Contents

List of Figures

List of Tables

Chapter 1

Phylogeography

1.1 Introduction

...

1.2 Methods

1.2.1 Sampling and DNA Isolation

I obtained tissue samples from museum tissue collections as well as from individuals that i collected from 2017 to 2020.

One *Rhinella marina* and one *Incillius nebulifer* were included to be used as outgroups for phylogenetic analyses.

I isolated DNA from tissues by first lysing a piece of tissue approximately the size of a grain of rice in 300 μL 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 μL 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/μL.

1.2.2 RADseq Library Preparation

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 μL of a solution with 1X CutSmart Buffer (New England Biolabs, USA; NEB), 10 units of XbaI, 10 units of EcoRI, 0.33 μM XbaI compatible adapter, 0.33 μM EcoRI compatible adapter, and nuclease free water with a 1 hour incubation at 37°C. I then immediately added 5 μL 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 μL 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 μL of TLE solution.

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 μL containing 1X Kapa Hifi Buffer (Kapa Biosystems, USA; Kapa), 0.3 μM iTru5-8N Primer, 0.3 mM dNTP, 1 unit Kapa HiFi DNA Polymerase, 10 μL 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 μL TLE. I added the remaining adapter and index sequences unique to each plate with four PCR reactions with a total volume of 50 μL containing 1X Kapa Hifi (Kapa), 0.3 μM iTru7 Primer, 0.3 μM P5 Primer, 0.3 mM dNTP, 1 unit of Kapa Hifi DNA Polymerase (Kapa), 10 μL 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 μL 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 μM P5 Primer, 0.3 μM P7 Primer, 0.3 mM dNTP, 1 unit of Kapa HiFi DNA Polymerase (Kapa), 10 μL 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 μL 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.

1.2.3 Data Processing

I demultiplexed plate indexes using the process_radtags command from Stacks2 v2.6.1 (Rochette et al., 2019) and allowed for 2 mismatches for rescuing reads. I demultiplexed

samples and trimmed, filtered, and assembled reads with ipyrad v0.9.9 (Eaton & Overcast, 2020). I specified the datatype as "pair3rad", used a clustering threshold of 80% sequence similarity, and allowed for two mismatches in the barcode sequence. The remaining parameters were left with the default settings. I used Ipyrad to filter loci that were absent in greater than 50% of samples and then filtered samples with fewer than 200 loci.

1.2.4 Maximum Likelihood

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.

1.2.5 Structure

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

1.2.6 Phycoeval

To see if there is evidence of shared divergence times and to estimate a species tree under multispecies coalescent Phycoeval v1.0.0 (**oaks2020**)

1.2.7 Dsuite

To look for evidence of past admixture. Dsuite (Malinsky et al., 2021)

1.3 Results

Average number of reads per individual Mean coverage per locus
Total loci and snps after filtering

1.3.1 Maximum Likelihood

Inferred a single well supported clade (>***%) for each recognized species.

1.4 Acknowledgments

...

1.5 Figures

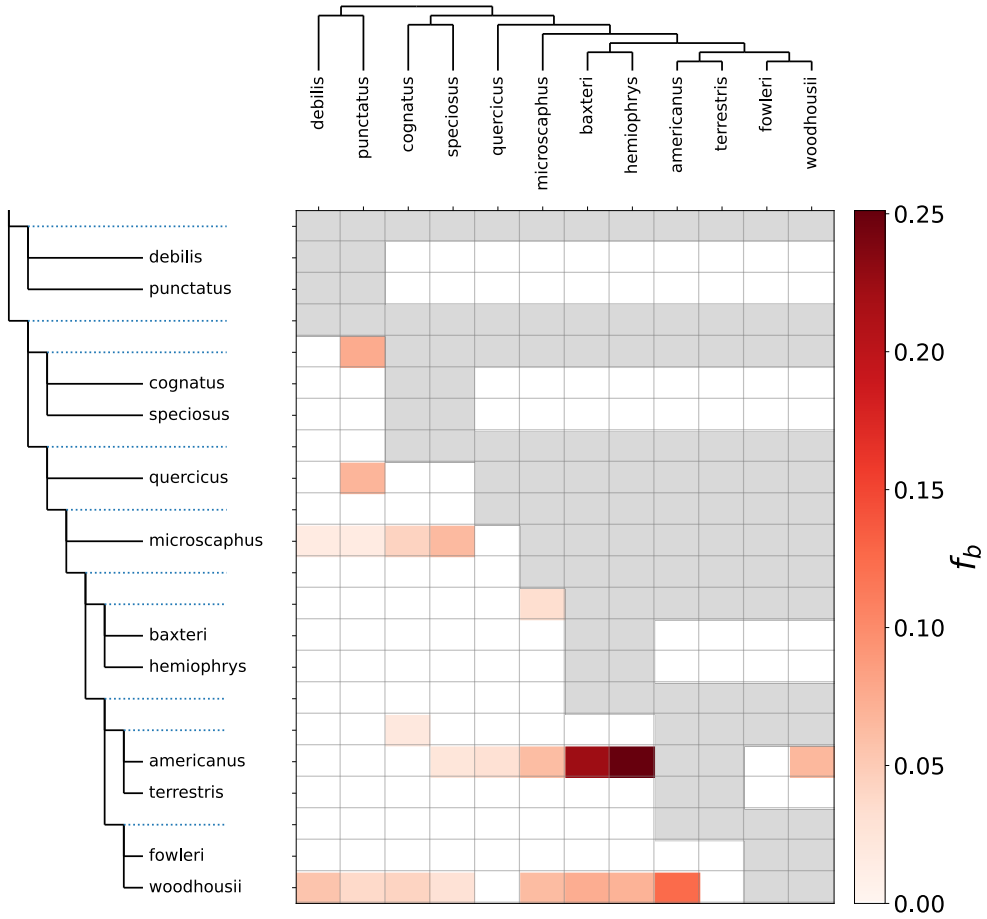

Figure 1.1. Historic admixture

1.6 Tables

Table 1.1. Samples collected for this study

Sample ID	Species	Latitude	Longitude
KAC 016	*Anaxyrus terrestris*	30.54819	-86.93067
KAC 053	*Anaxyrus fowleri*	32.78044	-86.73877
KAC 060	*Anaxyrus speciosus*	27.69185	-99.71955
KAC 062	*Anaxyrus punctatus*	29.43603	-103.50564
KAC 063	*Anaxyrus speciosus*	29.29522	-103.92916
KAC 064	*Anaxyrus speciosus*	29.29522	-103.92916
KAC 065	*Anaxyrus terrestris*	30.43282	-81.64088
KAC 066	*Anaxyrus terrestris*	30.43282	-81.64088
KAC 067	*Anaxyrus terrestris*	30.43282	-81.64088
KAC 070	*Anaxyrus americanus*	34.79963	-84.57678
KAC 074	*Anaxyrus terrestris*	30.77430	-85.22690
KAC 137	*Anaxyrus fowleri*	33.01461	-86.60953
KAC 157	*Anaxyrus fowleri*	32.43769	-85.63620
KAC 165	*Anaxyrus fowleri*	32.66356	-85.48498
KAC 166	*Anaxyrus fowleri*	32.66356	-85.48498
KAC 174	*Anaxyrus fowleri*	32.62938	-85.63828
KAC 175	*Anaxyrus fowleri*	32.64849	-85.64711
KAC 178	*Anaxyrus fowleri*	32.38644	-85.23561
KAC 179	*Anaxyrus fowleri*	32.38644	-85.23561
KAC 180	*Anaxyrus fowleri*	32.38644	-85.23561
KAC 186	*Anaxyrus fowleri*	32.38579	-85.23565
KAC 202	*Anaxyrus fowleri*	33.25104	-86.43850
KAC 203	*Anaxyrus fowleri*	32.62294	-85.49660
KAC 204	*Anaxyrus fowleri*	32.62294	-85.49660
KAC 226	*Anaxyrus fowleri*	32.48119	-85.79838
KAC 230	*Anaxyrus terrestris*	30.80933	-86.77686
KAC 232	*Anaxyrus terrestris*	30.80922	-86.78994
KAC 233	*Anaxyrus terrestris*	30.80922	-86.78994
KAC 234	*Anaxyrus terrestris*	30.80922	-86.78994
KAC 236	*Anaxyrus terrestris*	30.82632	-86.80258
KAC 237	*Anaxyrus terrestris*	30.83733	-86.77630
KAC 238	*Anaxyrus terrestris*	30.82433	-86.76284
KAC 239	*Anaxyrus terrestris*	30.80162	-86.76659
KAC 240	*Anaxyrus fowleri*	32.64328	-85.37114
KAC 241	*Anaxyrus fowleri*	32.64328	-85.37114
KAC 242	*Anaxyrus americanus*	34.50446	-85.63768
KAC 243	*Incillius nebulifer*	nan	nan
KAC 244	*Anaxyrus fowleri*	32.89261	-93.88756
KAC t1020	*Anaxyrus terrestris*	31.10783	-86.62247
KAC t2004	*Anaxyrus americanus*	33.58295	-85.73524
KAC t2015	*Anaxyrus americanus*	33.58435	-85.74064
KAC t2018-02-17-01	*Anaxyrus americanus*	33.55274	-85.82913
KAC t2018-02-17-04	*Anaxyrus americanus*	33.48548	-85.88857

Continued on next page

Table 1.1 – continued from previous page

Sample ID	Species	Latitude	Longitude
KAC t2018-03-10-2	*Anaxyrus fowleri*	32.93116	-86.08465
KAC t2018-08-18-1	*Anaxyrus terrestris*	30.66902	-81.44013
KAC t2018-08-18-2	*Anaxyrus terrestris*	30.66902	-81.44013
KAC t2018-08-18-3	*Anaxyrus terrestris*	30.43282	-81.64088
KAC t2018-08-18-4	*Anaxyrus terrestris*	30.66902	-81.44013
KAC t2019-08-25-2	*Anaxyrus fowleri*	34.21852	-87.36662
KAC t2040	*Anaxyrus americanus*	33.58295	-85.73539
KAC t3040	*Anaxyrus fowleri*	32.38644	-85.23561

