

Contents

1

Table of Contents	2	2
List of Figures	3	3
List of Tables	4	4
1 Phylogeography	5	5
1.1 Introduction	5	6
1.2 Methods	5	7
1.2.1 Sampling and DNA Isolation	5	8
1.2.2 RADseq Library Preparation	6	9
1.2.3 Data Processing	6	10
1.2.4 Maximum Likelihood	7	11
1.2.5 Structure	7	12
1.2.6 Phycoeval	7	13
1.2.7 Dsuite	7	14
1.3 Results	7	15
1.3.1 Maximum Likelihood	7	16
1.4 Discussion	7	17
1.5 Acknowledgments	8	18
1.6 Figures	9	19
1.7 Tables	10	20
2 Hybrid Zone	11	21
2.1 Introduction	11	22
2.2 Methods	11	23
2.2.1 Sampling and DNA Isolation	11	24
2.2.2 RADseq Library Preparation	12	25
2.2.3 Data Processing	13	26
2.2.4 Genetic Clustering & Ancestry Proportions	13	27
2.2.5 Genomic Cline Analysis	14	28
2.2.6 Genetic differentiation and Introgression	14	29
2.3 Results	14	30
2.3.1 Sampling and Data Processing	14	31
2.4 Discussion	15	32
2.5 Figures	16	33
2.6 Tables	21	34

3 Comparison of Linked versus Unlinked Character Models for Species Tree Inference	26	35
3.1 Introduction	26	36
3.2 Methods	28	38
3.2.1 Simulations of error-free data sets	28	39
3.2.2 Introducing Site-pattern Errors	28	40
3.2.3 Assessing Sensitivity to Errors	29	41
3.2.4 Project repository	29	42
3.3 Results	29	43
3.3.1 Behavior of linked (<i>StarBEAST2</i>) versus unlinked (<i>ecoevolity</i>) character models	29	44
3.3.2 Analyzing all sites versus SNPs with <i>ecoevolity</i>	30	46
3.3.3 Coverage of credible intervals	30	47
3.3.4 MCMC convergence and mixing	31	48
3.4 Discussion	31	49
3.4.1 Robustness to character-pattern errors	32	50
3.4.2 Relevance to empirical data sets	33	51
3.4.3 Recommendations for using unlinked-character models	33	52
3.4.4 Other complexities of empirical data in need of exploration	34	53
3.5 Acknowledgments	35	54
References	35	55
3.6 Figures	37	56

List of Figures

57

1.1. Dsuite	9	58
2.1. Evanno method for optimal value of K in <i>STRUCTURE</i>	16	59
2.2. <i>STRUCTURE</i> iterations	17	60
2.3. Summarized <i>STRUCTURE</i> results for each value of K.	18	61
2.4. Genetic evidence of hybridization between <i>A. americanus</i> and <i>A. terrestris</i>	19	62
2.5. Shape of genomic clines	20	63
3.1. Simulation model	37	64
3.2. Accuracy and precision of divergence-time estimates (in units of expected substitutions per site) with 1000 base pair loci	38	66
3.3. Accuracy and precision of divergence-time estimates (in units of expected substitutions per site) with 500 base pair loci	39	67
3.4. Accuracy and precision of divergence-time estimates (in units of expected substitutions per site) with 250 base pair loci	40	69
3.5. Accuracy and precision of estimates of root effective population size scaled by the mutation rate ($N_e^R \mu$) with 1000 base pair loci	41	71
3.6. Accuracy and precision of estimates of root effective population size scaled by the mutation rate ($N_e^R \mu$) with 500 base pair loci	42	73
3.7. Accuracy and precision of estimates of root effective population size scaled by the mutation rate ($N_e^R \mu$) with 250 base pair loci	43	75
3.8. 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	44	77
3.9. 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	45	80
3.10. 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	46	83
3.11. Performance of <i>ecoevolity</i> with data sets simulated with unlinked characters	47	86

List of Tables

88

2.1 Samples collected for this study	21	89
2.2 Samples loaned from museums	25	90

<h1>Chapter 1</h1>	91
<h2>Phyogeography</h2>	92
<h3>1.1 Introduction</h3>	93
...	94
<h3>1.2 Methods</h3>	95
<h4>1.2.1 Sampling and DNA Isolation</h4>	96
I obtained tissue samples museum tissue collections as well as from individuals that I collected from 2017 to 2020. I selected samples to represent as much of the range of each species of <i>Anaxyrus</i> as possible. I also included one <i>Rhinella marina</i> and one <i>Incilius nebulifer</i> for use as outgroups for phylogenetic analyses.	97 98 99 100
I isolated DNA from tissues by first lysing a piece of tissue approximately the size of a grain of rice in 300 μ L of a solution of 10mM Tris-HCL, 10mM EDTA, 1% SDS (w/v), and nuclease free water along with 6 mg Proteinase K that was incubated for 4-16 hours at 55°C in a 1.5 mL microcentrifuge tube. To purify the DNA and separate it from the lysis product, I mixed the lysis product with a 2X volume of SPRI bead solution containing 1 mM EDTA, 10 mM Tris-HCl, 1 M NaCl, 0.275% Tween-20 (v/v), 18% PEG 8000 (w/v), 2% Sera-Mag SpeedBeads (GE Healthcare PN 65152105050250) (v/v), and nuclease free water. I then incubated the samples at room temperature for 5 minutes, placed the beads on a magnetic rack, and discarded the supernatant once the beads had collected on the side of the tube. I then performed two ethanol washes by adding 1 mL of 70% ETOH to the beads while still placed in the magnet stand and allowing it to stand for 5 minutes before discarding the ethanol. After removing all ethanol from the second wash, I removed the tube from the magnet stand and allowed the sample to dry for 1 minute before mixing the beads with 100 μ L of TLE solution containing 10 mM Tris-HCL, 0.1 mm EDTA, and nuclease free water. After allowing the bead mixture to stand at room temperature for 5 minutes I returned the beads to the magnet stand, pipetted all of the TLE solution into another microcentrifuge tube, and discarded the beads. I quantified DNA with a Qubit fluorometer (Life Technologies, USA) and diluted samples with TLE solution to bring the concentration to 20 ng/ μ L.	101 102 103 104 105 106 107 108 109 110 111 112 113 114 115 116 117 118 119

1.2.2 RADseq Library Preparation	120
I prepared RADseq libraries using the 2RAD approach outlined by Bayona-Vásquez et al., 2019. On 96 well plates, I ligated 100 ng of sample DNA in 15 μ L of a solution with 1X CutSmart Buffer (New England Biolabs, USA; NEB), 10 units of XbaI, 10 units of EcoRI, 0.33 μ M XbaI compatible adapter, 0.33 μ M EcoRI compatible adapter, and nuclease free water with a 1 hour incubation at 37°C. I then immediately added 5 μ L of a solution with 1X Ligase Buffer (NEB), 0.75 mM ATP (NEB), 100 units DNA Ligase (NEB), and nuclease free water and incubated at 22°C for 20 min and 37°C for 10 min for two cycles, followed by 80°C for 20 min to stop enzyme activity. For each 96 well plate, I pooled 10 μ L of each sample and split this pool equally between two microcentrifuge tubes. I purified each pool of libraries with a 1X volume of SpeedBead solution followed by two ethanol washes as described in the previous section except that the DNA was resuspended in 25 μ L of TLE solution.	121 122 123 124 125 126 127 128 129 130 131 132
In order to be able to detect and remove PCR duplicates, I performed a single cycle of PCR with the iTru5-8N primer which adds a random 8 nucleotide barcode to each library construct. For each plate, I prepared four PCR reactions with a total volume of 50 μ L containing 1X Kapa Hifi Buffer (Kapa Biosystems, USA; Kapa), 0.3 μ M iTru5-8N Primer, 0.3 mM dNTP, 1 unit Kapa HiFi DNA Polymerase, 10 μ L of purified ligation product, and nuclease free water. I ran reactions through a single cycle of PCR on a thermocycler at 98°C for 2 min, 60°C for 30 s, and 72°C for 5 min. I pooled all of the PCR products for a plate into a single tube and purified the libraries with a 2X volume of SpeedBead solution as described before and resuspended in 25 μ L TLE. I added the remaining adapter and index sequences unique to each plate with four PCR reactions with a total volume of 50 μ L containing 1X Kapa Hifi (Kapa), 0.3 μ M iTru7 Primer, 0.3 μ M P5 Primer, 0.3 mM dNTP, 1 unit of Kapa Hifi DNA Polymerase (Kapa), 10 μ L purified iTru5-8N PCR product, and nuclease free water. I ran reactions on a thermocycler with an initial denaturation at 98°C for 2 min, followed by 6 cycles of 98°C for 20 s, 60°C for 15 s, 72°C for 30 s and a final extension of 72°C for 5 min. I pooled all of the PCR products for a plate into a single tube and purified the product with a 2X volume of SpeedBead solution as described before and resuspended in 45 μ L TLE.	133 134 135 136 137 138 139 140 141 142 143 144 145 146 147 148 149
I size selected the library DNA from each plate in the range of 450-650 base pairs using a BluePippin (Sage Science, USA) with a 1.5% dye free gel with internal R2 standards. To increase the final DNA concentrations I prepared four PCR reactions for each plate with 1X Kapa Hifi (Kapa), 0.3 μ M P5 Primer, 0.3 μ M P7 Primer, 0.3 mM dNTP, 1 unit of Kapa HiFi DNA Polymerase (Kapa), 10 μ L size selected DNA, and nuclease free water and used the same thermocycling conditions as the previous (P5-iTru7) amplification. I pooled all of the PCR products for a plate into a single tube and purified the product with a 2X volume of SpeedBead solution as before and resuspended in 20 μ L TLE. I quantified the DNA concentration for each plate with a Qubit fluorometer (Life Technologies, USA) then pooled each plate in equimolar amounts relative to the number of samples on the plate and diluted the pooled DNA to 5 nM with TLE solution. The pooled libraries were pooled with other projects and sequenced on an Illumina HiSeqX by Novogene (China) to obtain paired end, 150 base pair sequences.	150 151 152 153 154 155 156 157 158 159 160 161 162

1.2.3 Data Processing

 163 |

I demultiplexed the iTru7 indexes using the *process_radtags* command from Stacks2 v2.6.4 (Rochette et al., 2019) and allowed for two mismatches for rescuing reads. To

 164 165 |

remove PCR duplicates, I used the <i>clone_filter</i> command from <i>Stacks</i> . I demultiplexed inline sample barcodes, trimmed adapter sequence, and filtered reads with low quality scores as well as reads with any uncalled bases using the <i>process_radtags</i> command again and allowed for the rescue of restriction site sequence as well as barcodes with up to two mismatches.	166 167 168 169 170
1.2.4 Maximum Likelihood	171
Maximum likelihood using IQ-TREE v1.6.12 (Nguyen et al., 2015) with 1000 ultrafast bootstrap replicates (Hoang et al., 2018) under GTR substitution model.	172 173
1.2.5 Structure	174
Run for 100,000 iterations with 50,000 burnin	175
Run structure with all americanus group to make sure there aren't hybrids. Run structure with american toad, southern toad, woodhousii, and fowleri to see if there is population structure	176 177 178
Filtered loci not found in 75Filtered variants with minor allele count less than 3. Filtered sites with more than 2 alleles	179 180
1.2.6 Phycoeval	181
To see if there is evidence of shared divergence times and to estimate a species tree under multispecies coalescent Phycoeval v1.0.0 (Oaks et al., 2022)	182 183
1.2.7 Dsuite	184
To look for evidence of past admixture. Dsuite (Malinsky et al., 2021)	185
1.3 Results	186
Average number of reads per individual Mean coverage per locus	187
Total loci and snps after filtering	188
1.3.1 Maximum Likelihood	189
Inferred a single well supported clade (>***%) for each recognized species.	190
1.4 Discussion	191
Population structure in A. woodhousii with two overlapping mtDNA clades with one more associated with the Southwest and one more associated with the great planes (Masta et al., 2003)	192 193 194
Phylogeny of toads, doesn't place Americanus and Terrestris as sister, finds mitonuclear discordance and Finds two fowleri clades (Fontenot et al., 2011)	195 196

Inconsistent phylogeny (Masta et al., 2002)	197
Inconsistent phylogeny (Pramuk et al., 2007)	198
Inconsistent phylogeny (Graybeal, 1997)	199
Inconsistent phylogeny (Pyron & Wiens, 2011)	200
Idea: Given the appearance that there are many secondary contact zones. It seems probable that toad species have undergone range expansions. Following these range expansions, are there any barriers that are now reducing gene flow? We can test that by looking for population structure within species that aligns with possible biogeographic barriers.	201
	202
	203
	204

1.5 Acknowledgments

205

... 206

1.6 Figures

207

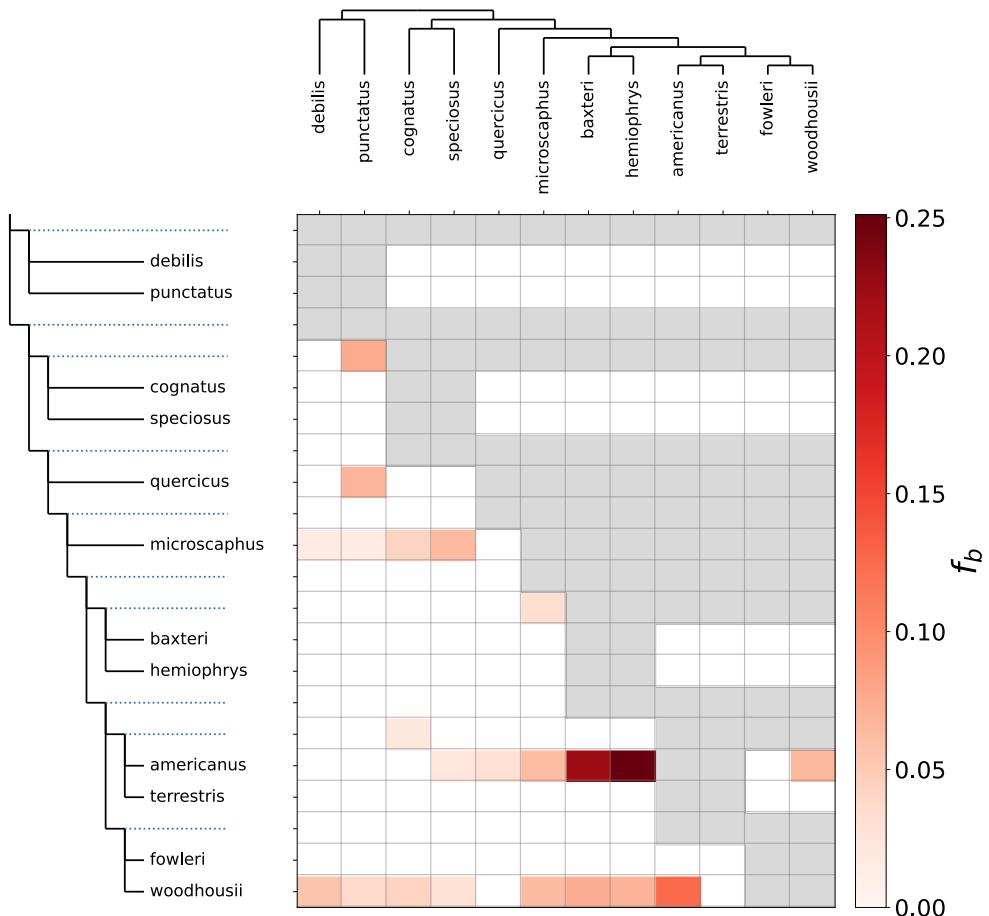


Figure 1.1. Historic admixture

1.7 Tables

208

Chapter 2

209

Hybrid Zone

210

2.1 Introduction

211

Hybrid zones have drawn the attention of biologist because they can teach us about the process of speciation and may also have important consequences in the evolution of hybridizing populations.

