Contents

List of Figures

List of Tables

Chapter 1

Phylogeography

1.1 Acknowledgments

...

1.2 Figures

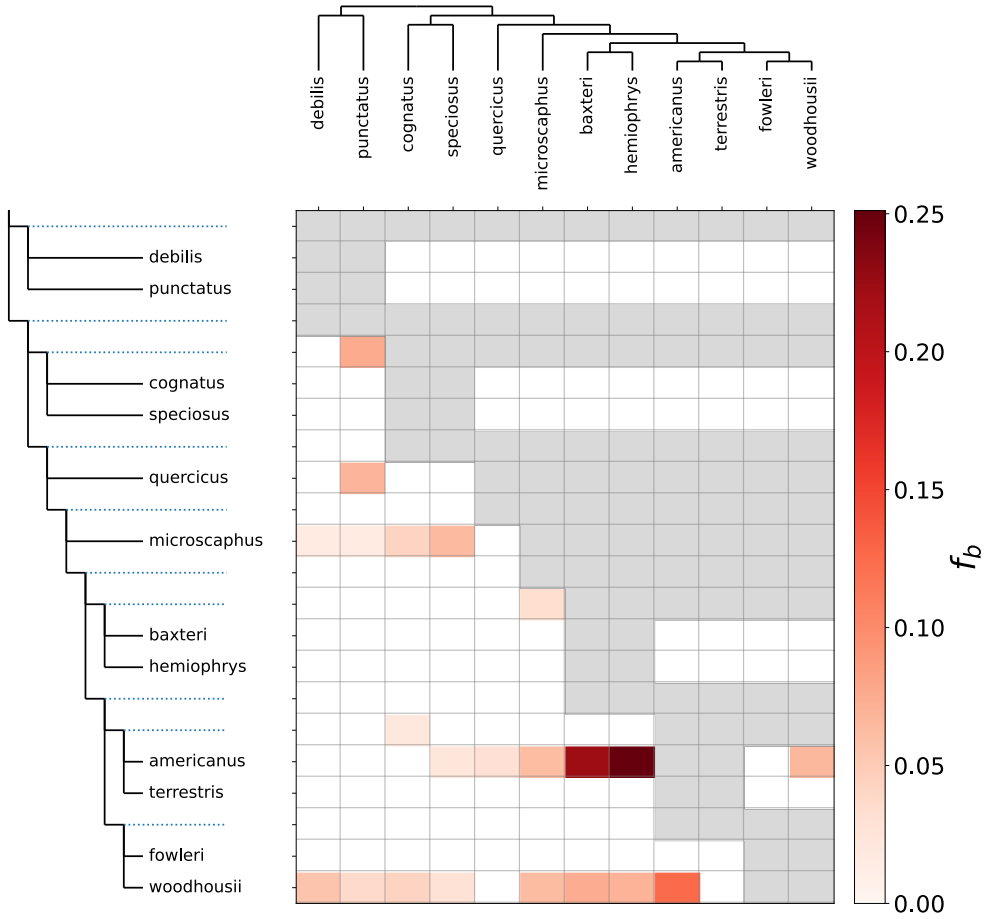

Figure 1.1. Historic admixture

1.3 Tables

94

Chapter 2

Hybrid Zone

2.1 Introduction

Speciation is the process by which genetic divergence leads to a reduction in the ability of populations of organisms to exchange genes. It is a gradual and continuous process during which the exchange of genetic variants through hybridization can occur (Mallet, 2008; Wu, 2001). Natural hybridization has become increasingly appreciated as a widespread phenomenon in recent years (Mallet, 2005; Moran et al., 2021). It is a phenomenon that can have important evolutionary consequences. Natural hybridization can be a source of adaptive variation (Hedrick, 2013). It can also introduce deleterious genetic load which persists long term within a population (Moran et al., 2021). When the rate of admixture is high, it can drive the evolution of traits that enhance assortative mating resulting in reinforcement of the barriers between divergent lineages (Servedio & Noor, 2003). If hybrids do not suffer any declines in fitness, hybridization could lead to the erosion of differences between divergent populations (Taylor et al., 2006). Potentially resulting in populations that are genetically distinct from either parent species which can themselves could eventually evolve to become isolated from those parent species (Moran et al., 2021).

Aside from having important evolutionary consequences which warrant investigation, hybridization can also be an excellent opportunity to investigate the processes that result in the evolution of reproductive incompatibility between divergent evolutionary lineages. Hybrid zones are particularly suitable due to the production of large numbers of recombinant genomes carrying many possible combinations of genomic elements from parent species resulting from many generations of hybridization and backcrossing that occur within hybrid zones (Rieseberg et al., 1999). It should therefore be possible to distinguish between the effects of closely linked genes (Rieseberg et al., 1999). The many generations and large number of individuals producing these genetic combinations are not feasible to produce experimentally in the vast majority of species (Rieseberg et al., 1999). Furthermore, the combination of genes produced are exposed to natural selection under natural conditions. This is important as the effect of hybrid incompatibilities can be dependent on on environmental conditions and can only truly understood in this context (Miller & Matute, 2016).

Despite being a fundamental evolutionary process, our understanding of speciation is far from complete (Butlin et al., 2011). Only a few loci, in a few species, have been pinpointed as the direct cause of reproductive incompatibility between species (Blackman, 2016; Nosil & Schluter, 2011). Consequently, our understanding of the processes

that drive the evolution of loci resulting in reproductive incompatibility is limited (Butlin et al., 2011). Studies of introgression within hybrid zones have identified highly variable rates of introgression among loci (Barton & Hewitt, 1985; Gompert et al., 2017). This heterogeneity can arise via genetic drift occurring within hybrid populations, but will also be caused by differences among loci in the strength of selection against them in a hybrid genomic background (Barton & Hewitt, 1985; Gompert et al., 2017). It has also been observed that the levels of genetic divergence between species are highly variable across the genome (Nosil et al., 2009). Much of this heterogeneity is the result of divergent selection acting in different ways in each species (Nosil et al., 2009). Regions with particularly high levels of divergence between species have been coined "genomic islands of divergence" (Wolf & Ellegren, 2017). When speciation occurs with gene flow, divergent selection can cause adaptive divergence in habitat use, phenology, or mating signals, and reduce the frequency or success of interspecific matings. When species diverge in geographic isolation, divergent selection and reproductive isolation could be decoupled. Whether loci under divergent selection between two species also contribute to reproductive isolation has not been widely explored. A small handful of studies have found evidence for a modest relationship between genetic divergence and selection against introgression (Gompert et al., 2012; Larson et al., 2013; Nikolakis et al., 2022; Parchman et al., 2013). How consistent and widespread this pattern is remains to be seen.

In this study I investigate hybridization between the American toad (*Anaxayrus americanus*)and Southern toad (*Anaxyrus terrestris*) at a suspected hybrid zone in the Southern United States to assess the extent of introgression between them and test for a relationship between introgression and genetic divergence. This suspected hybrid zone has not been investigated with genetic data previously. The ranges of these species meet but do not overlap at a long contact zone which closely corresponds with a prominent physiographic feature known as the "fall line" (Mount, 1975; Powell et al., 2016). The Fall line is the boundary between the Southern coastal plain to the South from the Appalachian Highlands to the North (Shankman & Hart, 2007). These regions differ in their underlying geology, topography, and elevation (Shankman & Hart, 2007). The distribution of *A. terrestris* is restricted to the coastal plain extending from the Mississippi River in the West to Virginia in the East Fig. 2.4. The distribution of the American Toad encompasses nearly all of the Eastern North American with the exception of the Southern coastal plain Fig. 2.4. The two species have only slight differences in male advertisement call and in morphological appearance (Cocroft & Ryan, 1995; Weatherby, 1982). They also differ slightly in the timing of their spawn (Mount, 1975) However, there is significant overlap in the spawning period and male Bufondidae are famously indiscriminate in their choice of mates (Đorđević & Simović, 2014; Weatherby, 1982). Analysis of morphological variation in central Alabama has suggested there is introgression between them (Weatherby, 1982).

The "true toads" in the family Bufonidae, to which *A. americanus* and *A. terrestris* belong, have been a prominent group of organisms in the literature on hybridization. W.F. Blair and colleagues performed a remarkable 1,934 separate experimental crosses to quantify the degree of reproductive incompatibility between species pairs within this family (Blair, 1972; Malone & Fontenot, 2008). These experiments demonstrated a high degree of compatibility between some closely related species pairs in which hybrids were capable of producing viable backcross or F_2 hybrid offspring (Blair, 1963). Furthermore, numerous cases of natural hybridization among toad species have been reported, including at several apparent or clear hybrid zones (Colliard et al., 2010; Green, 1996;

Van Riemsdijk et al., 2023; Weatherby, 1982). Despite the interest and appreciation for hybridization in Bufonidae, only a small amount of work has been done to understand patterns of introgression within hybrid zones. A clinal pattern of admixture at 26 allozyme loci has been show within the *Anaxyrus americnaus* X *Anaxyrus hemiophrys* hybrid zone in Ontario, Canada(Green, 1983). Almost no admixture was detected at 7 microsatellite loci within the *Bufo siculus* X *Bufo balearicus* hybrid zone in Sicily, Italy (Colliard et al., 2010). The most comprehensive study of introgression within a Bufonidae hybrid zone found significant levels of genome wide admixture, fitting a clinal pattern, at two separate transects at either end of the *Bufo bufo* x *Bufo spinosus* hybrid zone in Southern France (Van Riemsdijk et al., 2023).

[KAC comment: Planning to move this to intro chapter.] It is clear that significant levels of admixture occur within some Bufonidae hybrid zones which could yield insights into the evolution of reproductive incompatibility. There are a few qualities that make these hybrid zones particularly attractive for further investigation. One of these qualities is the ease with which the primary behavioral isolating mechanisms, spawning period and advertisement call, can be measured and quantified in order to understand the strength of prezygotic mating barriers and possible patterns consistent with reinforcement (Blair, 1974; Cocroft & Ryan, 1995; Kennedy, 1962). It has also been show that they can be readily bred in captivity (Blair, 1972). Many species produce thousands of offspring which are externally fertilized making a variety of embryological observations or manipulations possible (Blair, 1972). Breeding can be induced hormonally or performed in vitro, facilitating the planning and scheduling of experiments (Trudeau et al., 2010). Unlike many of the organisms which have undergone intensive study in the context of speciation such as *Drosophila*, *Mus*, and *Heliconius*, most Bufonidae have homomorphic sex chromosomes (Blair, 1972). This is an interesting contrast in light of the important roll of sex chromosomes have in the evolution of reproductive incompatibility and the roll of heterogamety in explaining evolutionary patterns such as Haldane's rule, faster male evolution, and faster-X evolution (Delph & Demuth, 2016). Furthermore, there is evidence of sex chromosome turnovers within Bufonidae which presents an opportunity to study differences in the evolution of reproductive incompatibility among closely related species with different sex determination systems (Dufresnes et al., 2020; Stöck et al., 2011). All of these qualities along with the near global distribution of a large 642 species radiation present an excellent opportunity to further our understanding of the evolution of reproductive incompatibility (AmphibiaWeb, 2023).

The suspected *A. americanus*, *A. terrestris* hybrid zone has great potential to expand our understanding of speciation if there is ongoing introgression between species near the center of the hybrid zone. In this study I use genome-wide sequence data to characterize patterns of introgression within the hybrid zone by using model based inference of admixture proportions, cline models, and measurement of parental population differentiation. With these approaches I specifically address the following questions: 1) Is there evidence of ongoing hybridization and admixture between the two species, 2) Do any loci have outstanding patterns of introgression consistent with them being linked to reproductive incompatibility loci?, and 3) Is there any relationship between patterns of introgression and levels of genetic differentiation between parental lineages?

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 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 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 μL 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 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, 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/μL.

