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

December 2, 2020

*Corresponding author: joaks@auburn.edu

1 Introduction

Phylogeneticists inferring species trees from reduced-representation genomic data sets are faced with a decision of whether to analyze all of their data or reduce it to only putatively unlinked SNPs. These data sets are becoming commonplace in phylogenetics (Leaché and Oaks, 2017), and usually comprise hundreds to thousands of loci as short as 50 nucleotides long and up to several thousand base pairs long. Full bayesian likelihood methods are an ideal tool for inferring species trees from these data when computationally feasible. These full likelihood methods 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 (Bryant et al., 2012). Unlinked-character models on the other hand are more tractable for a large number of loci, because estimating individual gene trees is avoided by integrating over all possible gene trees (Bryant et al., 2012). 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 this information.

Ideally, phylogeneticists would use a linked-character method that assumes sites within each locus evolved along a shared gene tree and uses all of the information contained at a locus. However this is not always computationally feasible and the model could be violated by intralocus recombination. Should investigators then remove all but a 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? Our goal is to use simulated data to help inform these decisions.

An important consideration for choosing a multi-species coalescent model and how to apply that model to data is 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, amplification of different genome regions can be highly variable resulting in uneven coverage across loci (Aird et al., 2011). 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). [KAC comment: Discuss biases inherent to reduced rep techniques?]

To avoid introducing sequencing errors into analyses performed with the data, it is not uncommon 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. This filtering can also introduce errors and biases which has been shown to have an effect on estimates derived from the assembled alignments (Harvey et al., 2015a; Linck and Battey, 2019). Furthermore, decisions have to be made during or after processing the raw sequence reads to avoid aligning paralogous sequences. This is often done by setting an upper threshold on the number of variable sites within a locus (Harvey et al., 2015b). Such a strategy will also filter out the most variable alignments , thus introducing an acquisition bias.

Given all of these potential problems throughout the data collection process, phylogeneticists should assume their high-throughput genomic data set suffers from errors and acquisition biases. Do linked and unlinked character models differ in their robustness to such errors? Linked-character models can leverage shared information among linked sites about each underlying gene tree. Thus, these models may 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 ([JRO comment: Figure showing this difference?]). 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. Our goal is to simulate data sets with varying degrees of errors and varying locus lengths to test this prediction that linked character models are more robust to the types of errors contained in reduced representation data sets. Our results support this prediction, but also show that region of parameter space where the differences between linked and unlinked character models is revealed is quite limited.

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 three different lengths—1000, 500, and 250 linked 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 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 MSC models (Bryant et al., 2012; Oaks, 2018) 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 with a birth rate of 10 using the Python package DendroPy (Version 4.40; XXXXX branch commit eb69003; Sukumaran and Holder, 2010). This is equivalent to the divergence time being Exponentially distributed with a rate of 20. 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, and 400 gene trees for the 1000, 500, and 250 locus length data sets 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.

ecoevolity makes the assumption that all characters in a dataset are unlinked. 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 was not being caused by violation of linkage assumptions, we simulated an additional 100 data sets of 100,000 biallelic characters as described above, except that all characters were unlinked—i.e. each character was simulated on a separate gene tree.

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 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 for each locus, and with a probability of 0.2 or 0.4 we randomly replaced one 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 Error

For each simulated data set, 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; dev branch commit a7e9bf2; Oaks, 2018) and a linked-character model using the StarBEAST2 (Version 0.15.1; Ogilvie et al., 2017) package 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 matches the model we used to

generate the data. The linkage of sites within loci of our simulated character data violates the unlinked-character model of ecoevolity (Bryant et al., 2012; Oaks, 2018). Therefore, we also analyzed each data set with ecoevolity after selecting at most, a single variable character from each locus. Loci without variable character sites were excluded.

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 (Gelman and Meng, 1998) to indicate adequate mixing of a chain. 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/XXXXXX, and includes all of the data and scripts necessary to produce our results.

3 Results

3.1 Analyzing all sites versus SNPs with ecoevolity

The unlinked character model implemented in ecoevolity assumes that ortholgous nucleotide sites evolve independently along separate gene tree. 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, this drastically reduces the amount of data. When analyzing the simulated data sets without errors, the precision of parameter estimates by ecoevolity was much greater with no less accuracy when all sites of the alignment were used relative to when a single SNP per locus was used despite violating the model (Figs. 3–10 and 12). This was true across the different lengths of loci. Analyzing only SNPs data 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). To rule out the possibility that sensitivity to error might be caused by violation of the unlinked character model, we inferred parameters from the dataset simulated under an unlinked character model whereby each orthologous nucleotide site evolved on a separate gene tree. We introduced singleton error into this dataset at a rate of 40%. We find the same bias in parameter estimates with this unlinked character data set (Fig. 13) as we did with the linked loci which violate the model of sequence evolution in ecoevolity (Figs. 3, 4 and 12).

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

The divergence time estimates of StarBEAST2 were very accurate and precise for all alignment lengths as well as types and degrees error.

For data sets without error (and when all characters are analyzed), the accuracy and precision of ecoevolity's divergence time estimates were comparable to StarBEAST2 (Figs. 3, 4 and 12). However when alignments contained errors, ecoevolity underestimated very recent divergence times with increasing severity as the frequency of errors increases (Figs. 3, 4 and 12); estimates of larger 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). StarBEAST2 consistently estimated the effective size of the ancestral population better than ecoevolity and was unaffected by errors in the data (Figs. 5–7); the precision of StarBEAST2's estimates of N_e^R increased with locus length.

The 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).

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) Fig. 13. The expected coverage behavior of StarBEAST2 and ecoevolity helps to confirm that the sequence data were simulated under the same model as that used for inference by these methods. As seen previously (Oaks, 2018), the coverage of ecoevolity is short when analyzing linked loci, due to the model violation.

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. 3–10 and 12; 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 StarBEAST2 root effective population size estimates with ESS values less than 200 was high across all analyses

(Figs. 5–7). Estimates of descendant effective population size had better ESS values across all analyses with the exception of estimates of small effective population sizes from 250 bp loci (Figs. 8–10).

4 Discussion

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. 3, 4 and 12). 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. 3, 4 and 12).

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 $8N_e\mu$. [JRO comment: We should plot div times in coalescent units versus error.] This pattern makes sense given that both types of character-pattern error reduce variation *within* the species. Thus, it is not too surprising that the unlinked-character model in ecoevolity struggles the most 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 estimating the descendant population sizes, the unlinked-character model of ecoevolity simultaneously reduces the divergence time and increases the effective size of the ancestral population. [JRO comment: Perhaps plot div times error versus ancestral pop size error.] 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. 3, 4 and 12)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 are fixed differences versus 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.

Unsurprisingly, MCMC performance in StarBEAST2 declines with decreasing locus length. There is less information contained in the shorter loci to inform gene tree estimation and it is expected that there would be more uncertainty in gene tree estimation. This uncertainly results in a wider posterior distribution of gene trees that must be sampled from. This poor 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.

4.2 Relevance to empirical data sets

It is reassuring to see the effect of character-pattern 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. The error rates of 40% that we simulated are likely higher than the rate of these types of errors induced during sample preparation and high-throughput sequencing. However, empirical alignments likely contain a mix of different sources of errors and biases from various steps in data collection process. Also, real data would not be generated under a 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 error from missing heterozygote haplotypes and missing singletons will be highly dependent on the method used to reduce representation of the genome, the depth of sequencing coverage (i.e., the number of overlapping sequence reads at a locus) and on how the data are processed. Coverage will vary across loci due to random chance and biases in PCR amplification and from sequencing. Most pipelines for processing sequence data processing pipelines set a minimum coverage threshold for variants or for alleles in order filter sequencing error. True variants or alleles are more likely to filtered when coverage is low. Filtering of very rare alleles that do not meet a minimum minor allele frequency threshold has been used as a way to eliminate the impact of sequencing errors. For example filtering alleles with counts of less than 3 would ensure that all alleles have been found in at least two diploid individuals (Rochette et al., 2019). However, filtering in this way can result in biased estimates of parameters that are sensitive to the frequencies of rare alleles (Linck and Battey, 2019). Such stringent filtering would introduce an even greater level of error than we evaluated with 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 much 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 simulated data sets with character-pattern errors are well beyond the prior distribution.