212

213

214

Hybridization in Bufonid toads has been appreciated for a long time.

215

Hybridization experiments in the laboratory have demonstrated a high degree of compatibility between many species pairs.

216

217

Numerous reports of hybridization have been reported.

218

Several hybrid zones have been studied to understand the patterns of introgression within them. (Give examples)

219

220

Bufo bufo x Bufo spinosus in France (Van Riemsdijk et al., 2023) Bufo siculus x Bufo balearicus in Italy (Colliard et al., 2010) Anaxyrus americanus x Anaxyrus hemiophrys in Canada (Green, 1983)

221

222

223

Here, I investigate a hybrid zone between two species of North American toad. The American toad (*Anaxyrus americanus*) and the southern toad (*Anaxyrus terrestris*). The ranges of these two species do not overlap but meet at a long zone of contact in the Southern United States. This contact zone corresponds closely with a prominent physiographic feature known as the fall line which is the boundary that separates the coastal plain to the South from the Appalachian Highlands to the North (Shankman & Hart, 2007). The two species have slight differences in male advertisement call and in morphological appearance. They differ somewhat in the timing of their spawn with some overlap in periods.

224

225

226

227

228

229

230

231

232

2.2 Methods

233

2.2.1 Sampling and DNA Isolation

234

I collected genetic samples from *A. americanus* and *A. terrestris* by driving roads during rainy nights between 2017 and 2020 in a region of central Alabama where hybridization has previously been inferred from the presence of morphological intermediates (Weatherby, 1982). I euthanized individuals with immersion in buffered MS-222. I removed liver and/or toes and preserved them in 100% ethanol and fixed specimens with XXX M (ask David how he makes Formalin) Formalin solution. Genetic samples

235

236

237

238

239

240

and formalin fixed specimens were deposited in the Auburn Museum of Natural History. Additional samples were also provided by museums (see Table 1).
I isolated DNA by lysing a small piece of liver or toe approximately the size of a grain of rice in 300 μ L of a solution of 10mM Tris-HCL, 10mM EDTA, 1% SDS (w/v), and nuclease free water along with 6 mg Proteinase K and incubating for 4-16 hours at 55°C. To purify the DNA and separate it from the lysis product, I mixed the lysis product with a 2X volume of SPRI bead solution containing 1 mM EDTA, 10 mM Tris-HCl, 1 M NaCl, 0.275% Tween-20 (v/v), 18% PEG 8000 (w/v), 2% Sera-Mag SpeedBeads (GE Healthcare PN 65152105050250) (v/v), and nuclease free water. I then incubated the samples at room temperature for 5 minutes, placed the beads on a magnetic rack, and discarded the supernatant once the beads had collected on the side of the tube. I then performed two ethanol washes by adding 1 mL of 70% ETOH to the beads while still placed in the magnet stand and allowing it to stand for 5 minutes before removing and discarding the ethanol. After removing all ethanol from the second wash, I removed the tube from the magnet stand and allowed the sample to dry for 1 minute before thoroughly mixing the beads with 100 μ L of TLE solution containing 10 mM Tris-HCL, 0.1 mM EDTA, and nuclease free water. After allowing the bead mixture to stand at room temperature for 5 minutes I returned the beads to the magnet stand, collected the TLE solution, and discarded the beads. I quantified DNA in the TLE solution with a Qubit fluorometer (Life Technologies, USA) and diluted samples with additional TLE solution to bring the concentration to 20 ng/ μ L.

2.2.2 RADseq Library Preparation

I prepared RADseq libraries using the 2RAD approach developed by Bayona-Vásquez et al., 2019. On 96 well plates, I ligated 100 ng of sample DNA in 15 μ L of a solution with 1X CutSmart Buffer (New England Biolabs, USA; NEB), 10 units of XbaI, 10 units of EcoRI, 0.33 μ M XbaI compatible adapter, 0.33 μ M EcoRI compatible adapter, and nuclease free water with a 1 hour incubation at 37°C. I then immediately added 5 μ L of a solution with 1X Ligase Buffer (NEB), 0.75 mM ATP (NEB), 100 units DNA Ligase (NEB), and nuclease free water and incubated at 22°C for 20 min and 37°C for 10 min for two cycles, followed by 80°C for 20 min to stop enzyme activity. For each 96 well plate, I pooled 10 μ L of each sample and split this pool equally between two microcentrifuge tubes. I purified each pool of libraries with a 1X volume of SpeedBead solution followed by two ethanol washes as described in the previous section except that the DNA was resuspended in 25 μ L of TLE solution and combined the two pools of cleaned ligation product.

In order to be able to detect and remove PCR duplicates, I performed a single cycle of PCR with the iTru5-8N primer which adds a random 8 nucleotide barcode to each library construct. For each plate, I prepared four PCR reactions with a total volume of 50 μ L containing 1X Kapa Hifi Buffer (Kapa Biosystems, USA; Kapa), 0.3 μ M iTru5-8N Primer, 0.3 mM dNTP, 1 unit Kapa HiFi DNA Polymerase, 10 μ L of purified ligation product, and nuclease free water. I ran reactions through a single cycle of PCR on a thermocycler at 98°C for 2 min, 60°C for 30 s, and 72°C for 5 min. I pooled all of the PCR products for a plate into a single tube and purified the libraries with a 2X volume of SpeedBead solution as described before and resuspended in 25 μ L TLE. I added the remaining adapter and index sequences which were unique to each plate with four PCR reactions with a total volume of 50 μ L containing 1X Kapa Hifi (Kapa), 0.3 μ M iTru7

Primer, 0.3 μ M P5 Primer, 0.3 mM dNTP, 1 unit of Kapa Hifi DNA Polymerase (Kapa), 10 μ L purified iTru5-8N PCR product, and nuclease free water. I ran reactions on a thermocycler with an initial denaturation at 98°C for 2 min, followed by 6 cycles of 98°C for 20 s, 60°C for 15 s, 72°C for 30 s and a final extension of 72°C for 5 min. I pooled all of the PCR products for a plate into a single tube and purified the product with a 2X volume of SpeedBead solution as described before and resuspended in 45 μ L TLE.

I size selected the library DNA from each plate in the range of 450-650 base pairs using a BluePippin (Sage Science, USA) with a 1.5% dye free gel with internal R2 standards. To increase the final DNA concentrations I prepared four PCR reactions for each plate with 1X Kapa Hifi (Kapa), 0.3 μ M P5 Primer, 0.3 μ M P7 Primer, 0.3 mM dNTP, 1 unit of Kapa HiFi DNA Polymerase (Kapa), 10 μ L size selected DNA, and nuclease free water and used the same thermocycling conditions as the previous (P5-iTru7) amplification. I pooled all of the PCR products for a plate into a single tube and purified the product with a 2X volume of SpeedBead solution as before and resuspended in 20 μ L TLE. I quantified the DNA concentration for each plate with a Qubit fluorometer (Life Technologies, USA) then pooled each plate in equimolar amounts relative to the number of samples on the plate and diluted the pooled DNA to 5 nM with TLE solution. The pooled libraries were pooled with other projects and sequenced on an Illumina HiSeqX by Novogene (China) to obtain paired end, 150 base pair sequences.

2.2.3 Data Processing

I demultiplexed the iTru7 indexes using the *process_radtags* command from *Stacks* v2.6.4 (Rochette et al., 2019) and allowed for two mismatches for rescuing reads. To remove PCR duplicates, I used the *clone_filter* command from *Stacks*. I demultiplexed inline sample barcodes, trimmed adapter sequence, and filtered reads with low quality scores as well as reads with any uncalled bases using the *process_radtags* command again and allowed for the rescue of restriction site sequence as well as barcodes with up to two mismatches. I built alignments from the processed reads using the *Stacks* pipeline. I allowed for 14 mismatches between alleles within, as well as between individuals (M and n parameters). This is equivalent to a sequence similarity threshold of 90% for the 140 bp length of reads post trimming. I also allowed for up to 7 gaps between alleles within and between individuals. I used the *populations* command from *Stacks* to filter loci missing in more than 5% of individuals, filter all sites with minor allele counts less than 3, filter any individuals with more than 90% missing loci, and randomly sample a single SNP from each locus.

2.2.4 Genetic Clustering & Ancestry Proportions

To cluster individuals and characterize patterns of genetic differentiation and admixture between clusters, I used the Bayesian inference program *STRUCTURE* v2.3.4 (Pritchard et al., 2000) with *STRUCTURE*'s admixture model which returns an estimate of ancestry proportions for each sample. To evaluate the assumption that samples are best modeled as inheriting their genetic variation just two groups corresponding to the species identification made in the field, I ran *STRUCTURE* under four different models, each with a different number of assumed clusters of individuals (K parameter) ranging from 1 to 4. For each value of K, I ran 20 iterations for 100,000 total steps with the first 50,000 as burnin. I used the R package *POPHelper* v2.3.1 (Francis, 2017) to

combine iterations for each value of K and to select the model producing the largest ΔK which is the the model that has the greatest increase in likelihood score from the previous model with K-1 as described by (Evanno et al., 2005). I also examined genetic clustering and evidence of admixture using a non-parametric approach with a principal component analysis (PCA) implemented in the R package *adegenet* v2.1.10 (Jombart, 2008). I visualized the relationship between the first principal component axis and the estimated admixture proportion for each individual to check for agreement between the parametric *STRUCTURE* analysis and the non-parametric PCA analysis.

2.2.5 Genomic Cline Analysis

To investigate patterns of introgression across the hybrid zone I used the bayesian genomic cline inference tool *BGC* v1.03 (Gompert & Buerkle, 2012) to infer parameters under a genomic cline model. I classified a sample as being admixed if it had an inferred admixture proportion of <95% for one of the parent species under the model with a K of two in the *STRUCTURE* analysis. I used *VCFtools* vX.XX.XX to filter all non-biallelic sites from the the VCF file produced by the *populations* program. I converted the VCF formatted data into the *BGC* format using *bgc_utils* v0.1.0 https://github.com/kerry Cobb/bgc_utils. I ran *BGC* with 5 independent chains, each for 1,000,000 steps and sampling every 1000. I discarded the first 10% of samples from the posterior, combined the independent chains, summarized the posterior samples, and identified exceptional loci with *bgc_utils*.

I classified loci as exceptional and therefore expected to be enriched for incompatibility loci with two different approaches, (1) if locus specific introgression differed from the genome-wide average and (2) if locus specific introgression is statistically unlikely relative to the genome-wide distribution of locus specific introgression.

More specifically, in the first approach I classified a locus as exceptional if the 90% highest posterior density interval (HPDI) for the alpha or beta parameter did not cover zero.

In the second approach, I classified a locus as exceptional if the median of the posterior sample for the α or β parameters for a locus were outside of the interval from 0.05 to 0.95 of the probability density functions $Normal(0, \tau_\alpha)$ or $Normal(0, \tau_\beta)$ respectively, where τ_α and τ_β are the median values from the posterior sample for the genome wide average

the median of the posterior probability distribution was not contained within the interval from 0.05 to 0.95 of the probability density function $Normal(0, \tau_\alpha)$ or $Normal(0, \tau_\beta)$ with τ_α and τ_β being the medians of the posterior samples for conditional random effect priors on τ_α and τ_β parameters as described by (Gompert & Buerkle, 2011).

2.2.6 Genetic differentiation and Introgression

I calculated F_{st} for each site used in the *BGC* analysis using *VCFtools*

2.3 Results

2.3.1 Sampling and Data Processing

I prepared reduced-representation sequencing libraries from 173 samples collected for this study (Table 2.1) and 19 samples available from existing collections (Table 2.2)).

After assembly and filtering, 43 samples were excluded from analyses. Of the remaining samples, 56 had been identified as most closely resembling *A. americanus* and 93 had been identified as most closely resembling *A. terrestis*. Stacks assembled reads into 432,336 loci with a mean length of 253.31 bp. Prior to filtering the mean coverage per sample was 32X. After filtering loci missing in greater than 5% of samples, filtering sites with minor allele counts less than 3, filtering individuals with greater than 90% missing loci, and randomly sampling a single SNP from each locus, 1194 sites remained.

A visual inspection of the *STRUCTURE* results shows that each iteration with same value for K converged on very similar results (Fig. 2.2).

Using the method described by (Evanno et al., 2005), I determined the best value for K to be 2 as expected Fig. 2.3.

XX pure americanus XX pure terrestis XX admixed individuals

The admixed individuals are concentrated near the contact zone Fig. 2.4

2.4 Discussion

With weak selection and low dispersal the false discovery rate will be higher (Gompert & Buerkle, 2011) Outliers loci are still expected to be enriched for genetic regions under selection. False discovery rate when patterns of introgression are affected by epistasis is quite high. Not all regions affecting introgression may be among outliers.

One limitation of this study is samples included in analysis are sampled from a fairly broad area and may be more appropriately treated as separate populations. They could be affected by different processes. This hybrid zone could serve as an excellent system to explore such things given the large extent. Many important environmental factors likely differ across the hybrid zone.

A reference genome for these species would greatly enhance the explanatory power of this analysis allowing for an exploration of the genomic architecture of the reproductive isolation between these lineages. Which regions of the genome are involved in isolation. Are they more associated with sex chromosomes? The availability of vertebrate genomes is increasing rapidly so it will likely be possible to map the outlier loci identified in this study to a reference genome for *A. americanus* or *A. terrestis* to reveal patterns in the distribution of these putative markers for genetic incompatibility.

Mitochondrial data would also be valuable. Would allow for the detection of mitonuclear incompatibilities that could be contributing to reproductive isolation. Could also be used to test for the effect of Haldane's rule which would predict that females would be more affected by hybridization. The estimated cline for mitochondria would be expected to be much steeper (carling2008).

With this study I have provided a characterization of this hybrid zone that can serve as a guide for future sampling. Sampling along transects. Could then apply geographic cline methods Would give spatial insights.