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 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 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 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 which were 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 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 model with one fewer populations 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. Explain a bit how *BGC* works??? ...I classified a sample as being admixed if it had an inferred admixture proportion of <95% for one species under the model with a K of two in the *STRUCTURE* analysis. I used *VCFtools* vXX.XX.XX to filter all non-biallelic sites from the the VCF file produced by the *populations* command in *Stacks*. I converted the VCF formatted data into the *BGC* format using *bgc_utils* v0.1.0, a *Python* package that I developed for this project github.com/kerrycobb/bgc_utils. I ran *BGC* with 5 independent chains, each for 1,000,000 steps and sampling every 1000. I visualized MCMC output, discarded samples from the posterior as burnin, combined the independent chains, summarized the posterior samples, and identified outlier loci with *bgc_utils*. A primary goal of *BGC* analysis is to identify loci which have exceptional patterns of introgression. These loci, or loci in close linkage to them, are expected to be enriched for genetic regions affected by selection due to reproductive incompatibility between the two species. I identified loci with exceptional patterns of introgression using two approaches described by Gompert and Buerkle, 2011. (1) If locus specific introgression differed from the genome-wide average which I will refer to as "excess ancestry" following Gompert and Buerkle, 2011. More specifically, I classified a locus as having excess ancestry if the 90% highest posterior density interval (HPDI) for the alpha or beta parameter did not cover zero. (2) If locus specific introgression is statistically unlikely relative to the genome-wide distribution of locus specific introgression which I will refer to as "outliers" following Gompert and Buerkle, 2011. I classified a locus as an outlier if the median of the posterior sample for the α or β parameters for a locus were not contained 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 conditional random effect priors on τ_α and τ_β. These conditional priors describe the genome-wide variation of locus specific α and β. I further classified outlier

α parameter estimates for a locus based on whether the median of the posterior sample was positive or negative. Positive estimates of α mean there is a greater probability of *A. americanus* ancestry in individuals at the locus relative to their hybrid index whereas negative estimates of α mean there is a greater probability of *A. terrestris* ancestry.

2.2.6 Genetic differentiation and Introgression

To test for a relationship between patterns of introgression and genetic divergence I used *VCFtools* to calculate the Weir and Cockerham, 1984 F_{ST} between each species using only the samples inferred through the *STRUCTURE* analysis to have >95% ancestry for one species under the model with a K of two (Danecek et al., 2011). The Weir and Cockerham F_{ST} is calculated per site and I calculated the per site F_{ST} for the same sites as those used in the *BGC* analysis. To determine if patterns of introgression are correlated with population differentiation a performed a Pearson Correlation to test if F_{ST} correlates with either the α or β parameters. I ran the correlation test with the absolute value of the median of the posterior sample for the α parameter and the median of the posterior sample for the β parameter. I binned loci based on their status as α parameter outliers. I categorized loci as being α outliers greater expected *A. americanus* ancestry, α outliers with greater expected *A. terrestris* ancestyr, and estimates of α that are not outliers. I performed a Kruskal-Wallis test using *SciPy* v1.10.1 to test whether there were significant differences in values of F_{ST} at each locus between these groups (Virtanen et al., 2020). I then performed Mann-Whitney tests between all pairs of groups using *scikit-posthocs* to test which groups differ significantly from each other github.com/maximtrp/scikit-posthocs.

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)). *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 from 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 and 43 samples were excluded from further analyses leaving a total of 149. For the included samples, 56 had been identified as most closely resembling *A. americanus* and 93 had been identified as most closely resembling *A. terrestris*.

2.3.2 Genetic Clustering & Ancestry Proportions

A visual inspection of the *STRUCTURE* results shows that each iteration with same value for K converged on very similar results (Fig. 2.2). The *STRUCTURE* model with the largest ΔK was the model with a K of two Fig. 2.1. Furthermore, individuals are inferred as having ancestry derived largely from only two ancestral groups even for K values of three and four. For these values of K, only a small amount of ancestry is attributed to the third or fourth ancestral groups for any individual sample. Fig. 2.3 Using a 95% estimated ancestry proportion as a cutoff for considering individuals to have

pure ancestry, 36 samples were classified as pure *A. americanus*, 75 as pure *A. terrestris*, and 38 as being admixed. The proportions of admixture among the samples shows a clear gradient between 0 and 1 which is consistent with many individuals being the product of advanced generation hybrids beyond the F_1 generation. The transition of admixture proportions from one species to the other increase with distance from the locations of pure individuals with proportions closest to 0.5 being found in the center of this transition Fig. 2.4.

2.3.3 Patterns of Introgression

Visualization of the MCMC output with trace plots and histograms of each parameter indicated that each of the five chains run in *BGC* converged on the same parameter space and that each chain quickly reached stationarity. I conservatively discarded the first 10% of samples as burnin. The median of the posterior sample for α ranged from -0.525-0.494. The β parameter was less variable and ranged from-0.158-0.220. I identified 16 loci with excess ancestry for the α parameter relative to the genome wide average; i.e., the 90% HDPI does not cover 0. Of these, the median of the posterior sample for 5 of these loci was negative and for 11 loci was positive. Negative values represent a greater probability of *A. americanus* ancestry at a locus relative to the hybrid index whereas positive values represent a greater probability of *A. terrestris* ancestry. I did not identify any loci for which the estimates of β were outliers relative to the genome-wide average. I identified 116 loci as outliers for the α parameter relative to the genome-wide distribution of locus specific introgression. Of these, the median of the posterior sample for 24 of these loci was negative and for 92 loci was positive Fig. 2.5. I did not identify any loci for which the estimates of β were outliers relative to the genome-wide distribution of locus specific introgression. All 16 of the loci identified as having excess ancestry for the α parameter relative to the genome-wide average were also identified as outliers relative to the genome-wide distribution of locus specific introgression.

2.3.4 Genomic Differentiation

Genetic differentiation between *A. americanus* and *A. terrestris* was highly variable Fig. 2.6. Locus-specific F_{ST} between non-admixed *A. americanus* and *A. terrestris* had a mean of 0.07. F_{ST} values for 249 loci were 0. Only a single locus had fixed differences between species with an F_{ST} of 1.0. There is no obvious relationship between α and F_{ST} when looking at the plotted data Fig. 2.7. However, higher α estimates had non-zero values for F_{ST}. The Pearson correlation test estimates a weak correlation between α and F_{ST} (r=0.29, p=1.62e-23). There is a more apparent relationship between F_{ST} and β when looking at plotted data Fig. 2.7 but a Pearson correlation test estimates only a slightly greater correlation between F_{ST} and β, albeit with a much smaller p-value (r=0.32, $p = 8.28 \times 10^{-30}$). The result of the Kruskal-Wallis test are consistent with there being significant differences between the F_{ST} values of loci with outlier α estimates and non-outlier α estimates on average ($p = 1.32 \times 10^{-40}$) Fig. 2.6. The results of the post hoc pairwise Mann-Whitney tests are consistent with both categories of loci with outlier α estimates having great F_{ST} values on average than the non-outlier estimates of α. The difference between non-outlier loci and loci with greater probability of *A. americanus* ancestry was slightly higher ($p = 2.72 \times 10^{-38}$) than the difference between non-outlier loci and loci with greater *A. terrestris* ancestry ($p = 8.16 \times 10^{-6}$).

2.4 Discussion

In this study, I investigated the patterns of genetic variation across putative a hybrid zone between two species of toads using genome wide sequence data collected from samples distributed across the hybrid zone.

2.4.1 Evidence for ongoing hybridization

Even in the absence of the genetic data presented in this study, the contact zone of *A. americanus* and *A. terrestris* seems to bear the hallmarks of a "tension zone" (Barton & Hewitt, 1985). A tension zones is a hybrid zone model which is characterized by abrupt genotypic clines maintained by a balance between dispersal and selection against certain hybrid genotypes (Barton & Hewitt, 1985). Tension zones are expected to correspond with natural features that reduce dispersal or density of the parent species. The ranges of *A. americanus* and *A. terrestris* abut with an abrupt transition and no apparent overlap. Furthermore, the boundary between these species forms a long, smooth arc from Louisianna to Virginia with a position closely corresponding to the fall line, a prominent physiographic transition between the Piedmont Plateau in the North and the southern coastal plain in the South Fig. 2.4. Such a sudden transition is difficult to explain except that it is the result of the processes characteristic of tension zones. For there to be no mutually hospitable areas permitting some range overlap is implausible without there being an extreme level of competition or extreme degree of adaptation by each species to their respective environments. The similarity of male advertisement calls, overlap in spawning period, and laboratory crossing experiments demonstrating reproductive compatibility suggest barriers to hybridization are weak. Previous analysis of morphological variation by Weatherby, 1982 showing the presence of morphological intermediates across a large region of central Alabama suggested that introgression is occurring.

With the genome-wide sequence data obtained in this study, I find evidence of substantial gene flow across the hybrid zone of these two species. The *STRUCTURE* analysis inferred 38 out of 149 samples as having a proportion of ancestry of at least 5% of sites attributable to admixture Fig. 2.4. The admixture coefficients inferred in the *STRUCTURE* analysis span the entire range from 0.05%-0.5% which would be expected if hybrids are viable, fertile, and capable of backcrossing over multiple generations **[CITATION NEEDED!]** Fig. 2.4. When backcrossing occurs over multiple generations in combination with migration of hybrid progeny and selection against introgressing alleles, a cline will form across the hybrid zone with introgressing alleles becoming more uncommon with distance from the cline center (Barton & Hewitt, 1985). The results of the *STRUCTURE* analysis are largely consistent with this. Inferred admixture coefficients decrease and approach zero with distance from the center of the hybrid zone Fig. 2.4. The results of the *STRUCTURE* analysis indicate that the width of the hybrid zone is quite wide. At least as wide as the sampling area as there are samples with appreciable admixture proportions near the Northeastern and Southwestern edges of the sampling area Fig. 2.4.

The width of a hybrid zone is largely determined by selection and dispersal.

The reason for a low level of admixture observed even in very distance samples could be due to adaptive introgression.

It is also possible that some of this inferred admixture is the result of a statistical artifact or due to error. Some reassurance is provided by the result of the PCA which

is consistent with the *STRUCTURE* results although it is possible that they could be affected by the same bias or error introduced at by data collection and processing Fig. 2.4 **[CITATION NEEDED!]**.

2.4.2 Variability of introgression

Used genomic cline instead of geographic cline. Theory and evidence point to a balance between migration and selection as the cause for the creation and persistence of hybrid zones . Organisms migrate and move into one another's ranges.

Hybrid zones are thought to persist because of a balance between selection and migration. Incompatible combinations of alleles arise through the process of recombination happening from backcrossing (Barton & Hewitt, 1985; Orr, 1996)

A reference genome for these species would greatly enhance this analysis Sampled from what might be considered multiple populations which could be a problem Treated a fairly broad area as a single population.

2.4.3 Relationship between introgression and differentiation

Imperfect association between divergence and introgression is consistent with secondary contact. If speciation happened with gene flow, a tight linkage between the two would be expected. Ancestry inference is dependent on allele frequency differences (Gompert et al., 2017) Association between divergence and introgression is hard to explain from neutral processes and suggests that highly differentiated loci affect hybrid fitness depending on the genomic background (Gompert et al., 2017)

2.4.4 Conclusion

References

AmphibiaWeb. (2023).

Barton, N. H., & Hewitt, G. M. (1985). Analysis of Hybrid Zones. *Annual Review*, *16*, 113–148.

Bayona-Vásquez, N. J., Glenn, T. C., Kieran, T. J., Pierson, T. W., Hoffberg, S. L., Scott, P. A., Bentley, K. E., Finger, J. W., Louha, S., Troendle, N., Diaz-Jaimes, P., Mauricio, R., & Faircloth, B. C. (2019). Adapterama III: Quadruple-indexed, double/triple-enzyme RADseq libraries (2RAD/3RAD). *PeerJ*, *7*, e7724. https://doi.org/10.7717/peerj.7724

Blackman, B. (2016). Speciation Genes. *Encyclopedia of Evolutionary Biology* (pp. 166–175). Elsevier. https://doi.org/10.1016/B978-0-12-800049-6.00066-4

Blair, W. F. (1963). Intragroup genetic compatibility in the Bufo americanus species group of toads. *The Texas Journal of Science*, *13*, 15–34.

Blair, W. F. (1972). *Evolution in the genus Bufo*. University of Texas Press.

Blair, W. F. (1974). Character Displacement in Frogs. *American Zoologist*, *14*(4), 1119–1125. https://doi.org/10.1093/icb/14.4.1119

Butlin, R., Debelle, A., Kerth, C., Snook, R. R., Beukeboom, L. W., RF, C. C., Diao, W., Maan, M. E., Paolucci, S., Weissing, F. J., et al. (2011). What do we need to know about speciation? *Trends in ecology & evolution*, *27*(1), 27–39.