When using unlinked-character models with empirical high-throughput data sets, it might also help to perform the analysis 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 improbably 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, 2018; Oaks et al., 2018), 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 model of unlinked characters implemented in ecoevolity does not use information about linkage among sites (Bryant et al., 2012; Oaks, 2018), 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, 2018; Oaks et al., 2018). 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 Future directions

In this study we used a simple two species model with a small number of gene copies sampled from each to minimize the effect of differences in algorithms for searching species and gene tree space on the performance of the linked and unlinked character models. Our goal was to compare the theoretical performance of these two models, not their current software implementations. [KAC comment: worth noting what differences there are?] Nonetheless, exploring how character-pattern errors and biases affect the inference of larger species trees would be informative. From our simulations, we saw that their was little information in the data to update the prior distribution on the effective size of the ancestral population. Exploring larger trees will determine whether estimates of the population size of most internal branches show this behavior, or if it will be confined to the to only the most basal branches and root of the species tree. Also, the tree topology is frequently a parameter of great interest and it would therefore be interesting to know how character-pattern errors and biases affect estimation of the species-tree topology.

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 violations of our models, 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 MSC 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 MSC 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 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 XXXX of the Auburn University Museum of Natural History.

References

Aird, D., M. G. Ross, W.-S. Chen, M. Danielsson, T. Fennell, C. Russ, D. B. Jaffe, C. Nusbaum, and A. Gnirke. 2011. Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. Genome Biology 12:R18.

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.

Gelman, A. and X.-L. Meng. 1998. Statistical Science 13:163–185.

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. 2015a. Similarity thresholds used in DNA sequence assembly from short reads can reduce the comparability of population histories across species. PeerJ 3:e895.

Harvey, M. G., C. D. Judy, G. F. Seeholzer, J. M. Maley, G. R. Graves, and R. T. Brumfield. 2015b. 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. 2018. Full Bayesian comparative phylogeography from genomic data. Systematic Biology .

Oaks, J. R., C. D. Siler, and R. M. Brown. 2018. The comparative biogeography of geckos challenges predictions from a paradigm of climate-driven vicariant diversification across an island archipelago. bioRxiv .

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.

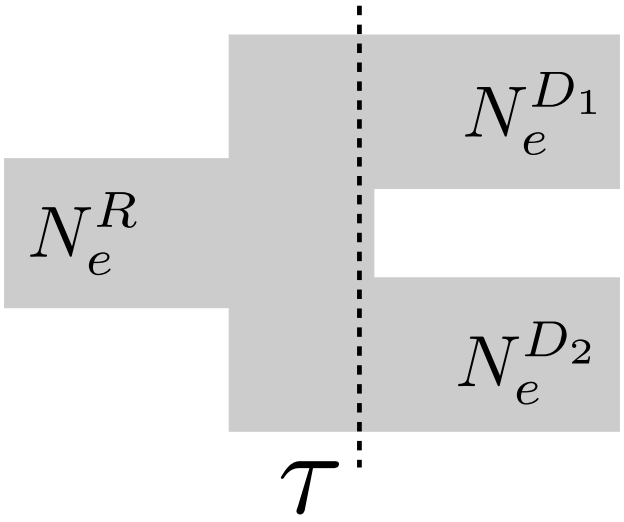

Figure 1. The model

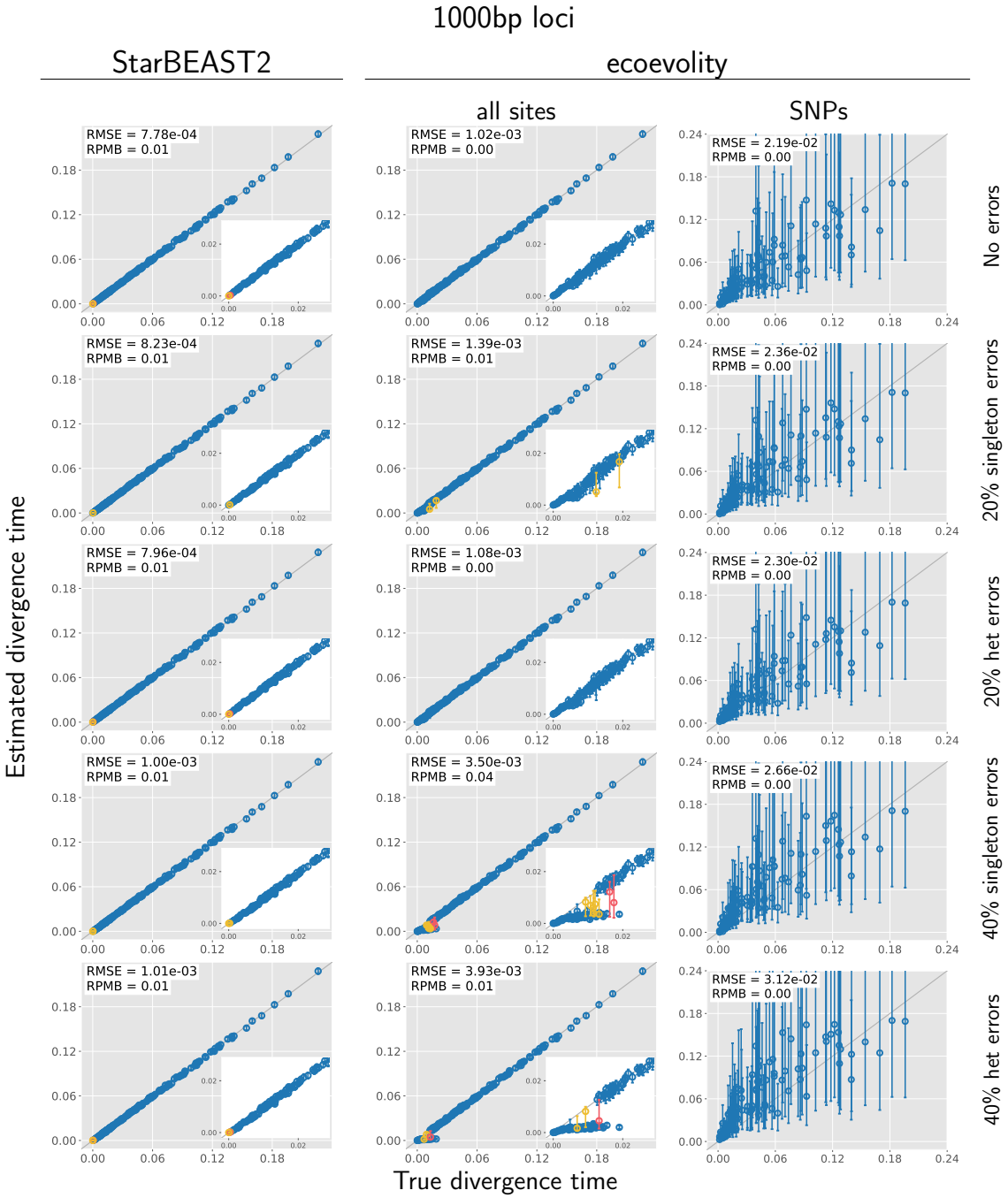

Figure 2. Accuracy and precision of divergence-time estimates (in units of expected substitutions per site) with 1000 base pair loci. The left column shows estimates from StarBEAST2. The center column shows estimates from ecoevolity using all sites and the right column shows estimates from ecoevolity using a single SNP per locus. The top row are estimates from 200 data sets simulated without character-pattern errors. Rows labelled 20% and 40% singleton errors are 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 are estimates from the same alignments after one copy of randomly paired gene copies within each species was replaced with the other with probabilities 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. Inset plots magnify estimates of most recent divergence times. We generated the plots using matplotlib Version 3.1.1 (Hunter, 2007).