2.5 Figures

410

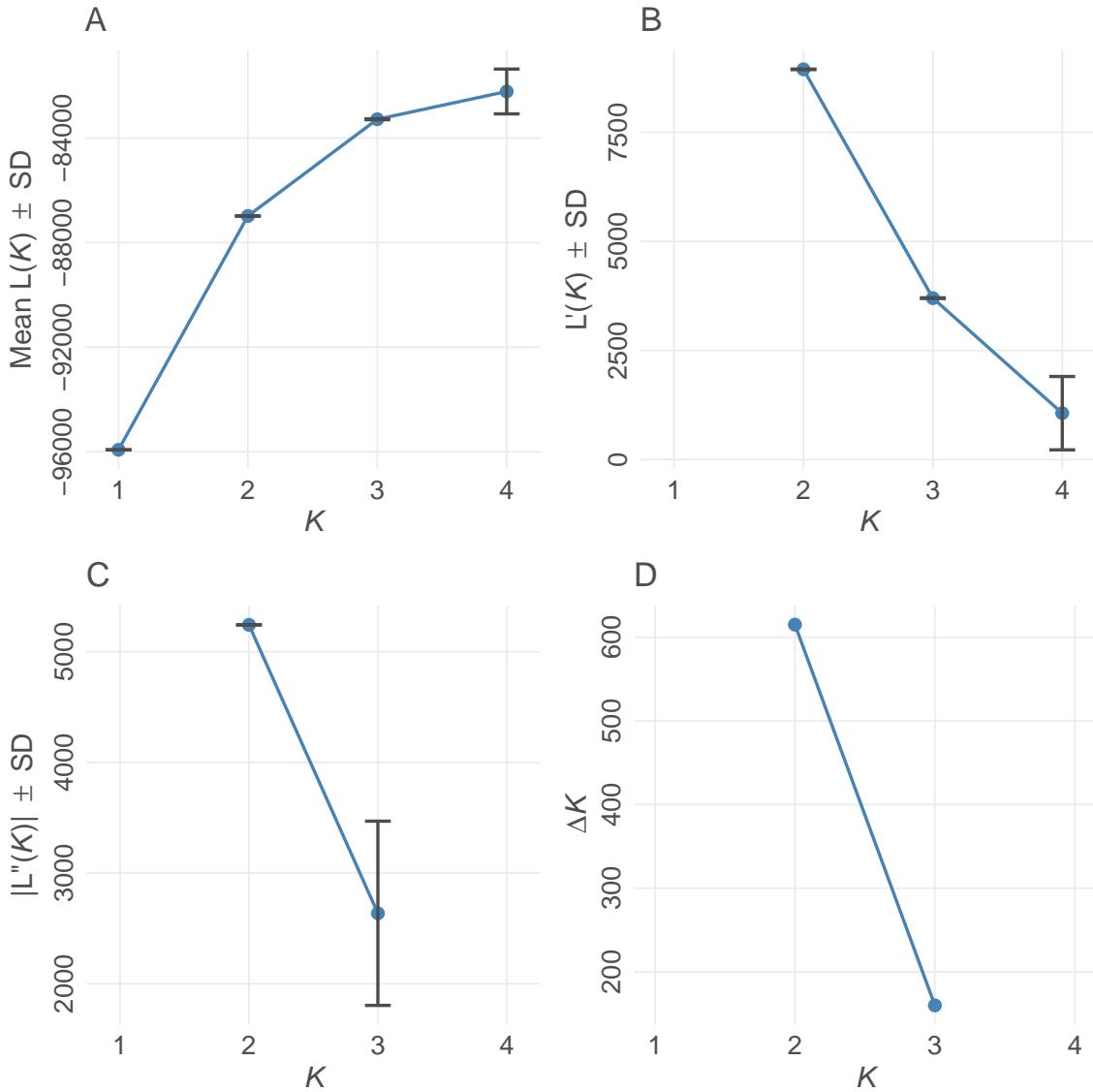


Figure 2.1. Evanno method for optimal value for K in *STRUCTURE* (Evanno et al., 2005). K refers to the number of populations for each of the different *STRUCTURE* models examined. (A) Mean estimated \ln probability of data over 10 iterations for each value of $K \pm SD$. (B) Rate of change of the likelihood distribution (mean $\pm SD$) (C) Absolute values of the second order rate of change of the likelihood distribution (mean $\pm SD$) (D) ΔK . The modal value of this distribution is considered the true value of K for the data. Plot created using *POPHELPER* (Francis, 2017).

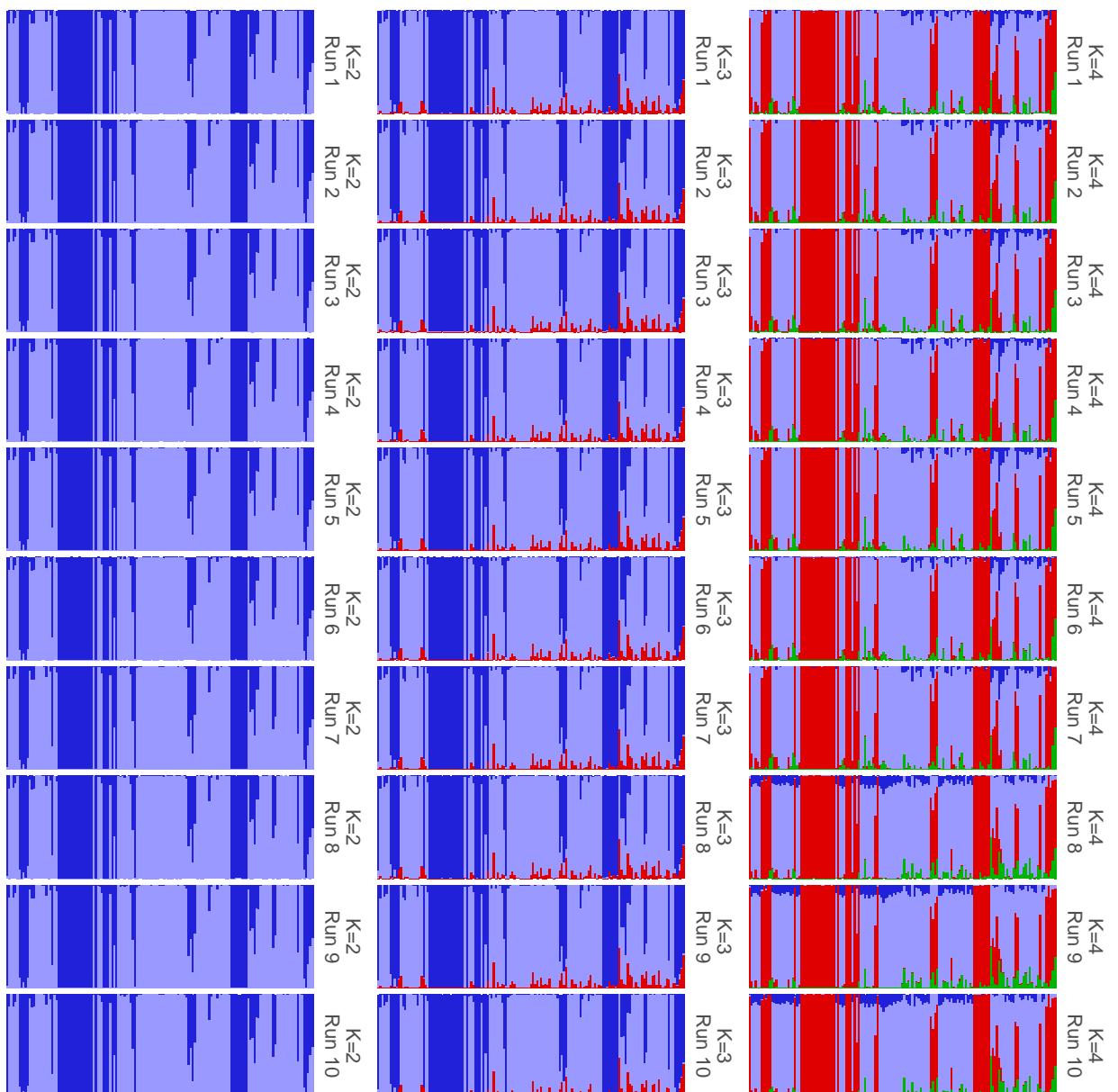


Figure 2.2. Results of each iteration of *STRUCTURE* showing convergence among iterations within runs having the same value for K . Plot was created with *POPHELPER* (Francis, 2017).

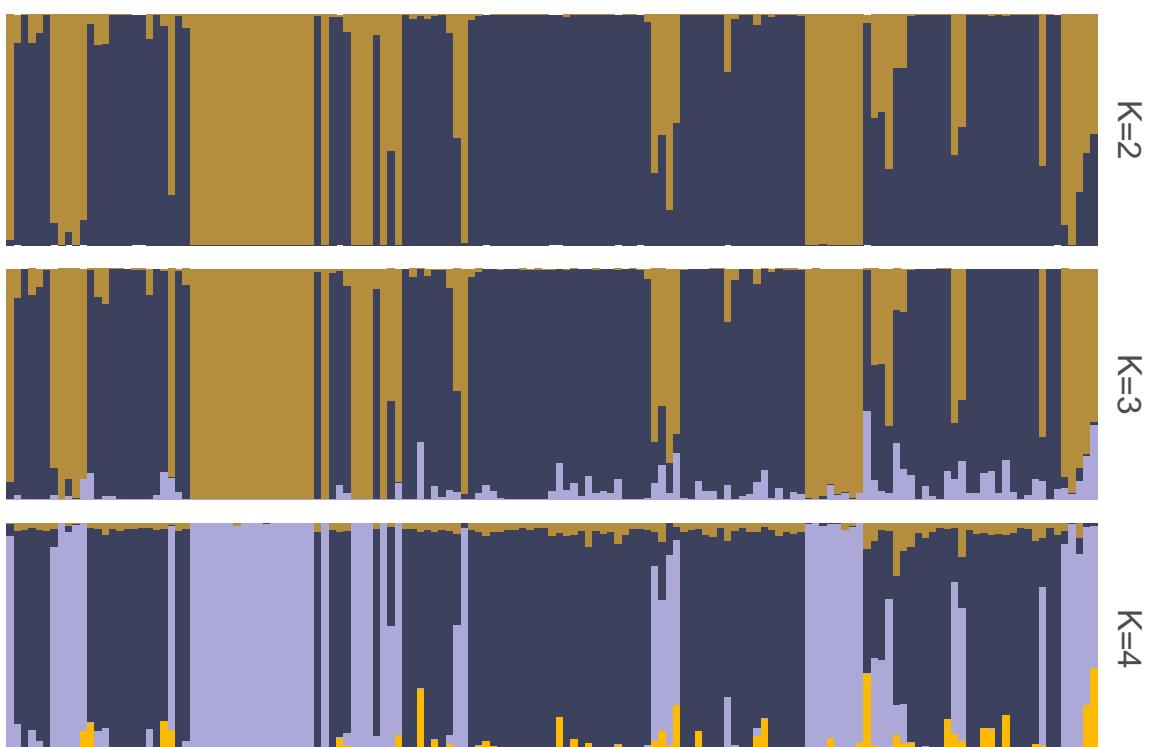


Figure 2.3. Summarized *STRUCTURE* results for each value of K. Ancestry proportions shown are the mean of ancestry proportions across all iterations. Summarization and plotting done using *POPHELP* (Francis, 2017).

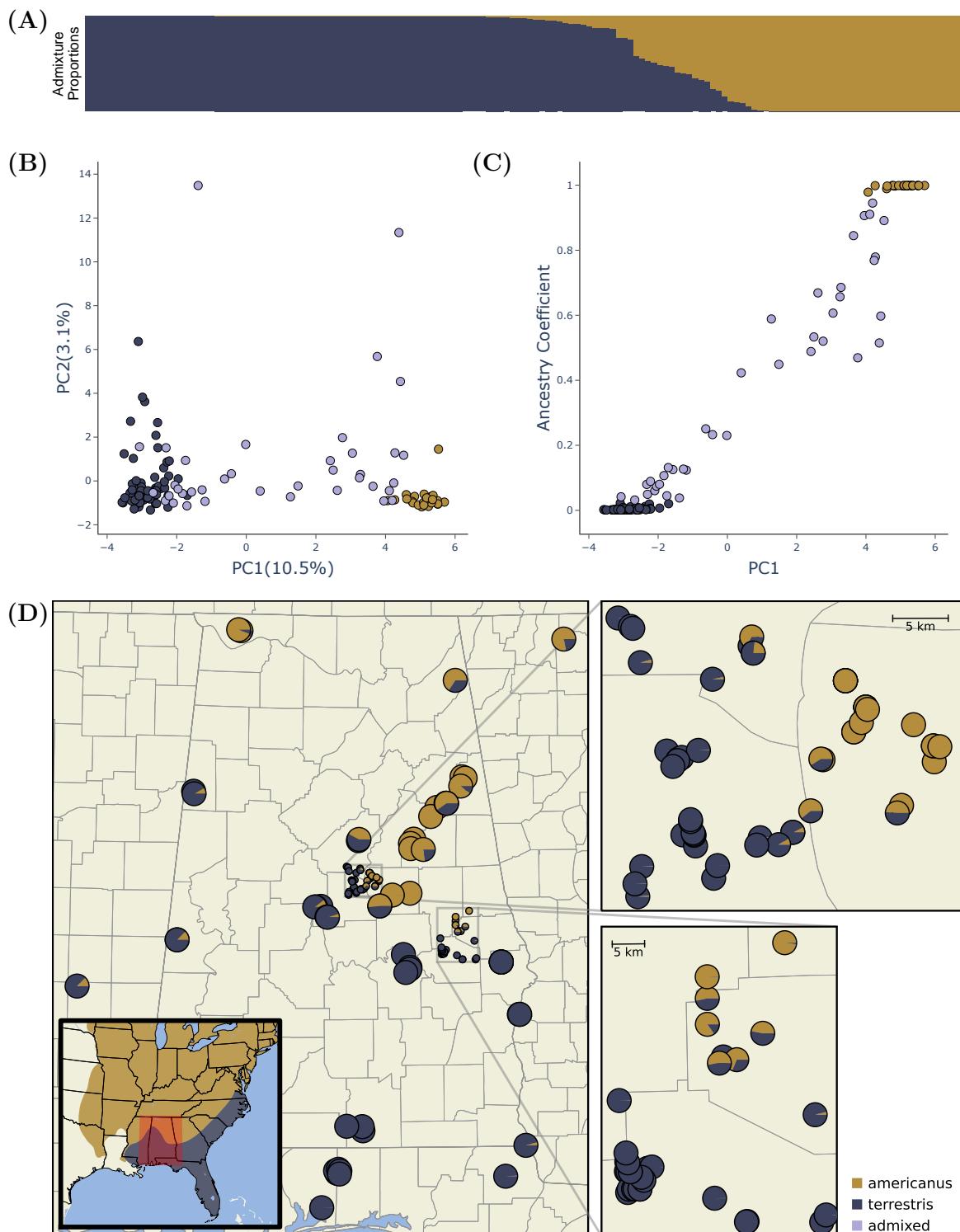


Figure 2.4. Genetic evidence of hybridization between *A. americanus* and *A. terrestris*. (A) *STRUCTURE* plot showing estimated ancestry proportions. (B) Summary of population genetic structure based on the principle component axes one (PC1) and two (PC2). These axes explain 10.5% (PC1) and 3.1% (PC2) of the genetic variation among individuals. (C) Relationship between the first principal component axis and the admixture proportions estimated with *STRUCTURE*. (D) Sample map showing the sampling location and estimated ancestry proportion of each sample. The inset map shows the approximate ranges of each species and the study area highlighted in red. Figure created using *POPHelper* (Francis, 2017) and *Matplotlib* (Hunter, 2007)