Cocroft, R. B., & Ryan, M. J. (1995). Patterns of advertisement call evolution in toads and chorus frogs. *Animal Behaviour*, *49*(2), 283–303. https://doi.org/10.1006/anbe.1995.0043

Colliard, C., Sicilia, A., Turrisi, G. F., Arculeo, M., Perrin, N., & Stöck, M. (2010). Strong reproductive barriers in a narrow hybrid zone of West-Mediterranean green toads (Bufo viridissubgroup) with Plio-Pleistocene divergence. *BMC Evolutionary Biology*, *10*(1), 232. https://doi.org/10.1186/1471-2148-10-232

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., Durbin, R., & 1000 Genomes Project Analysis Group. (2011). The variant call format and VCFtools. *Bioinformatics*, *27*(15), 2156–2158. https://doi.org/10.1093/bioinformatics/btr330

Delph, L. F., & Demuth, J. P. (2016). Haldane's Rule: Genetic Bases and Their Empirical Support. *Journal of Heredity*, *107*(5), 383–391. https://doi.org/10.1093/jhered/esw026

Đorđević, S., & Simović, A. (2014). STRANGE AFFECTION: MALE BUFO BUFO (ANURA: BUFONIDAE) PASSIONATELY EMBRACING A BULGE OF MUD. *Ecologica Montenegrina*, *1*(1), 15–17. https://doi.org/10.37828/em.2014.1.4

Dufresnes, C., Litvinchuk, S. N., Rozenblut-Kościsty, B., Rodrigues, N., Perrin, N., Crochet, P.-A., & Jeffries, D. L. (2020). Hybridization and introgression between toads with different sex chromosome systems. *Evolution Letters*, *4*(5), 444–456. https://doi.org/10.1002/evl3.191

Evanno, G., Regnaut, S., & Goudet, J. (2005). Detecting the number of clusters of individuals using the software structure: A simulation study. *Molecular Ecology*, *14*(8), 2611–2620. https://doi.org/10.1111/j.1365-294X.2005.02553.x

Francis, R. M. (2017). POPHELPER: An R package and web app to analyse and visualize population structure. *Molecular Ecology Resources*, *17*(1), 27–32. https://doi.org/10.1111/1755-0998.12509

Gompert, Z., & Buerkle, C. A. (2012). Bgc: Software for Bayesian estimation of genomic clines. *Molecular Ecology Resources*, *12*(6), 1168–1176. https://doi.org/10.1111/1755-0998.12009.x

Gompert, Z., & Buerkle, C. A. (2011). Bayesian estimation of genomic clines: BAYESIAN GENOMIC CLINES. *Molecular Ecology*, *20*(10), 2111–2127. https://doi.org/10.1111/j.1365-294X.2011.05074.x

Gompert, Z., Lucas, L. K., Nice, C. C., Fordyce, J. A., Forister, M. L., & Buerkle, C. A. (2012). GENOMIC REGIONS WITH A HISTORY OF DIVERGENT SELECTION AFFECT FITNESS OF HYBRIDS BETWEEN TWO BUTTERFLY SPECIES: GENOMICS OF SPECIATION. *Evolution*, *66*(7), 2167–2181. https://doi.org/10.1111/j.1558-5646.2012.01587.x

Gompert, Z., Mandeville, E. G., & Buerkle, C. A. (2017). Analysis of Population Genomic Data from Hybrid Zones. *Annual Review of Ecology, Evolution, and Systematics*, *48*(1), 207–229. https://doi.org/10.1146/annurev-ecolsys-110316-022652

Green, D. M. (1983). Allozyme Variation through a Clinal Hybrid Zone between the Toads Bufo americanus and B. hemiophrys in Southeastern Manitoba. *Herpetologica*, *39*(1), 28–40.

Green, D. M. (1996). The bounds of species: Hybridization in the Bufo americanus group of North American toads. *Israel Journal of Zoology*, *42*, 95–109.

Hedrick, P. W. (2013). Adaptive introgression in animals: Examples and comparison to new mutation and standing variation as sources of adaptive variation. *Molecular Ecology*, *22*(18), 4606–4618. https://doi.org/10.1111/mec.12415

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

Jombart, T. (2008). Adegenet : A R package for the multivariate analysis of genetic markers. *Bioinformatics*, *24*(11), 1403–1405. https://doi.org/10.1093/bioinformatics/btn129

Kennedy, J. P. (1962). Spawning Season and Experimental Hybridization of the Houston Toad, Bufo houstonensis. *Herpetologica*, *17*(4), 239–245.

Larson, E. L., Andrés, J. A., Bogdanowicz, S. M., & Harrison, R. G. (2013). DIFFERENTIAL INTROGRESSION IN A MOSAIC HYBRID ZONE REVEALS CANDIDATE BARRIER GENES. *Evolution*, *67*(12), 3653–3661. https://doi.org/10.1111/evo.12205

Mallet, J. (2005). Hybridization as an invasion of the genome. *Trends in Ecology & Evolution*, *20*(5), 229–237. https://doi.org/10.1016/j.tree.2005.02.010

Mallet, J. (2008). Hybridization, ecological races and the nature of species: Empirical evidence for the ease of speciation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *363*(1506), 2971–2986. https://doi.org/10.1098/rstb.2008.0081

Malone, J. H., & Fontenot, B. E. (2008). Patterns of Reproductive Isolation in Toads (R. DeSalle, Ed.). *PLoS ONE*, *3*(12), e3900. https://doi.org/10.1371/journal.pone.0003900

Miller, C. J. J., & Matute, D. R. (2016). The Effect of Temperature on Drosophila Hybrid Fitness. *G3: Genes/Genomes/Genetics*, *7*(2), 377–385. https://doi.org/10.1534/g3.116.034926

Moran, B. M., Payne, C., Langdon, Q., Powell, D. L., Brandvain, Y., & Schumer, M. (2021). The genomic consequences of hybridization (P. J. Wittkopp, Ed.). *eLife*, *10*, e69016. https://doi.org/10.7554/eLife.69016

Mount, R. H. (1975). *The Reptiles and Amphibians of Alabama*. The University of Alabama Press.

Nikolakis, Z. L., Schield, D. R., Westfall, A. K., Perry, B. W., Ivey, K. N., Orton, R. W., Hales, N. R., Adams, R. H., Meik, J. M., Parker, J. M., Smith, C. F., Gompert, Z., Mackessy, S. P., & Castoe, T. A. (2022). Evidence that genomic incompatibilities and other multilocus processes impact hybrid fitness in a rattlesnake hybrid zone. *Evolution*, *76*(11), 2513–2530. https://doi.org/10.1111/evo.14612

Nosil, P., Funk, D. J., & Ortiz-Barrientos, D. (2009). Divergent selection and heterogeneous genomic divergence. *Molecular Ecology*, *18*(3), 375–402. https://doi.org/10.1111/j.1365-294X.2008.03946.x

Nosil, P., & Schluter, D. (2011). The genes underlying the process of speciation. *Trends in Ecology & Evolution*, *26*(4), 160–167. https://doi.org/10.1016/j.tree.2011.01.001

Orr, H. A. (1996). Dobzhansky, Bateson, and the Genetics of Speciation. *Genetics*, *144*(4), 1331–1335.

Parchman, T. L., Gompert, Z., Braun, M. J., Brumfield, R. T., McDonald, D. B., Uy, J. a. C., Zhang, G., Jarvis, E. D., Schlinger, B. A., & Buerkle, C. A. (2013). The genomic consequences of adaptive divergence and reproductive isolation between species of manakins. *Molecular Ecology*, *22*(12), 3304–3317. https://doi.org/10.1111/mec.12201

Powell, R., Conant, R., & Collins, J. T. (2016). *A Field Guide to Reptiles & Amphibians: Eastern and Central North America* (4th ed.). Houghton Mifflin Harcourt.

Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of Population Structure Using Multilocus Genotype Data. *Genetics*, *155*(2), 945–959. https://doi.org/10.1093/genetics/155.2.945

Rieseberg, L. H., Whitton, J., & Gardner, K. (1999). Hybrid zones and the genetic architecture of a barrier to gene flow between two sunflower species. *Genetics*, *152*(2), 713–727.

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

Servedio, M. R., & Noor, M. A. (2003). The Role of Reinforcement in Speciation: Theory and Data. *Annual Review of Ecology, Evolution, and Systematics*, *34*(1), 339–364. https://doi.org/10.1146/annurev.ecolsys.34.011802.132412

Shankman, D., & Hart, J. L. (2007). The Fall Line: A Physiographic-Forest Vegetation Boundary. *Geographical Review*, *97*(4), 502–519. https://doi.org/10.1111/j.1931-0846.2007.tb00409.x

Stöck, M., Croll, D., Dumas, Z., Biollay, S., Wang, J., & Perrin, N. (2011). A cryptic heterogametic transition revealed by sex-linked DNA markers in Palearctic green toads: Heterogametic transition in Bufonidae. *Journal of Evolutionary Biology*, *24*(5), 1064–1070. https://doi.org/10.1111/j.1420-9101.2011.02239.x

Taylor, E. B., Boughman, J. W., Groenenboom, M., Sniatynski, M., Schluter, D., & Gow, J. L. (2006). Speciation in reverse: Morphological and genetic evidence of the collapse of a three-spined stickleback (Gasterosteus aculeatus) species pair. *Molecular Ecology*, *15*(2), 343–355. https://doi.org/10.1111/j.1365-294X.2005.02794.x

Trudeau, V. L., Somoza, G. M., Natale, G. S., Pauli, B., Wignall, J., Jackman, P., Doe, K., & Schueler, F. W. (2010). Hormonal induction of spawning in 4 species of frogs by coinjection with a gonadotropin-releasing hormone agonist and a dopamine antagonist. *Reproductive Biology and Endocrinology*, *8*(1), 36. https://doi.org/10.1186/1477-7827-8-36

Van Riemsdijk, I., Arntzen, J. W., Bucciarelli, G. M., McCartney-Melstad, E., Rafajlović, M., Scott, P. A., Toffelmier, E., Shaffer, H. B., & Wielstra, B. (2023). Two transects reveal remarkable variation in gene flow on opposite ends of a European toad hybrid zone. *Heredity*. https://doi.org/10.1038/s41437-023-00617-6

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., … SciPy 1.0 Contributors. (2020). SciPy 1.0: Fundamental algorithms for scientific computing in python. *Nature Methods*, *17*, 261–272. https://doi.org/10.1038/s41592-019-0686-2

Weatherby, C. A. (1982). INTROGRESSION BETWEEN THE AMERICAN TOAD, BUFO AMERICANUS, AND THE SOUTHERN TOAD, B. TERRESTRIS, IN ALABAMA.

Weir, B. S., & Cockerham, C. C. (1984). Estimating F-Statistics for the Analysis of Population Structure. *Evolution*, *38*(6), 1358–1370. https://doi.org/10.2307/2408641

Wolf, J. B. W., & Ellegren, H. (2017). Making sense of genomic islands of differentiation in light of speciation. *Nature Reviews Genetics*, *18*(2), 87–100. https://doi.org/10.1038/nrg.2016.133

Wu, C.-I. (2001). The genic view of the process of speciation. *Journal of Evolutionary Biology*, *14*(6), 851–865. https://doi.org/10.1046/j.1420-9101.2001.00335.x

