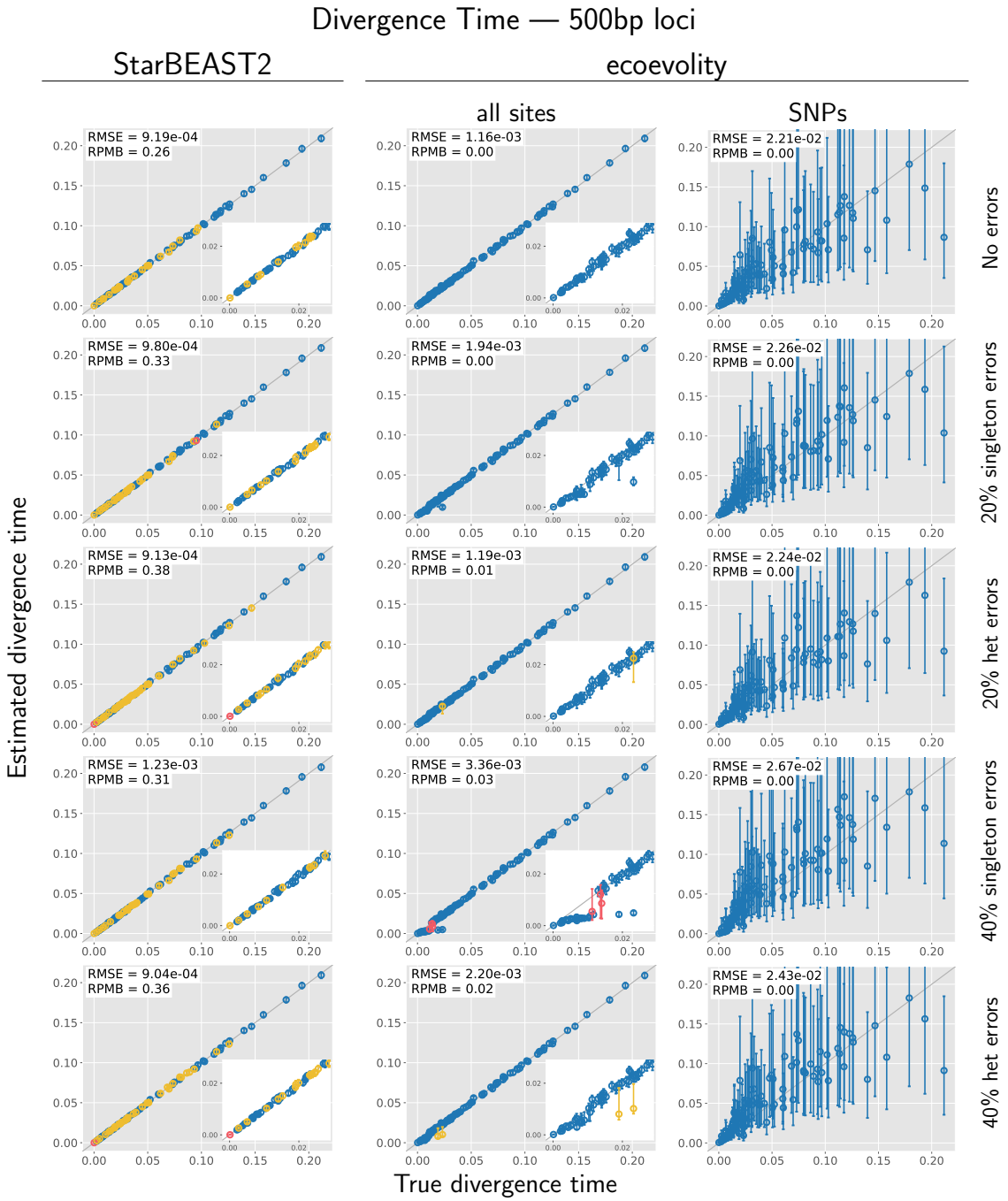

Figure 3. **Accuracy and precision of divergence-time estimates (in units of expected substitutions per site) with 500 base pair loci.** The left column shows estimates from StarBEAST2, and the center and right column shows estimates from ecoevolity using all sites and (at most) one SNP per locus. The top row shows estimates from 200 data sets simulated without character-pattern errors. Rows labelled 20% and 40% singleton errors show estimates from the same alignments after singleton site patterns were changed to invariant sites with probabilities 0.2 and 0.4, respectively. Rows labelled 20% and 40% het errors show estimates from the same (error-free) alignments after we randomly paired gene copies within each species into two diploid genotypes, and for each genotype we randomly replaced one allele with the other with probability 0.2 and 0.4, respectively. Each plotted circle and associated error bars represent the posterior mean and 95% credible interval. Circles and error bars are colored yellow if the effective sample size (ESS) of the estimate was less than 200, red if the potential scale reduction factor (PSRF) was greater than 1.2, and green if both conditions were true. The root mean square error (RMSE) and rate of poor MCMC behavior (RPMB) is given for each plot, the latter of which is the proportion of estimates with ESS < 200 and/or PSRF > 1.2. We generated the plots using matplotlib Version 3.1.1 (Hunter, 2007).