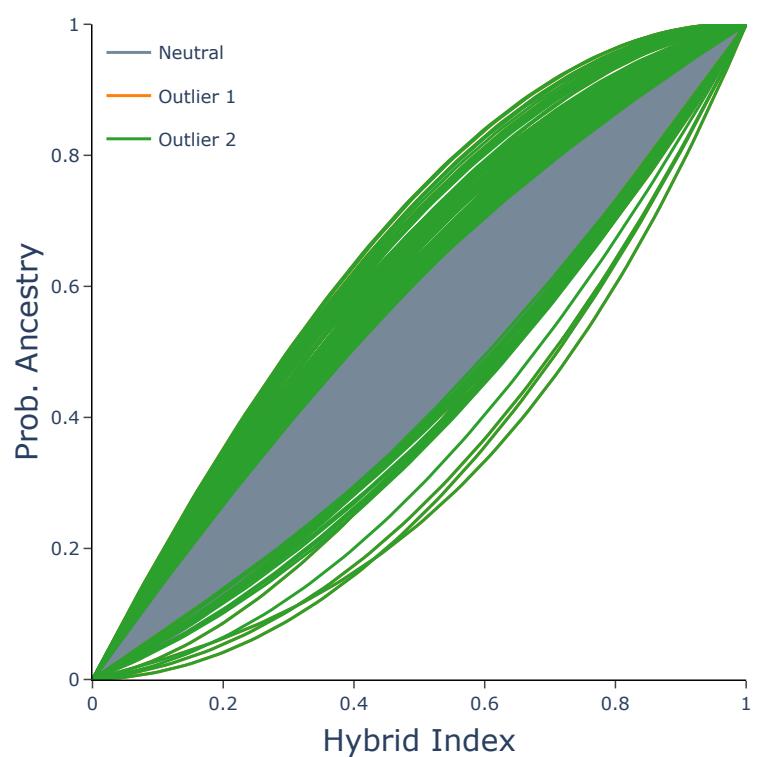


Figure 2.5. Shape of genomic clines estimated for each locus with *BGC*. Outliers are highlighted with XX.

2.6 Tables

411

Table 2.1. Samples collected for this study

Sample ID	Species	Latitude	Longitude	Passed Filtering
KAC 016	<i>terrestris</i>	30.54819	-86.93067	X
KAC 038	<i>terrestris</i>	32.81470	-86.93968	X
KAC 039	<i>terrestris</i>	32.81094	-86.98967	X
KAC 040	<i>terrestris</i>	32.80985	-86.99795	X
KAC 042	<i>terrestris</i>	32.82406	-86.99314	
KAC 043	<i>terrestris</i>	32.82406	-86.99314	
KAC 044	<i>terrestris</i>	32.80450	-87.03078	
KAC 045	<i>terrestris</i>	32.76703	-87.07073	
KAC 046	<i>terrestris</i>	32.76592	-87.07184	
KAC 047	<i>terrestris</i>	32.78932	-86.90850	
KAC 048	<i>terrestris</i>	32.73575	-86.88149	X
KAC 049	<i>terrestris</i>	32.73291	-86.87707	X
KAC 050	<i>terrestris</i>	32.74822	-86.79806	
KAC 051	<i>terrestris</i>	32.78742	-86.75847	
KAC 052	<i>terrestris</i>	32.78044	-86.73877	
KAC 070	<i>americanus</i>	34.79963	-84.57678	X
KAC 071	<i>terrestris</i>	32.43478	-85.64630	
KAC 074	<i>terrestris</i>	30.77430	-85.22690	X
KAC 075	<i>terrestris</i>	32.94778	-86.63224	X
KAC 076	<i>terrestris</i>	32.94970	-86.52687	
KAC 077	<i>terrestris</i>	32.94970	-86.52687	
KAC 078	<i>americanus</i>	33.00267	-86.38960	X
KAC 079	<i>americanus</i>	33.01205	-86.47872	
KAC 080	<i>americanus</i>	33.04456	-86.45547	
KAC 081	<i>americanus</i>	33.04456	-86.45547	X
KAC 082	<i>americanus</i>	33.04456	-86.45547	X
KAC 083	<i>americanus</i>	33.04456	-86.45547	X
KAC 084	<i>americanus</i>	33.04456	-86.45547	X
KAC 085	<i>americanus</i>	33.04456	-86.45547	
KAC 086	<i>americanus</i>	33.04456	-86.45547	X
KAC 087	<i>americanus</i>	33.01484	-86.39040	X
KAC 089	<i>americanus</i>	33.01484	-86.39040	X
KAC 090	<i>americanus</i>	33.06472	-86.47496	X
KAC 091	<i>americanus</i>	33.06472	-86.47496	X
KAC 092	<i>americanus</i>	33.06472	-86.47496	
KAC 093	<i>americanus</i>	33.06472	-86.47496	X
KAC 094	<i>americanus</i>	33.06472	-86.47496	X
KAC 095	<i>americanus</i>	33.06472	-86.47496	X
KAC 096	<i>americanus</i>	33.06472	-86.47496	X
KAC 097	<i>americanus</i>	33.06472	-86.47496	X
KAC 098	<i>americanus</i>	33.02572	-86.46711	X
KAC 099	<i>americanus</i>	33.02572	-86.46711	X
KAC 100	<i>terrestris</i>	32.92374	-86.67199	X

Continued on next page

Table 2.1 – continued from previous page

Sample ID	Species	Latitude	Longitude	Passed Filtering
KAC 101	<i>americanus</i>	33.03283	-86.45975	X
KAC 102	<i>terrestris</i>	32.94544	-86.55777	X
KAC 103	<i>terrestris</i>	32.94947	-86.52630	X
KAC 104	<i>terrestris</i>	32.94947	-86.52630	X
KAC 105	<i>americanus</i>	33.04278	-86.45377	X
KAC 106	<i>americanus</i>	33.00464	-86.49692	X
KAC 107	<i>americanus</i>	33.01416	-86.38417	X
KAC 108	<i>terrestris</i>	32.94013	-86.54004	X
KAC 109	<i>terrestris</i>	32.94173	-86.55787	
KAC 110	<i>americanus</i>	33.03099	-86.40941	X
KAC 111	<i>americanus</i>	33.00518	-86.49895	X
KAC 112	<i>terrestris</i>	32.95011	-86.53723	
KAC 113	<i>americanus</i>	33.00528	-86.38897	
KAC 114	<i>americanus</i>	33.01617	-86.40318	
KAC 115	<i>americanus</i>	32.98218	-86.40488	
KAC 116	<i>americanus</i>	32.96964	-86.42137	X
KAC 117	<i>terrestris</i>	32.97146	-86.52901	
KAC 121	<i>terrestris</i>	32.44120	-85.65386	X
KAC 122	<i>terrestris</i>	32.85411	-86.76619	
KAC 123	<i>terrestris</i>	32.90084	-86.67587	X
KAC 124	<i>terrestris</i>	32.91060	-86.67850	X
KAC 125	<i>terrestris</i>	32.91715	-86.68208	
KAC 126	<i>terrestris</i>	32.92717	-86.67407	
KAC 127	<i>terrestris</i>	32.97159	-86.62516	
KAC 128	<i>terrestris</i>	33.00585	-86.63703	
KAC 129	<i>terrestris</i>	33.00797	-86.64210	
KAC 130	<i>terrestris</i>	33.00818	-86.64333	
KAC 131	<i>terrestris</i>	33.01508	-86.64937	
KAC 132	<i>terrestris</i>	33.02034	-86.66651	
KAC 133	<i>terrestris</i>	33.01163	-86.64759	X
KAC 134	<i>terrestris</i>	33.00537	-86.63652	X
KAC 135	<i>terrestris</i>	33.00644	-86.63368	X
KAC 136	<i>terrestris</i>	33.00673	-86.63316	X
KAC 138	<i>americanus</i>	32.70224	-85.66196	X
KAC 139	<i>americanus</i>	32.73042	-85.66173	X
KAC 140	<i>terrestris</i>	32.62553	-85.63684	X
KAC 141	<i>terrestris</i>	32.41032	-85.60107	X
KAC 142	<i>terrestris</i>	32.57011	-85.80888	X
KAC 143	<i>terrestris</i>	32.47773	-85.79824	X
KAC 144	<i>terrestris</i>	32.47707	-85.79577	X
KAC 145	<i>terrestris</i>	32.48128	-85.76354	X
KAC 146	<i>terrestris</i>	32.48291	-85.75622	X
KAC 147	<i>terrestris</i>	32.45001	-85.79652	X
KAC 148	<i>terrestris</i>	32.45420	-85.79408	X
KAC 149	<i>terrestris</i>	32.45449	-85.78664	X

Continued on next page

Table 2.1 – continued from previous page

Sample ID	Species	Latitude	Longitude	Passed Filtering
KAC 150	<i>terrestris</i>	32.45449	-85.78664	X
KAC 151	<i>terrestris</i>	32.45451	-85.78416	X
KAC 152	<i>terrestris</i>	32.45423	-85.77634	X
KAC 153	<i>terrestris</i>	32.45423	-85.77634	X
KAC 154	<i>terrestris</i>	32.46574	-85.76977	X
KAC 155	<i>terrestris</i>	32.46961	-85.77369	X
KAC 156	<i>terrestris</i>	32.47709	-85.79175	X
KAC 158	<i>terrestris</i>	32.47709	-85.79175	X
KAC 159	<i>terrestris</i>	32.49000	-85.79741	X
KAC 160	<i>terrestris</i>	32.40809	-85.47857	X
KAC 161	<i>terrestris</i>	32.41744	-85.47117	X
KAC 162	<i>terrestris</i>	32.35417	-86.09838	X
KAC 163	<i>terrestris</i>	32.33994	-86.09946	X
KAC 164	<i>terrestris</i>	32.31562	-86.13789	X
KAC 167	<i>terrestris</i>	33.06620	-86.60328	X
KAC 172	<i>americanus</i>	32.62171	-85.61467	X
KAC 173	<i>americanus</i>	32.61751	-85.64335	X
KAC 176	<i>americanus</i>	32.66836	-85.66233	X
KAC 177	<i>americanus</i>	32.65571	-85.57134	X
KAC 181	<i>terrestris</i>	32.38644	-85.23561	X
KAC 182	<i>terrestris</i>	32.38579	-85.23565	X
KAC 183	<i>terrestris</i>	32.38579	-85.23565	X
KAC 184	<i>terrestris</i>	32.38579	-85.23565	X
KAC 185	<i>terrestris</i>	32.38579	-85.23565	X
KAC 187	<i>americanus</i>	32.64548	-85.55135	
KAC 188	<i>terrestris</i>	32.40976	-85.60208	X
KAC 189	<i>terrestris</i>	33.09152	-86.56686	X
KAC 190	<i>terrestris</i>	33.11298	-86.69434	X
KAC 191	<i>terrestris</i>	33.10659	-86.68228	X
KAC 192	<i>terrestris</i>	33.10509	-86.68014	X
KAC 193	<i>terrestris</i>	33.07896	-86.67286	X
KAC 194	<i>terrestris</i>	32.93933	-86.62008	X
KAC 195	<i>terrestris</i>	32.94745	-86.62146	X
KAC 196	<i>terrestris</i>	32.94829	-86.62190	X
KAC 197	<i>terrestris</i>	32.94929	-86.62241	X
KAC 198	<i>terrestris</i>	32.95077	-86.62306	
KAC 199	<i>terrestris</i>	32.95794	-86.62477	X
KAC 200	<i>terrestris</i>	32.95940	-86.62489	X
KAC 205	<i>terrestris</i>	32.54852	-85.48692	X
KAC 206	<i>americanus</i>	33.30759	-86.58201	X
KAC 207	<i>americanus</i>	33.31685	-86.57596	X
KAC 208	<i>americanus</i>	33.09829	-86.56529	X
KAC 209	<i>terrestris</i>	33.08600	-86.56394	X
KAC 210	<i>terrestris</i>	33.08600	-86.56394	X
KAC 211	<i>terrestris</i>	33.01464	-86.60995	

Continued on next page

Table 2.1 – continued from previous page

Sample ID	Species	Latitude	Longitude	Passed Filtering
KAC 212	<i>terrestris</i>	33.01208	-86.61707	X
KAC 213	<i>terrestris</i>	33.00435	-86.63710	X
KAC 214	<i>terrestris</i>	32.99991	-86.64181	X
KAC 215	<i>terrestris</i>	32.99605	-86.64526	
KAC 216	<i>terrestris</i>	33.01346	-86.60960	
KAC 217	<i>terrestris</i>	32.91470	-86.60270	X
KAC 218	<i>terrestris</i>	32.92432	-86.59895	X
KAC 219	<i>terrestris</i>	32.93987	-86.56113	X
KAC 220	<i>americanus</i>	32.96579	-86.50892	X
KAC 221	<i>americanus</i>	32.96389	-86.42549	X
KAC 223	<i>terrestris</i>	32.53362	-85.79839	
KAC 224	<i>terrestris</i>	32.48869	-85.79555	X
KAC 225	<i>terrestris</i>	32.50159	-85.79860	X
KAC 230	<i>terrestris</i>	30.80933	-86.77686	X
KAC 232	<i>terrestris</i>	30.80922	-86.78994	X
KAC 233	<i>terrestris</i>	30.80922	-86.78994	X
KAC 234	<i>terrestris</i>	30.80922	-86.78994	X
KAC 236	<i>terrestris</i>	30.82632	-86.80258	X
KAC 237	<i>terrestris</i>	30.83733	-86.77630	X
KAC 238	<i>terrestris</i>	30.82433	-86.76284	X
KAC 239	<i>terrestris</i>	30.80162	-86.76659	X
KAC 242	<i>americanus</i>	34.50446	-85.63768	X
KAC t1020	<i>terrestris</i>	31.10783	-86.62247	
KAC t1030	<i>terrestris</i>	31.99042	-85.07423	X
KAC t1040	<i>terrestris</i>	31.99016	-85.07046	X
KAC t2004	<i>americanus</i>	33.58295	-85.73524	X
KAC t2015	<i>americanus</i>	33.58435	-85.74064	X
KAC t2018-02-17-01	<i>americanus</i>	33.55274	-85.82913	X
KAC t2018-02-17-04	<i>americanus</i>	33.48548	-85.88857	X
KAC t2018-02-17-05	<i>americanus</i>	33.31649	-86.05293	X
KAC t2018-02-17-06	<i>americanus</i>	33.28443	-86.08443	X
KAC t2018-02-17-07	<i>americanus</i>	33.24576	-86.08168	X
KAC t2018-03-10-1	<i>americanus</i>	32.91057	-86.09272	X
KAC t2018-03-10-3	<i>americanus</i>	32.95104	-86.14539	
KAC t2018-03-10-4	<i>americanus</i>	32.89787	-86.26061	X
KAC t2018-03-10-5	<i>americanus</i>	32.81642	-86.38018	X
KAC t2019-08-25-1	<i>americanus</i>	34.21852	-87.36662	
KAC t2020	<i>americanus</i>	33.23853	-85.96270	X
KAC t2040	<i>americanus</i>	33.58295	-85.73539	X
KAC t2043	<i>americanus</i>	32.81642	-86.38018	X