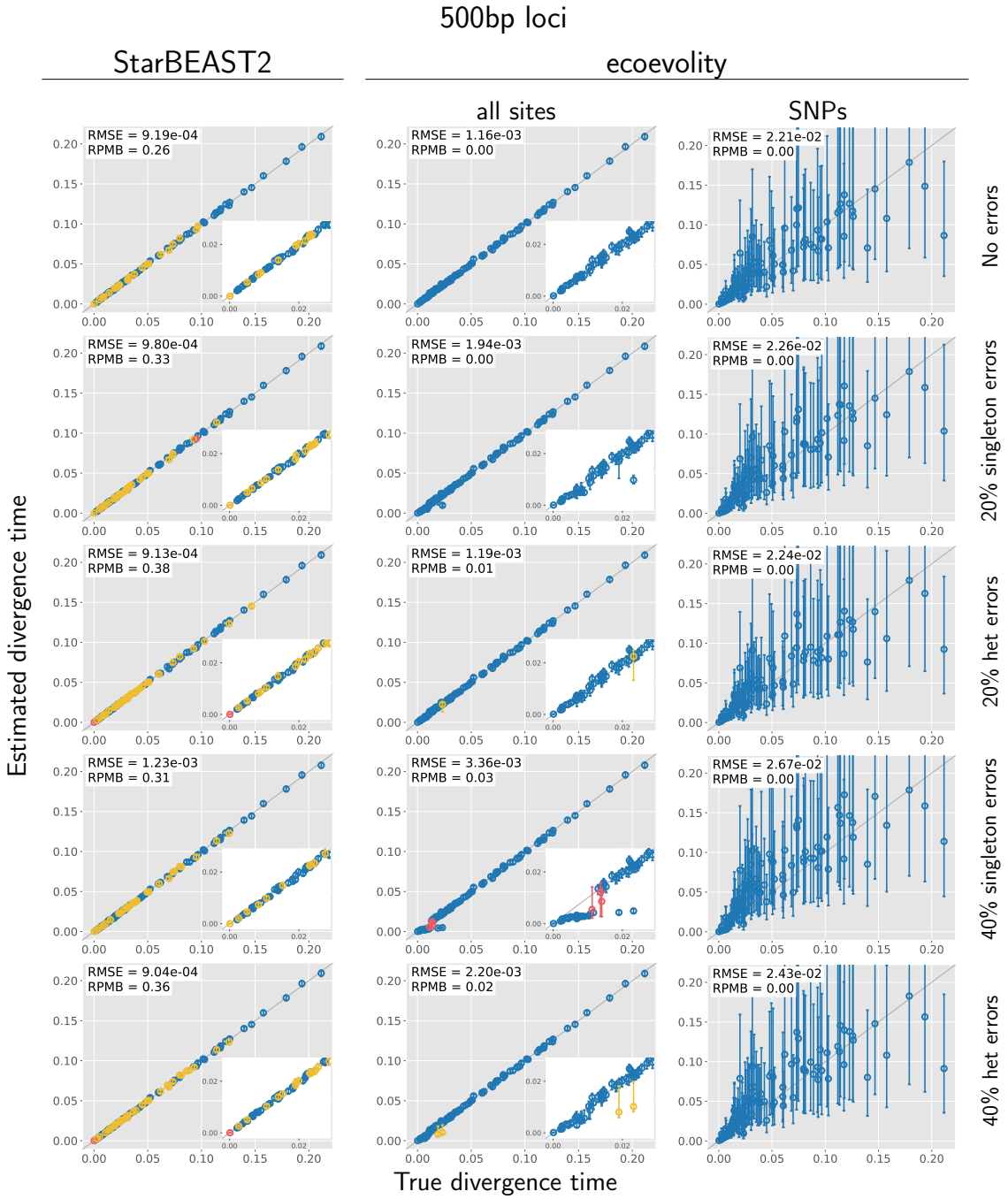

Figure 3. Accuracy and precision of divergence-time estimates (in units of expected substitutions per site) with 500 base pair loci. The left column shows estimates from StarBEAST2. The center column shows estimates from ecoevolity using all sites and the right column shows estimates from ecoevolity using a single SNP per locus. The top row are estimates from 200 data sets simulated without character-pattern errors. Rows labelled 20% and 40% singleton errors are 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 are estimates from the same alignments after one copy of randomly paired gene copies within each species was replaced with the other with probabilities 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. Inset plots magnify estimates of most recent divergence times. We generated the plots using matplotlib Version 3.1.1 (Hunter, 2007).

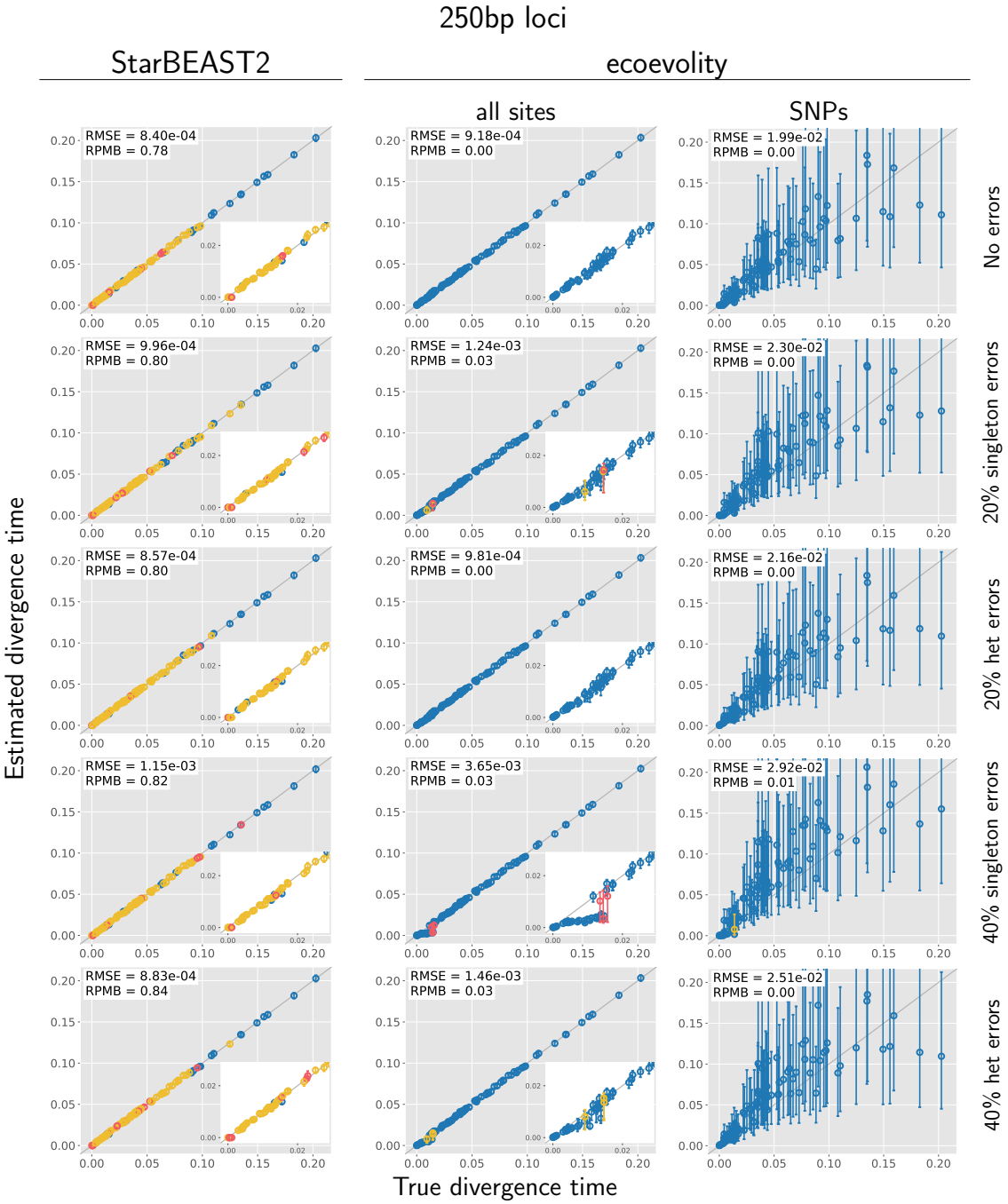

Figure 4. Accuracy and precision of divergence-time estimates (in units of expected substitutions per site) with 250 base pair loci. The left column shows estimates from StarBEAST2. The center column shows estimates from ecoevolity using all sites and the right column shows estimates from ecoevolity using a single SNP per locus. The top row are estimates from 200 data sets simulated without character-pattern errors. Rows labelled 20% and 40% singleton errors are 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 are estimates from the same alignments after one copy of randomly paired gene copies within each species was replaced with the other with probabilities 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. Inset plots magnify estimates of most recent divergence times. We generated the plots using matplotlib Version 3.1.1 (Hunter, 2007).