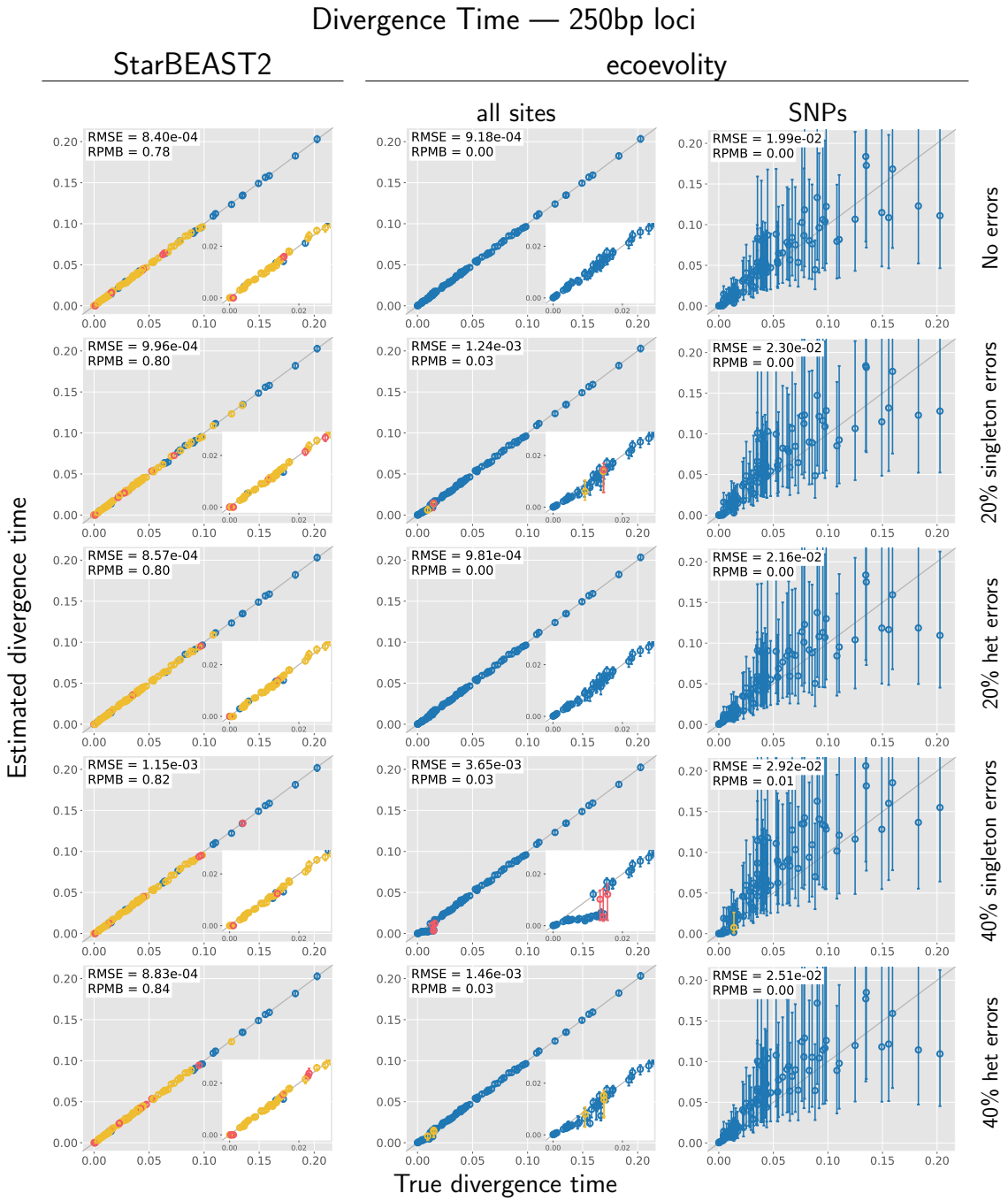

Figure 4. **Accuracy and precision of divergence-time estimates (in units of expected substitutions per site) with 250 base pair loci.** The left column shows estimates from StarBEAST2, and the center and right column shows estimates from ecoevolity using all sites and (at most) one SNP per locus. The top row shows estimates from 200 data sets simulated without character-pattern errors. Rows labelled 20% and 40% singleton errors show estimates from the same alignments after singleton site patterns were changed to invariant sites with probabilities 0.2 and 0.4, respectively. Rows labelled 20% and 40% het errors show estimates from the same (error-free) alignments after we randomly paired gene copies within each species into two diploid genotypes, and for each genotype we randomly replaced one allele with the other with probability 0.2 and 0.4, respectively. Each plotted circle and associated error bars represent the posterior mean and 95% credible interval. Circles and error bars are colored yellow if the effective sample size (ESS) of the estimate was less than 200, red if the potential scale reduction factor (PSRF) was greater than 1.2, and green if both conditions were true. The root mean square error (RMSE) and rate of poor MCMC behavior (RPMB) is given for each plot, the latter of which is the proportion of estimates with ESS < 200 and/or PSRF > 1.2. We generated the plots using matplotlib Version 3.1.1 (Hunter, 2007).