2.5 Figures

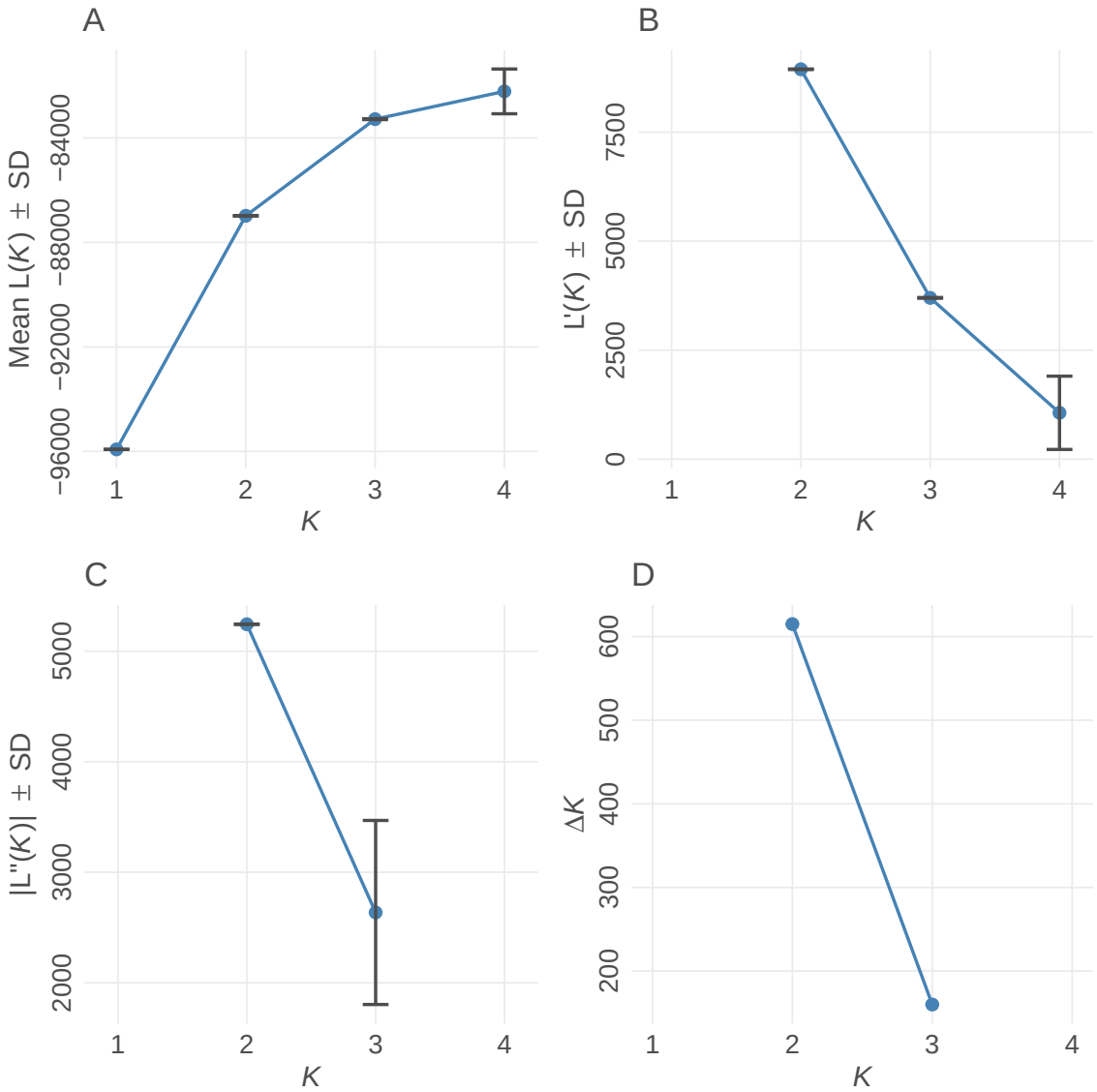

Figure 2.1. Evanno method for optimal value for K in *STRUCTURE* (Evanno et al., 2005). K refers to the number of populations for each of the different *STRUCTURE* models examined. (A) Mean estimated $\ln$ probability of data over 10 iterations for each value of $K \pm SD$. (B) Rate of change of the 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).

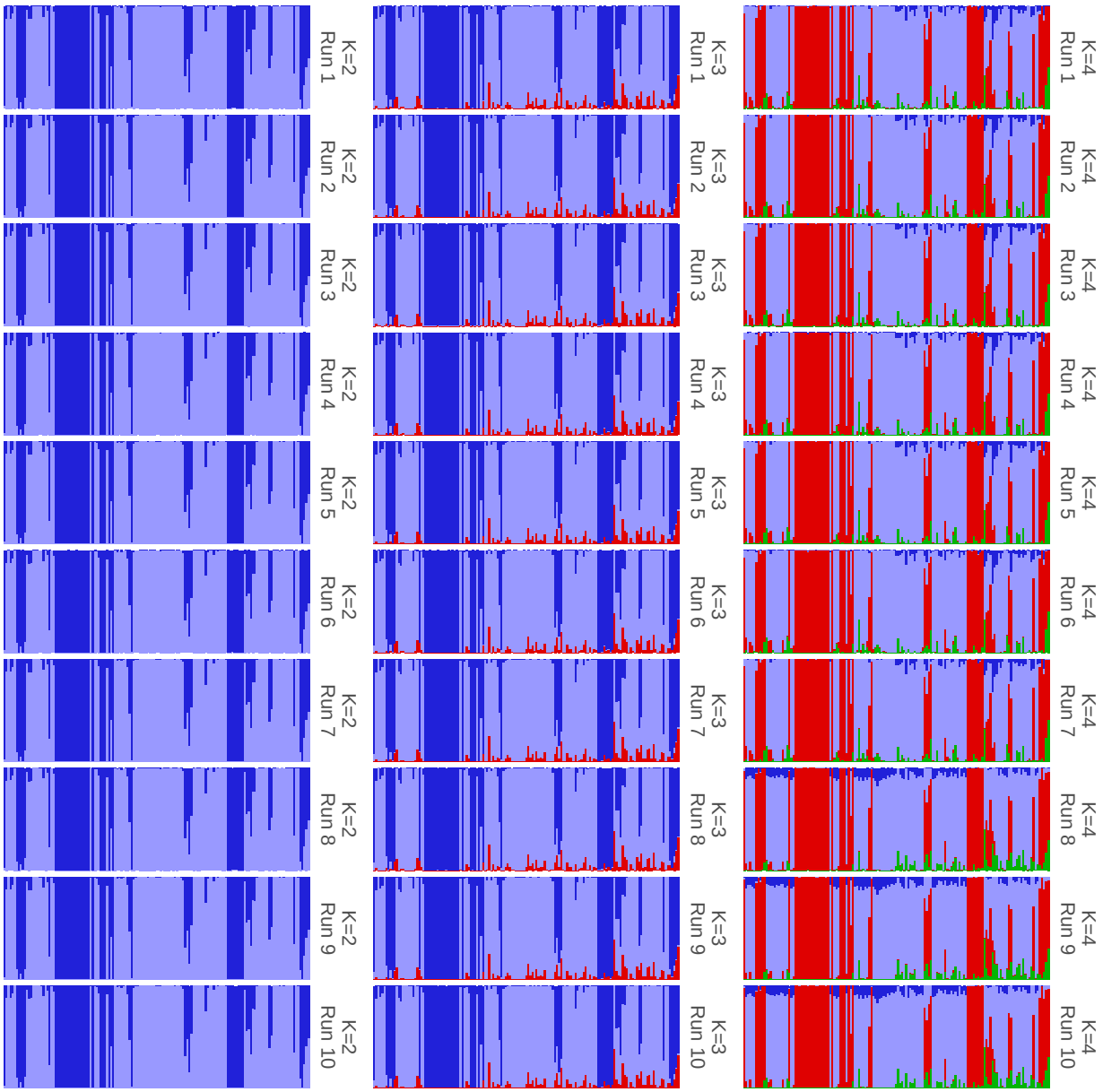

Figure 2.2. Results of each iteration of *STRUCTURE* showing convergence among iterations within runs having the same value for K. Plot was created with *POPHELPER* (Francis, 2017).

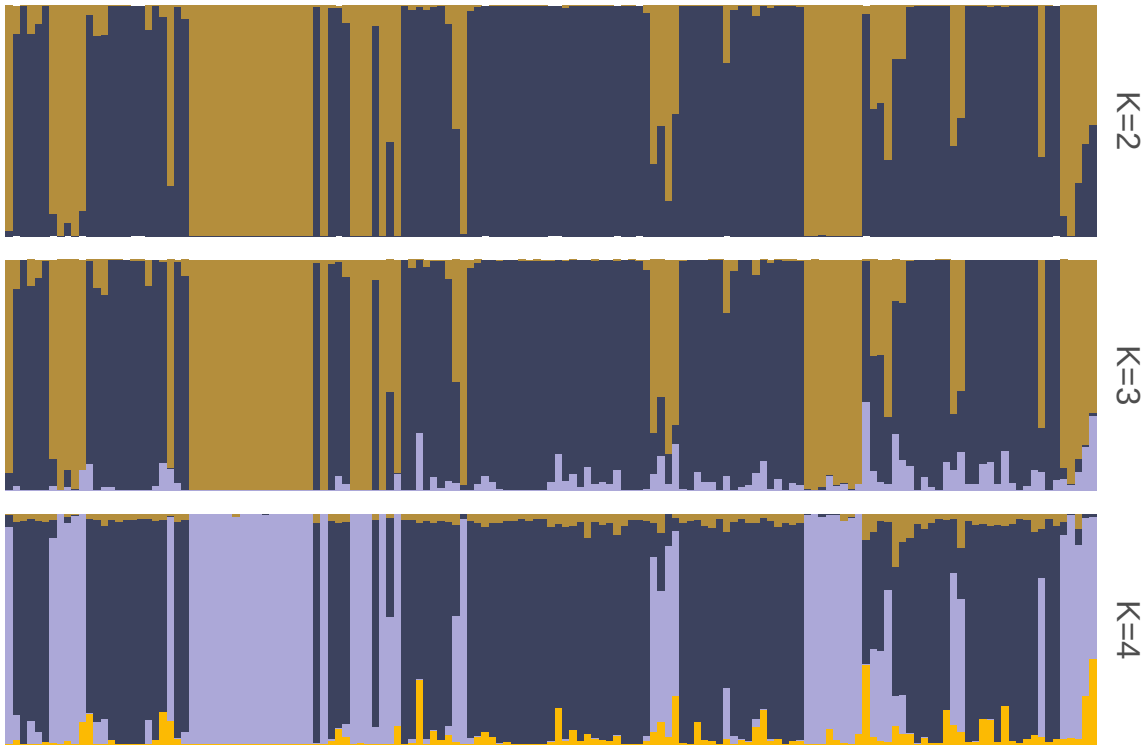

Figure 2.3. Summarized *STRUCTURE* results for each value of K. Ancestry proportions shown are the mean of ancestry proportions across all iterations. Summarization and plotting done using *POPHELPER* (Francis, 2017).

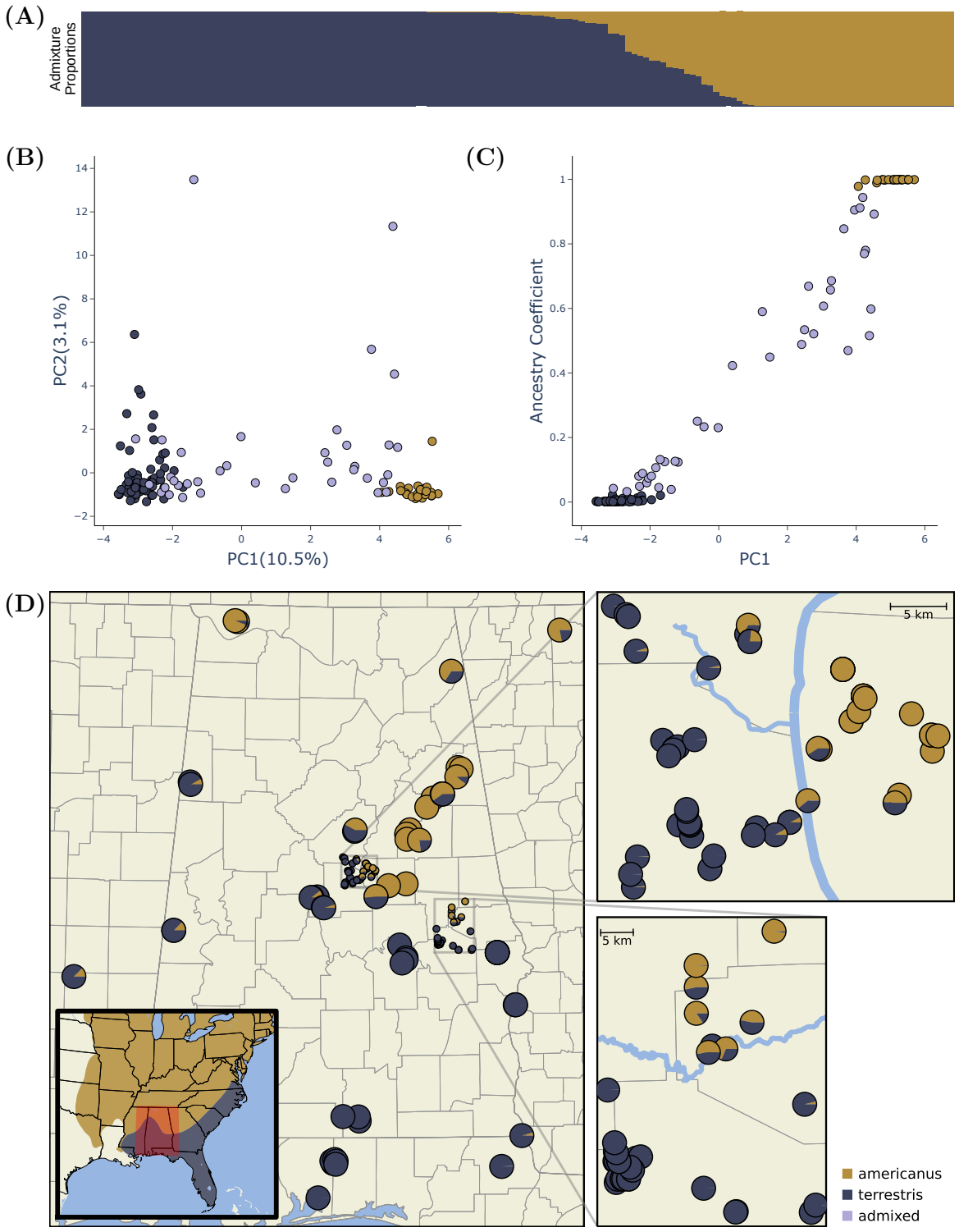

Figure 2.4. Genetic evidence of hybridization between *A. americanus* and *A. terrestris*. (A) *STRUCTURE* plot showing estimated ancestry proportions. (B) Summary of poulation genetic structure based on the principle component axes one (PC1) and two (PC2). These axes explain 10.5% (PC1) and 3.1% (PC2) 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)