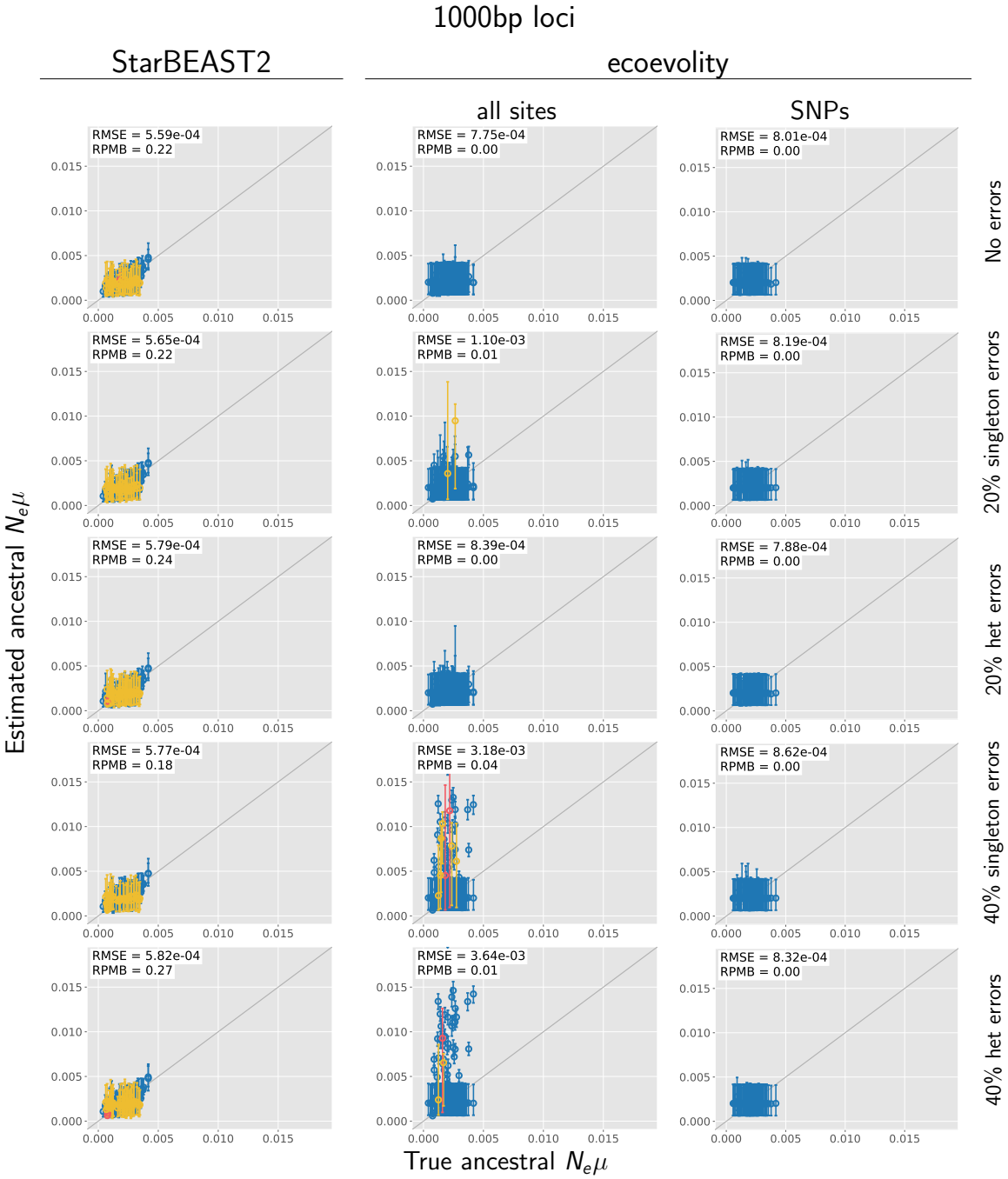

Figure 5. Accuracy and precision of estimates of root effective population size scaled by the mutation rate ($N_e^R\mu$) with 1000 base pair loci. The left column shows estimates from StarBEAST2. The center column shows estimates from ecoevolity using all sites and the right column shows estimates from ecoevolity using a single SNP per locus. The top row are estimates from 200 data sets simulated without character-pattern errors. Rows labelled 20% and 40% singleton errors are 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 are estimates from the same alignments after one copy of randomly paired gene copies within each species was replaced with the other with probabilities 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).