Table 1.2. Samples loaned from museums

Sample ID	Species	Latitude	Longitude
AHT 2544	*Anaxyrus quercicus*	30.99523	-86.23332
AHT 2564	*Anaxyrus terrestris*	31.55752	-84.04267
AHT 3413	*Anaxyrus fowleri*	33.36940	-88.12941
AHT 3428	*Anaxyrus terrestris*	31.12679	-86.54755
AHT 3459	*Anaxyrus americanus*	34.88028	-87.71849
AHT 3460	*Anaxyrus americanus*	33.78013	-85.58421
AHT 3461	*Anaxyrus americanus*	34.88779	-87.74103
AHT 3462	*Anaxyrus americanus*	33.77001	-85.55434
AHT 3463	*Anaxyrus americanus*	33.71125	-85.59762
AHT 3658	*Anaxyrus fowleri*	32.85842	-86.39697
AHT 3665	*Anaxyrus fowleri*	32.81220	-86.17698
AHT 3813	*Anaxyrus terrestris*	31.13854	-86.53906
AHT 3833	*Anaxyrus terrestris*	31.00422	-85.03427
AHT 4373	*Anaxyrus americanus*	38.94913	-95.39818
AHT 5276	*Anaxyrus terrestris*	31.55613	-86.82514
AHT 5277	*Anaxyrus terrestris*	31.15830	-86.55430
AHT 5278	*Anaxyrus terrestris*	31.16105	-86.69868
HERA 10025	*Anaxyrus fowleri*	37.11151	-84.11812
HERA 10233	*Anaxyrus americanus*	39.86453	-85.01037
HERA 10239	*Anaxyrus americanus*	38.99151	-92.31078
HERA 10248	*Anaxyrus americanus*	41.27319	-73.38974
HERA 10255	*Anaxyrus americanus*	37.11151	-84.11812
HERA 10350	*Anaxyrus americanus*	45.51396	-69.95928
HERA 10372	*Anaxyrus americanus*	42.22795	-79.36759
HERA 10396	*Anaxyrus fowleri*	41.80663	-72.73281
HERA 10484	*Rhinella marina*	25.61296	-80.56606
HERA 10493	*Anaxyrus fowleri*	39.08588	-75.56844
HERA 11976	*Anaxyrus americanus*	43.51819	-71.42336
HERA 13722	*Anaxyrus fowleri*	36.55514	-89.18929
HERA 14196	*Anaxyrus retiformis*	33.34906	-112.49010
HERA 14926	*Anaxyrus microscaphus*	33.73033	-113.98078
HERA 15787	*Anaxyrus americanus*	38.88546	-95.29399
HERA 20415	*Anaxyrus woodhousii*	34.31743	-92.94602
HERA 20514	*Anaxyrus fowleri*	33.95140	-83.36715
INHS 16273	*Anaxyrus americanus*	42.30245	-89.55950
INHS 17016	*Anaxyrus americanus*	37.46121	-88.18728
INHS 19127	*Anaxyrus fowleri*	41.58247	-88.07273
INHS 21799	*Anaxyrus americanus*	46.01258	-94.26710
MSB 100793	*Anaxyrus microscaphus*	37.27154	-114.46478
MSB 100800	*Anaxyrus woodhousii*	36.73612	-114.21972
MSB 100913	*Anaxyrus microscaphus*	33.28038	-108.08868
MSB 104548	*Anaxyrus woodhousii*	36.49094	-103.20838
MSB 104570	*Anaxyrus fowleri*	34.00087	-95.38229
MSB 104571	*Anaxyrus americanus*	34.00917	-95.38058
MSB 104608	*Anaxyrus americanus*	34.00367	-94.82670

Continued on next page

Table 1.2 – continued from previous page

Sample ID	Species	Latitude	Longitude
MSB 104644	*Anaxyrus americanus*	36.95124	-94.27782
MSB 104677	*Anaxyrus cognatus*	46.39834	-97.20927
MSB 104681	*Anaxyrus hemiophrys*	46.47076	-97.04604
MSB 104731	*Anaxyrus woodhousii*	42.61091	-100.65607
MSB 75646	*Anaxyrus woodhousii*	33.36365	-104.34282
MSB 92689	*Anaxyrus baxteri*	41.21182	-105.82558
MSB 92691	*Anaxyrus baxteri*	41.21182	-105.82558
MSB 92692	*Anaxyrus baxteri*	41.21182	-105.82558
MSB 96528	*Anaxyrus debilis*	32.58239	-107.46348
MSB 98058	*Anaxyrus woodhousii*	32.83360	-108.60900
MSB 98065	*Anaxyrus cognatus*	32.63240	-108.73800
UTEP 18705	*Anaxyrus woodhousii*	32.45198	-106.88317
UTEP 19941	*Anaxyrus fowleri*	34.79137	-88.95715
UTEP 19943	*Anaxyrus fowleri*	33.81998	-88.29533
UTEP 19947	*Anaxyrus terrestris*	31.22432	-88.77548
UTEP 20105	*Anaxyrus woodhousii*	33.62853	-103.08198
UTEP 20482	*Anaxyrus woodhousii*	32.90708	-94.74945
UTEP 20921	*Anaxyrus americanus*	35.55405	-91.83443
UTEP 21284	*Anaxyrus debilis*	31.25968	-105.33402
UTEP 21286	*Anaxyrus speciosus*	31.70140	-105.47958
UTEP 21724	*Anaxyrus speciosus*	31.26087	-104.60168
UTEP 21881	*Anaxyrus cognatus*	35.53600	-100.44035
UTEP 21884	*Anaxyrus speciosus*	32.75472	-101.43208
UTEP 21885	*Anaxyrus speciosus*	32.20195	-100.34345
UTEP 21886	*Anaxyrus woodhousii*	35.07800	-100.43392

Chapter 2

Hybrid Zone

2.1 Introduction

...

2.2 Methods

2.2.1 Sampling and DNA Isolation

I collected genetic samples from *A. americanus* and *A. terrestris* by driving roads during rainy nights between 2017 and 2020 in an 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 MS-222. I removed liver and/or toes and preserved them in 100% ethanol. Samples and formalin fixed specimens were deposited in the Auburn Museum of Natural History. Additional samples were also provided by museums. I isolated DNA by first lysing a small piece of liver or toe approximately the size of a grain of rice in 300 μL 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 μL 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/μL.

2.2.2 RADseq Library Preparation

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 μL of a solution with 1X CutSmart Buffer (New England Biolabs, USA; NEB), 10 units of XbaI, 10 units of EcoRI, 0.33 μM XbaI compatible adapter, 0.33 μM EcoRI compatible adapter, and nuclease free water with a 1 hour incubation at 37°C. I then immediately added 5 μL 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 μL 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 μL of TLE solution.

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 μL containing 1X Kapa Hifi Buffer (Kapa Biosystems, USA; Kapa), 0.3 μM iTru5-8N Primer, 0.3 mM dNTP, 1 unit Kapa HiFi DNA Polymerase, 10 μL 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 μL TLE. I added the remaining adapter and index sequences unique to each plate with four PCR reactions with a total volume of 50 μL containing 1X Kapa Hifi (Kapa), 0.3 μM iTru7 Primer, 0.3 μM P5 Primer, 0.3 mM dNTP, 1 unit of Kapa Hifi DNA Polymerase (Kapa), 10 μL 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 μL 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 μM P5 Primer, 0.3 μM P7 Primer, 0.3 mM dNTP, 1 unit of Kapa HiFi DNA Polymerase (Kapa), 10 μL 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 μL 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 plate indexes using the process_radtags command from Stacks2 v2.6.1 (Rochette et al., 2019) and allowed for 2 mismatches for rescuing reads. I demultiplexed

samples and trimmed, filtered, and assembled reads with ipyrad v0.9.84 (Eaton & Overcast, 2020). I specified the datatype as "pair3rad", used a clustering threshold of 90% sequence similarity, and allowed for two mismatches in the barcode sequence. The remaining parameters were left with the default settings. I randomly sampled a single SNP from each locus from each locus using a custom python script relying on the scikit-allel v1.3.5 package. Using vcftools v0.1.17 (Danecek et al., 2011), I dropped any individuals that had more than 25% missing sites and filtered individuals that had more than 35% missing data after the missing sites filter was applied.

2.2.4 Ancestry Proportions

I used the program STRUCTURE v2.3.4 (Pritchard et al., 2000) to infer population structure and ancestry proportions of sampled individuals with STRUCTURE's admixture model . I ran structure with values of K ranging from 1 to 4. For each value of K I ran 20 iterations for 250,000 steps with 50,000 burnin steps. I used the program POPHELPER v2.3.1 (Francis, 2017) to combine iterations for each value of K and to compute ΔK as described by (Evanno et al., 2005) I considered a sample to be a putative hybrid if it had an ancestry coefficient $<$95% for one of the parent species under the K=2 STRUCTURE model.

2.2.5 Genomic Cline Analysis

I excluded ** samples from cline analysis that were inferred as putative hybrids but were from locations not in proximity to the hybrid zone. I used bgc v1.03 (Gompert & Buerkle, 2012)

2.3 Results

2.3.1 Sampling and Data Processing

...

2.4 Figures

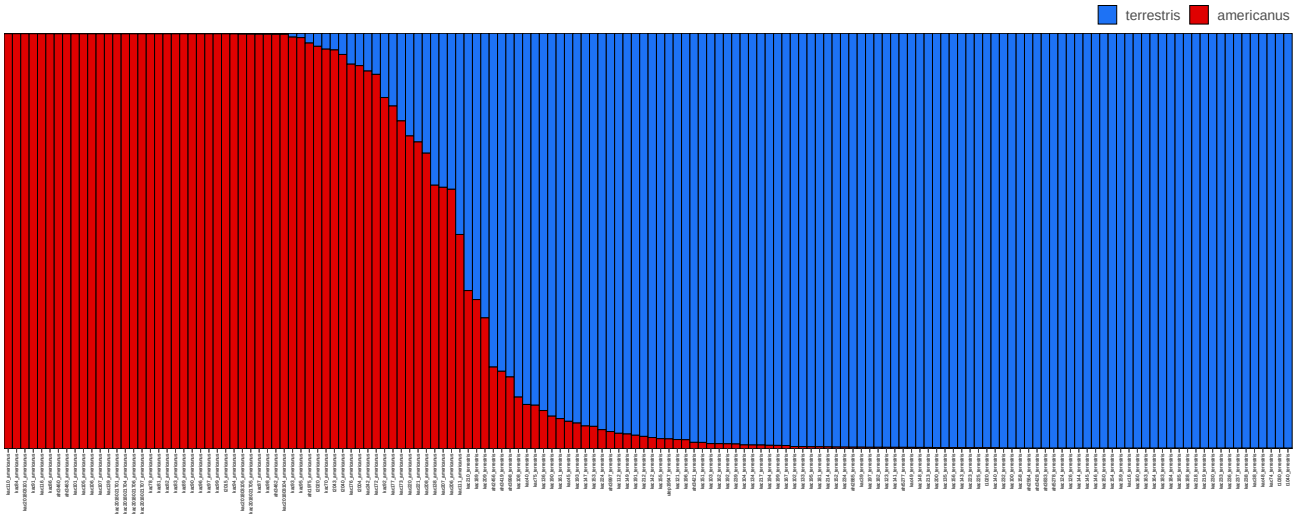

Figure 2.1. Structure plot for K = 2. Red is ???? species and blue is ???? species.