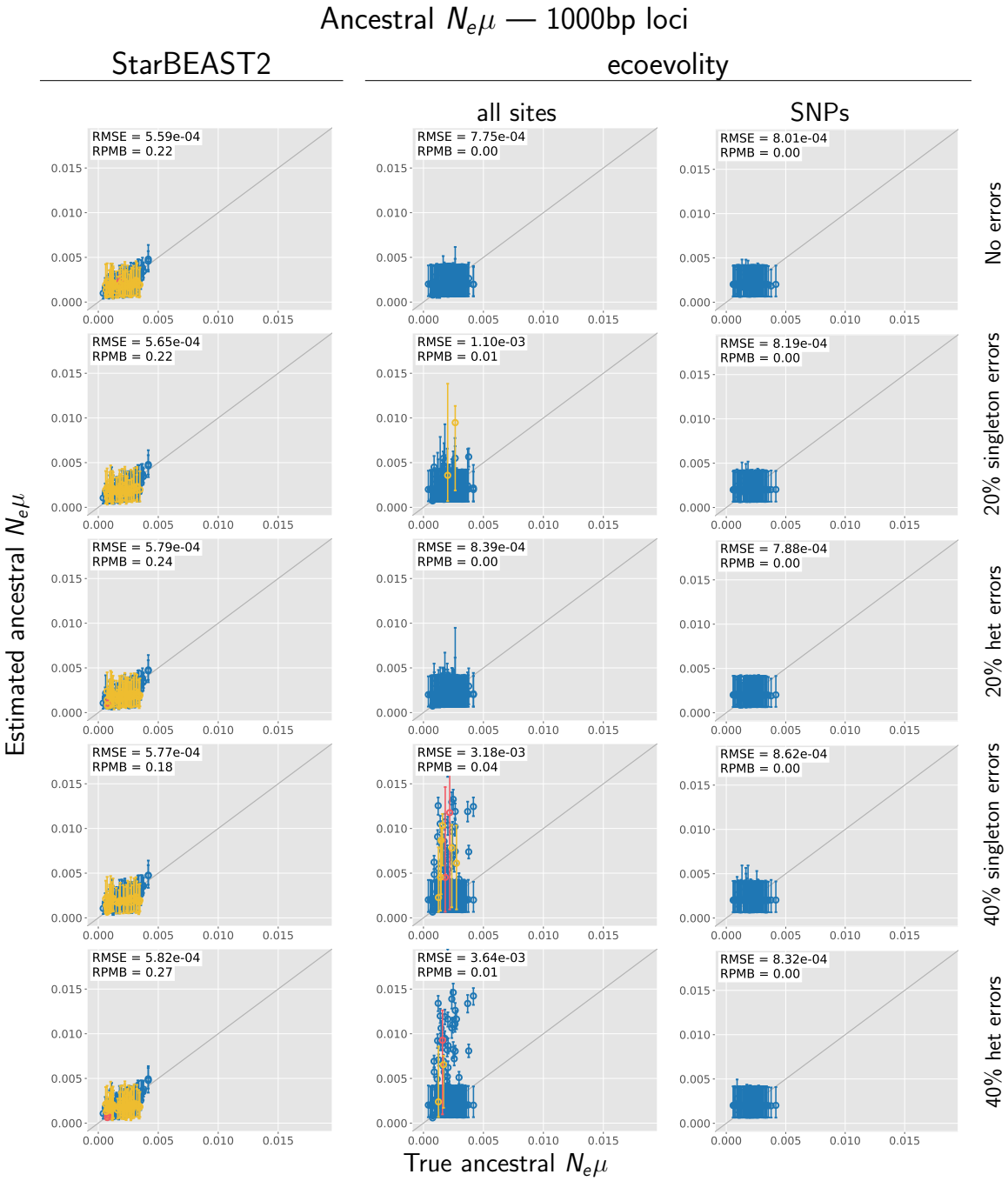

Figure 5. **Accuracy and precision of estimates of root effective population size scaled by the mutation rate ($N_e^R\mu$) with 1000 base pair loci.** The left column shows estimates from StarBEAST2, and the center and right column shows estimates from ecoevolity using all sites and (at most) one SNP per locus. The top row shows estimates from 200 data sets simulated without character-pattern errors. Rows labelled 20% and 40% singleton errors show estimates from the same alignments after singleton site patterns were changed to invariant sites with probabilities 0.2 and 0.4, respectively. Rows labelled 20% and 40% het errors show estimates from the same (error-free) alignments after we randomly paired gene copies within each species into two diploid genotypes, and for each genotype we randomly replaced one allele with the other with probability 0.2 and 0.4, respectively. Each plotted circle and associated error bars represent the posterior mean and 95% credible interval. Circles and error bars are colored yellow if the effective sample size (ESS) of the estimate was less than 200, red if the potential scale reduction factor (PSRF) was greater than 1.2, and green if both conditions were true. The root mean square error (RMSE) and rate of poor MCMC behavior (RPMB) is given for each plot, the latter of which is the proportion of estimates with ESS < 200 and/or PSRF > 1.2. We generated the plots using matplotlib Version 3.1.1 (Hunter, 2007).