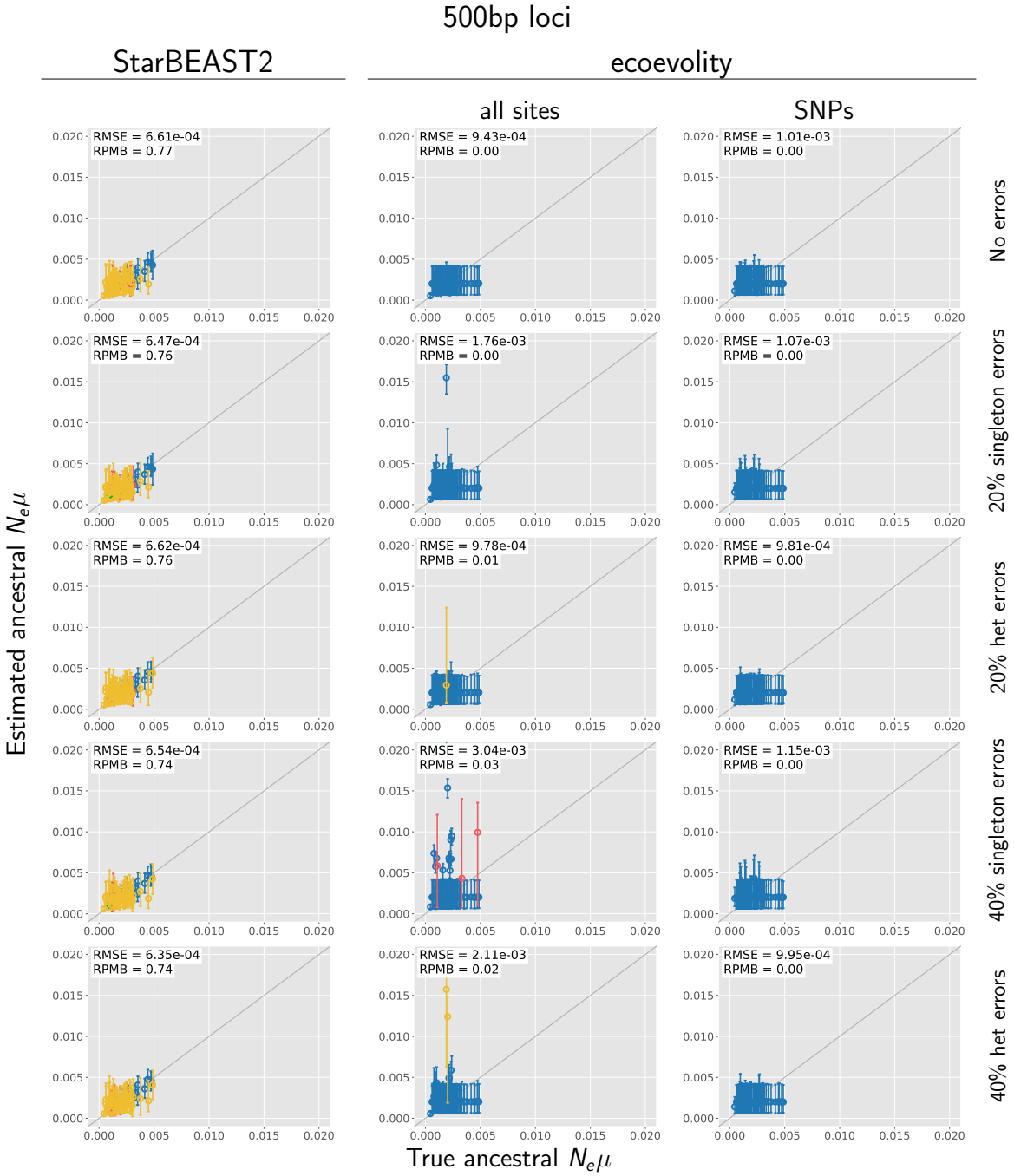

Figure 6. Accuracy and precision of estimates of root effective population size scaled by the mutation rate ($N_e^R\mu$) with 500 base pair loci. The left column shows estimates from StarBEAST2. The center column shows estimates from ecoevolity using all sites and the right column shows estimates from ecoevolity using a single SNP per locus. The top row are estimates from 200 data sets simulated without character-pattern errors. Rows labelled 20% and 40% singleton errors are 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 are estimates from the same alignments after one copy of randomly paired gene copies within each species was replaced with the other with probabilities 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).

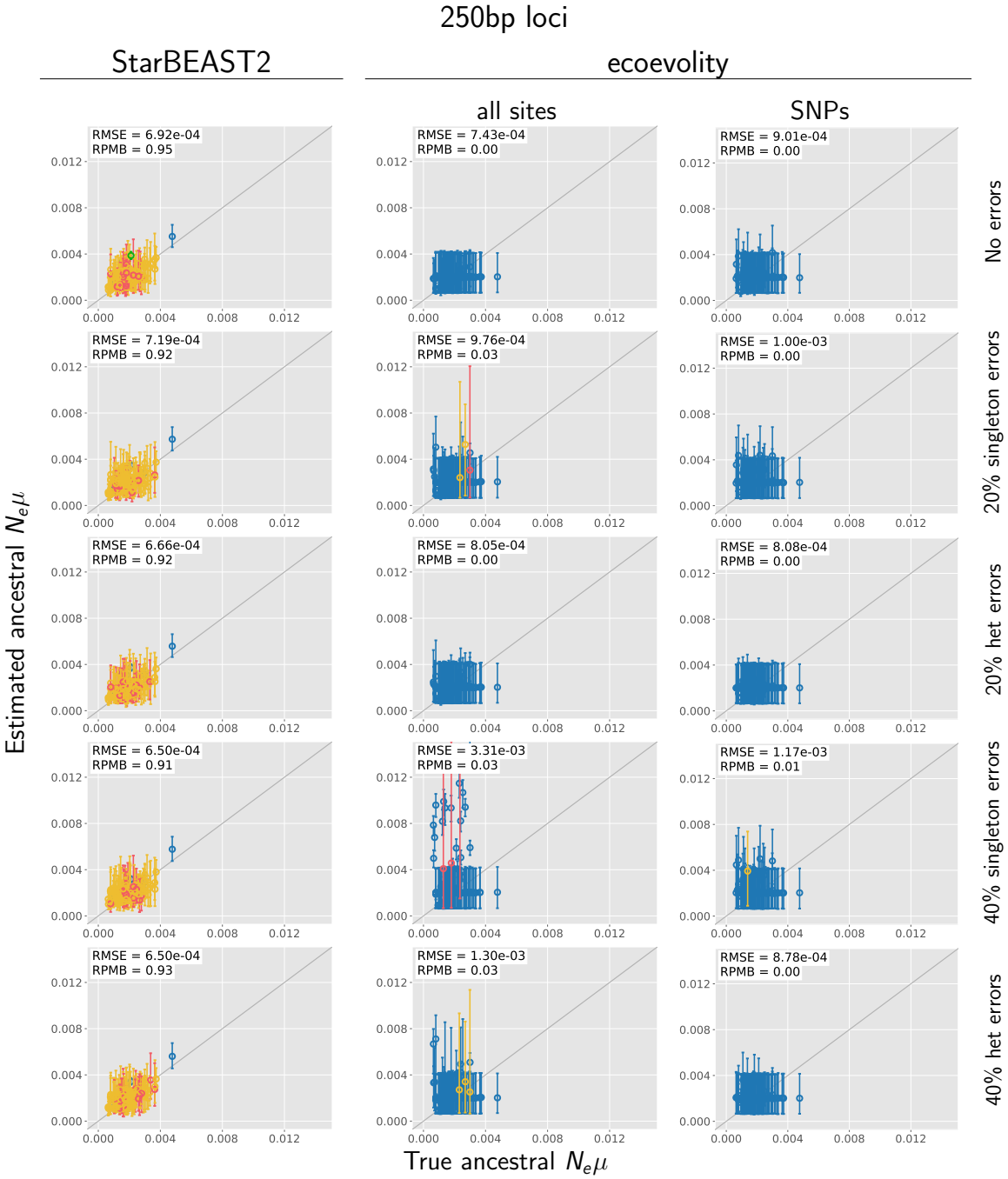

Figure 7. Accuracy and precision of estimates of root effective population size scaled by the mutation rate ($N_e^R\mu$) with 250 base pair loci. The left column shows estimates from StarBEAST2. The center column shows estimates from ecoevolity using all sites and the right column shows estimates from ecoevolity using a single SNP per locus. The top row are estimates from 200 data sets simulated without character-pattern errors. Rows labelled 20% and 40% singleton errors are 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 are estimates from the same alignments after one copy of randomly paired gene copies within each species was replaced with the other with probabilities 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).