Table 2.2. Samples loaned from museums

Sample ID	Species	Latitude	Longitude	Passed Filtering
AHT 1975	<i>americanus</i>	32.77356	-85.53325	X
AHT 2456	<i>terrestris</i>	32.19494	-89.23629	X
AHT 2885	<i>terrestris</i>	32.45090	-86.15934	X
AHT 3419	<i>terrestris</i>	33.67290	-88.16068	X
AHT 3421	<i>terrestris</i>	33.65420	-88.15580	X
AHT 3428	<i>terrestris</i>	31.12679	-86.54755	X
AHT 3459	<i>americanus</i>	34.88028	-87.71849	X
AHT 3460	<i>americanus</i>	33.78013	-85.58421	X
AHT 3461	<i>americanus</i>	34.88779	-87.74103	X
AHT 3462	<i>americanus</i>	33.77001	-85.55434	X
AHT 3463	<i>americanus</i>	33.71125	-85.59762	X
AHT 3813	<i>terrestris</i>	31.13854	-86.53906	
AHT 3833	<i>terrestris</i>	31.00422	-85.03427	X
AHT 3997	<i>terrestris</i>	32.55607	-88.29975	X
AHT 3998	<i>terrestris</i>	32.55607	-88.29975	X
AHT 5276	<i>terrestris</i>	31.55613	-86.82514	
AHT 5277	<i>terrestris</i>	31.15830	-86.55430	X
AHT 5278	<i>terrestris</i>	31.16105	-86.69868	X
UTEP 19947	<i>terrestris</i>	31.22432	-88.77548	

Chapter 3

412

Comparison of Linked versus Unlinked Character Models for Species Tree Inference

413

414

415

3.1 Introduction

416

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.

417

418

419

420

421

422

423

424

Reduced-representation genomic data sets acquired from high-throughput instruments are becoming commonplace in phylogenetics (Leaché & 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; De Maio et al., 2015), or (2) contiguous, linked sites evolved along a shared gene tree (Heled & Drummond, 2010; Liu & Pearl, 2007; 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 (Ogilvie et al., 2017; Yang, 2015). 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; De 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.

425

426

427

428

429

430

431

432

433

434

435

436

437

438

439

440

441

442

443

444

Investigators are thus faced with decisions about how best to use their data to in-

fer 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. Im-
portant 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 us-
ing polymerase chain reaction (PCR) which can introduce mutational error at a rate of
up to 1.5×10^{-5} substitutions per base (Potapov & 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 (Linck &
Battey, 2019; Rochette et al., 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 & Knowles, 2016; Linck & 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.

3.2 Methods 491

3.2.1 Simulations of error-free data sets 492

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 descendant populations (terminal branches) with constant effective sizes of N_e^{D1} and N_e^{D2} (Fig. 3.1). 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 intra-locus 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 & 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 & 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 & 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.

3.2.2 Introducing Site-pattern Errors 522

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.

3.2.3 Assessing Sensitivity to Errors	535
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 <i>ecoevolity</i> (Version 0.3.2, Commit a7e9bf2; Oaks, 2019) and a linked-character model using the <i>StarBEAST2</i> package (Version 0.15.1; Ogilvie et al., 2017) in <i>BEAST2</i> (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 <i>ecoevolity</i> was parameterized to be relative to the mean effective size of the descendant populations. We added an option to <i>ecoevolity</i> to compile a version where the prior is specified as the absolute effective size of the root population, which matches the model in <i>StarBEAST2</i> 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 <i>ecoevolity</i> (Bryant et al., 2012; Oaks, 2019). Therefore, we also analyzed each data set with <i>ecoevolity</i> after selecting, at most, one variable character from each locus; loci without variable sites were excluded.	536 537 538 539 540 541 542 543 544 545 546 547 548 549 550 551
We analyzed the data sets simulated with 1-character per locus (i.e., unlinked data) with <i>ecoevolity</i> . Our goal with these analyses was to verify that the generative model of our simulation pipeline matched the underlying model of <i>ecoevolity</i> , and to confirm that any behavior of the method with the other simulated data sets was not being caused by the linkage violation.	552 553 554 555 556
For <i>ecoevolity</i> , we ran four independent Markov chain Monte Carlo (MCMC) analyses with 75,000 steps and a sample frequency of 50 steps. For <i>StarBEAST2</i> , we ran two independent MCMC analyses with 20 million steps and a sample frequency of 5000 steps. To assess convergence and mixing of the <i>ecoevolity</i> and <i>StarBEAST2</i> MCMC chains, we computed the effective sample size (ESS; Gong & Flegal, 2016) and potential scale reduction factor (PSRF; the square root of Equation 1.1 in Brooks & Gelman, 1998) from the samples of each parameter, and considered an ESS value greater than 200 and PSRF less than 1.2 (Brooks & 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 <i>ecoevolity</i> and <i>StarBEAST2</i> , leaving 4000 and 7600 posterior samples for each data set, respectively.	557 558 559 560 561 562 563 564 565 566 567
3.2.4 Project repository	568
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.	569 570 571
3.3 Results	572
3.3.1 Behavior of linked (<i>StarBEAST2</i>) versus unlinked (<i>ecoevolity</i>) character models	573 574
The divergence times estimated by the linked-character method, <i>StarBEAST2</i> , were very accurate and precise for all alignment lengths and types and degrees errors, despite	575 576

poor MCMC mixing (i.e., low ESS values) for shorter loci (Figs. 3.2–3.4). 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. 3.2–3.4). However when alignments contained errors, *ecoevolity* underestimated very recent divergence times with increasing severity as the frequency of errors increased (Figs. 3.2–3.4); 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. 3.5–3.7). When analyzing the alignments without errors, *ecoevolity* essentially returned the prior distribution on the effective size of the ancestral population (Figs. 3.5–3.7). 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. 3.5–3.7), 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. 3.8–3.10). 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. 3.8–3.10).

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. 3.11) as we did with the linked loci (Figs. 3.2–3.4). 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.

3.3.2 Analyzing all sites versus SNPs with *ecoevolity*

The unlinked character model implemented in *ecoevolity* assumes that orthologous 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. 3.2–3.10). 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. 3.8–3.10).

3.3.3 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 un-linked characters (i.e., no linked sites). This coverage behavior is expected, and helps to 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.

3.3.4 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. 3.2–3.10; 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. 3.5–3.7). 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. 3.8–3.10).

3.4 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.

3.4.1 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. 3.2–3.4). 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. 3.2–3.4). 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. 3.8–3.10). 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. 3.8–3.10), 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. 3.2–3.4) and ancestral population size (Figs. 3.5–3.7).

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 errors. This pattern makes sense in light of how we generated these errors. We introduced singleton errors persist and heterozygosity errors per-locus. Thus, the same per-locus rate of heterozygosity 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.

3.4.2 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 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 & Knowles, 2016; Linck & Battey, 2019). If the thresholds for such processing steps are stringent, it could introduce levels of error greater than our simulations.

3.4.3 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
753 data might indicate that the model is being biased by errors or acquisition biases present
754 in the character patterns.
755

Consistent with what has been shown in previous work (Oaks, 2019; Oaks et al., 2019),
756 *ecoevolity* performed better when all sites were utilized despite violating the assumption
757 that all sites are unlinked. This suggests that investigators might obtain better estimates
758 by analyzing all their data under unlinked-character models, rather than discarding much
759 of it to avoid violating an assumption of the model. Given that the model of unlinked
760 characters implemented in *ecoevolity* does not use information about linkage among sites
761 (Bryant et al., 2012; Oaks, 2019), it is not surprising that this model violation does not
762 introduce a bias. Linkage among sites does not change the gene trees and site patterns
763 that are expected under the model, but it does reduce the variance of the those patterns
764 due to them evolving along fewer gene trees. As a result, the accuracy of the parameter
765 estimates is not affected by the linkage among sites within loci, but the credible intervals
766 become too narrow as the length of loci increase (Oaks, 2019; Oaks et al., 2019). However,
767 it remains to be seen whether the robustness of the model’s accuracy to linked sites holds
768 true for larger species trees.
769

3.4.4 Other complexities of empirical data in need of exploration

770

Our goal was to compare the theoretical performance of linked and unlinked character
771 models, not their current software implementations. Accordingly, to minimize differences
772 in performance that are due to differences in algorithms for exploring the space of gene
773 and species trees, we restricted our simulations to two species model and a small number
774 of individuals. Nonetheless, exploring how character-pattern errors and biases affect the
775 inference of larger species trees would be informative. The species tree topology is usually
776 a parameter of great interest to biologists, so it would be interesting to know whether
777 the linked model continues to be more robust to errors than the unlinked model as the
778 number of species increases. We saw the MCMC performance of *StarBEAST2* decline
779 concomitantly with locus length in our simulations due to greater uncertainty in gene
780 trees. Given that data sets frequently contain loci shorter than 250 bp, it is important
781 to know whether good sampling of the posterior of linked-character models becomes
782 prohibitive for larger trees. Also, *ecoevolity* greatly overestimated the effective size of
783 the ancestral population in the face of high rates of errors in the data. Exploring larger
784 trees will also determine whether this behavior is limited to the root population or is a
785 potential problem for all internal branches of the specie tree.
786

Exploring other types of errors and biases would also be informative. To generate
787 alignments of orthologous loci from high-throughput data, sequences are matched to a
788 similar portion of a reference sequence or clustered together based on similarity. To avoid
789 aligning paralogous sequences it is necessary to establish a minimum level of similarity for
790 establishing orthology between sequences. This can lead to an acquisition bias due to the
791 exclusion of more variable loci or alleles from the alignment (Huang & Knowles, 2016).
792 Furthermore, when a reference sequence is used, this data filtering will not be random
793 with respect to the species, but rather there will be a bias towards filtering loci and alleles
794 with greater sequence divergence from the reference. Simulations exploring the affect of
795 these types of data acquisition biases would complement the errors we explored here.
796

In our analyses, there was no model misspecification other than the introduced er-
797 rors (except for the linked sites violating the unlinked-character model). With empirical
798

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.

3.5 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., Heled, J., Kühnert, D., Vaughan, T., Wu, C.-H., Xie, D., Suchard, M. A., Rambaut, A., & Drummond, A. J. (2014). BEAST 2: A Software Platform for Bayesian Evolutionary Analysis (A. Prlic, Ed.). *PLoS Computational Biology*, 10(4), e1003537. <https://doi.org/10.1371/journal.pcbi.1003537>
- Brooks, S. P., & Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of computational and graphical statistics*, 7(4), 434–455.
- Bryant, D., Bouckaert, R., Felsenstein, J., Rosenberg, N. A., & RoyChoudhury, A. (2012). Inferring Species Trees Directly from Biallelic Genetic Markers: Bypassing Gene Trees in a Full Coalescent Analysis. *Molecular Biology and Evolution*, 29(8), 1917–1932. <https://doi.org/10.1093/molbev/mss086>
- De Maio, N., Schrempf, D., & Kosiol, C. (2015). PoMo: An Allele Frequency-Based Approach for Species Tree Estimation. *Systematic Biology*, 64(6), 1018–1031. <https://doi.org/10.1093/sysbio/syv048>
- Gong, L., & Flegal, J. M. (2016). A Practical Sequential Stopping Rule for High-Dimensional Markov Chain Monte Carlo. *Journal of Computational and Graphical Statistics*, 25(3), 684–700. <https://doi.org/10.1080/10618600.2015.1044092>
- Harvey, M. G., Judy, C. D., Seeholzer, G. F., Maley, J. M., Graves, G. R., & Brumfield, R. T. (2015). Similarity thresholds used in DNA sequence assembly from short reads can reduce the comparability of population histories across species. *PeerJ*, 3, e895. <https://doi.org/10.7717/peerj.895>
- Heled, J., & Drummond, A. J. (2010). Bayesian Inference of Species Trees from Multilocus Data. *Molecular Biology and Evolution*, 27(3), 570–580. <https://doi.org/10.1093/molbev/msp274>

Huang, H., & Knowles, L. L. (2016). Unforeseen Consequences of Excluding Missing Data from Next-Generation Sequences: Simulation Study of RAD Sequences. <i>Systematic Biology</i> , 65(3), 357–365. https://doi.org/10.1093/sysbio/syu046	841
	842
	843
Hunter, J. D. (2007). Matplotlib: A 2D Graphics Environment. <i>Computing in Science & Engineering</i> , 9(3), 90–95. https://doi.org/10.1109/MCSE.2007.55	844
	845
Leaché, A. D., & Oaks, J. R. (2017). The Utility of Single Nucleotide Polymorphism (SNP) Data in Phylogenetics. <i>Annual Review of Ecology, Evolution, and Systematics</i> , 48(1), 69–84. https://doi.org/10.1146/annurev-ecolsys-110316-022645	846
	847
Linck, E., & Battey, C. J. (2019). Minor allele frequency thresholds strongly affect population structure inference with genomic data sets. <i>Molecular Ecology Resources</i> , 19(3), 639–647. https://doi.org/10.1111/1755-0998.12995	849
	850
Liu, L., & Pearl, D. K. (2007). Species Trees from Gene Trees: Reconstructing Bayesian Posterior Distributions of a Species Phylogeny Using Estimated Gene Tree Distributions (T. Buckley, Ed.). <i>Systematic Biology</i> , 56(3), 504–514. https://doi.org/10.1080/10635150701429982	852
	853
	854
Oaks, J. R. (2019). Full Bayesian Comparative Phylogeography from Genomic Data (L. Kubatko, Ed.). <i>Systematic Biology</i> , 68(3), 371–395. https://doi.org/10.1093/sysbio/syy063	855
	856
Oaks, J. R., Siler, C. D., & Brown, R. M. (2019). The comparative biogeography of Philippine geckos challenges predictions from a paradigm of climate-driven vicariant diversification across an island archipelago. <i>Evolution</i> , 73(6), 1151–1167. https://doi.org/10.1111/evo.13754	859
	860
Ogilvie, H. A., Bouckaert, R. R., & Drummond, A. J. (2017). StarBEAST2 Brings Faster Species Tree Inference and Accurate Estimates of Substitution Rates. <i>Molecular Biology and Evolution</i> , 34(8), 2101–2114. https://doi.org/10.1093/molbev/msx126	861
	862
Pfeiffer, F., Gröber, C., Blank, M., Händler, K., Beyer, M., Schultze, J. L., & Mayer, G. (2018). Systematic evaluation of error rates and causes in short samples in next-generation sequencing. <i>Scientific Reports</i> , 8(1), 10950. https://doi.org/10.1038/s41598-018-29325-6	863
	864
Potapov, V., & Ong, J. L. (2017). Examining Sources of Error in PCR by Single-Molecule Sequencing (R. Kalender, Ed.). <i>PLOS ONE</i> , 12(1), e0169774. https://doi.org/10.1371/journal.pone.0169774	865
	866
Rambaut, A., & Grass, N. C. (1997). Seq-Gen: An application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. <i>Bioinformatics</i> , 13(3), 235–238. https://doi.org/10.1093/bioinformatics/13.3.235	871
	872
Rochette, N. C., Rivera-Colón, A. G., & Catchen, J. M. (2019). Stacks 2: Analytical methods for paired-end sequencing improve RADseq-based population genomics. <i>Molecular Ecology</i> , 28(21), 4737–4754. https://doi.org/10.1111/mec.15253	873
	874
Sukumaran, J., & Holder, M. T. (2010). DendroPy: A Python library for phylogenetic computing. <i>Bioinformatics</i> , 26(12), 1569–1571. https://doi.org/10.1093/bioinformatics/btq228	875
	876
Yang, Z. (2015). The BPP program for species tree estimation and species delimitation. <i>Current Zoology</i> , 61(5), 854–865. https://doi.org/10.1093/czoolo/61.5.854	877
	878
Yule, G. U. (1925). A mathematical theory of evolution, based on the conclusions of Dr. JC Willis, FR S. <i>Philosophical transactions of the Royal Society of London. Series B, containing papers of a biological character</i> , 213(402-410), 21–87.	879
	880
	881
	882
	883
	884
	885
	886
	887