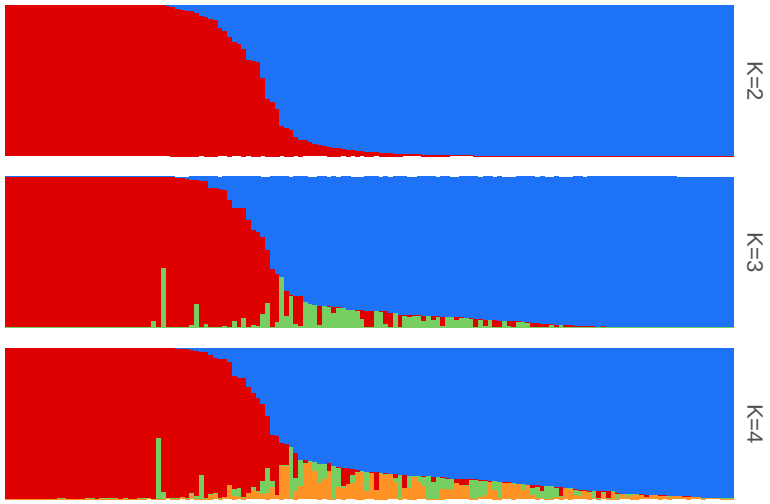

Figure 2.2. Structure plot for K = 2, 3, 4. Red is ???? species and blue is ???? species.

2.5 Tables

Table 2.1. Samples collected for this study

Sample ID	Species	Latitude	Longitude
KAC 016	*Anaxyrus terrestris*	30.54819	-86.93067
KAC 038	*Anaxyrus terrestris*	32.81470	-86.93968
KAC 039	*Anaxyrus terrestris*	32.81094	-86.98967
KAC 040	*Anaxyrus terrestris*	32.80985	-86.99795
KAC 042	*Anaxyrus terrestris*	32.82406	-86.99314
KAC 043	*Anaxyrus terrestris*	32.82406	-86.99314
KAC 044	*Anaxyrus terrestris*	32.80450	-87.03078
KAC 045	*Anaxyrus terrestris*	32.76703	-87.07073
KAC 046	*Anaxyrus terrestris*	32.76592	-87.07184
KAC 047	*Anaxyrus terrestris*	32.78932	-86.90850
KAC 048	*Anaxyrus terrestris*	32.73575	-86.88149
KAC 049	*Anaxyrus terrestris*	32.73291	-86.87707
KAC 050	*Anaxyrus terrestris*	32.74822	-86.79806
KAC 051	*Anaxyrus terrestris*	32.78742	-86.75847
KAC 052	*Anaxyrus terrestris*	32.78044	-86.73877
KAC 065	*Anaxyrus terrestris*	30.43282	-81.64088
KAC 066	*Anaxyrus terrestris*	30.43282	-81.64088
KAC 067	*Anaxyrus terrestris*	30.43282	-81.64088
KAC 070	*Anaxyrus americanus*	34.79963	-84.57678
KAC 071	*Anaxyrus terrestris*	32.43478	-85.64630
KAC 074	*Anaxyrus terrestris*	30.77430	-85.22690
KAC 075	*Anaxyrus terrestris*	32.94778	-86.63224
KAC 076	*Anaxyrus terrestris*	32.94970	-86.52687
KAC 077	*Anaxyrus terrestris*	32.94970	-86.52687
KAC 078	*Anaxyrus americanus*	33.00267	-86.38960
KAC 079	*Anaxyrus americanus*	33.01205	-86.47872
KAC 080	*Anaxyrus americanus*	33.04456	-86.45547
KAC 081	*Anaxyrus americanus*	33.04456	-86.45547
KAC 082	*Anaxyrus americanus*	33.04456	-86.45547
KAC 083	*Anaxyrus americanus*	33.04456	-86.45547
KAC 084	*Anaxyrus americanus*	33.04456	-86.45547
KAC 085	*Anaxyrus americanus*	33.04456	-86.45547
KAC 086	*Anaxyrus americanus*	33.04456	-86.45547
KAC 087	*Anaxyrus americanus*	33.01484	-86.39040
KAC 089	*Anaxyrus americanus*	33.01484	-86.39040
KAC 090	*Anaxyrus americanus*	33.06472	-86.47496
KAC 091	*Anaxyrus americanus*	33.06472	-86.47496
KAC 092	*Anaxyrus americanus*	33.06472	-86.47496
KAC 093	*Anaxyrus americanus*	33.06472	-86.47496
KAC 094	*Anaxyrus americanus*	33.06472	-86.47496
KAC 095	*Anaxyrus americanus*	33.06472	-86.47496
KAC 096	*Anaxyrus americanus*	33.06472	-86.47496
KAC 097	*Anaxyrus americanus*	33.06472	-86.47496

Continued on next page

Table 2.1 – continued from previous page

Sample ID	Species	Latitude	Longitude
KAC 098	*Anaxyrus americanus*	33.02572	-86.46711
KAC 099	*Anaxyrus americanus*	33.02572	-86.46711
KAC 100	*Anaxyrus terrestris*	32.92374	-86.67199
KAC 101	*Anaxyrus americanus*	33.03283	-86.45975
KAC 102	*Anaxyrus terrestris*	32.94544	-86.55777
KAC 103	*Anaxyrus terrestris*	32.94947	-86.52630
KAC 104	*Anaxyrus terrestris*	32.94947	-86.52630
KAC 105	*Anaxyrus americanus*	33.04278	-86.45377
KAC 106	*Anaxyrus americanus*	33.00464	-86.49692
KAC 107	*Anaxyrus americanus*	33.01416	-86.38417
KAC 108	*Anaxyrus terrestris*	32.94013	-86.54004
KAC 109	*Anaxyrus terrestris*	32.94173	-86.55787
KAC 110	*Anaxyrus americanus*	33.03099	-86.40941
KAC 111	*Anaxyrus americanus*	33.00518	-86.49895
KAC 112	*Anaxyrus terrestris*	32.95011	-86.53723
KAC 113	*Anaxyrus americanus*	33.00528	-86.38897
KAC 114	*Anaxyrus americanus*	33.01617	-86.40318
KAC 115	*Anaxyrus americanus*	32.98218	-86.40488
KAC 116	*Anaxyrus americanus*	32.96964	-86.42137
KAC 117	*Anaxyrus terrestris*	32.97146	-86.52901
KAC 121	*Anaxyrus terrestris*	32.44120	-85.65386
KAC 122	*Anaxyrus terrestris*	32.85411	-86.76619
KAC 123	*Anaxyrus terrestris*	32.90084	-86.67587
KAC 124	*Anaxyrus terrestris*	32.91060	-86.67850
KAC 125	*Anaxyrus terrestris*	32.91715	-86.68208
KAC 126	*Anaxyrus terrestris*	32.92717	-86.67407
KAC 127	*Anaxyrus terrestris*	32.97159	-86.62516
KAC 128	*Anaxyrus terrestris*	33.00585	-86.63703
KAC 129	*Anaxyrus terrestris*	33.00797	-86.64210
KAC 130	*Anaxyrus terrestris*	33.00818	-86.64333
KAC 131	*Anaxyrus terrestris*	33.01508	-86.64937
KAC 132	*Anaxyrus terrestris*	33.02034	-86.66651
KAC 133	*Anaxyrus terrestris*	33.01163	-86.64759
KAC 134	*Anaxyrus terrestris*	33.00537	-86.63652
KAC 135	*Anaxyrus terrestris*	33.00644	-86.63368
KAC 136	*Anaxyrus terrestris*	33.00673	-86.63316
KAC 138	*Anaxyrus americanus*	32.70224	-85.66196
KAC 139	*Anaxyrus americanus*	32.73042	-85.66173
KAC 140	*Anaxyrus terrestris*	32.62553	-85.63684
KAC 141	*Anaxyrus terrestris*	32.41032	-85.60107
KAC 142	*Anaxyrus terrestris*	32.57011	-85.80888
KAC 143	*Anaxyrus terrestris*	32.47773	-85.79824
KAC 144	*Anaxyrus terrestris*	32.47707	-85.79577
KAC 145	*Anaxyrus terrestris*	32.48128	-85.76354
KAC 146	*Anaxyrus terrestris*	32.48291	-85.75622

Continued on next page

Table 2.1 – continued from previous page

Sample ID	Species	Latitude	Longitude
KAC 147	*Anaxyrus terrestris*	32.45001	-85.79652
KAC 148	*Anaxyrus terrestris*	32.45420	-85.79408
KAC 149	*Anaxyrus terrestris*	32.45449	-85.78664
KAC 150	*Anaxyrus terrestris*	32.45449	-85.78664
KAC 151	*Anaxyrus terrestris*	32.45451	-85.78416
KAC 152	*Anaxyrus terrestris*	32.45423	-85.77634
KAC 153	*Anaxyrus terrestris*	32.45423	-85.77634
KAC 154	*Anaxyrus terrestris*	32.46574	-85.76977
KAC 155	*Anaxyrus terrestris*	32.46961	-85.77369
KAC 156	*Anaxyrus terrestris*	32.47709	-85.79175
KAC 158	*Anaxyrus terrestris*	32.47709	-85.79175
KAC 159	*Anaxyrus terrestris*	32.49000	-85.79741
KAC 160	*Anaxyrus terrestris*	32.40809	-85.47857
KAC 161	*Anaxyrus terrestris*	32.41744	-85.47117
KAC 162	*Anaxyrus terrestris*	32.35417	-86.09838
KAC 163	*Anaxyrus terrestris*	32.33994	-86.09946
KAC 164	*Anaxyrus terrestris*	32.31562	-86.13789
KAC 167	*Anaxyrus terrestris*	33.06620	-86.60328
KAC 172	*Anaxyrus americanus*	32.62171	-85.61467
KAC 173	*Anaxyrus americanus*	32.61751	-85.64335
KAC 176	*Anaxyrus americanus*	32.66836	-85.66233
KAC 177	*Anaxyrus americanus*	32.65571	-85.57134
KAC 181	*Anaxyrus terrestris*	32.38644	-85.23561
KAC 182	*Anaxyrus terrestris*	32.38579	-85.23565
KAC 183	*Anaxyrus terrestris*	32.38579	-85.23565
KAC 184	*Anaxyrus terrestris*	32.38579	-85.23565
KAC 185	*Anaxyrus terrestris*	32.38579	-85.23565
KAC 187	*Anaxyrus americanus*	32.64548	-85.55135
KAC 188	*Anaxyrus terrestris*	32.40976	-85.60208
KAC 189	*Anaxyrus terrestris*	33.09152	-86.56686
KAC 190	*Anaxyrus terrestris*	33.11298	-86.69434
KAC 191	*Anaxyrus terrestris*	33.10659	-86.68228
KAC 192	*Anaxyrus terrestris*	33.10509	-86.68014
KAC 193	*Anaxyrus terrestris*	33.07896	-86.67286
KAC 194	*Anaxyrus terrestris*	32.93933	-86.62008
KAC 195	*Anaxyrus terrestris*	32.94745	-86.62146
KAC 196	*Anaxyrus terrestris*	32.94829	-86.62190
KAC 197	*Anaxyrus terrestris*	32.94929	-86.62241
KAC 198	*Anaxyrus terrestris*	32.95077	-86.62306
KAC 199	*Anaxyrus terrestris*	32.95794	-86.62477
KAC 200	*Anaxyrus terrestris*	32.95940	-86.62489
KAC 205	*Anaxyrus terrestris*	32.54852	-85.48692
KAC 206	*Anaxyrus americanus*	33.30759	-86.58201
KAC 207	*Anaxyrus americanus*	33.31685	-86.57596
KAC 208	*Anaxyrus americanus*	33.09829	-86.56529