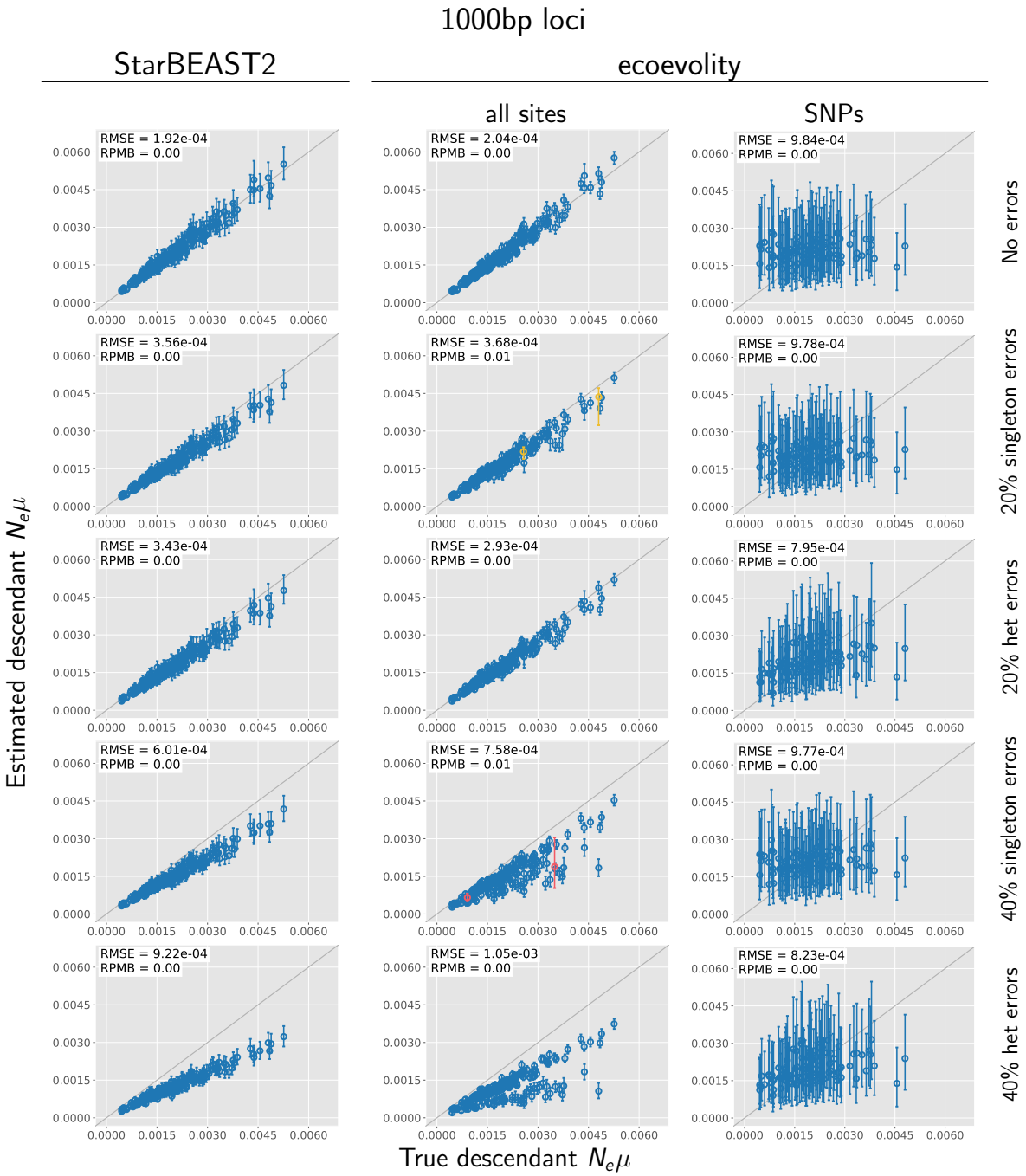

Figure 8. Accuracy and precision of estimates of effective population sizes of the descendant branches of the tree scaled by the mutation rate ($N_e^D\mu$) with 1000 base pair loci. The left column shows estimates from StarBEAST2. The center column shows estimates from ecoevolity using all sites and the right column shows estimates from ecoevolity using a single SNP per locus. The top row are estimates from 200 data sets simulated without character-pattern errors. Rows labelled 20% and 40% singleton errors are 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 are estimates from the same alignments after one copy of randomly paired gene copies within each species was replaced with the other with probabilities 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).

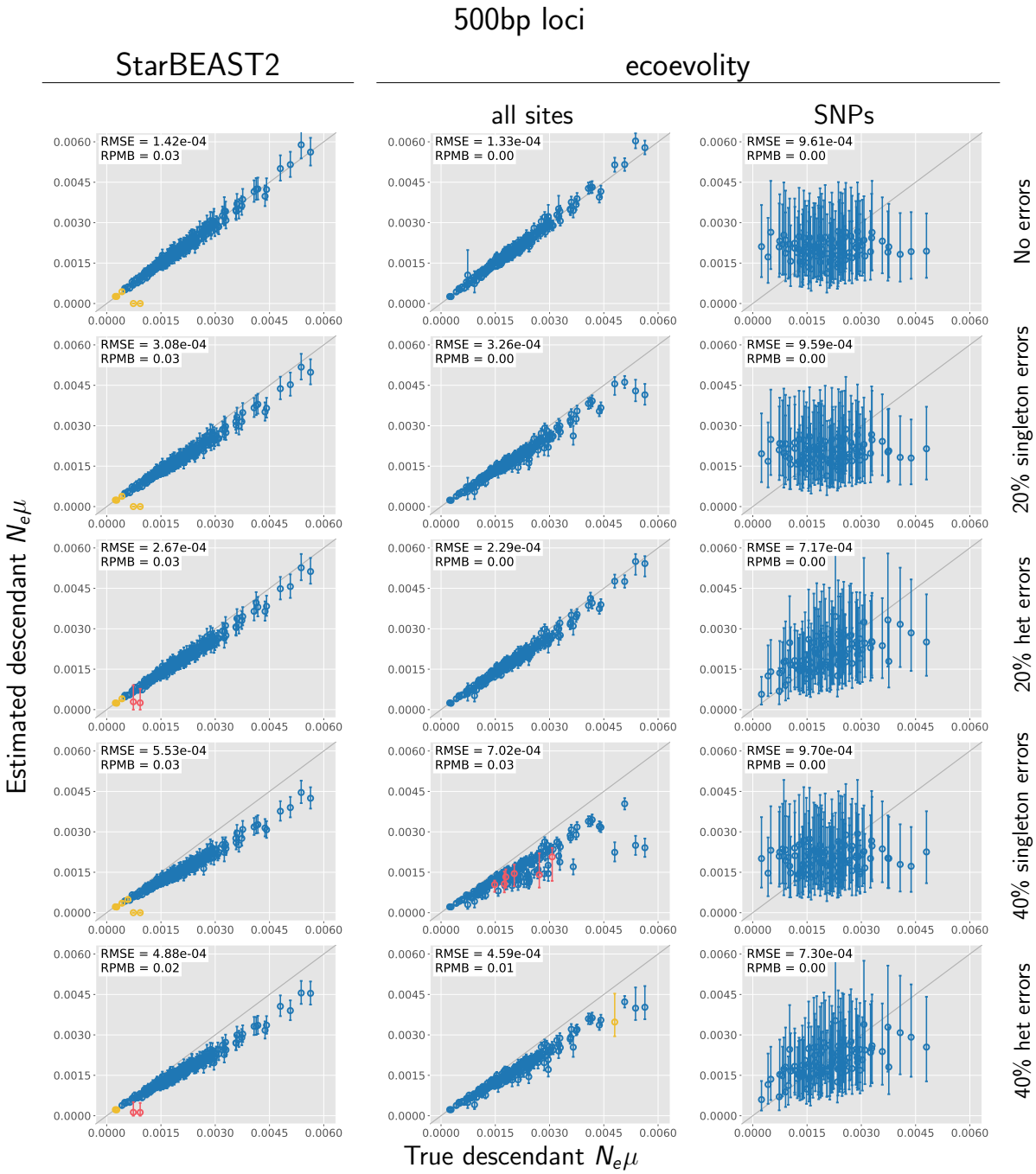

Figure 9. Accuracy and precision of estimates of effective population sizes of the descendant branches of the tree scaled by the mutation rate ($N_e^D\mu$) with 500 base pair loci. The left column shows estimates from StarBEAST2. The center column shows estimates from ecoevolity using all sites and the right column shows estimates from ecoevolity using a single SNP per locus. The top row are estimates from 200 data sets simulated without character-pattern errors. Rows labelled 20% and 40% singleton errors are 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 are estimates from the same alignments after one copy of randomly paired gene copies within each species was replaced with the other with probabilities 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).