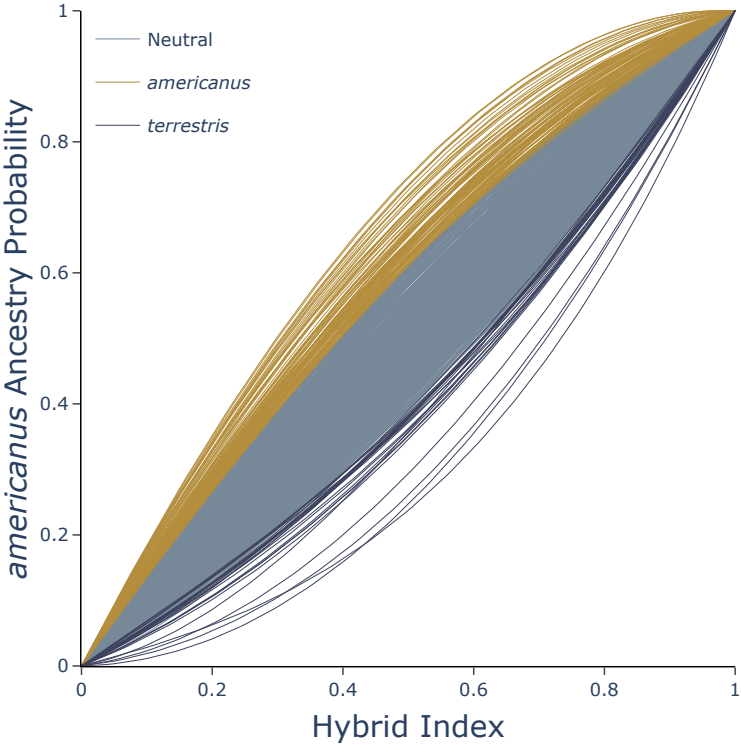

Figure 2.5. Shape of genomic clines estimated for each locus with *BGC*. Outliers are highlighted with XX.

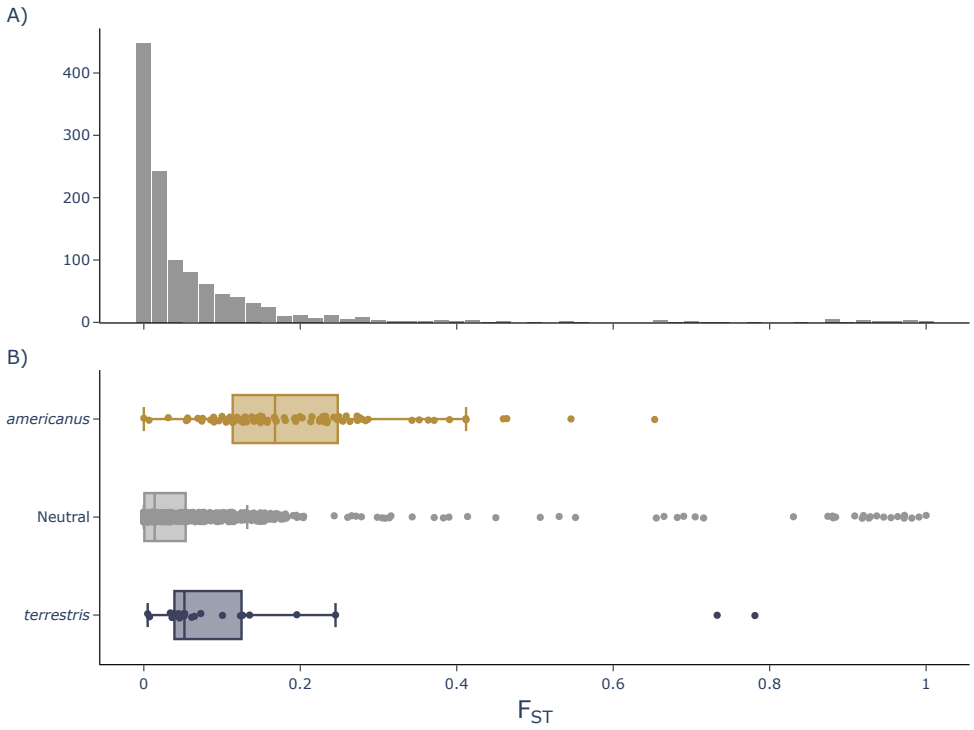

Figure 2.6. Relationship between genetic divergence and introgression. Write more stuff . . .

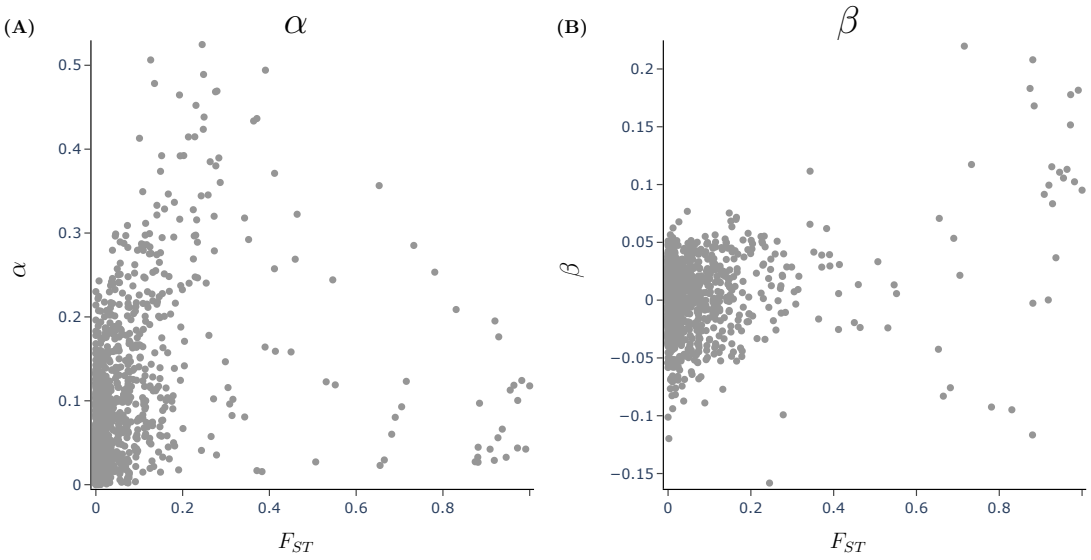

Figure 2.7. Relationship between genetic divergence and introgression. Write more stuff ...

2.6 Tables

Table 2.1. Samples collected for this study

Sample ID	Species	Latitude	Longitude	Passed Filtering
KAC 016	*terrestris*	30.54819	-86.93067	X
KAC 038	*terrestris*	32.81470	-86.93968	X
KAC 039	*terrestris*	32.81094	-86.98967	X
KAC 040	*terrestris*	32.80985	-86.99795	X
KAC 042	*terrestris*	32.82406	-86.99314	
KAC 043	*terrestris*	32.82406	-86.99314	
KAC 044	*terrestris*	32.80450	-87.03078	
KAC 045	*terrestris*	32.76703	-87.07073	
KAC 046	*terrestris*	32.76592	-87.07184	
KAC 047	*terrestris*	32.78932	-86.90850	
KAC 048	*terrestris*	32.73575	-86.88149	X
KAC 049	*terrestris*	32.73291	-86.87707	X
KAC 050	*terrestris*	32.74822	-86.79806	
KAC 051	*terrestris*	32.78742	-86.75847	
KAC 052	*terrestris*	32.78044	-86.73877	
KAC 070	*americanus*	34.79963	-84.57678	X
KAC 071	*terrestris*	32.43478	-85.64630	
KAC 074	*terrestris*	30.77430	-85.22690	X
KAC 075	*terrestris*	32.94778	-86.63224	X
KAC 076	*terrestris*	32.94970	-86.52687	
KAC 077	*terrestris*	32.94970	-86.52687	
KAC 078	*americanus*	33.00267	-86.38960	X
KAC 079	*americanus*	33.01205	-86.47872	
KAC 080	*americanus*	33.04456	-86.45547	
KAC 081	*americanus*	33.04456	-86.45547	X
KAC 082	*americanus*	33.04456	-86.45547	X
KAC 083	*americanus*	33.04456	-86.45547	X
KAC 084	*americanus*	33.04456	-86.45547	X
KAC 085	*americanus*	33.04456	-86.45547	
KAC 086	*americanus*	33.04456	-86.45547	X
KAC 087	*americanus*	33.01484	-86.39040	X
KAC 089	*americanus*	33.01484	-86.39040	X
KAC 090	*americanus*	33.06472	-86.47496	X
KAC 091	*americanus*	33.06472	-86.47496	X
KAC 092	*americanus*	33.06472	-86.47496	
KAC 093	*americanus*	33.06472	-86.47496	X
KAC 094	*americanus*	33.06472	-86.47496	X
KAC 095	*americanus*	33.06472	-86.47496	X
KAC 096	*americanus*	33.06472	-86.47496	X
KAC 097	*americanus*	33.06472	-86.47496	X
KAC 098	*americanus*	33.02572	-86.46711	X
KAC 099	*americanus*	33.02572	-86.46711	X
KAC 100	*terrestris*	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	*americanus*	33.03283	-86.45975	X
KAC 102	*terrestris*	32.94544	-86.55777	X
KAC 103	*terrestris*	32.94947	-86.52630	X
KAC 104	*terrestris*	32.94947	-86.52630	X
KAC 105	*americanus*	33.04278	-86.45377	X
KAC 106	*americanus*	33.00464	-86.49692	X
KAC 107	*americanus*	33.01416	-86.38417	X
KAC 108	*terrestris*	32.94013	-86.54004	X
KAC 109	*terrestris*	32.94173	-86.55787	
KAC 110	*americanus*	33.03099	-86.40941	X
KAC 111	*americanus*	33.00518	-86.49895	X
KAC 112	*terrestris*	32.95011	-86.53723	
KAC 113	*americanus*	33.00528	-86.38897	
KAC 114	*americanus*	33.01617	-86.40318	
KAC 115	*americanus*	32.98218	-86.40488	
KAC 116	*americanus*	32.96964	-86.42137	X
KAC 117	*terrestris*	32.97146	-86.52901	
KAC 121	*terrestris*	32.44120	-85.65386	X
KAC 122	*terrestris*	32.85411	-86.76619	
KAC 123	*terrestris*	32.90084	-86.67587	X
KAC 124	*terrestris*	32.91060	-86.67850	X
KAC 125	*terrestris*	32.91715	-86.68208	
KAC 126	*terrestris*	32.92717	-86.67407	
KAC 127	*terrestris*	32.97159	-86.62516	
KAC 128	*terrestris*	33.00585	-86.63703	
KAC 129	*terrestris*	33.00797	-86.64210	
KAC 130	*terrestris*	33.00818	-86.64333	
KAC 131	*terrestris*	33.01508	-86.64937	
KAC 132	*terrestris*	33.02034	-86.66651	
KAC 133	*terrestris*	33.01163	-86.64759	X
KAC 134	*terrestris*	33.00537	-86.63652	X
KAC 135	*terrestris*	33.00644	-86.63368	X
KAC 136	*terrestris*	33.00673	-86.63316	X
KAC 138	*americanus*	32.70224	-85.66196	X
KAC 139	*americanus*	32.73042	-85.66173	X
KAC 140	*terrestris*	32.62553	-85.63684	X
KAC 141	*terrestris*	32.41032	-85.60107	X
KAC 142	*terrestris*	32.57011	-85.80888	X
KAC 143	*terrestris*	32.47773	-85.79824	X
KAC 144	*terrestris*	32.47707	-85.79577	X
KAC 145	*terrestris*	32.48128	-85.76354	X
KAC 146	*terrestris*	32.48291	-85.75622	X
KAC 147	*terrestris*	32.45001	-85.79652	X
KAC 148	*terrestris*	32.45420	-85.79408	X
KAC 149	*terrestris*	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	*terrestris*	32.45449	-85.78664	X
KAC 151	*terrestris*	32.45451	-85.78416	X
KAC 152	*terrestris*	32.45423	-85.77634	X
KAC 153	*terrestris*	32.45423	-85.77634	X
KAC 154	*terrestris*	32.46574	-85.76977	X
KAC 155	*terrestris*	32.46961	-85.77369	X
KAC 156	*terrestris*	32.47709	-85.79175	X
KAC 158	*terrestris*	32.47709	-85.79175	X
KAC 159	*terrestris*	32.49000	-85.79741	X
KAC 160	*terrestris*	32.40809	-85.47857	X
KAC 161	*terrestris*	32.41744	-85.47117	X
KAC 162	*terrestris*	32.35417	-86.09838	X
KAC 163	*terrestris*	32.33994	-86.09946	X
KAC 164	*terrestris*	32.31562	-86.13789	X
KAC 167	*terrestris*	33.06620	-86.60328	X
KAC 172	*americanus*	32.62171	-85.61467	X
KAC 173	*americanus*	32.61751	-85.64335	X
KAC 176	*americanus*	32.66836	-85.66233	X
KAC 177	*americanus*	32.65571	-85.57134	X
KAC 181	*terrestris*	32.38644	-85.23561	X
KAC 182	*terrestris*	32.38579	-85.23565	X
KAC 183	*terrestris*	32.38579	-85.23565	X
KAC 184	*terrestris*	32.38579	-85.23565	X
KAC 185	*terrestris*	32.38579	-85.23565	X
KAC 187	*americanus*	32.64548	-85.55135	
KAC 188	*terrestris*	32.40976	-85.60208	X
KAC 189	*terrestris*	33.09152	-86.56686	X
KAC 190	*terrestris*	33.11298	-86.69434	X
KAC 191	*terrestris*	33.10659	-86.68228	X
KAC 192	*terrestris*	33.10509	-86.68014	X
KAC 193	*terrestris*	33.07896	-86.67286	X
KAC 194	*terrestris*	32.93933	-86.62008	X
KAC 195	*terrestris*	32.94745	-86.62146	X
KAC 196	*terrestris*	32.94829	-86.62190	X
KAC 197	*terrestris*	32.94929	-86.62241	X
KAC 198	*terrestris*	32.95077	-86.62306	
KAC 199	*terrestris*	32.95794	-86.62477	X
KAC 200	*terrestris*	32.95940	-86.62489	X
KAC 205	*terrestris*	32.54852	-85.48692	X
KAC 206	*americanus*	33.30759	-86.58201	X
KAC 207	*americanus*	33.31685	-86.57596	X
KAC 208	*americanus*	33.09829	-86.56529	X
KAC 209	*terrestris*	33.08600	-86.56394	X
KAC 210	*terrestris*	33.08600	-86.56394	X
KAC 211	*terrestris*	33.01464	-86.60995	