Continued on next page

Table 2.1 – continued from previous page

Sample ID	Species	Latitude	Longitude
KAC 209	*Anaxyrus terrestris*	33.08600	-86.56394
KAC 210	*Anaxyrus terrestris*	33.08600	-86.56394
KAC 211	*Anaxyrus terrestris*	33.01464	-86.60995
KAC 212	*Anaxyrus terrestris*	33.01208	-86.61707
KAC 213	*Anaxyrus terrestris*	33.00435	-86.63710
KAC 214	*Anaxyrus terrestris*	32.99991	-86.64181
KAC 215	*Anaxyrus terrestris*	32.99605	-86.64526
KAC 216	*Anaxyrus terrestris*	33.01346	-86.60960
KAC 217	*Anaxyrus terrestris*	32.91470	-86.60270
KAC 218	*Anaxyrus terrestris*	32.92432	-86.59895
KAC 219	*Anaxyrus terrestris*	32.93987	-86.56113
KAC 220	*Anaxyrus americanus*	32.96579	-86.50892
KAC 221	*Anaxyrus americanus*	32.96389	-86.42549
KAC 223	*Anaxyrus terrestris*	32.53362	-85.79839
KAC 224	*Anaxyrus terrestris*	32.48869	-85.79555
KAC 225	*Anaxyrus terrestris*	32.50159	-85.79860
KAC 230	*Anaxyrus terrestris*	30.80933	-86.77686
KAC 232	*Anaxyrus terrestris*	30.80922	-86.78994
KAC 233	*Anaxyrus terrestris*	30.80922	-86.78994
KAC 234	*Anaxyrus terrestris*	30.80922	-86.78994
KAC 236	*Anaxyrus terrestris*	30.82632	-86.80258
KAC 237	*Anaxyrus terrestris*	30.83733	-86.77630
KAC 238	*Anaxyrus terrestris*	30.82433	-86.76284
KAC 239	*Anaxyrus terrestris*	30.80162	-86.76659
KAC 242	*Anaxyrus americanus*	34.50446	-85.63768
KAC t1020	*Anaxyrus terrestris*	31.10783	-86.62247
KAC t1030	*Anaxyrus terrestris*	31.99042	-85.07423
KAC t1040	*Anaxyrus terrestris*	31.99016	-85.07046
KAC t2004	*Anaxyrus americanus*	33.58295	-85.73524
KAC t2015	*Anaxyrus americanus*	33.58435	-85.74064
KAC t2018-02-17-01	*Anaxyrus americanus*	33.55274	-85.82913
KAC t2018-02-17-04	*Anaxyrus americanus*	33.48548	-85.88857
KAC t2018-02-17-05	*Anaxyrus americanus*	33.31649	-86.05293
KAC t2018-02-17-06	*Anaxyrus americanus*	33.28443	-86.08443
KAC t2018-02-17-07	*Anaxyrus americanus*	33.24576	-86.08168
KAC t2018-03-10-1	*Anaxyrus americanus*	32.91057	-86.09272
KAC t2018-03-10-3	*Anaxyrus americanus*	32.95104	-86.14539
KAC t2018-03-10-4	*Anaxyrus americanus*	32.89787	-86.26061
KAC t2018-03-10-5	*Anaxyrus americanus*	32.81642	-86.38018
KAC t2018-08-18-1	*Anaxyrus terrestris*	30.66902	-81.44013
KAC t2018-08-18-2	*Anaxyrus terrestris*	30.66902	-81.44013
KAC t2018-08-18-3	*Anaxyrus terrestris*	30.43282	-81.64088
KAC t2018-08-18-4	*Anaxyrus terrestris*	30.66902	-81.44013
KAC t2019-08-25-1	*Anaxyrus americanus*	34.21852	-87.36662
KAC t2020	*Anaxyrus americanus*	33.23853	-85.96270

Continued on next page

Table 2.1 – continued from previous page

Sample ID	Species	Latitude	Longitude
KAC t2040	*Anaxyrus americanus*	33.58295	-85.73539
KAC t2043	*Anaxyrus americanus*	32.81642	-86.38018

Table 2.2. Samples loaned from museums

Sample ID	Species	Latitude	Longitude
AHT 1975	*Anaxyrus americanus*	32.77356	-85.53325
AHT 2456	*Anaxyrus terrestris*	32.19494	-89.23629
AHT 2564	*Anaxyrus terrestris*	31.55752	-84.04267
AHT 2885	*Anaxyrus terrestris*	32.45090	-86.15934
AHT 3419	*Anaxyrus terrestris*	33.67290	-88.16068
AHT 3421	*Anaxyrus terrestris*	33.65420	-88.15580
AHT 3428	*Anaxyrus terrestris*	31.12679	-86.54755
AHT 3459	*Anaxyrus americanus*	34.88028	-87.71849
AHT 3460	*Anaxyrus americanus*	33.78013	-85.58421
AHT 3461	*Anaxyrus americanus*	34.88779	-87.74103
AHT 3462	*Anaxyrus americanus*	33.77001	-85.55434
AHT 3463	*Anaxyrus americanus*	33.71125	-85.59762
AHT 3813	*Anaxyrus terrestris*	31.13854	-86.53906
AHT 3833	*Anaxyrus terrestris*	31.00422	-85.03427
AHT 3997	*Anaxyrus terrestris*	32.55607	-88.29975
AHT 3998	*Anaxyrus terrestris*	32.55607	-88.29975
AHT 5276	*Anaxyrus terrestris*	31.55613	-86.82514
AHT 5277	*Anaxyrus terrestris*	31.15830	-86.55430
AHT 5278	*Anaxyrus terrestris*	31.16105	-86.69868
UTEP 19947	*Anaxyrus terrestris*	31.22432	-88.77548

Chapter 3

Comparison of Linked versus Unlinked Character Models for Species Tree Inference

3.1 Introduction

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

Reduced-representation genomic data sets acquired from high-throughput instruments are becoming commonplace in phylogenetics (Leaché & 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.

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. Important considerations in such decisions include the sources of error and bias that result from reduced-representation protocols, high-throughput sequencing technologies, and the processing of these data.

Most reduced-representation sequencing workflows employ amplification of DNA using polymerase chain reaction (PCR) which can introduce mutational error at a rate of up to 1.5×10^{-5} substitutions per base (Potapov & 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

3.2.1 Simulations of error-free data sets

For our simulations, we assumed a simple two-tipped species tree with one ancestral population with a constant effective size of N_e^R that diverged at time τ into two descendent populations (terminal branches) with constant effective sizes of N_e^{D1} and N_e^{D2} (Fig. 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

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

For each simulated data set with loci of 250, 500, and 1000 characters, we approximated the posterior distribution of the divergence time (τ) and effective population sizes (N_e^R , N_e^{D1} , and N_e^{D2}) under an unlinked-character model using ecoevolity (Version 0.3.2, Commit a7e9bf2; Oaks, 2019) and a linked-character model using the StarBEAST2 package (Version 0.15.1; Ogilvie et al., 2017) in BEAST2 (Version 2.5.2; Bouckaert et al., 2014). For both methods, we specified a CTMC model of character evolution and prior distributions that matched the model and distributions from which the data were generated. The prior on the effective size of the root population in the original implementation of ecoevolity was parameterized to be relative to the mean effective size of the descendant populations. We added an option to ecoevolity to compile a version where the prior is specified as the absolute effective size of the root population, which matches the model in StarBEAST2 and the model we used to generate the data. The linkage of sites within loci of our simulated data violates the unlinked-character model of ecoevolity (Bryant et al., 2012; Oaks, 2019). Therefore, we also analyzed each data set with ecoevolity after selecting, at most, one variable character from each locus; loci without variable sites were excluded.

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

For ecoevolity, we ran four independent Markov chain Monte Carlo (MCMC) analyses with 75,000 steps and a sample frequency of 50 steps. For StarBEAST2, we ran two independent MCMC analyses with 20 million steps and a sample frequency of 5000 steps. To assess convergence and mixing of the ecoevolity and StarBEAST2 MCMC chains, we computed the effective sample size (ESS; Gong & 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 ecoevolity and StarBEAST2, leaving 4000 and 7600 posterior samples for each data set, respectively.

3.2.4 Project repository

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

3.3 Results

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

The divergence times estimated by the linked-character method, StarBEAST2, were very accurate and precise for all alignment lengths and types and degrees errors, despite

poor MCMC mixing (i.e., low ESS values) for shorter loci (Figs. 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 unlinked characters (i.e., no linked sites). This coverage behavior is expected, and helps to confirm confirm that our simulation pipeline generated data under the same model used for inference by StarBEAST2 and ecoevolity. As seen previously (Oaks, 2019), analyzing longer linked loci causes the coverage of ecoevolity to be lower, due to the violation of the model's assumption that the sites are unlinked.

3.3.4 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 per-site 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 be generated under a known model with no prior mis-specification. 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 data might indicate that the model is being biased by errors or acquisition biases present in the character patterns.

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

3.4.4 Other complexities of empirical data in need of exploration

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

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

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

data, there are likely many model violations, and our prior distributions will never match the distributions that generated the data. Introducing other model violations and misspecified prior distributions would thus help to better understand how species-tree models behave on real data sets. Of particular concern is whether misspecified priors will amplify the effect of character-pattern errors or biases.

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

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. *Systematic Biology*, *65*(3), 357–365. https://doi.org/10.1093/sysbio/syu046

Hunter, J. D. (2007). Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering*, *9*(3), 90–95. https://doi.org/10.1109/MCSE.2007.55

Leaché, A. D., & Oaks, J. R. (2017). The Utility of Single Nucleotide Polymorphism (SNP) Data in Phylogenetics. *Annual Review of Ecology, Evolution, and Systematics*, *48*(1), 69–84. https://doi.org/10.1146/annurev-ecolsys-110316-022645

Linck, E., & Battey, C. J. (2019). Minor allele frequency thresholds strongly affect population structure inference with genomic data sets. *Molecular Ecology Resources*, *19*(3), 639–647. https://doi.org/10.1111/1755-0998.12995

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.). *Systematic Biology*, *56*(3), 504–514. https://doi.org/10.1080/10635150701429982

Oaks, J. R. (2019). Full Bayesian Comparative Phylogeography from Genomic Data (L. Kubatko, Ed.). *Systematic Biology*, *68*(3), 371–395. https://doi.org/10.1093/sysbio/syy063

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. *Evolution*, *73*(6), 1151–1167. https://doi.org/10.1111/evo.13754

Ogilvie, H. A., Bouckaert, R. R., & Drummond, A. J. (2017). StarBEAST2 Brings Faster Species Tree Inference and Accurate Estimates of Substitution Rates. *Molecular Biology and Evolution*, *34*(8), 2101–2114. https://doi.org/10.1093/molbev/msx126

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. *Scientific Reports*, *8*(1), 10950. https://doi.org/10.1038/s41598-018-29325-6

Potapov, V., & Ong, J. L. (2017). Examining Sources of Error in PCR by Single-Molecule Sequencing (R. Kalendar, Ed.). *PLOS ONE*, *12*(1), e0169774. https://doi.org/10.1371/journal.pone.0169774

Rambaut, A., & Grass, N. C. (1997). Seq-Gen: An application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Bioinformatics*, *13*(3), 235–238. https://doi.org/10.1093/bioinformatics/13.3.235

Rochette, N. C., Rivera-Colón, A. G., & Catchen, J. M. (2019). Stacks 2: Analytical methods for paired-end sequencing improve RADseq-based population genomics. *Molecular Ecology*, *28*(21), 4737–4754. https://doi.org/10.1111/mec.15253

Sukumaran, J., & Holder, M. T. (2010). DendroPy: A Python library for phylogenetic computing. *Bioinformatics*, *26*(12), 1569–1571. https://doi.org/10.1093/bioinformatics/btq228

Yang, Z. (2015). The BPP program for species tree estimation and species delimitation. *Current Zoology*, *61*(5), 854–865. https://doi.org/10.1093/czoolo/61.5.854

Yule, G. U. (1925). A mathematical theory of evolution, based on the conclusions of Dr. JC Willis, FR S. *Philosophical transactions of the Royal Society of London. Series B, containing papers of a biological character*, *213*(402-410), 21–87.