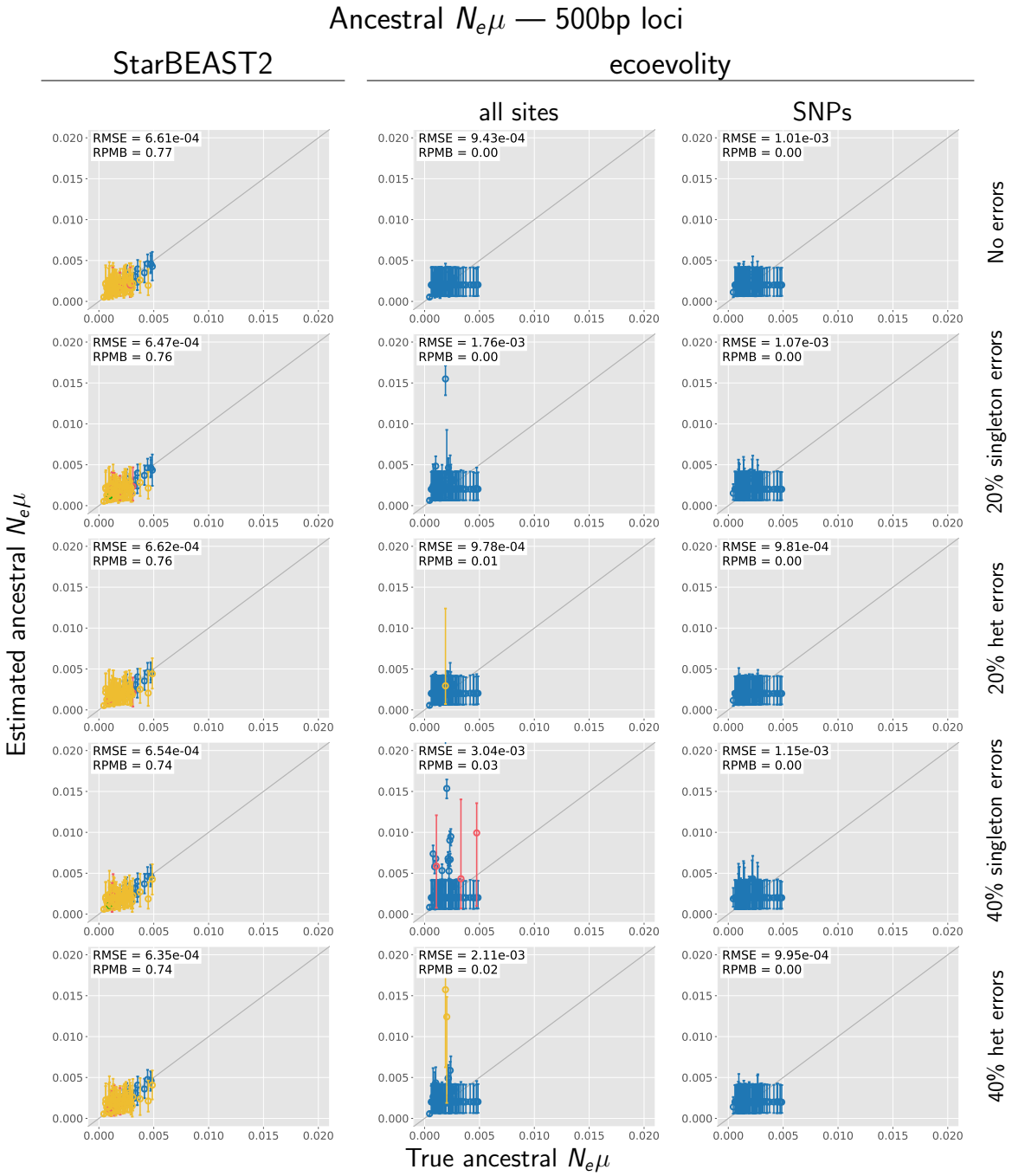

Figure 6. **Accuracy and precision of estimates of root effective population size scaled by the mutation rate ($N_e^R\mu$) with 500 base pair loci.** The left column shows estimates from StarBEAST2, and the center and right column shows estimates from ecoevolity using all sites and (at most) one SNP per locus. The top row shows estimates from 200 data sets simulated without character-pattern errors. Rows labelled 20% and 40% singleton errors show estimates from the same alignments after singleton site patterns were changed to invariant sites with probabilities 0.2 and 0.4, respectively. Rows labelled 20% and 40% het errors show estimates from the same (error-free) alignments after we randomly paired gene copies within each species into two diploid genotypes, and for each genotype we randomly replaced one allele with the other with probability 0.2 and 0.4, respectively. Each plotted circle and associated error bars represent the posterior mean and 95% credible interval. Circles and error bars are colored yellow if the effective sample size (ESS) of the estimate was less than 200, red if the potential scale reduction factor (PSRF) was greater than 1.2, and green if both conditions were true. The root mean square error (RMSE) and rate of poor MCMC behavior (RPMB) is given for each plot, the latter of which is the proportion of estimates with ESS < 200 and/or PSRF > 1.2. We generated the plots using matplotlib Version 3.1.1 (Hunter, 2007).