Continued on next page

Table 2.1 – continued from previous page

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

Table 2.2. Samples loaned from museums

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

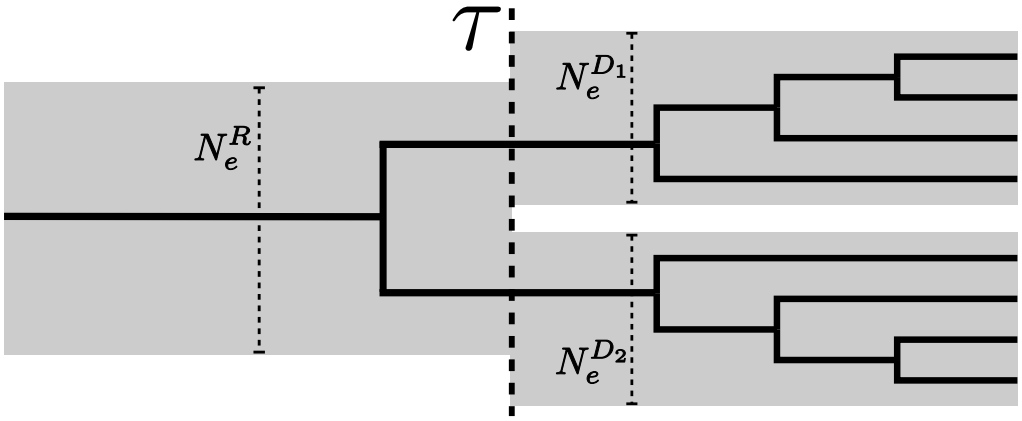

Figure 3.1. An illsustration of the species-tree model we used to simulate data. N_e^R , N_e^{D1} , and N_e^{D2} represent the constant effective population sizes of the root, and each of the two terminal populations. τ reperesents the instantaneous separation of the ancestral population into two descendant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

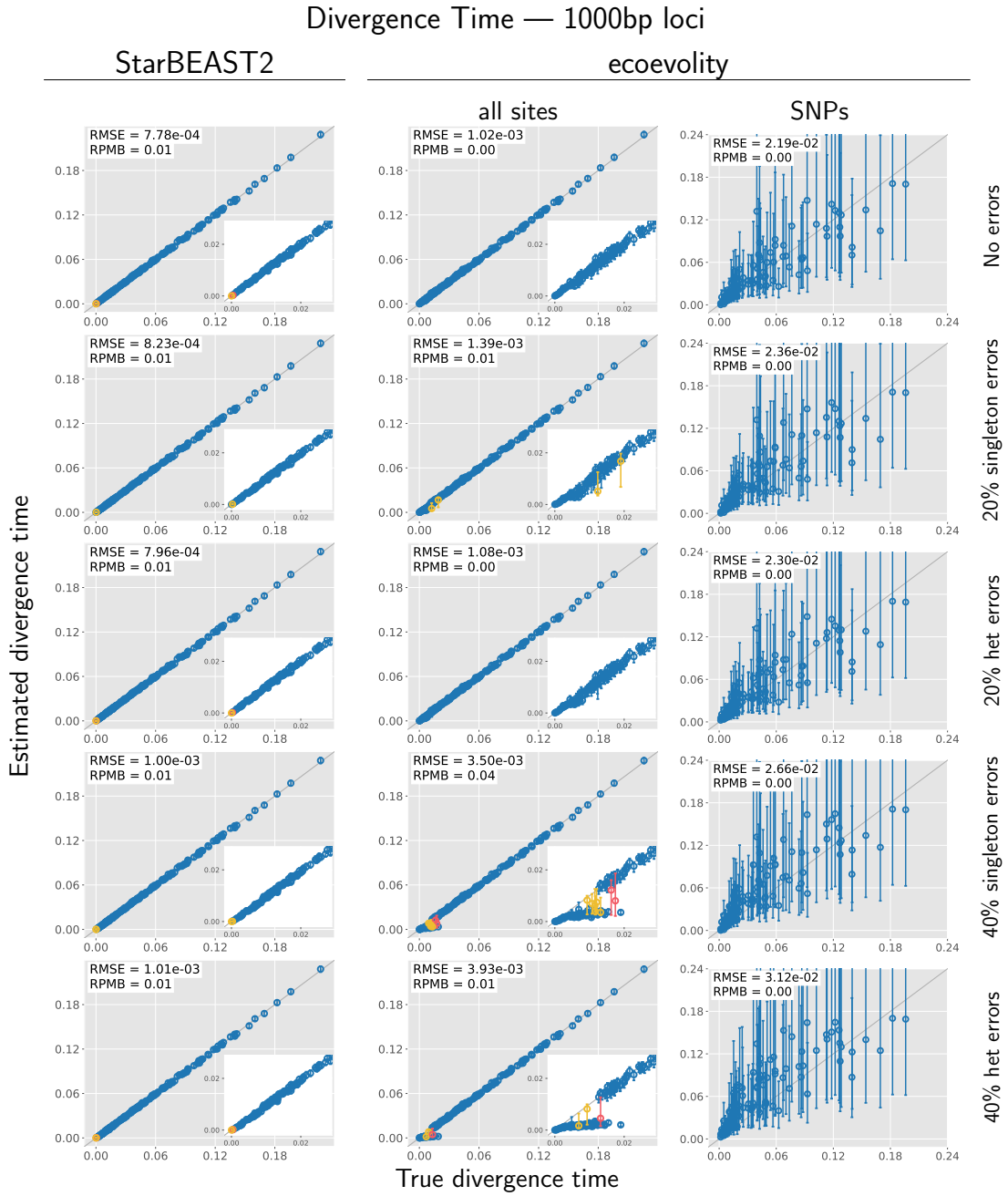

Figure 3.2. **Accuracy and precision of divergence-time estimates (in units of expected substitutions per site) with 1000 base pair loci.** The left column shows estimates from *StarBEAST2*, and the center and right column shows estimates from *ecoevolity* using all sites and (at most) one SNP per locus.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

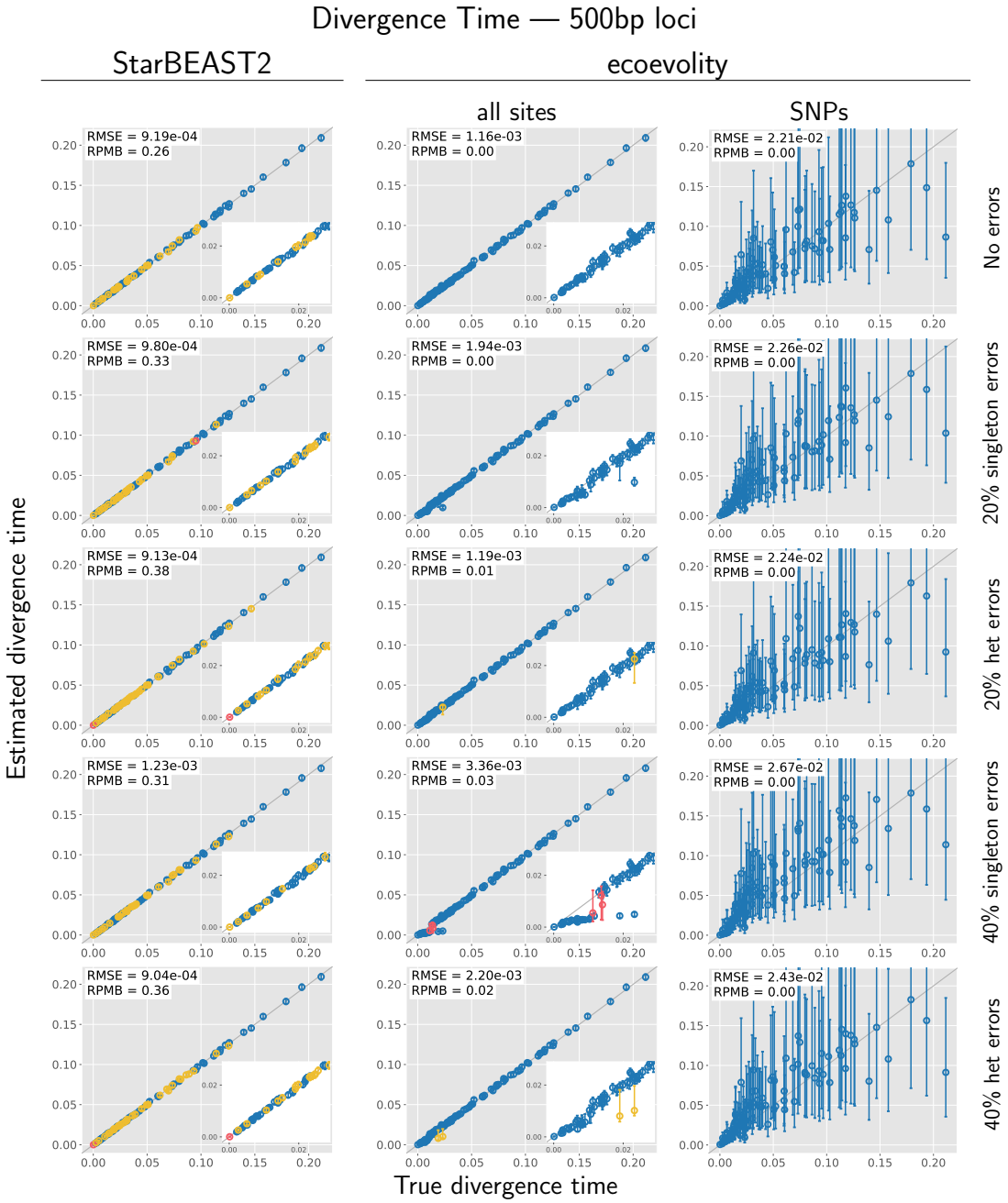

Figure 3.3. **Accuracy and precision of divergence-time estimates (in units of expected substitutions per site) with 500 base pair loci.** The left column shows estimates from *StarBEAST2*, and the center and right column shows estimates from *ecoevolity* using all sites and (at most) one SNP per locus.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