3.6 Figures

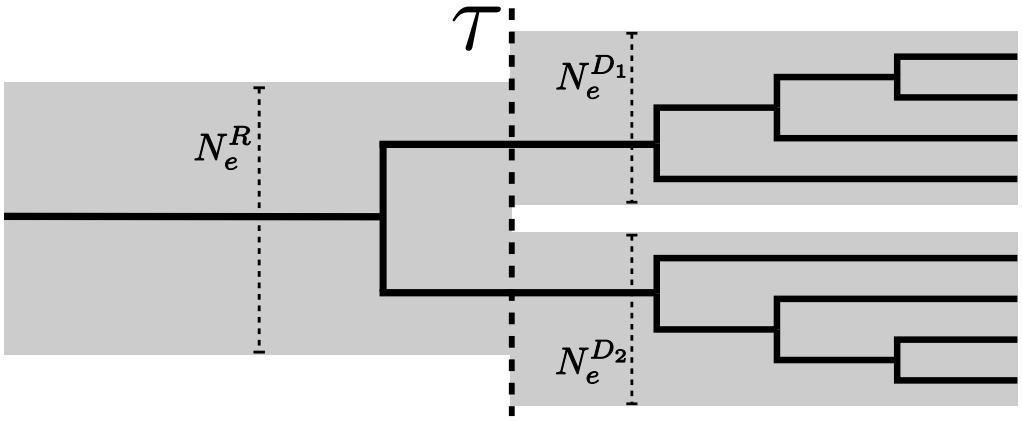

Figure 3.1. An illsustration of the species-tree model we used to simulate data. N_e^R , N_e^{D1} , and N_e^{D2} represent the constant effective population sizes of the root, and each of the two terminal populations. τ reperesents the instantaneous separation of the ancestral population into two descendant populations. One hypothetical gene tree is shown to illustrate the gene trees simulated under a contained coalescent process for 4 haploid gene copies sampled from each of the terminal branches of the species tree.

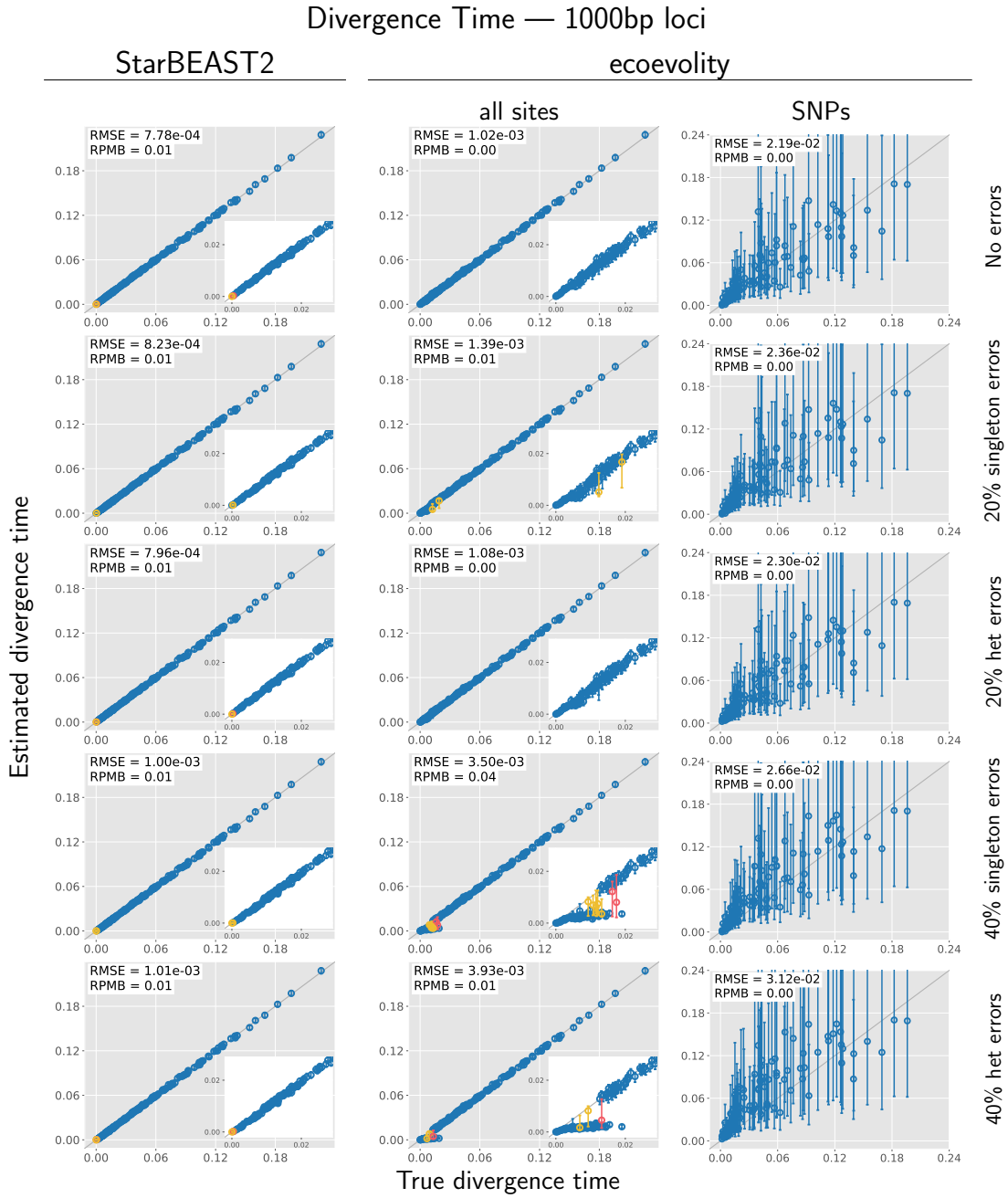

Figure 3.2. **Accuracy and precision of divergence-time estimates (in units of expected substitutions per site) with 1000 base pair loci.** The left column shows estimates from StarBEAST2, and the center and right column shows estimates from ecoevolity using all sites and (at most) one SNP per locus.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

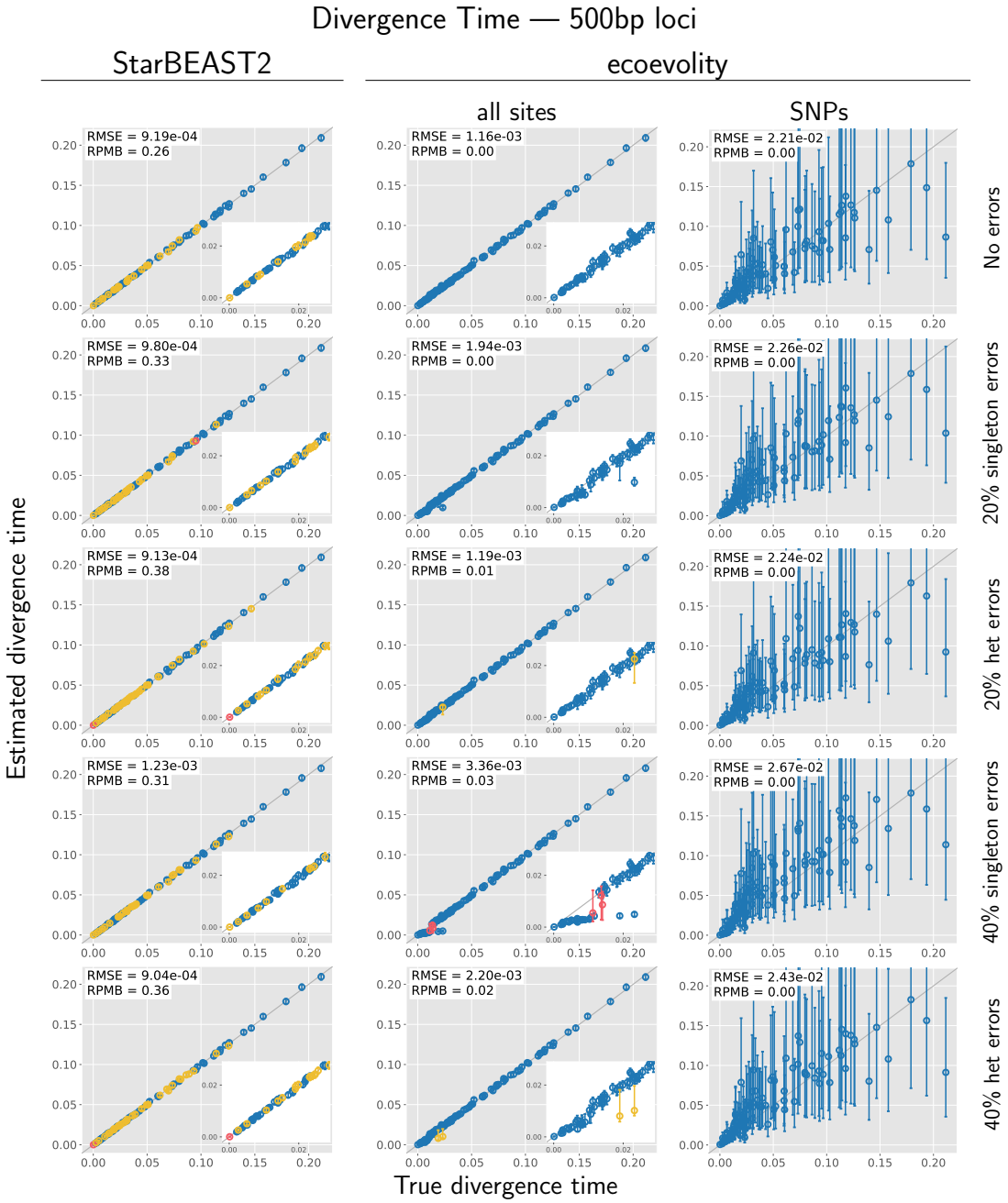

Figure 3.3. **Accuracy and precision of divergence-time estimates (in units of expected substitutions per site) with 500 base pair loci.** The left column shows estimates from StarBEAST2, and the center and right column shows estimates from ecoevolity using all sites and (at most) one SNP per locus.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