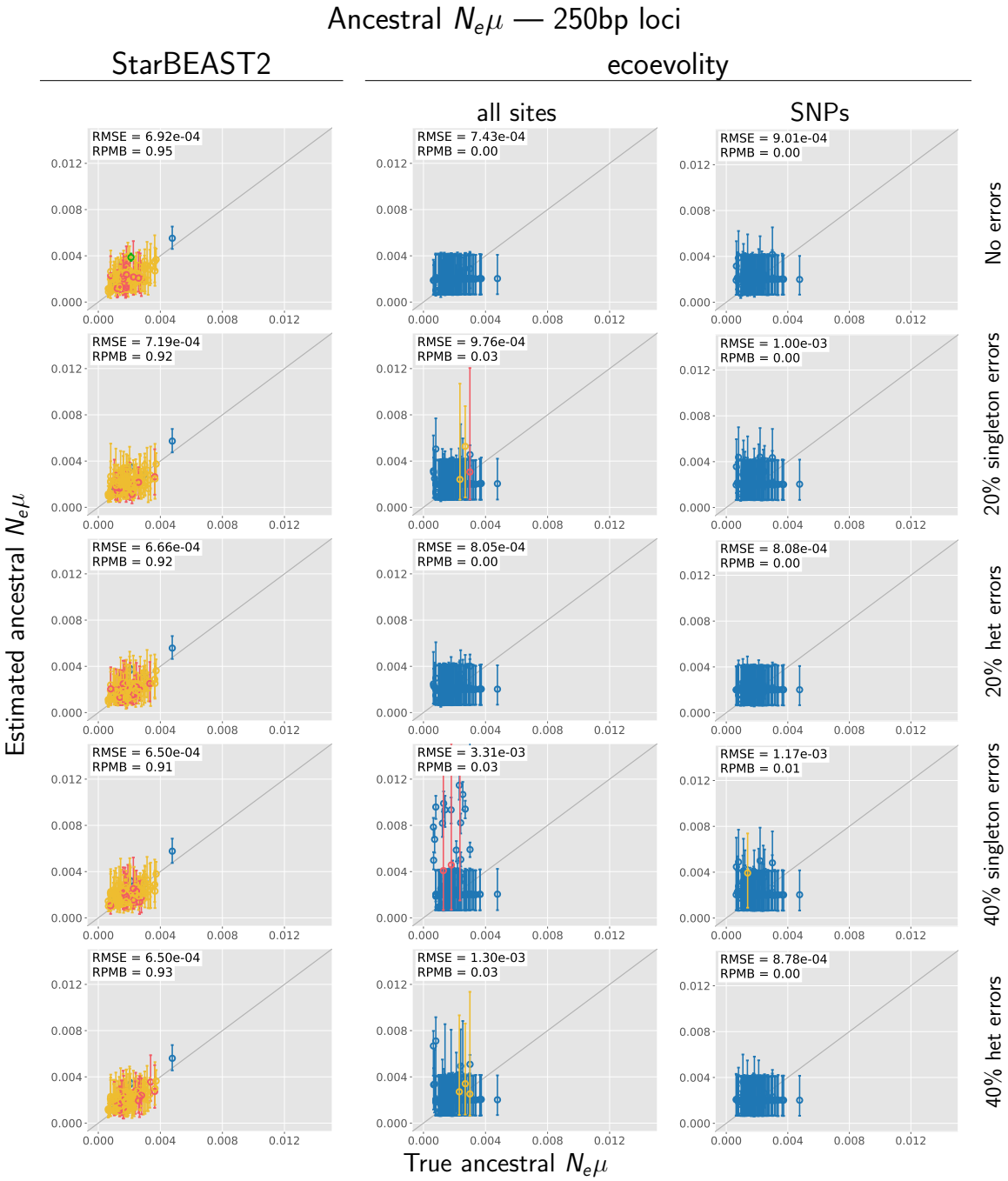

Figure 7. **Accuracy and precision of estimates of root effective population size scaled by the mutation rate ($N_e^R\mu$) with 250 base pair loci.** The left column shows estimates from StarBEAST2, and the center and right column shows estimates from ecoevolity using all sites and (at most) one SNP per locus. The top row shows estimates from 200 data sets simulated without character-pattern errors. Rows labelled 20% and 40% singleton errors show estimates from the same alignments after singleton site patterns were changed to invariant sites with probabilities 0.2 and 0.4, respectively. Rows labelled 20% and 40% het errors show estimates from the same (error-free) alignments after we randomly paired gene copies within each species into two diploid genotypes, and for each genotype we randomly replaced one allele with the other with probability 0.2 and 0.4, respectively. Each plotted circle and associated error bars represent the posterior mean and 95% credible interval. Circles and error bars are colored yellow if the effective sample size (ESS) of the estimate was less than 200, red if the potential scale reduction factor (PSRF) was greater than 1.2, and green if both conditions were true. The root mean square error (RMSE) and rate of poor MCMC behavior (RPMB) is given for each plot, the latter of which is the proportion of estimates with ESS < 200 and/or PSRF > 1.2. We generated the plots using matplotlib Version 3.1.1 (Hunter, 2007).