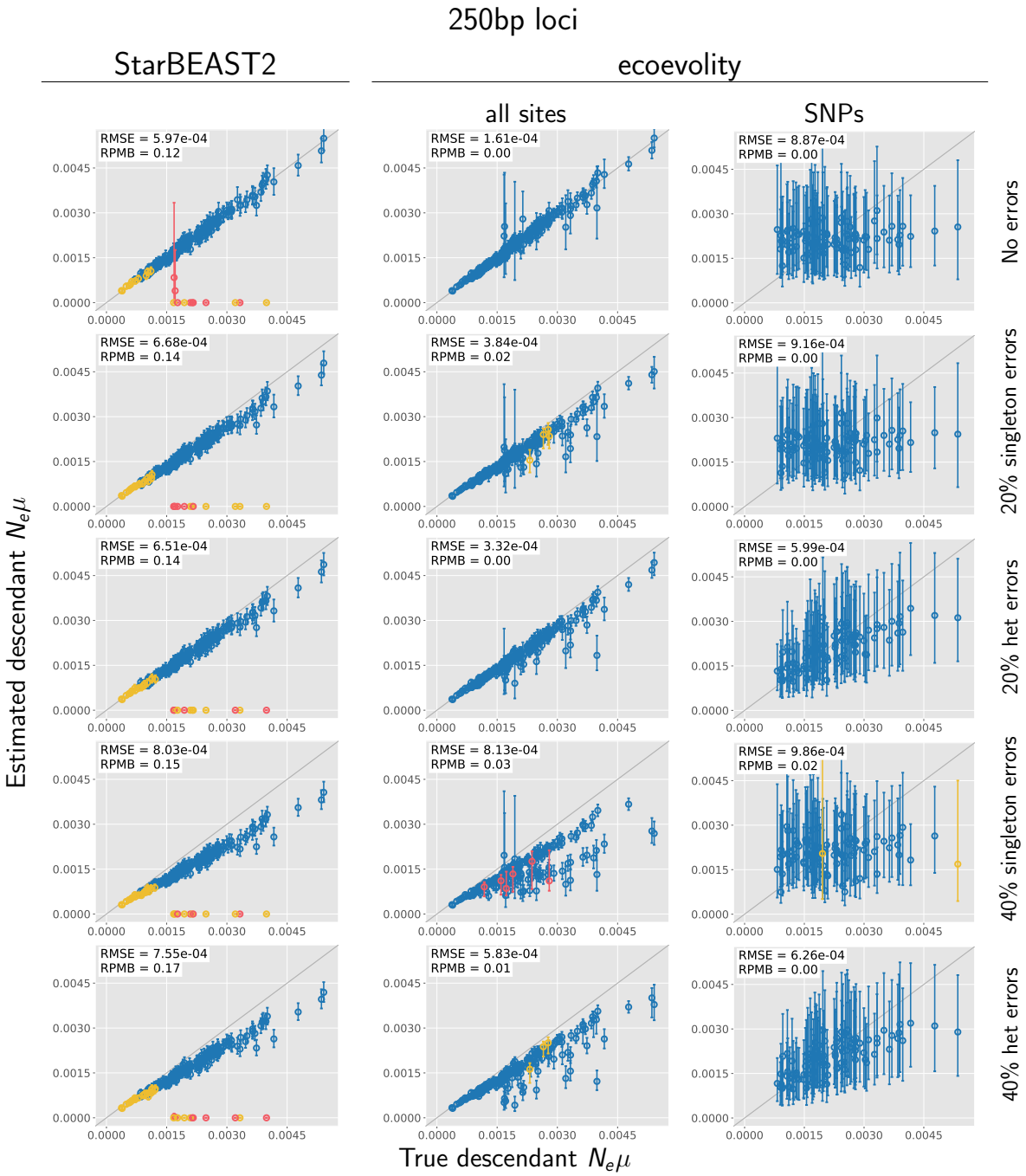

Figure 10. Accuracy and precision of estimates of effective population sizes of the descendant branches of the tree scaled by the mutation rate ($N_e^D\mu$) with 250 base pair loci. The left column shows estimates from StarBEAST2. The center column shows estimates from ecoevolity using all sites and the right column shows estimates from ecoevolity using a single SNP per locus. The top row are estimates from 200 data sets simulated without character-pattern errors. Rows labelled 20% and 40% singleton errors are 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 are estimates from the same alignments after one copy of randomly paired gene copies within each species was replaced with the other with probabilities 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).

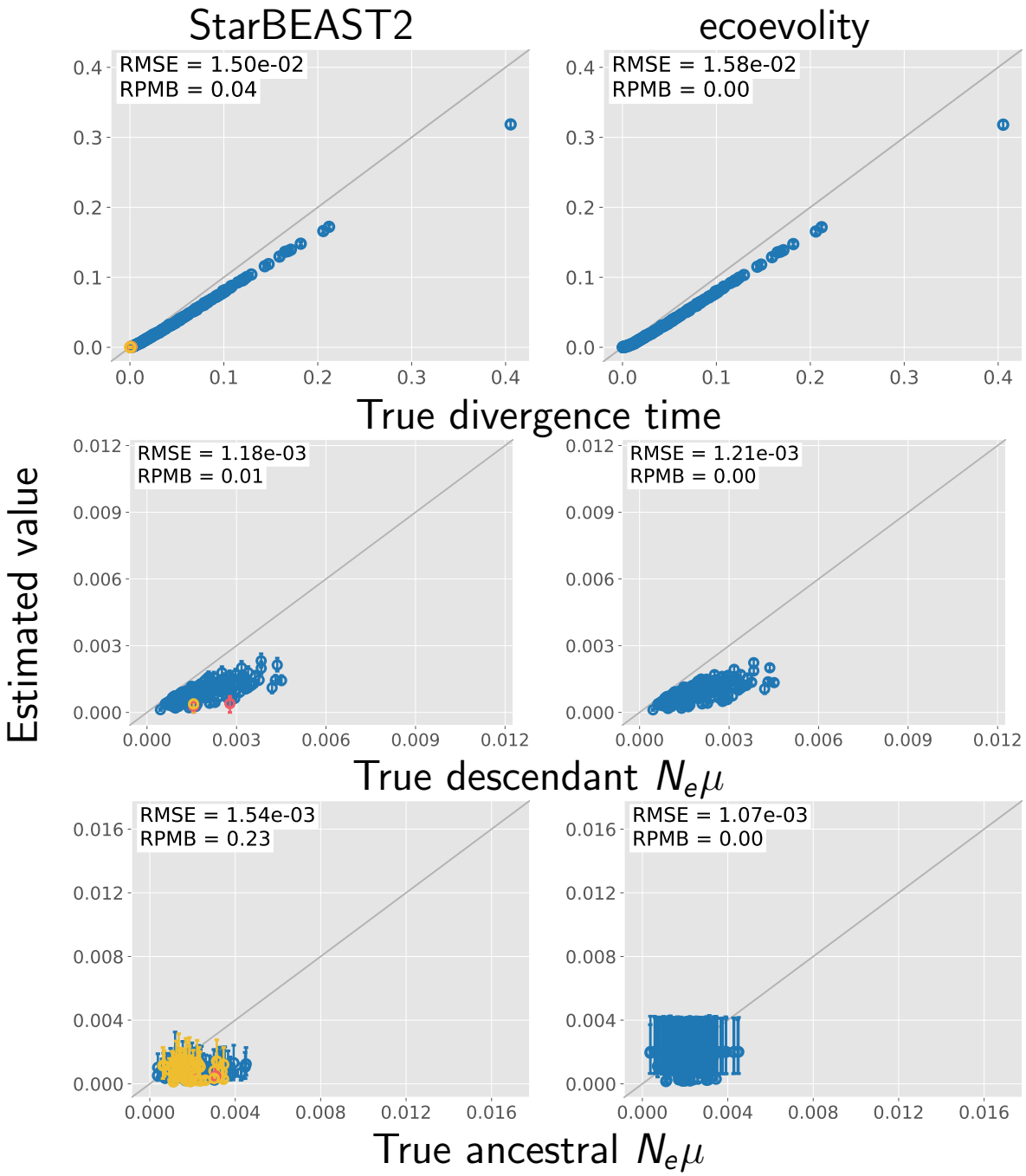

Figure 11. Accuracy and precision of divergence-time estimates (in units of expected substitutions per site) with 1000 base pair loci. Circles and error bars are colored yellow if the effective sample size (ESS) of the estimate was less than 200, red if the potential scale reduction factor (PSRF) was 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).