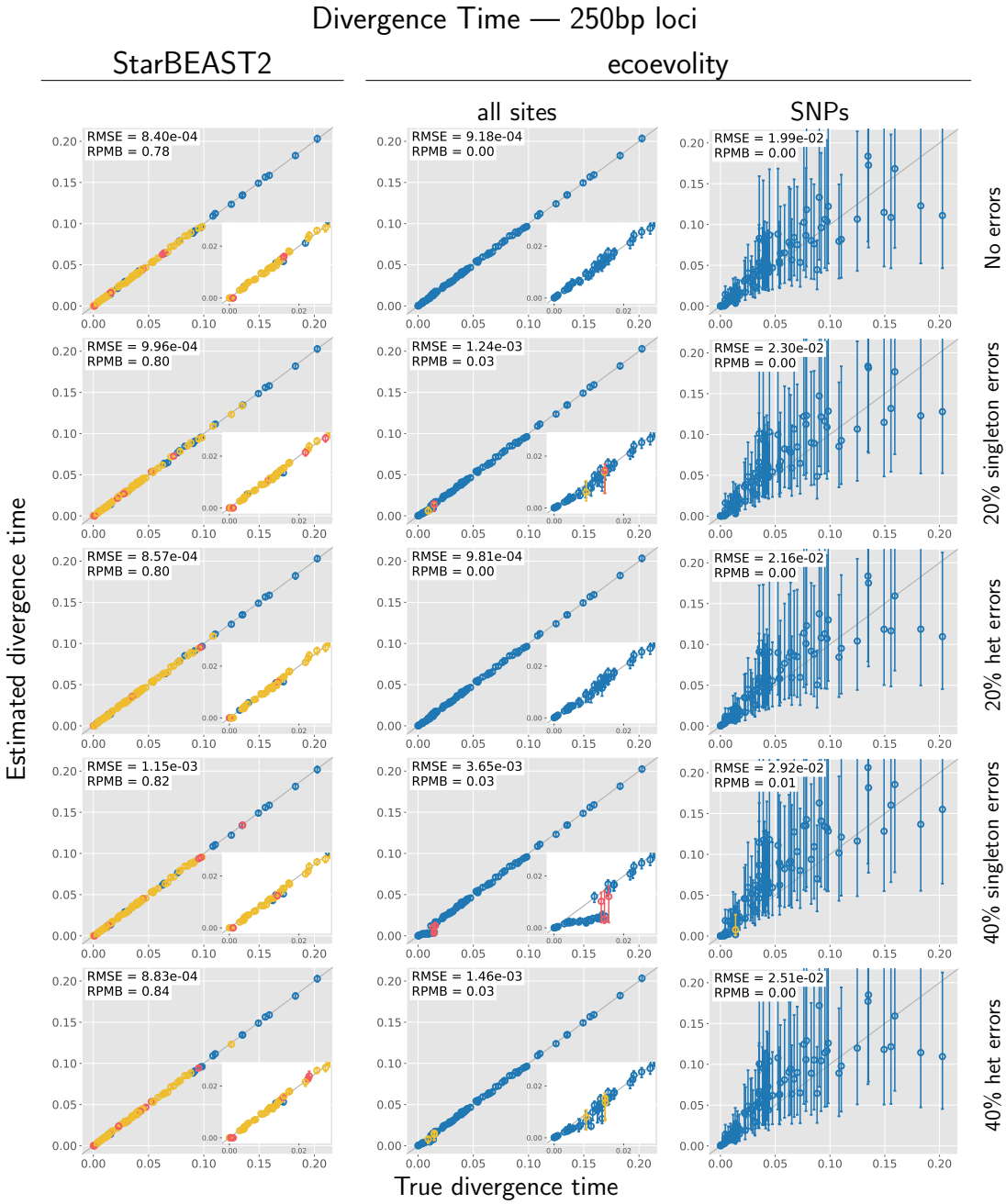

Figure 3.4. **Accuracy and precision of divergence-time estimates (in units of expected substitutions per site) with 250 base pair loci.** The left column shows estimates from StarBEAST2, and the center and right column shows estimates from ecoevolity using all sites and (at most) one SNP per locus.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

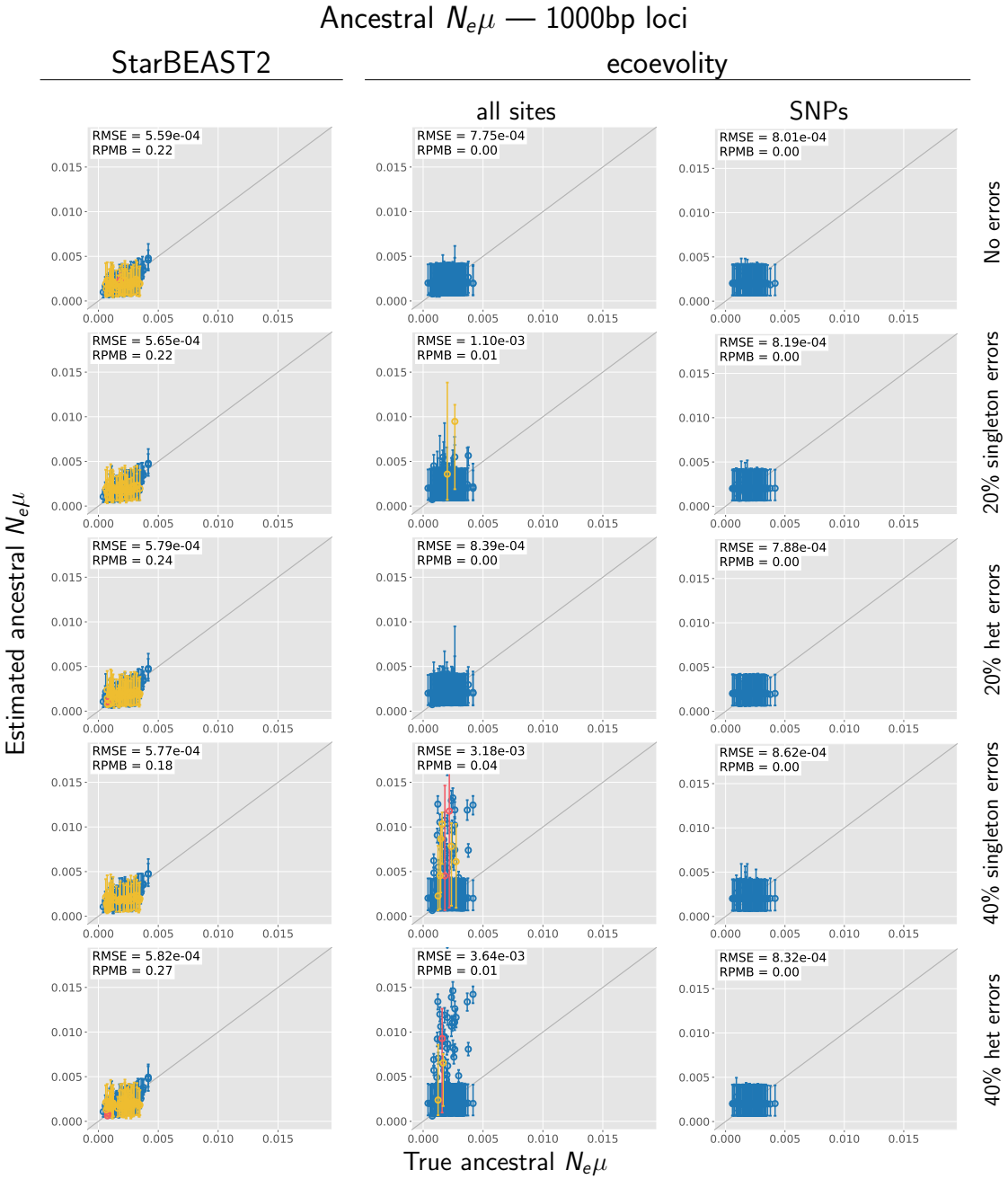

Figure 3.5. **Accuracy and precision of estimates of root effective population size scaled by the mutation rate ($N_e^R\mu$) with 1000 base pair loci.** The left column shows estimates from StarBEAST2, and the center and right column shows estimates from ecoevolity using all sites and (at most) one SNP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

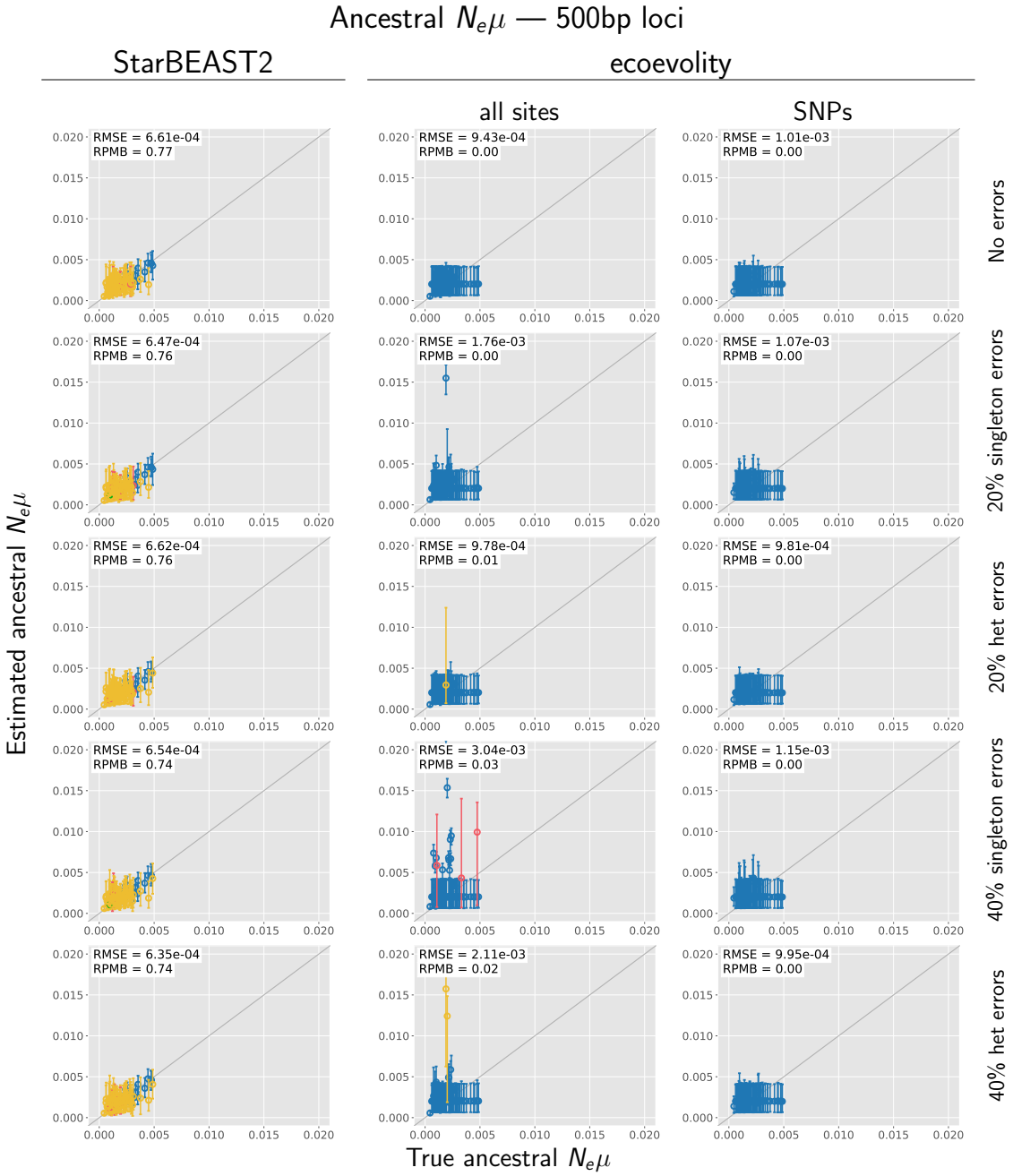

Figure 3.6. **Accuracy and precision of estimates of root effective population size scaled by the mutation rate ($N_e^R\mu$) with 500 base pair loci.** The left column shows estimates from StarBEAST2, and the center and right column shows estimates from ecoevolity using all sites and (at most) one SNP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