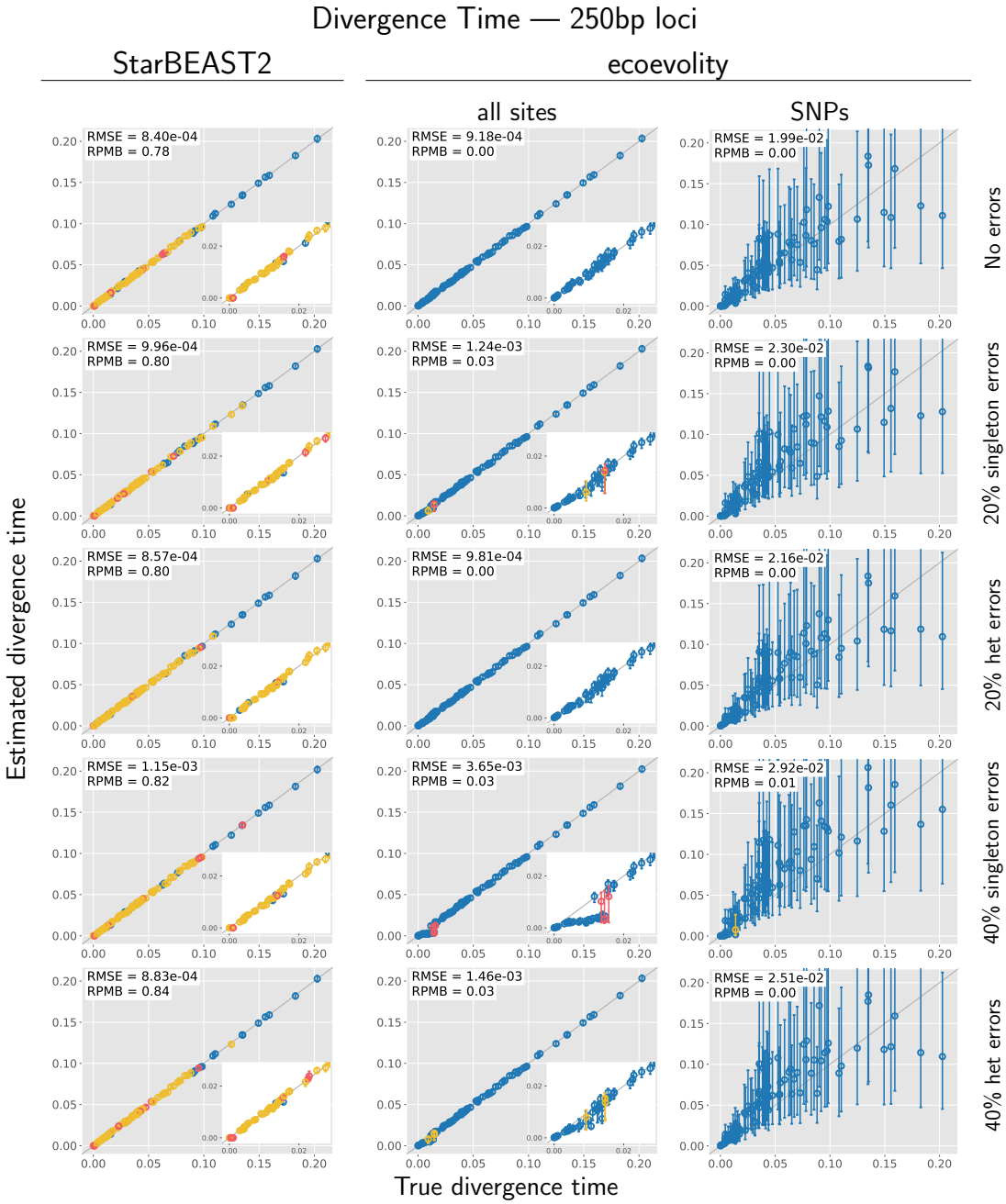

Figure 3.4. **Accuracy and precision of divergence-time estimates (in units of expected substitutions per site) with 250 base pair loci.** The left column shows estimates from *StarBEAST2*, and the center and right column shows estimates from *ecoevolity* using all sites and (at most) one SNP per locus.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

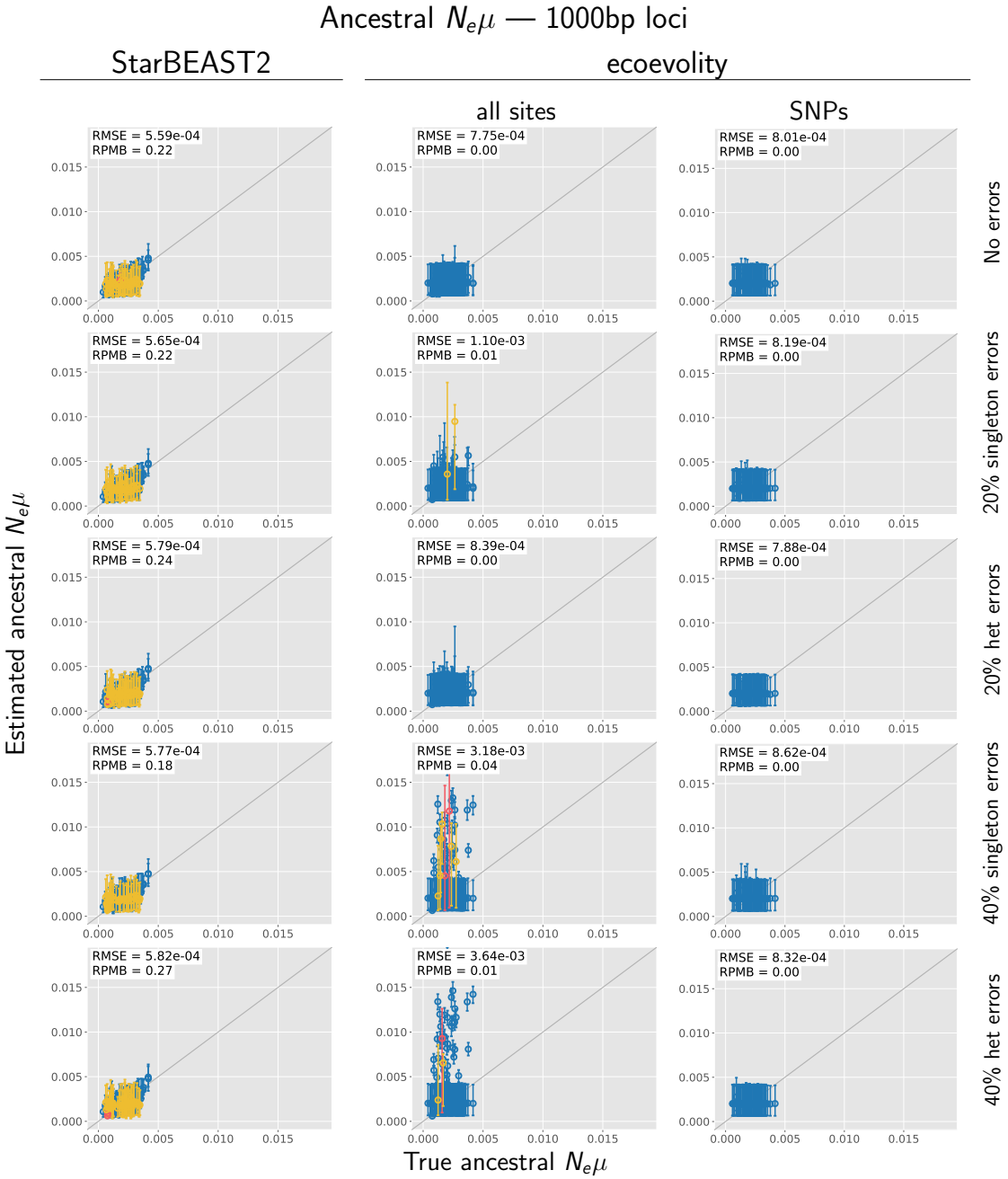

Figure 3.5. **Accuracy and precision of estimates of root effective population size scaled by the mutation rate ($N_e^R\mu$) with 1000 base pair loci.** The left column shows estimates from *StarBEAST2*, and the center and right column shows estimates from *ecoevolity* using all sites and (at most) one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

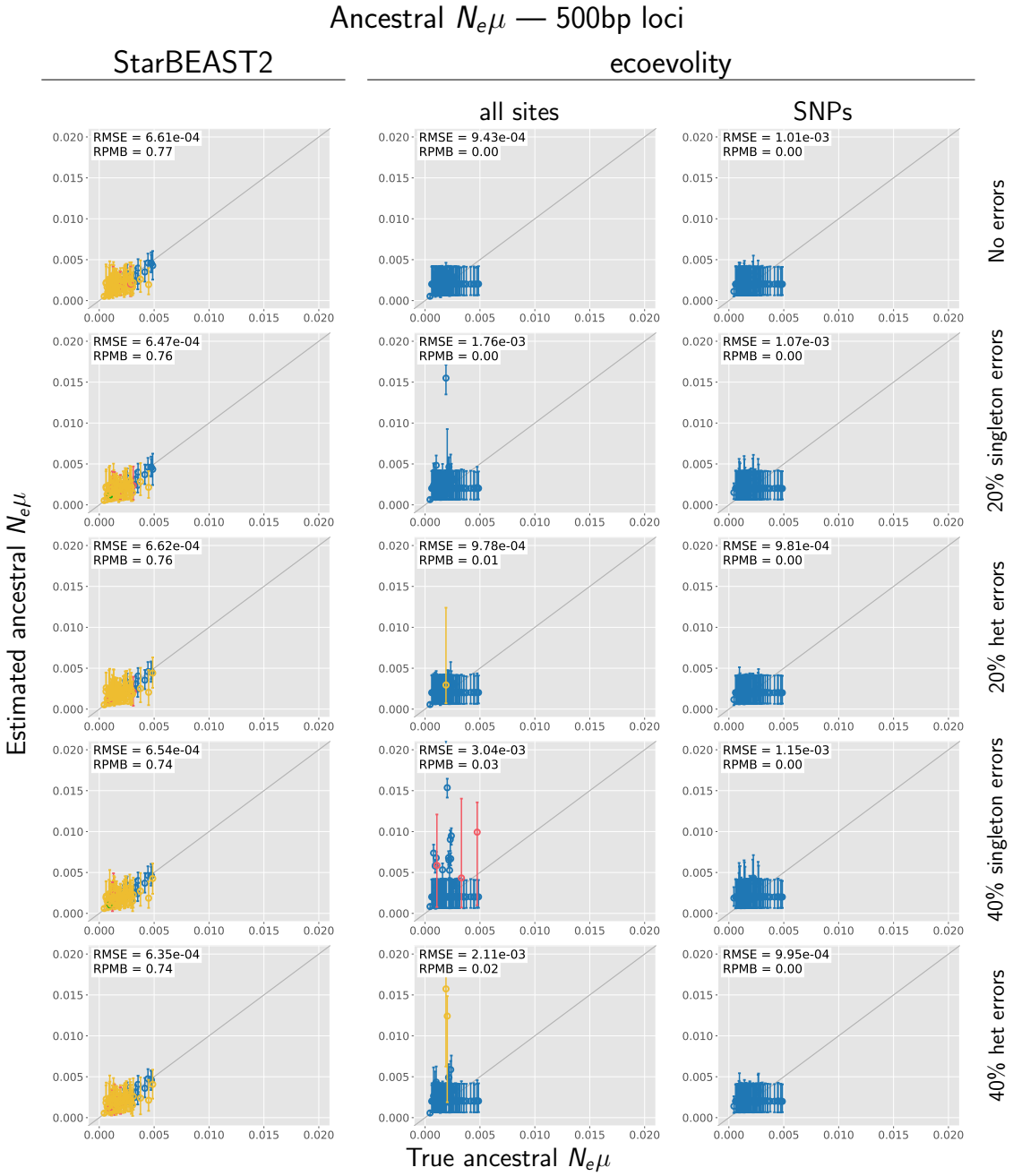

Figure 3.6. **Accuracy and precision of estimates of root effective population size scaled by the mutation rate ($N_e^R\mu$) with 500 base pair loci.** The left column shows estimates from *StarBEAST2*, and the center and right column shows estimates from *ecoevolity* using all sites and (at most) one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