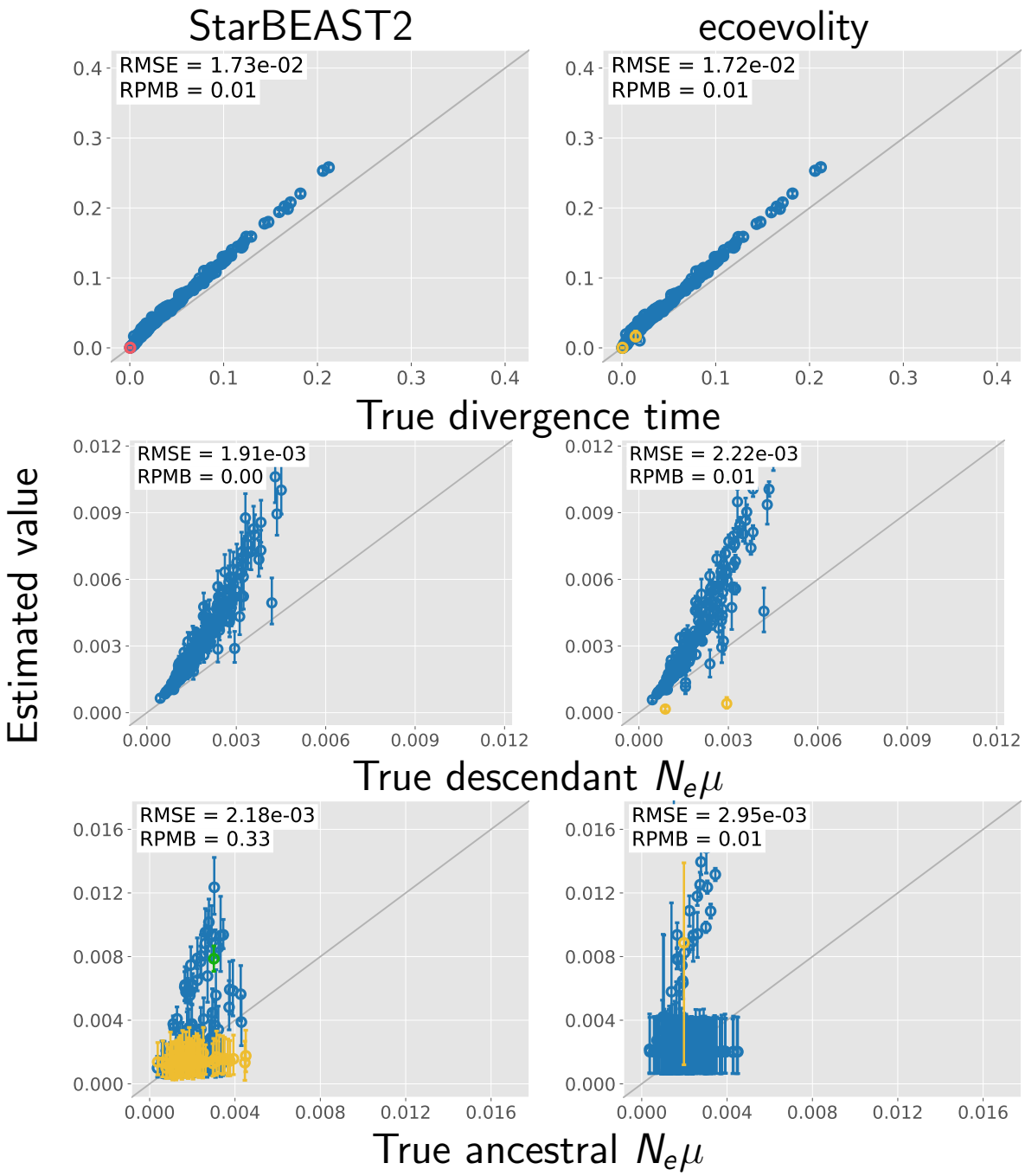

Figure 12. Circles and error bars are colored yellow if the effective sample size (ESS) of the estimate was less than 200, red if the potential scale reduction factor (PSRF) was greater than 1.2, and green if both conditions were true. The root mean square error (RMSE) and rate of poor MCMC behavior (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).

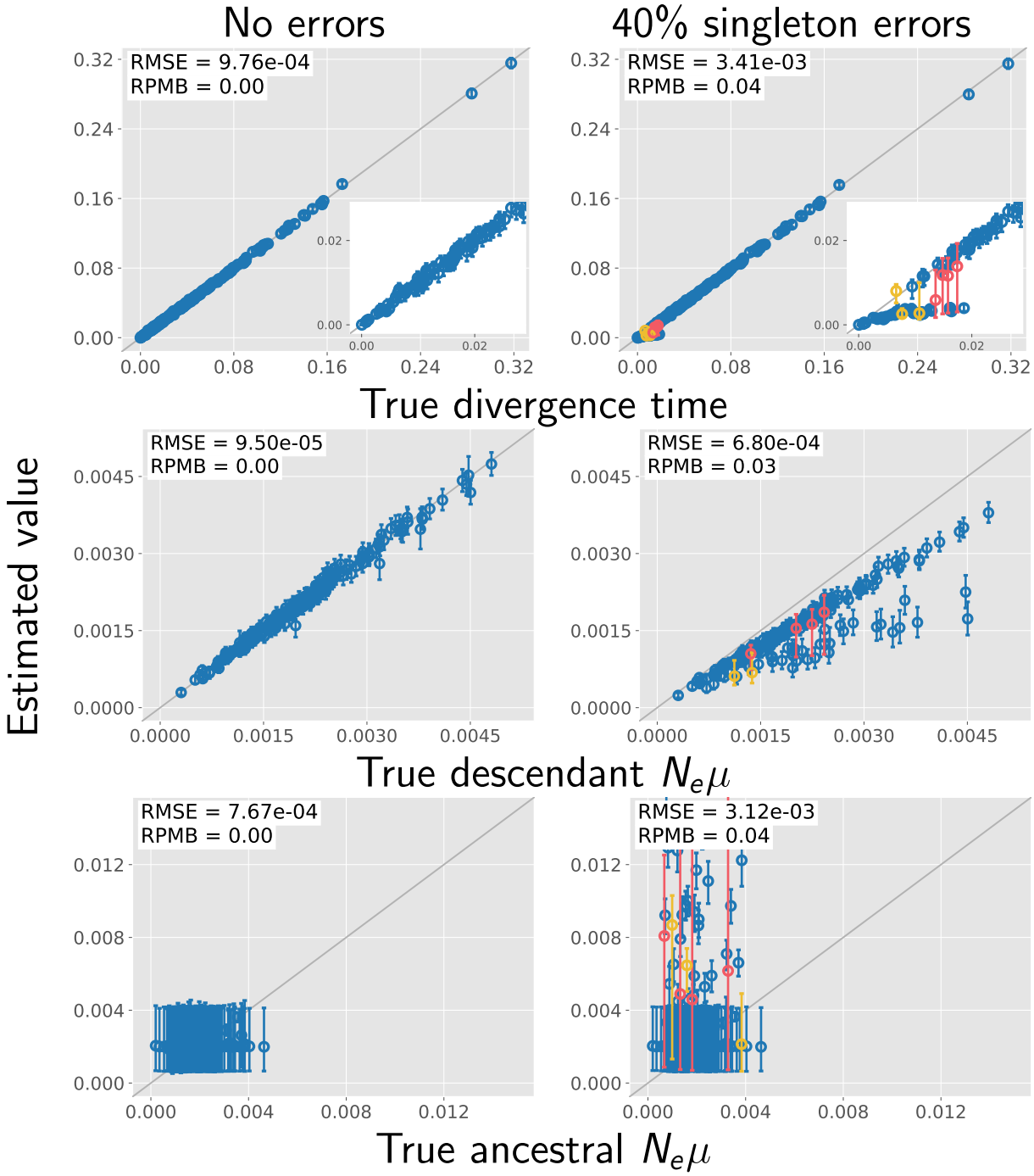

Figure 13. The performance of ecoevolity with data sets simulated with unlinked characters.