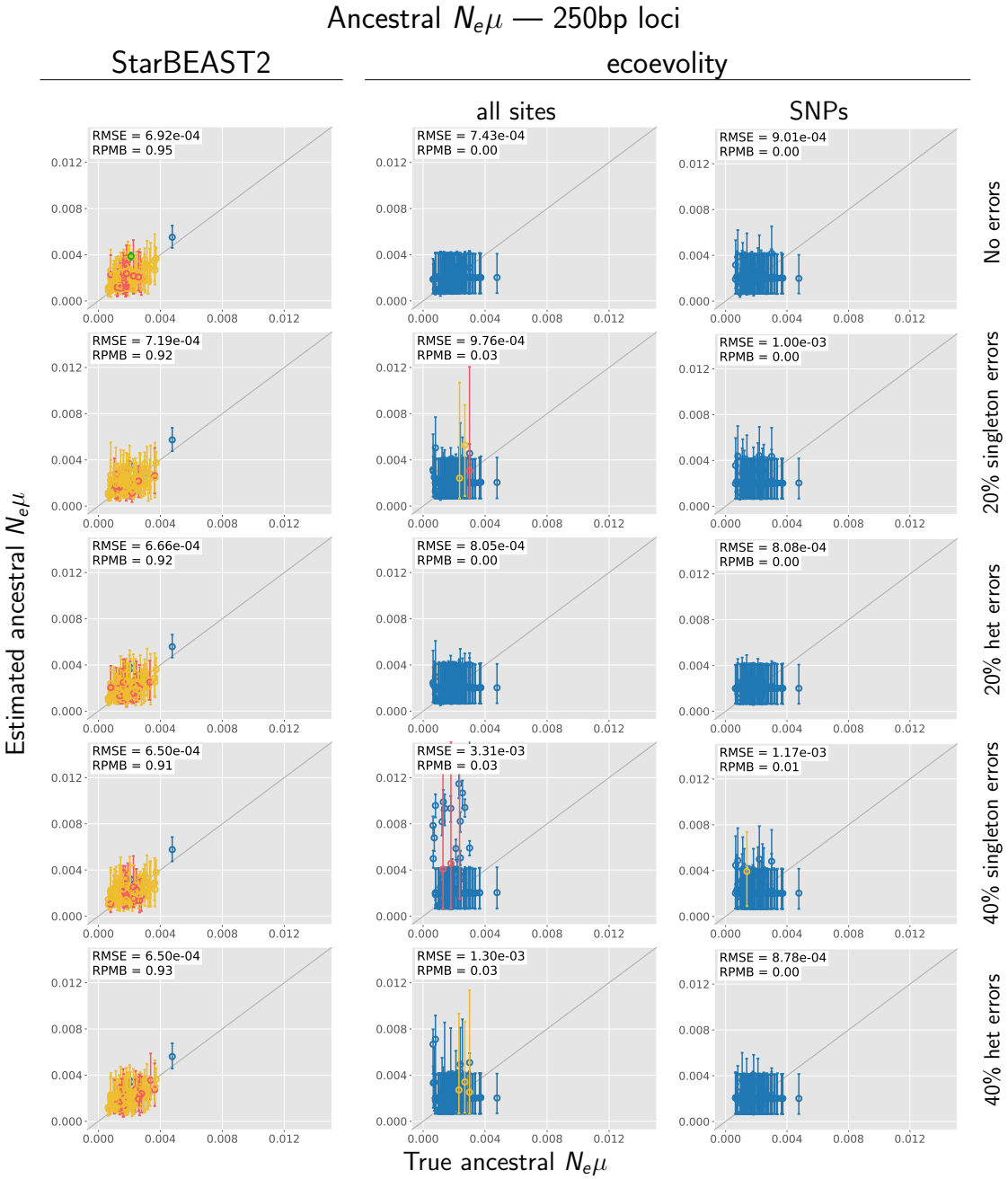

Figure 3.7. **Accuracy and precision of estimates of root effective population size scaled by the mutation rate ($N_e^R\mu$) with 250 base pair loci.** The left column shows estimates from StarBEAST2, and the center and right column shows estimates from ecoevolity using all sites and (at most) one SNP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

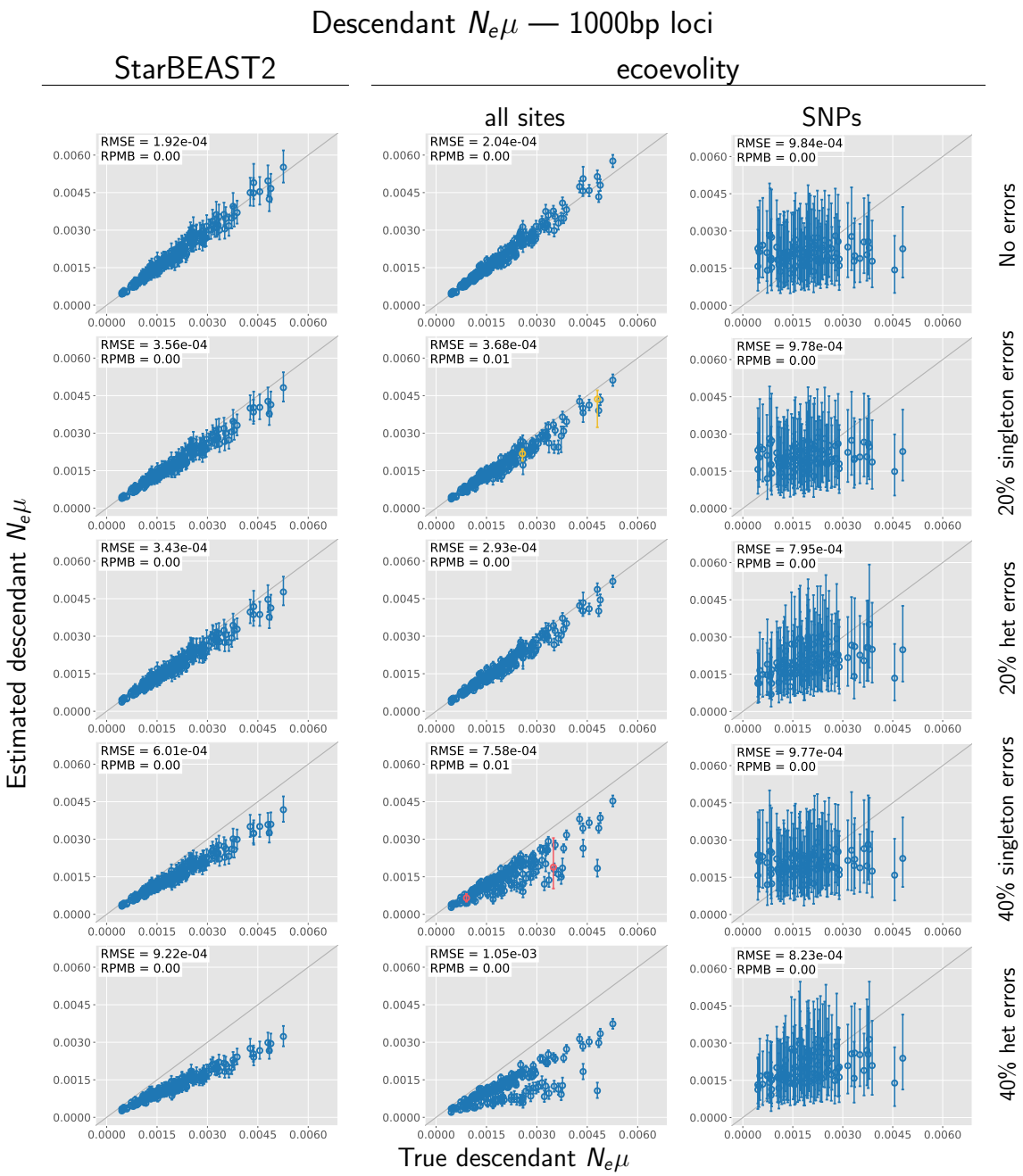

Figure 3.8. **Accuracy and precision of estimates of effective population sizes of the descendant branches of the tree scaled by the mutation rate ($N_e^D\mu$) with 1000 base pair loci.** The left column shows estimates from StarBEAST2, and the center and right column shows estimates from ecoevolity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