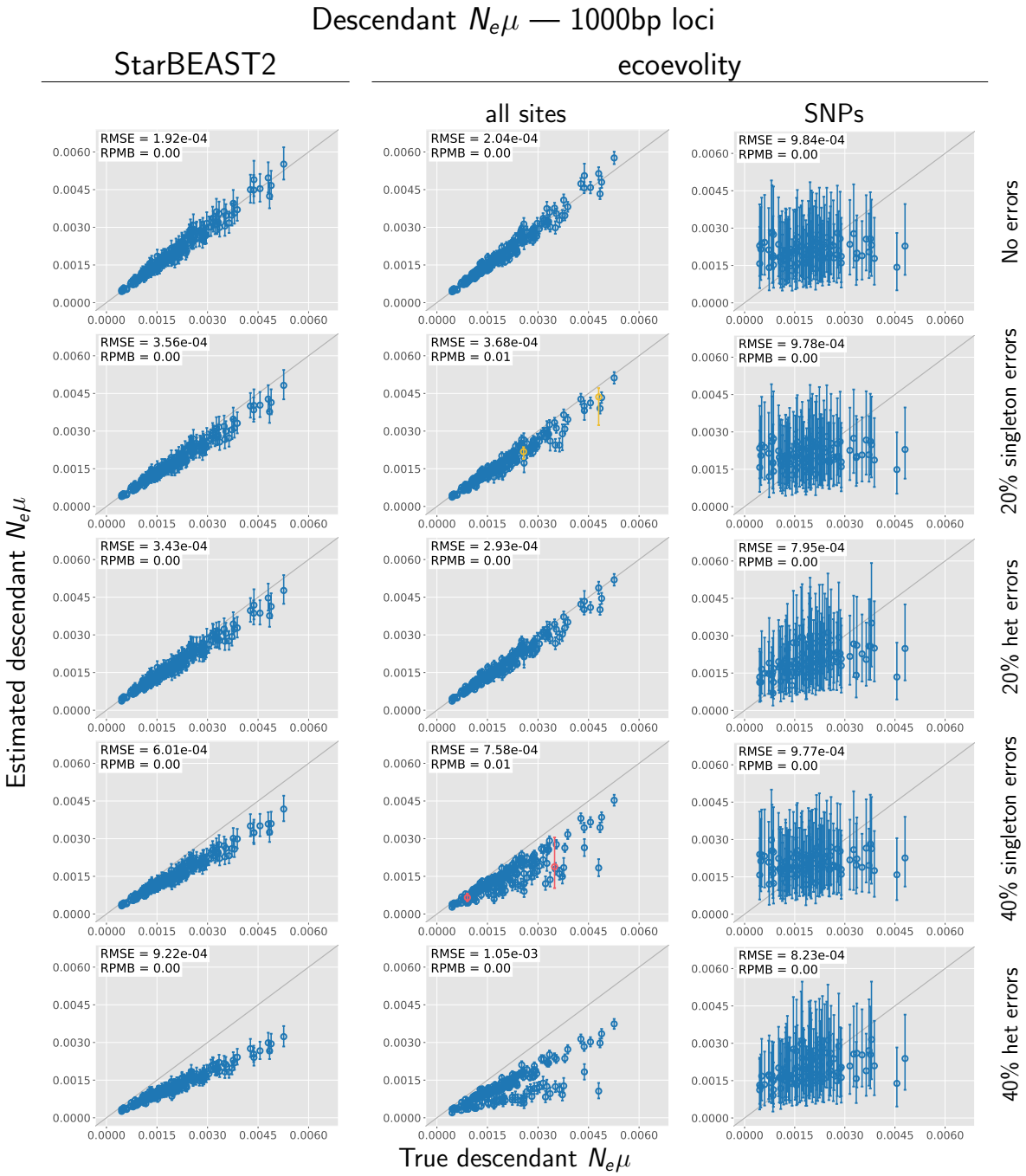

Figure 8. **Accuracy and precision of estimates of effective population sizes of the descendant branches of the tree scaled by the mutation rate ($N_e^D\mu$) with 1000 base pair loci.** The left column shows estimates from StarBEAST2, and the center and right column shows estimates from ecoevolity using all sites and (at most) one SNP per locus. The top row shows estimates from 200 data sets simulated without character-pattern errors. Rows labelled 20% and 40% singleton errors show estimates from the same alignments after singleton site patterns were changed to invariant sites with probabilities 0.2 and 0.4, respectively. Rows labelled 20% and 40% het errors show estimates from the same (error-free) alignments after we randomly paired gene copies within each species into two diploid genotypes, and for each genotype we randomly replaced one allele with the other with probability 0.2 and 0.4, respectively. Each plotted circle and associated error bars represent the posterior mean and 95% credible interval. Circles and error bars are colored yellow if the effective sample size (ESS) of the estimate was less than 200, red if the potential scale reduction factor (PSRF) was greater than 1.2, and green if both conditions were true. The root mean square error (RMSE) and rate of poor MCMC behavior (RPMB) is given for each plot, the latter of which is the proportion of estimates with ESS $<$ 200 and/or PSRF $>$ 1.2. We generated the plots using matplotlib Version 3.1.1 (Hunter, 2007).