3.6 Figures

888

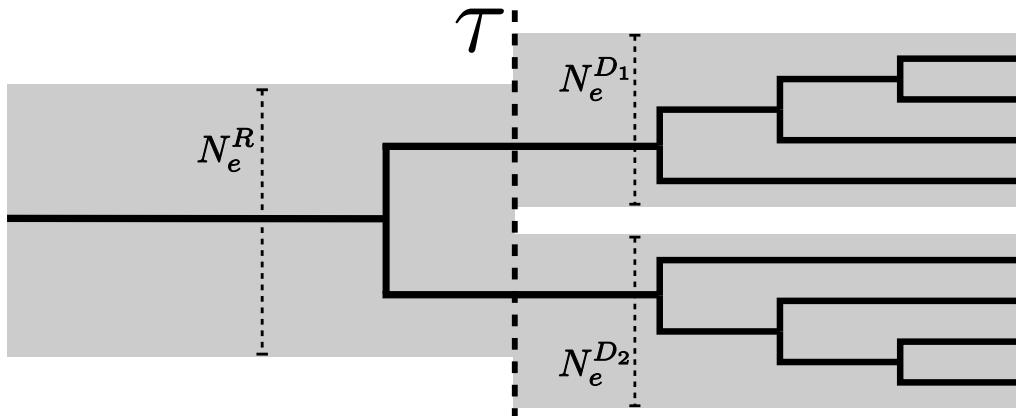


Figure 3.1. An illustration of the species-tree model we used to simulate data. N_e^R , N_e^{D1} , and N_e^{D2} represent the constant effective population sizes of the root, and each of the two terminal populations. τ represents the instantaneous separation of the ancestral population into two descendant 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.

Divergence Time — 1000bp loci

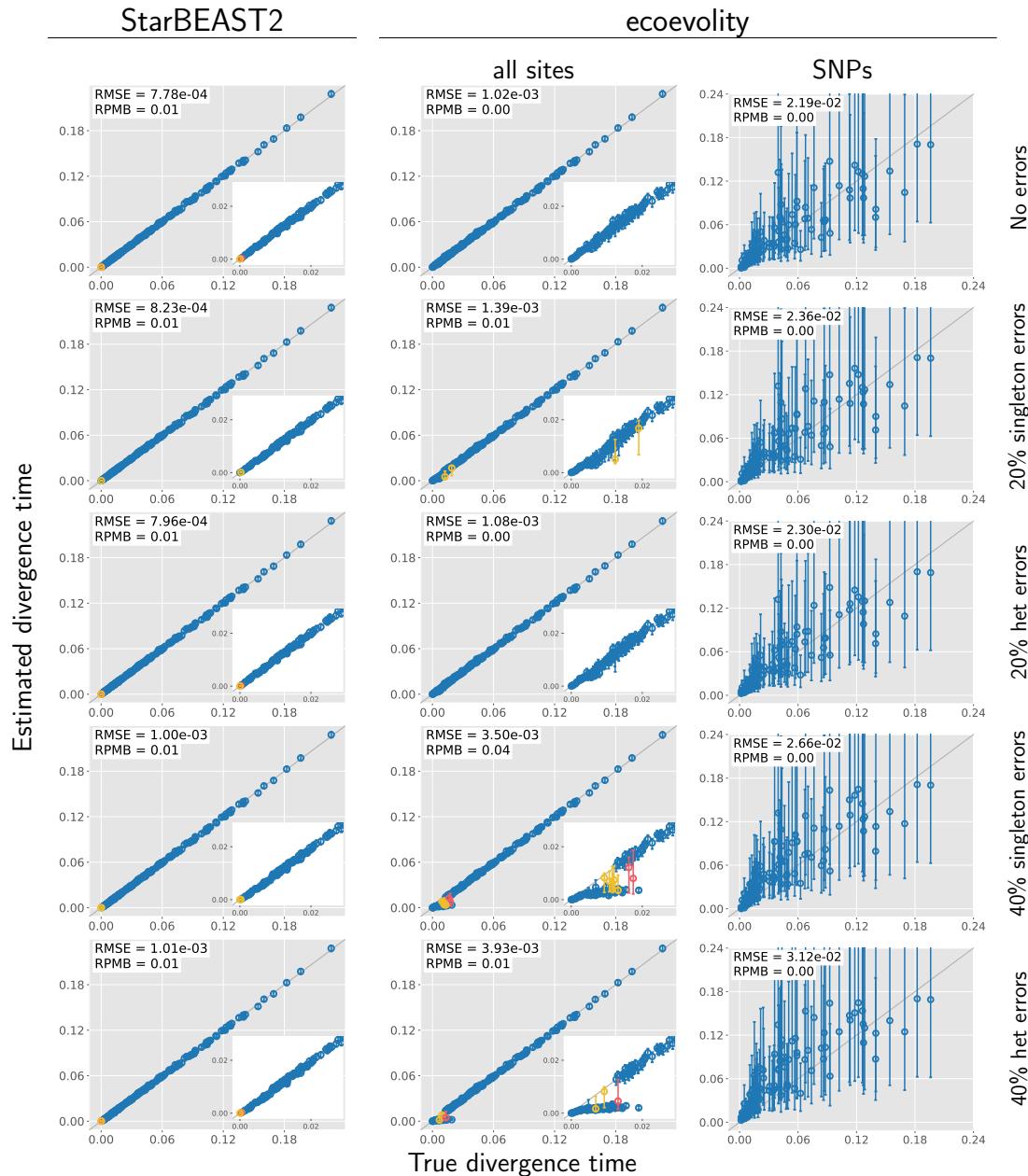


Figure 3.2. Accuracy and precision of divergence-time estimates (in units of expected substitutions per site) with 1000 base pair loci. The left column shows estimates from *StarBEAST2*, and the center and right column shows estimates from *ecoevolity* using all sites and (at most) one SNP per locus. 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).

Divergence Time — 500bp loci

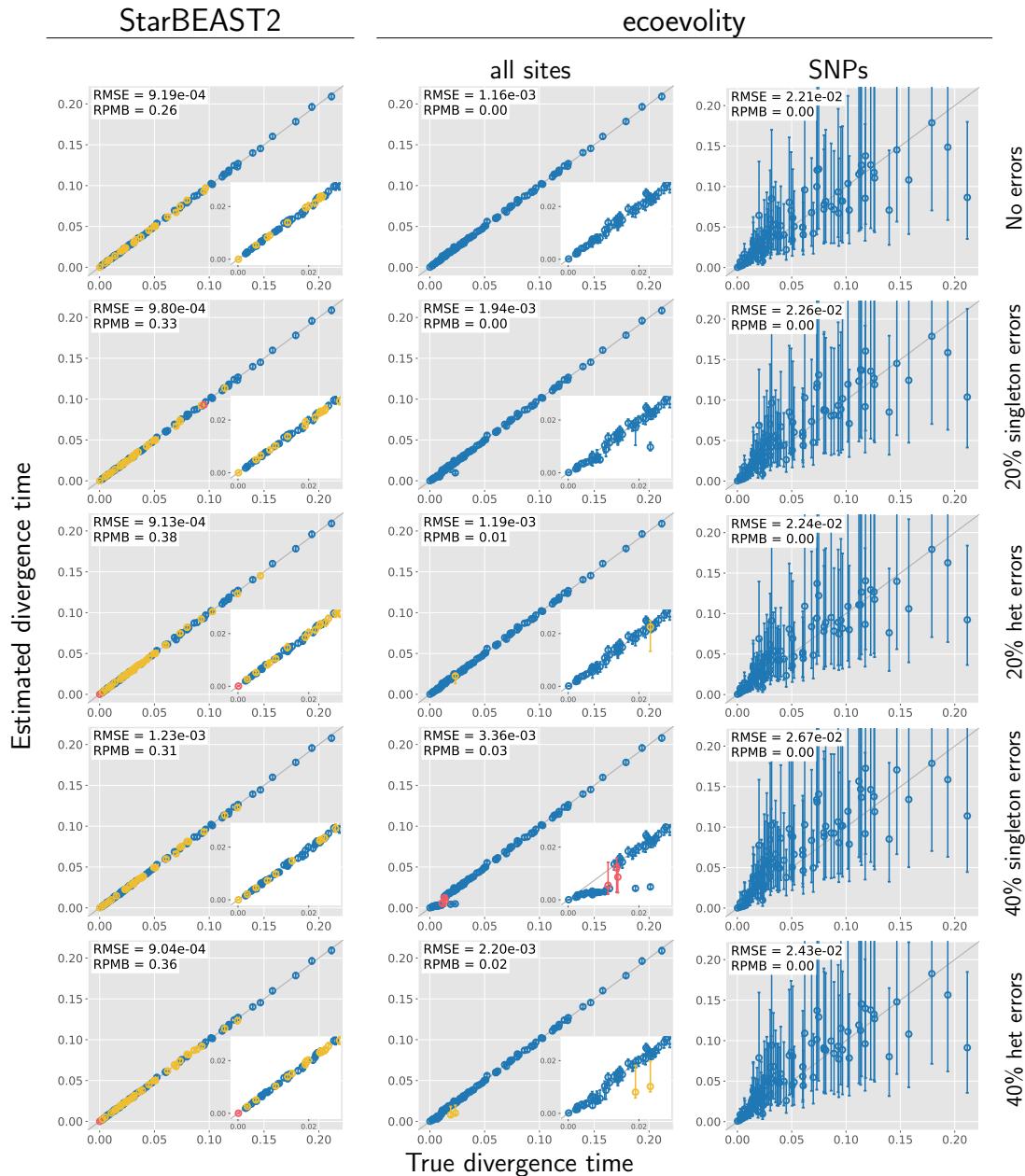


Figure 3.3. Accuracy and precision of divergence-time estimates (in units of expected substitutions per site) with 500 base pair loci. The left column shows estimates from *StarBEAST2*, and the center and right column shows estimates from *ecoevolity* using all sites and (at most) one SNP per locus. 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).

Divergence Time — 250bp loci

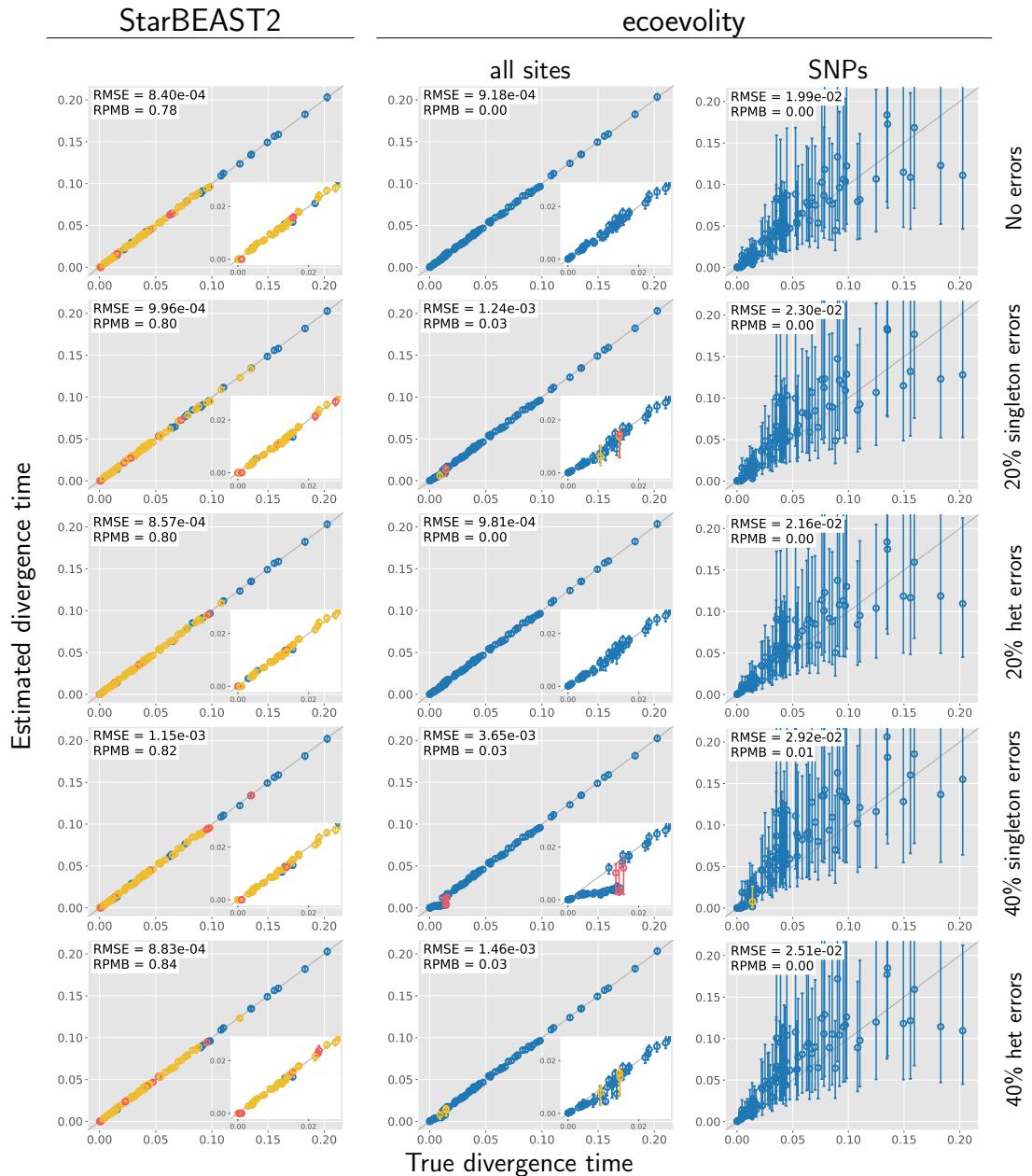


Figure 3.4. Accuracy and precision of divergence-time estimates (in units of expected substitutions per site) with 250 base pair loci. The left column shows estimates from *StarBEAST2*, and the center and right column shows estimates from *ecoevolity* using all sites and (at most) one SNP per locus. 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).