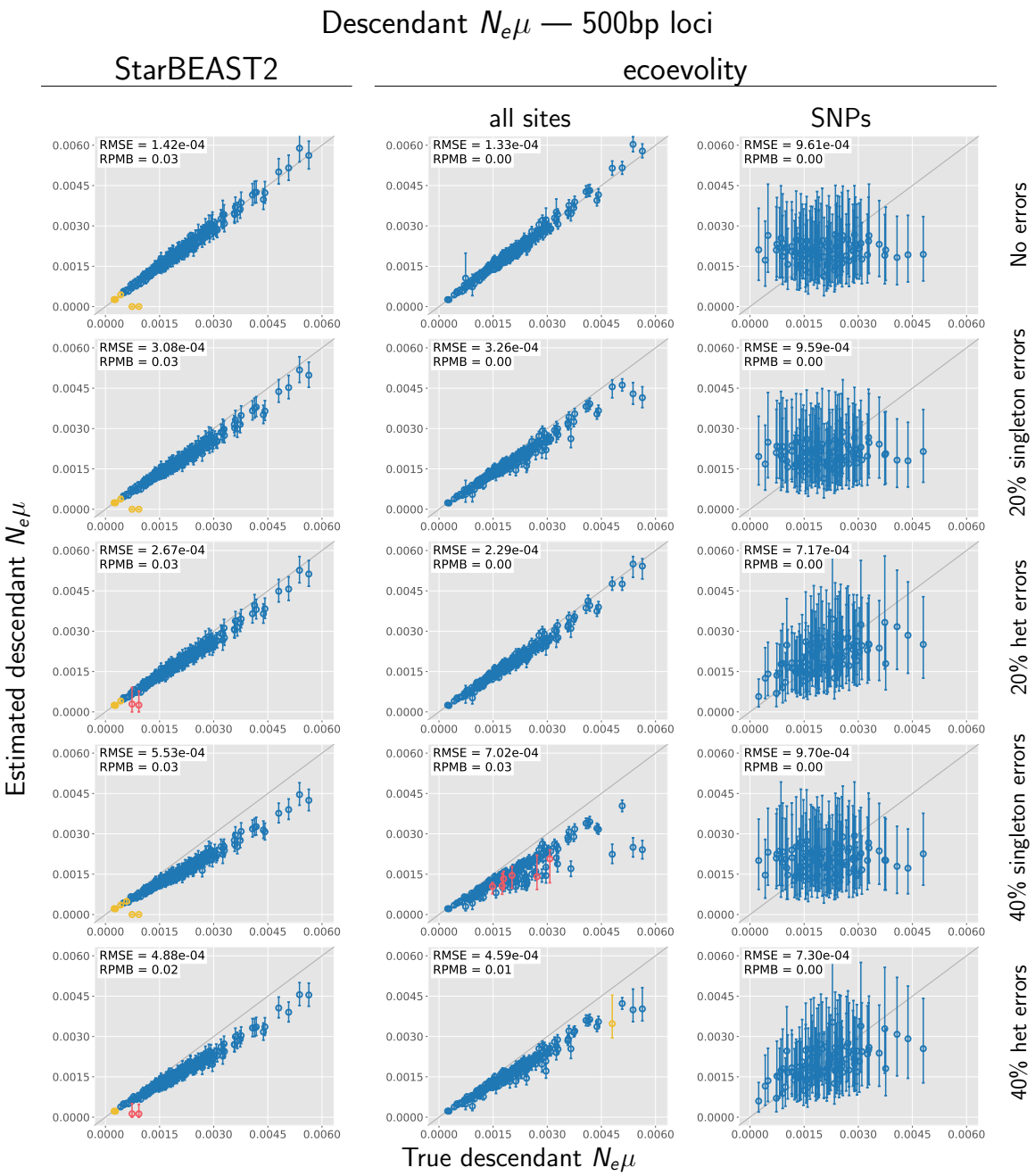

Figure 3.9. **Accuracy and precision of estimates of effective population sizes of the descendant branches of the tree scaled by the mutation rate ($N_e^D\mu$) with 500 base pair loci.** The left column shows estimates from StarBEAST2, and the center and right column shows estimates from ecoevolity using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

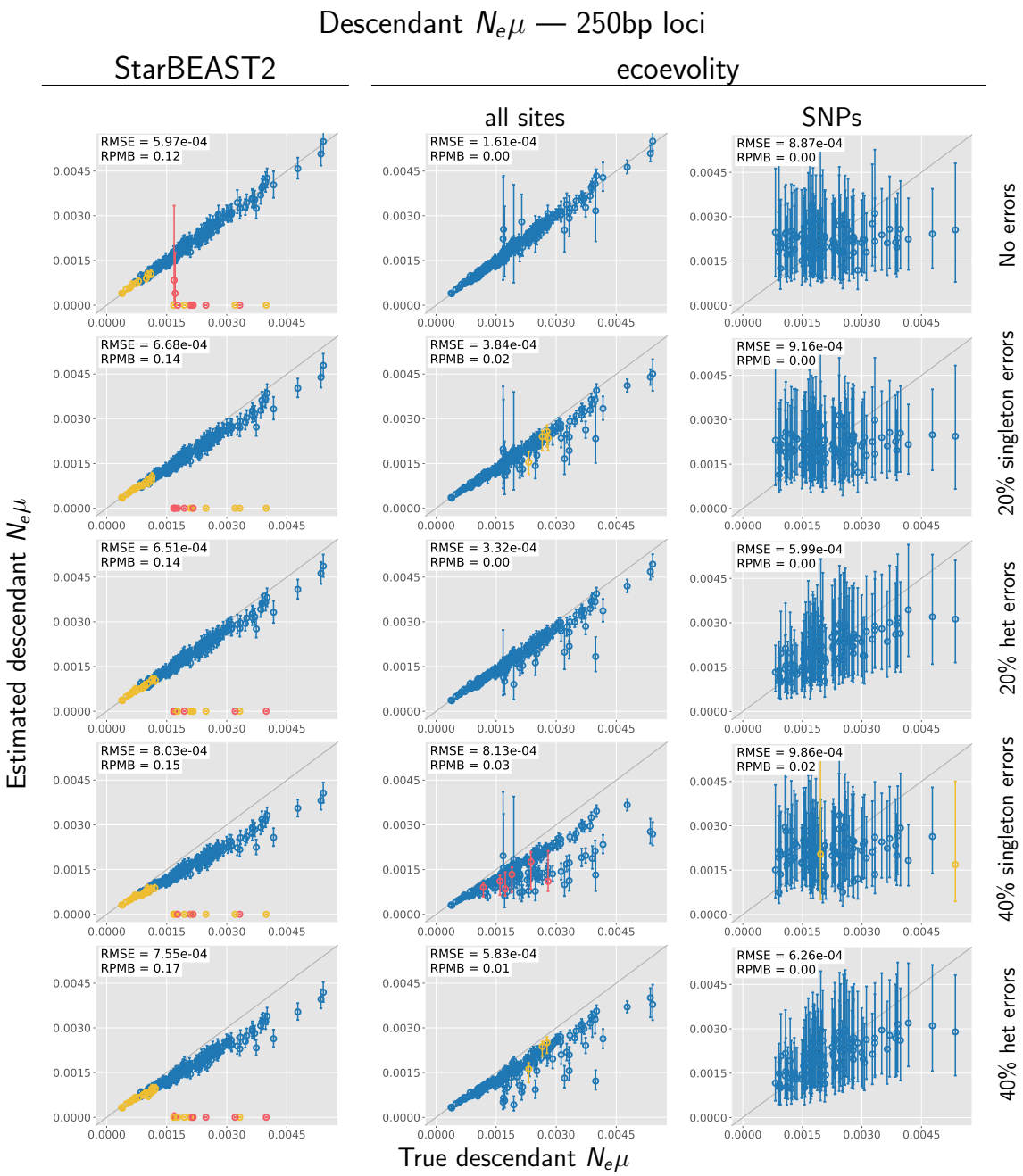

Figure 3.10. **Accuracy and precision of estimates of effective population sizes of the descendant branches of the tree scaled by the mutation rate ($N_e^D\mu$) with 250 base pair loci.** The left column shows estimates from StarBEAST2, and the center and right column shows estimates from ecoevolity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

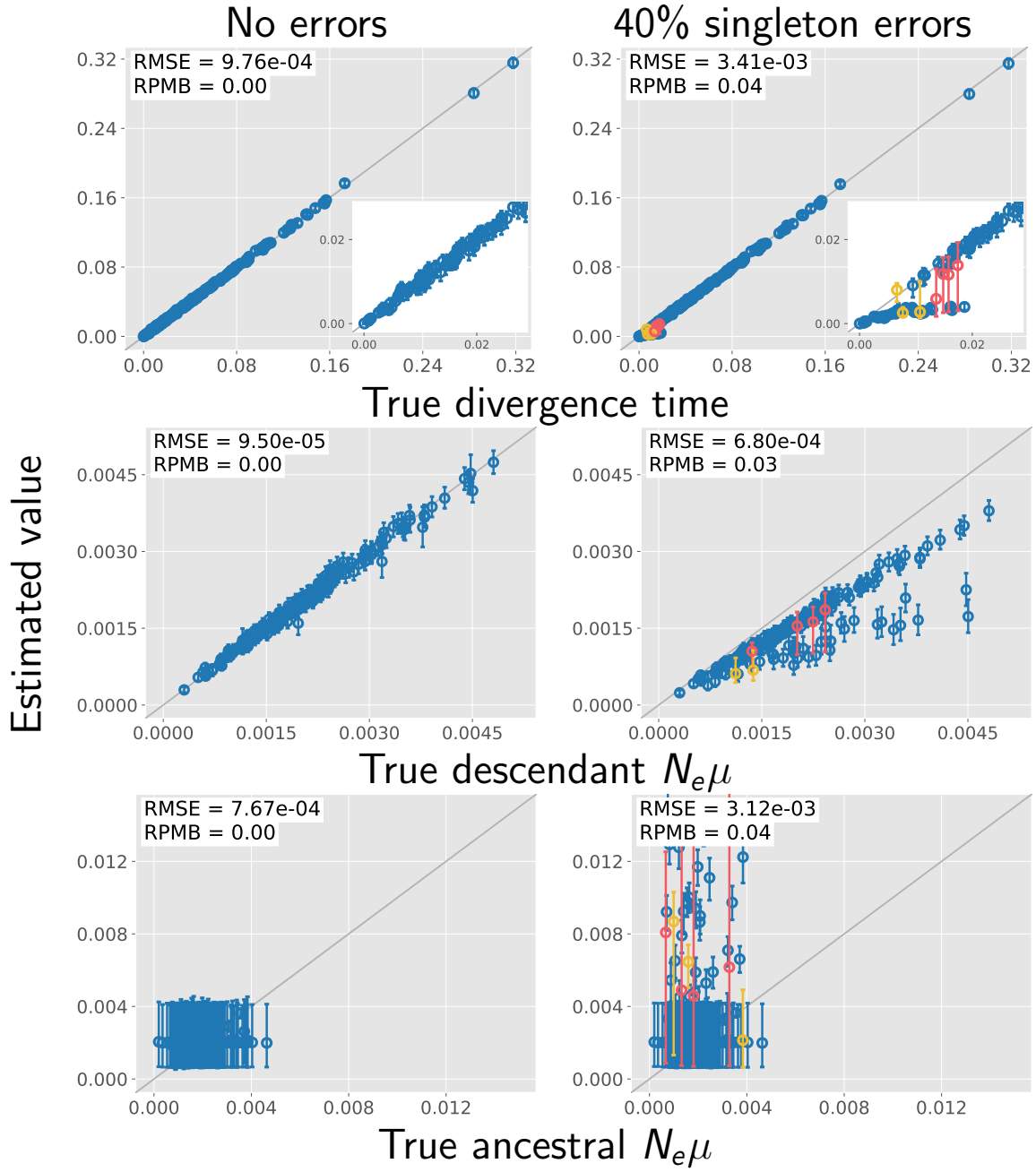

Figure 3.11. The performance of ecoevolity with data sets simulated with unlinked characters. Each plotted circle and associated error bars represent the posterior mean and 95% credible interval. Circles and error bars are colored yellow if the effective sample size (ESS) of the estimate was less than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