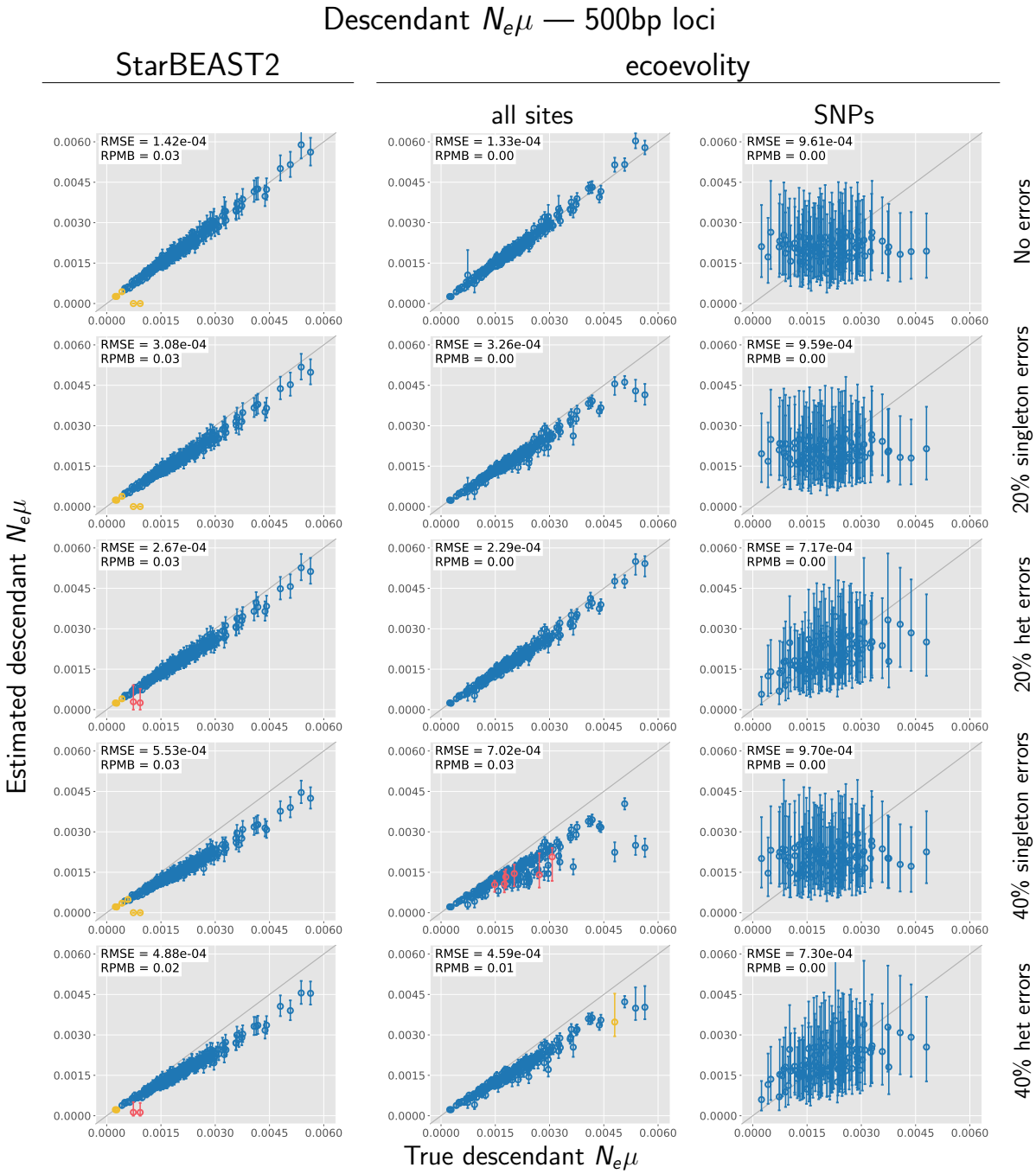

Figure 9. **Accuracy and precision of estimates of effective population sizes of the descendant branches of the tree scaled by the mutation rate ($N_e^D\mu$) with 500 base pair loci.** The left column shows estimates from StarBEAST2, and the center and right column shows estimates from ecoevolity using all sites and (at most) one SNP per locus. The top row shows estimates from 200 data sets simulated without character-pattern errors. Rows labelled 20% and 40% singleton errors show estimates from the same alignments after singleton site patterns were changed to invariant sites with probabilities 0.2 and 0.4, respectively. Rows labelled 20% and 40% het errors show estimates from the same (error-free) alignments after we randomly paired gene copies within each species into two diploid genotypes, and for each genotype we randomly replaced one allele with the other with probability 0.2 and 0.4, respectively. Each plotted circle and associated error bars represent the posterior mean and 95% credible interval. Circles and error bars are colored yellow if the effective sample size (ESS) of the estimate was less than 200, red if the potential scale reduction factor (PSRF) was greater than 1.2, and green if both conditions were true. The root mean square error (RMSE) and rate of poor MCMC behavior (RPMB) is given for each plot, the latter of which is the proportion of estimates with ESS < 200 and/or PSRF > 1.2. We generated the plots using matplotlib Version 3.1.1 (Hunter, 2007).