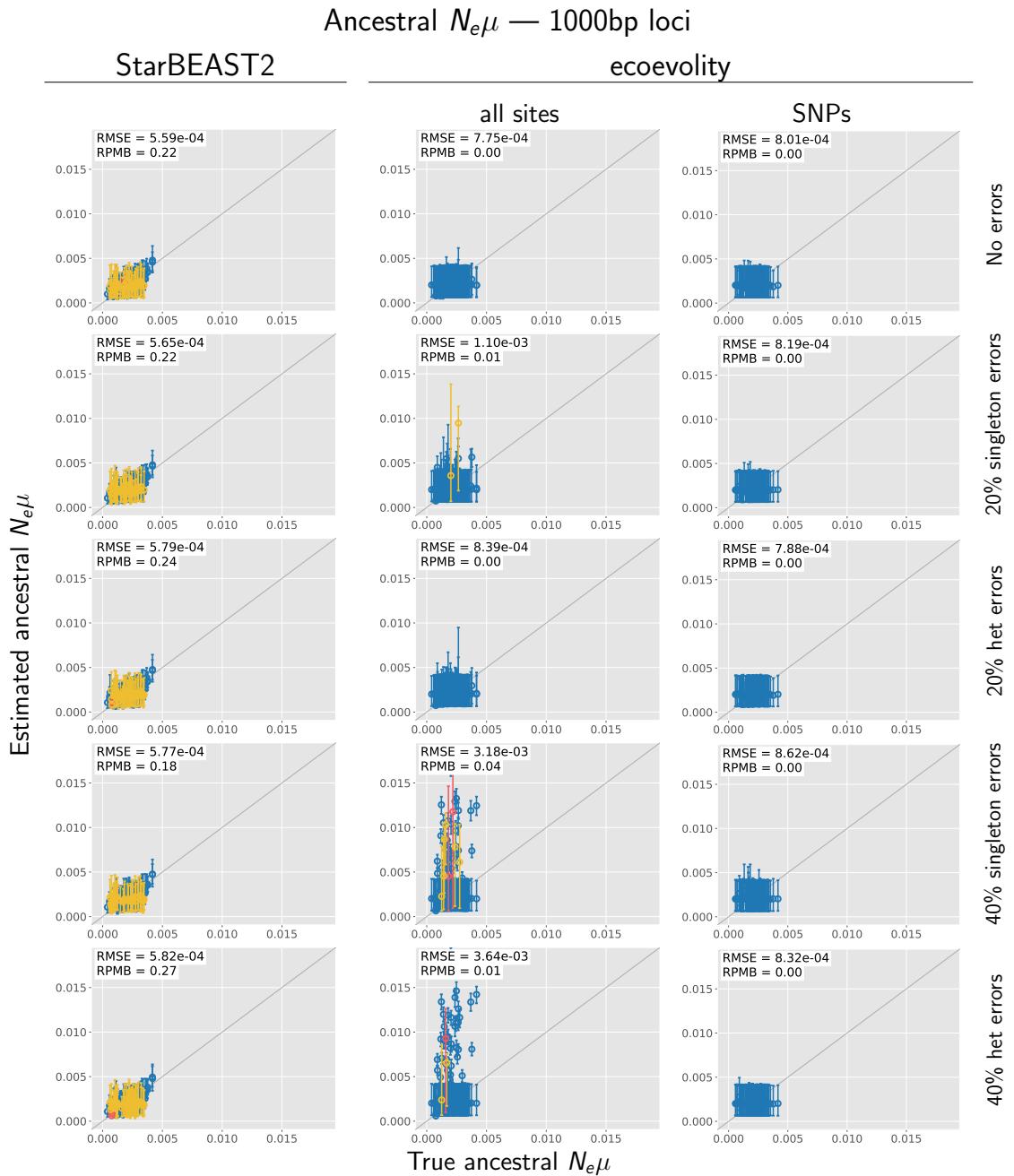


Figure 3.5. Accuracy and precision of estimates of root effective population size scaled by the mutation rate ($N_e^R \mu$) with 1000 base pair loci. The left column shows estimates from *StarBEAST2*, and the center and right column shows estimates from *ecoevolity* using all sites and (at most) one 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).

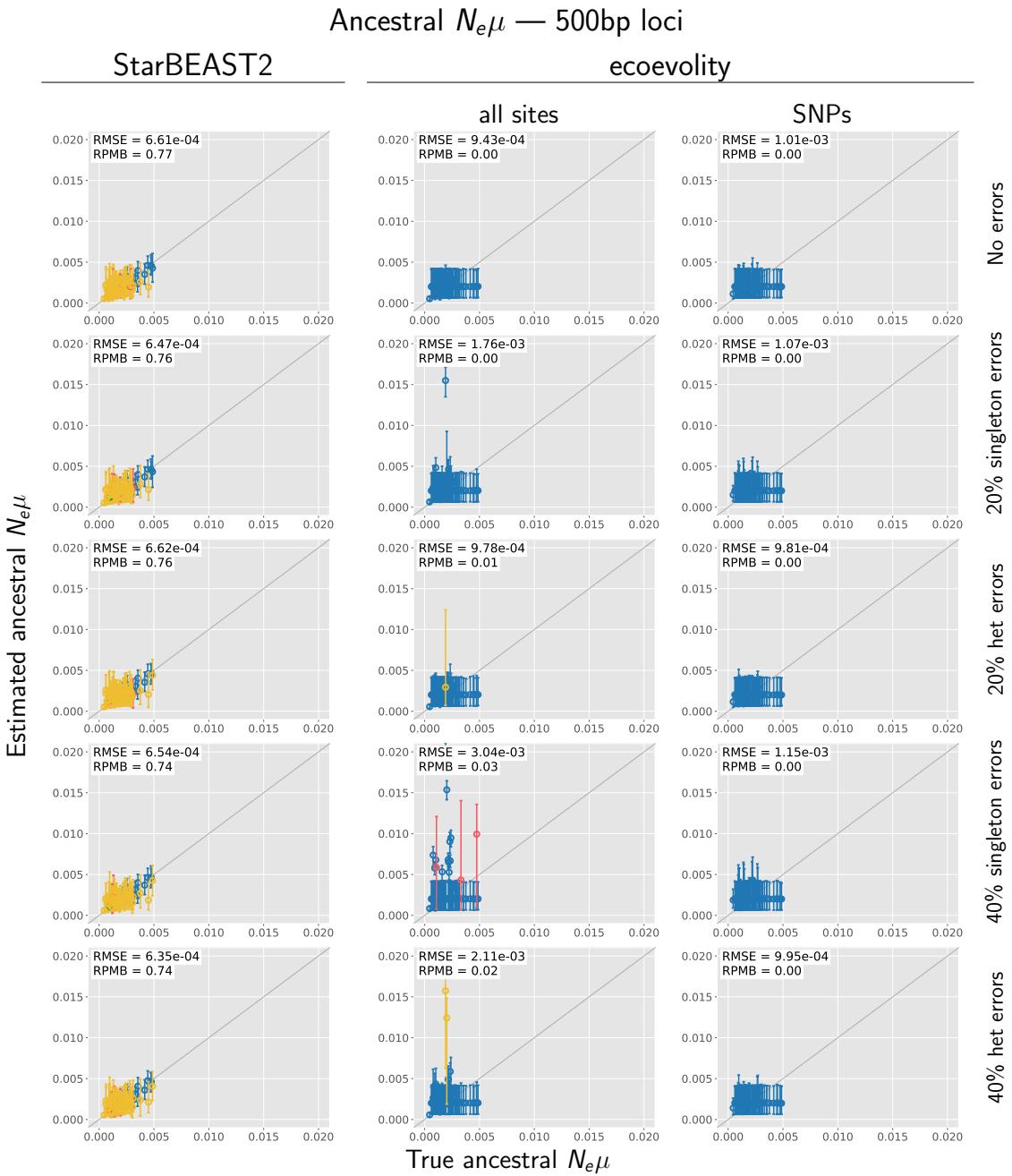


Figure 3.6. Accuracy and precision of estimates of root effective population size scaled by the mutation rate ($N_e^R\mu$) with 500 base pair loci. The left column shows estimates from *StarBEAST2*, and the center and right column shows estimates from *ecoevolity* using all sites and (at most) one SNP 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).

Ancestral $N_e\mu$ — 250bp loci

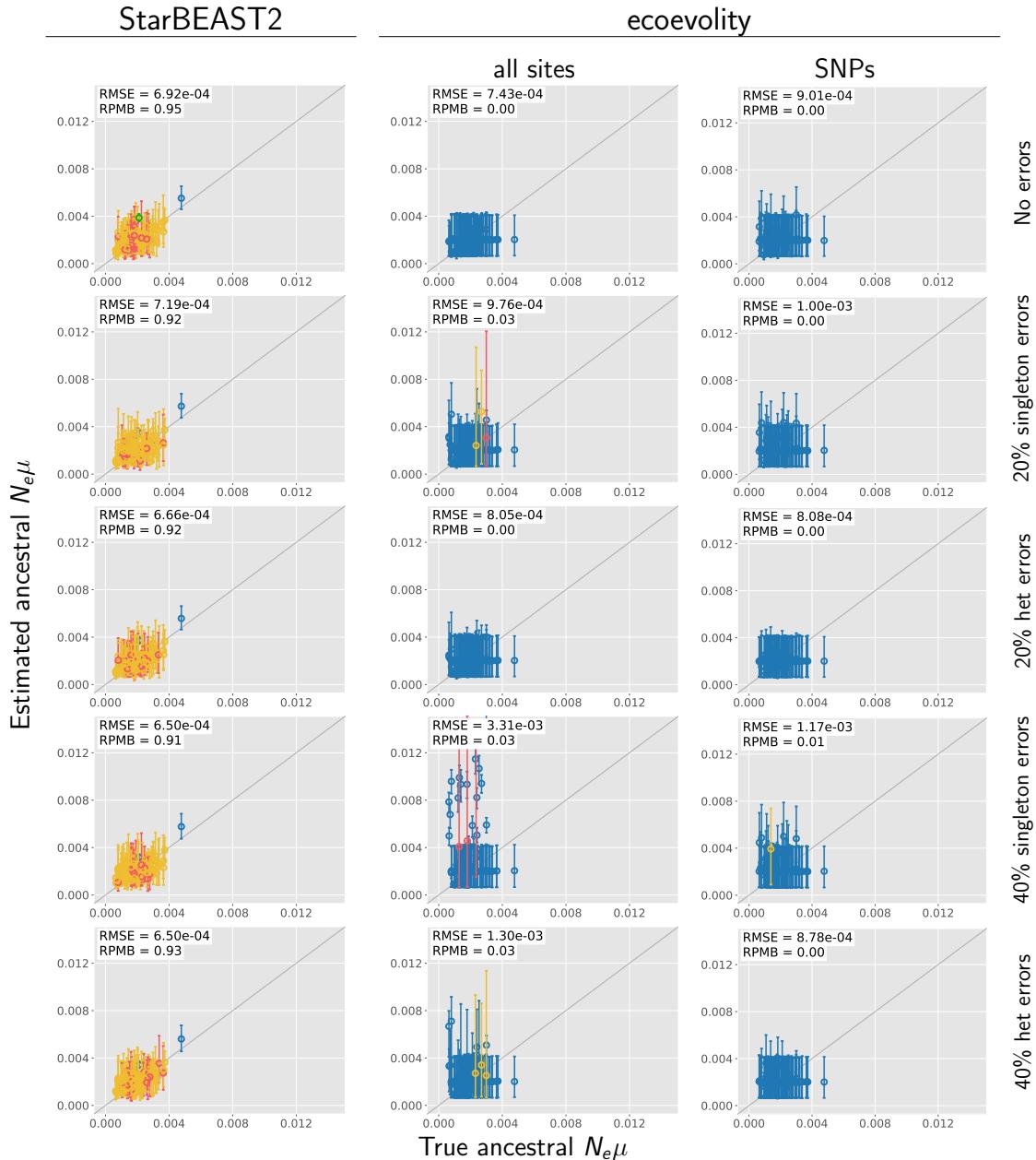


Figure 3.7. Accuracy and precision of estimates of root effective population size scaled by the mutation rate ($N_e^R \mu$) with 250 base pair loci. The left column shows estimates from *StarBEAST2*, and the center and right column shows estimates from *ecoevolity* using all sites and (at most) one SNP 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).