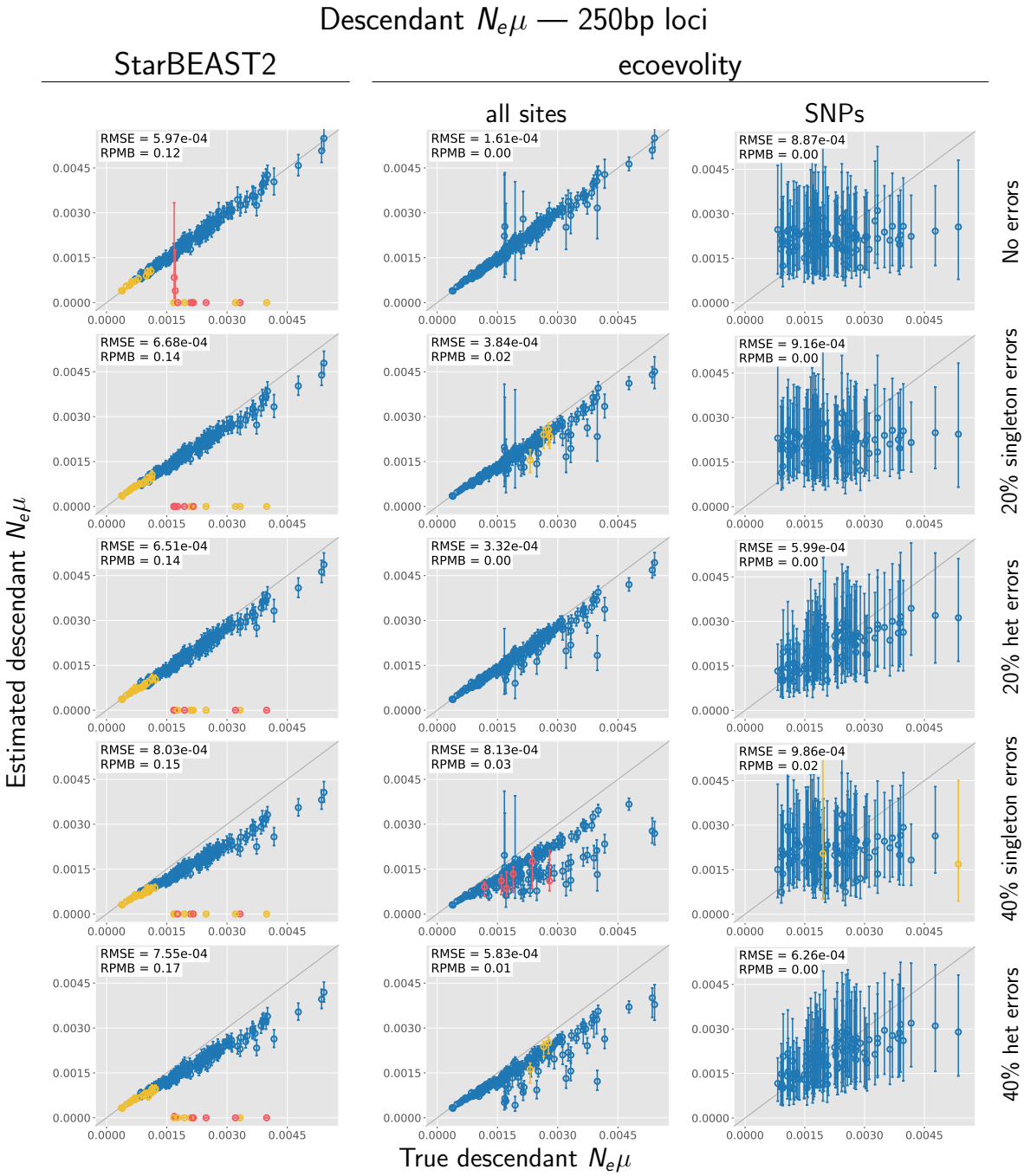

Figure 10. **Accuracy and precision of estimates of effective population sizes of the descendant branches of the tree scaled by the mutation rate ($N_e^D\mu$) with 250 base pair loci.** The left column shows estimates from StarBEAST2, and the center and right column shows estimates from ecoevolity using all sites and (at most) one SNP per locus. The top row shows estimates from 200 data sets simulated without character-pattern errors. Rows labelled 20% and 40% singleton errors show estimates from the same alignments after singleton site patterns were changed to invariant sites with probabilities 0.2 and 0.4, respectively. Rows labelled 20% and 40% het errors show estimates from the same (error-free) alignments after we randomly paired gene copies within each species into two diploid genotypes, and for each genotype we randomly replaced one allele with the other with probability 0.2 and 0.4, respectively. Each plotted circle and associated error bars represent the posterior mean and 95% credible interval. Circles and error bars are colored yellow if the effective sample size (ESS) of the estimate was less than 200, red if the potential scale reduction factor (PSRF) was greater than 1.2, and green if both conditions were true. The root mean square error (RMSE) and rate of poor MCMC behavior (RPMB) is given for each plot, the latter of which is the proportion of estimates with ESS $<$ 200 and/or PSRF $>$ 1.2. We generated the plots using matplotlib Version 3.1.1 (Hunter, 2007).

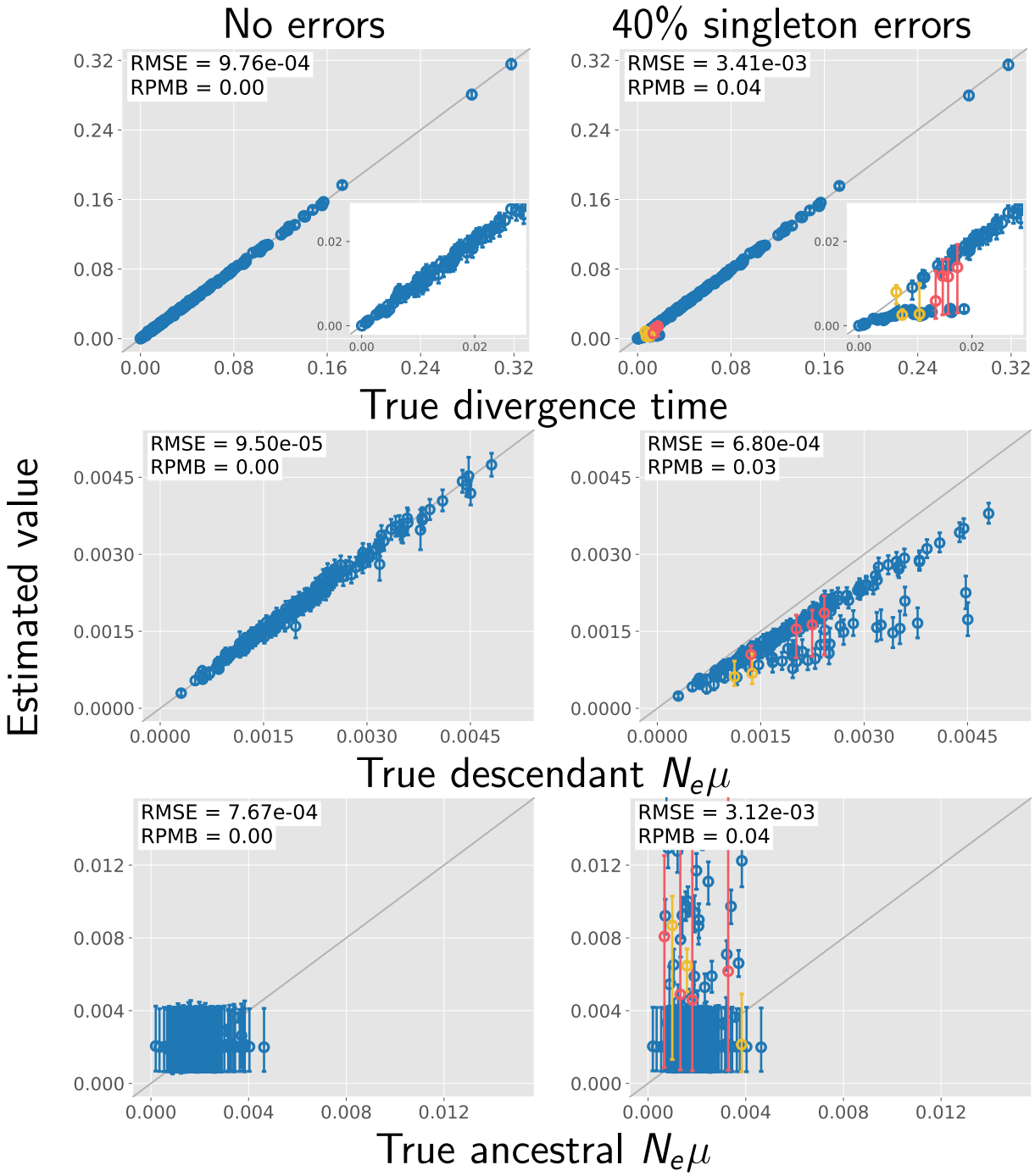

Figure 11. The performance of ecoevolity with data sets simulated with unlinked characters. Each plotted circle and associated error bars represent the posterior mean and 95% credible interval. Circles and error bars are colored yellow if the effective sample size (ESS) of the estimate was less than 200, red if the potential scale reduction factor (PSRF) was greater than 1.2, and green if both conditions were true. The root mean square error (RMSE) and rate of poor MCMC behavior (RPMB) is given for each plot, the latter of which is the proportion of estimates with ESS $<$ 200 and/or PSRF $>$ 1.2. Inset plots magnify estimates of most recent divergence times. We generated the plots using matplotlib Version 3.1.1 (Hunter, 2007).