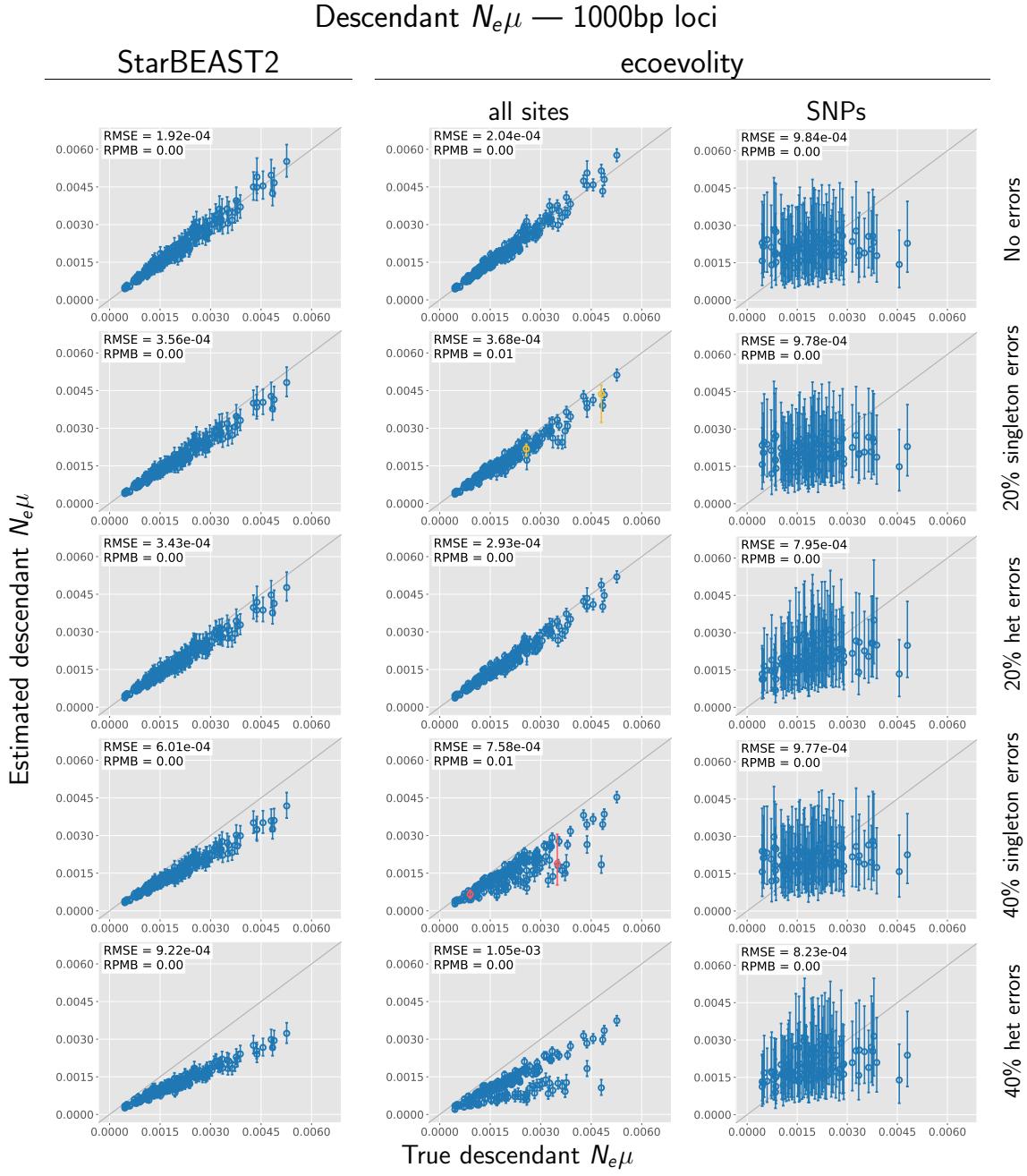


Figure 3.8. Accuracy and precision of estimates of effective population sizes of the descendant branches of the tree scaled by the mutation rate ($N_e^D\mu$) with 1000 base pair loci. The left column shows estimates from *StarBEAST2*, and the center and right column shows estimates from *ecoevolity* 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).

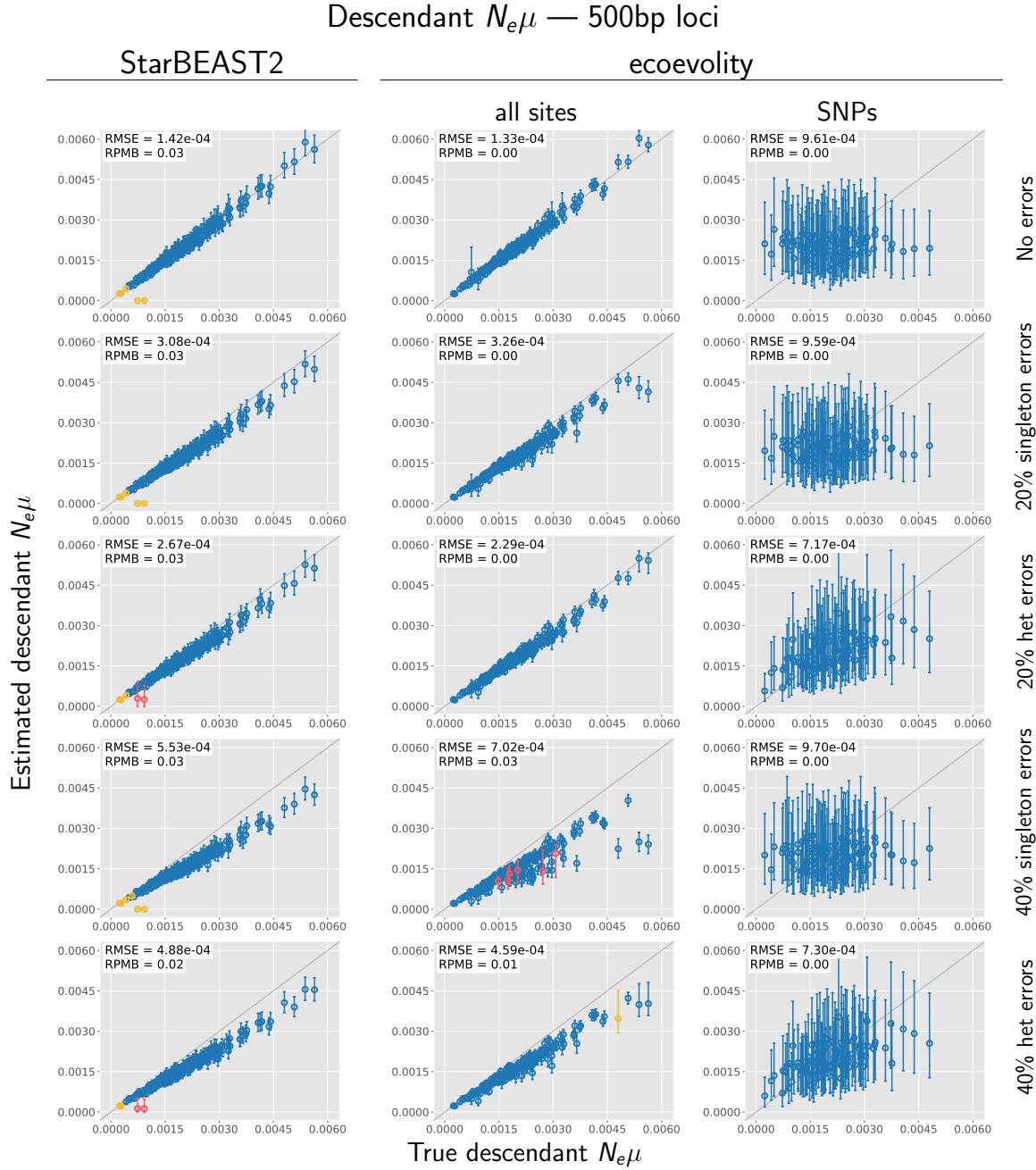


Figure 3.9. Accuracy and precision of estimates of effective population sizes of the descendant branches of the tree scaled by the mutation rate ($N_e^D\mu$) with 500 base pair loci. The left column shows estimates from *StarBEAST2*, and the center and right column shows estimates from *ecoevolity* using 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).

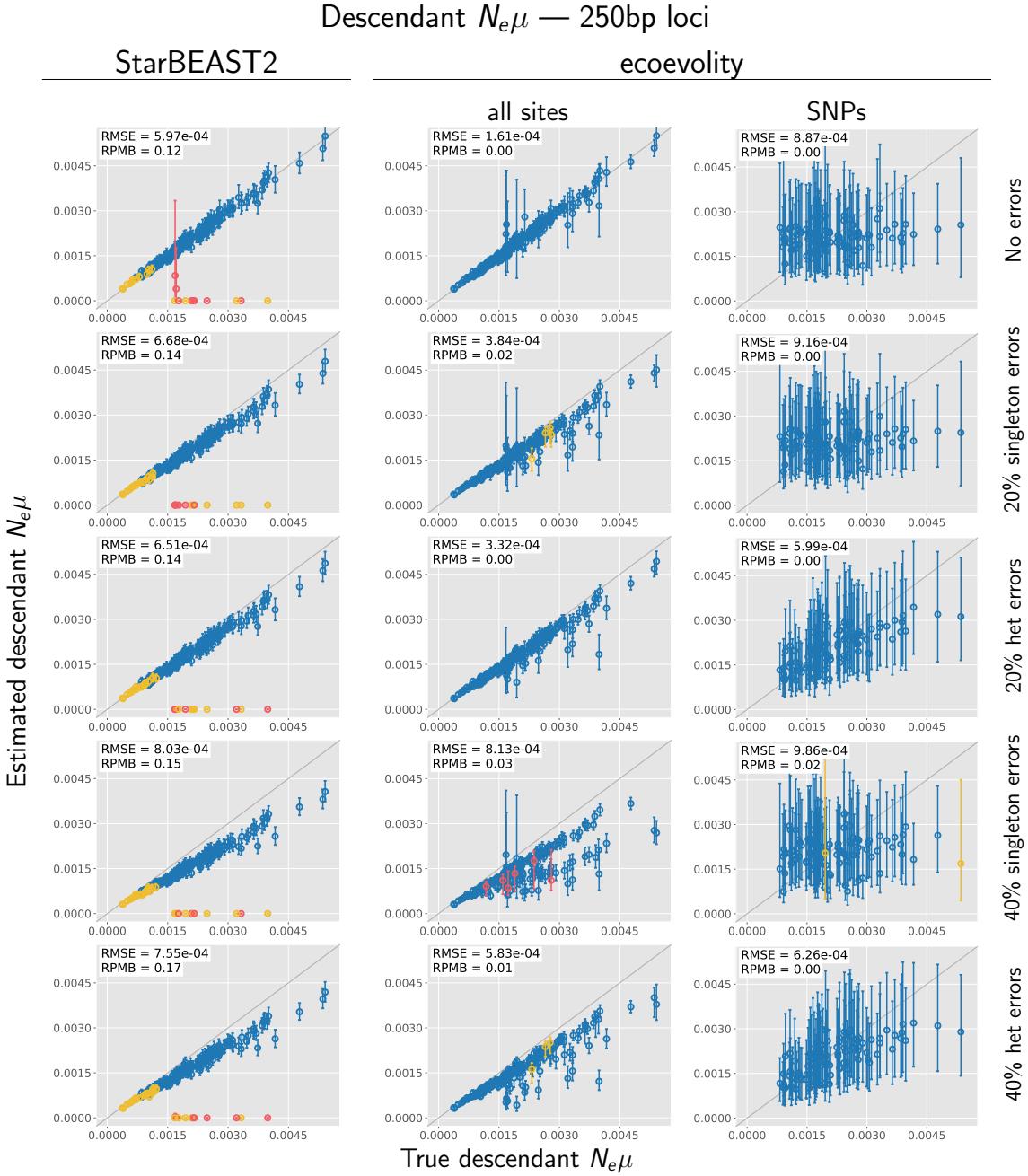


Figure 3.10. Accuracy and precision of estimates of effective population sizes of the descendant branches of the tree scaled by the mutation rate ($N_e^D\mu$) with 250 base pair loci. The left column shows estimates from *StarBEAST2*, and the center and right column shows estimates from *ecoevolity* 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).

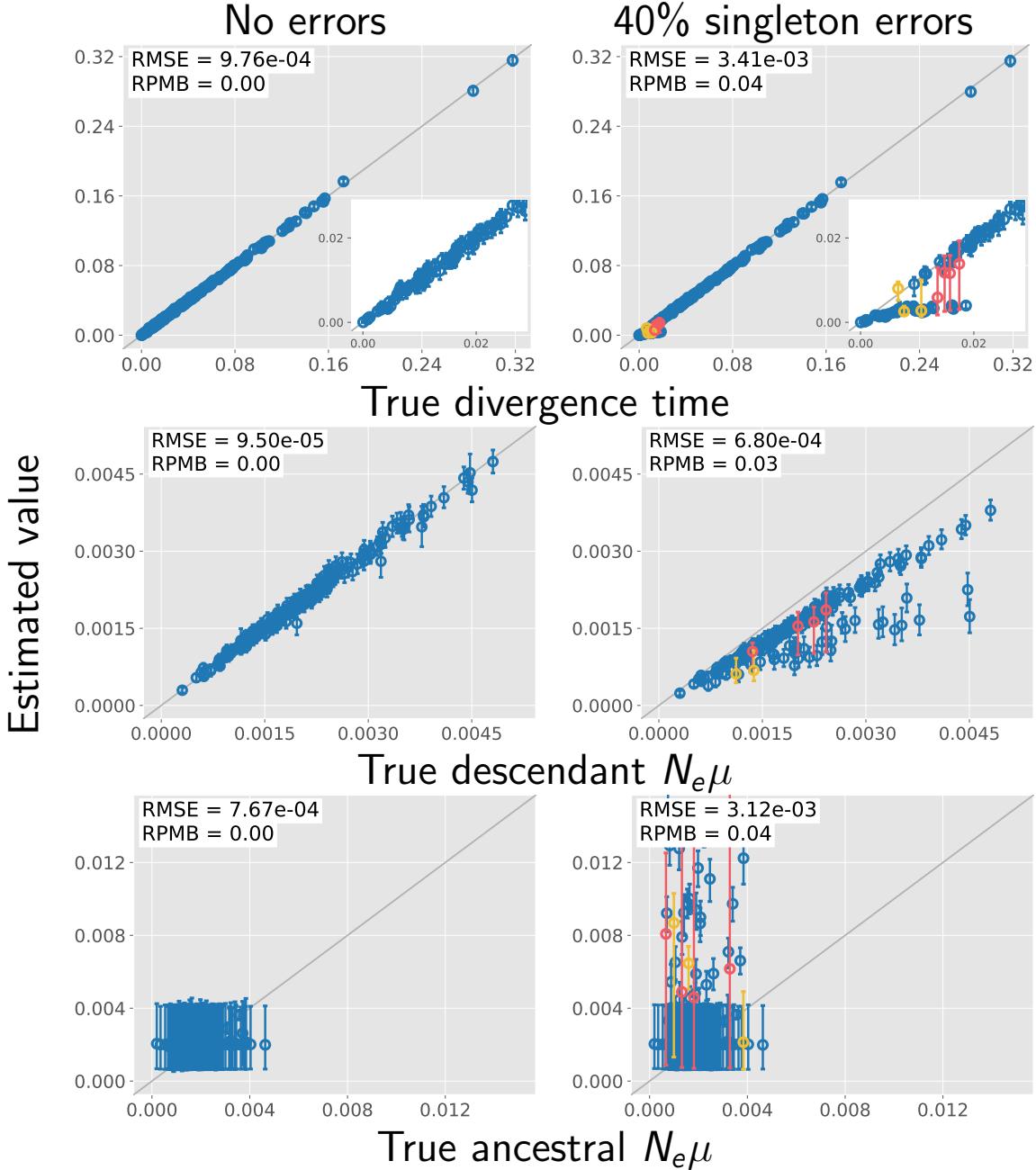


Figure 3.11. The performance of *ecoevolity* with data sets simulated with unlinked characters. Each plotted circle and associated error bars represent the posterior mean and 95% credible interval. Circles and error bars are colored yellow if the effective sample size (ESS) of the estimate was less than 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 $\text{ESS} < 200$ and/or $\text{PSRF} > 1.2$. Inset plots magnify estimates of most recent divergence times. We generated the plots using matplotlib Version 3.1.1 (Hunter, 2007).