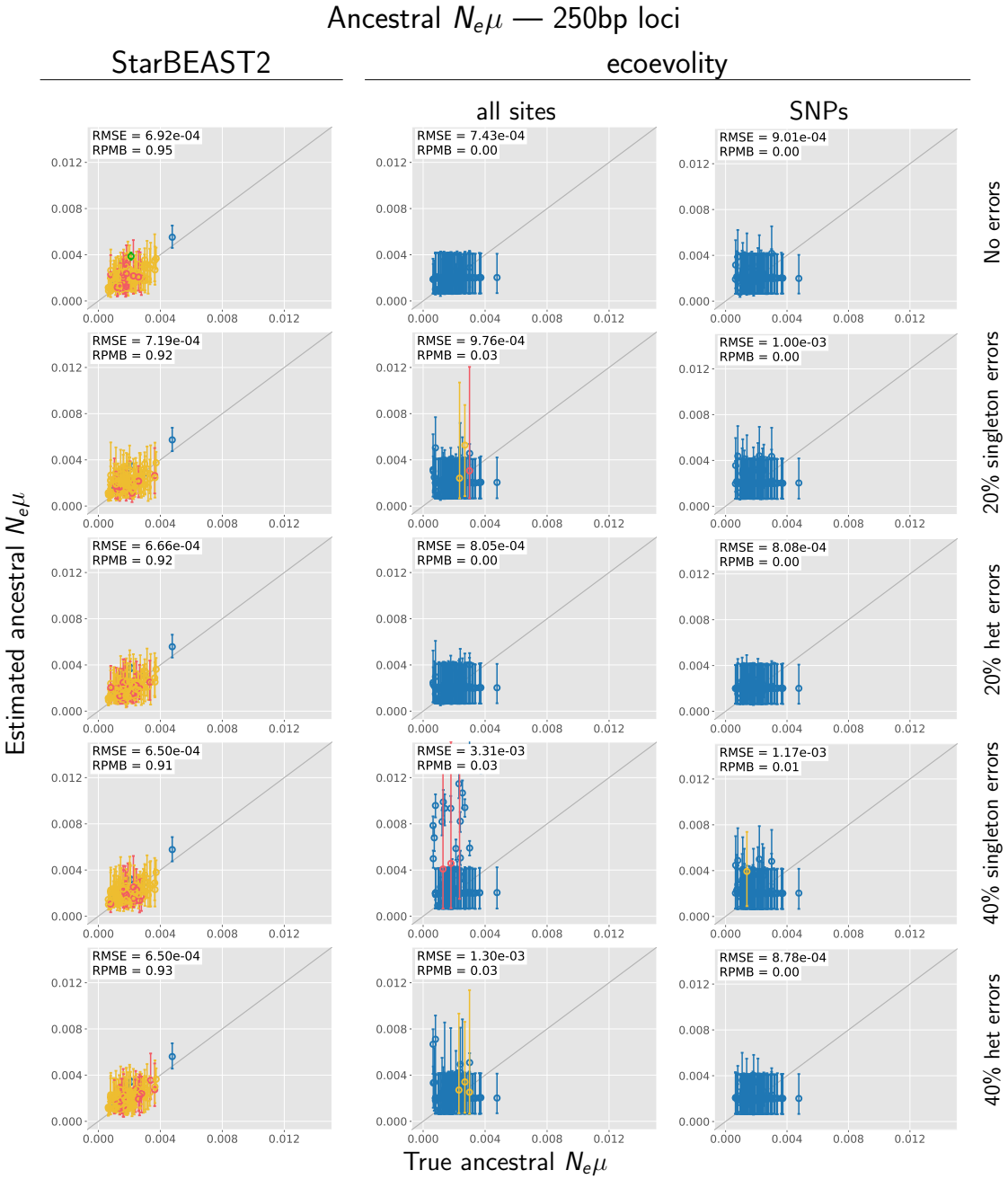

Figure 3.7. **Accuracy and precision of estimates of root effective population size scaled by the mutation rate ($N_e^R\mu$) with 250 base pair loci.** The left column shows estimates from *StarBEAST2*, and the center and right column shows estimates from *ecoevolity* using all sites and (at most) one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

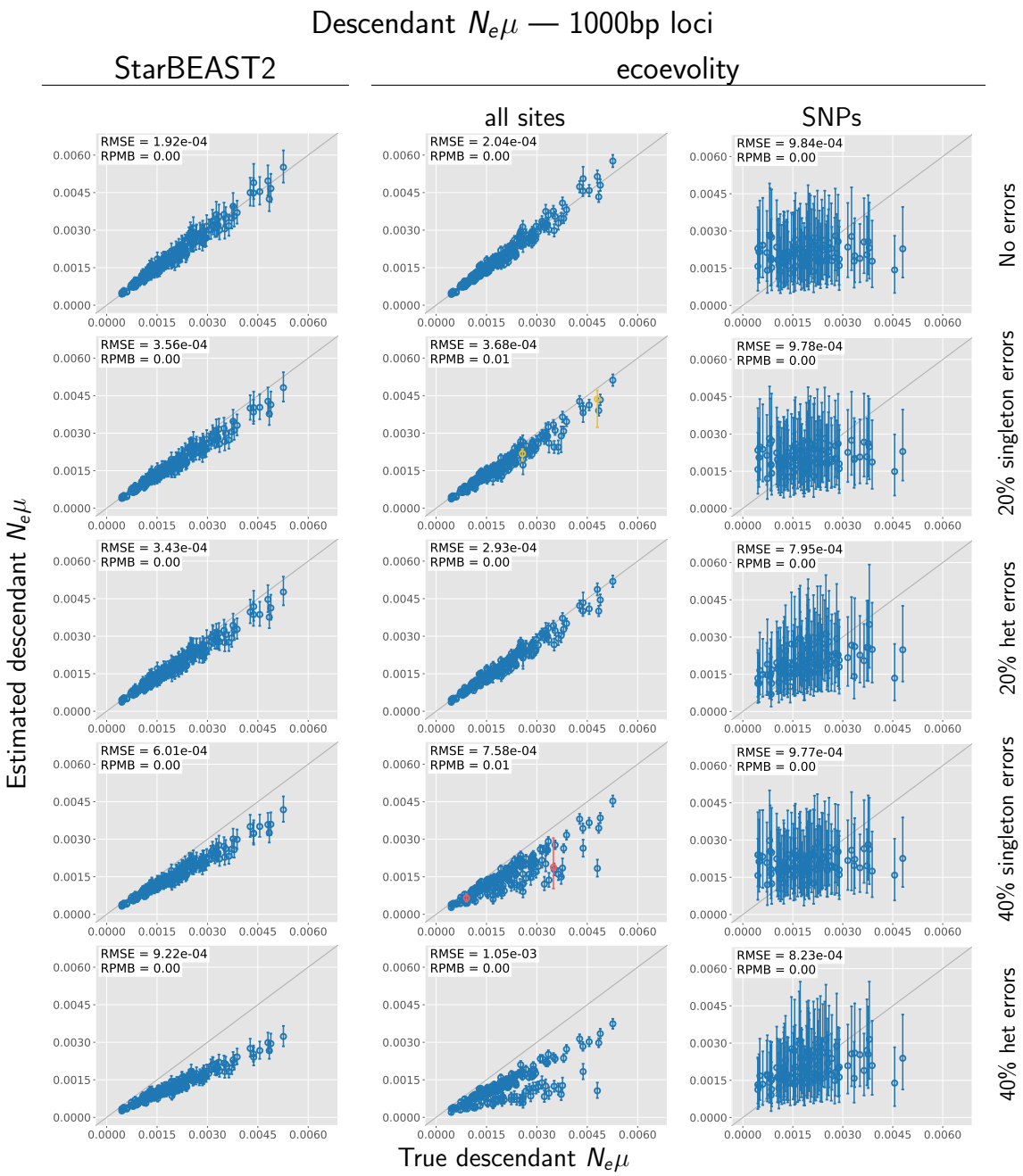

Figure 3.8. **Accuracy and precision of estimates of effective population sizes of the descendant branches of the tree scaled by the mutation rate ($N_e^D\mu$) with 1000 base pair loci.** The left column shows estimates from *StarBEAST2*, and the center and right column shows estimates from *ecoevolity*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

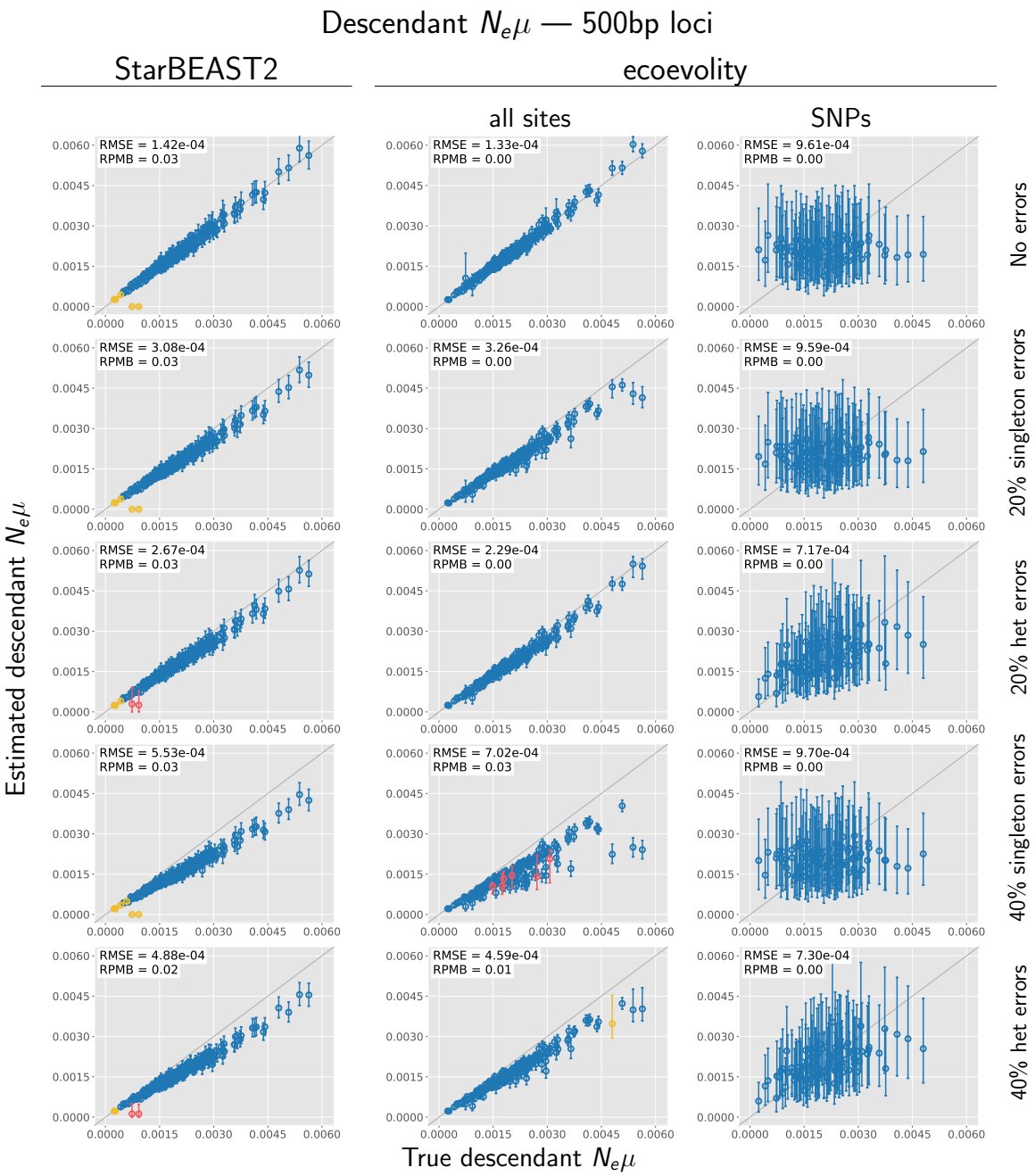

Figure 3.9. **Accuracy and precision of estimates of effective population sizes of the descendant branches of the tree scaled by the mutation rate ($N_e^D\mu$) with 500 base pair loci.** The left column shows estimates from *StarBEAST2*, and the center and right column shows estimates from *ecoevolity*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

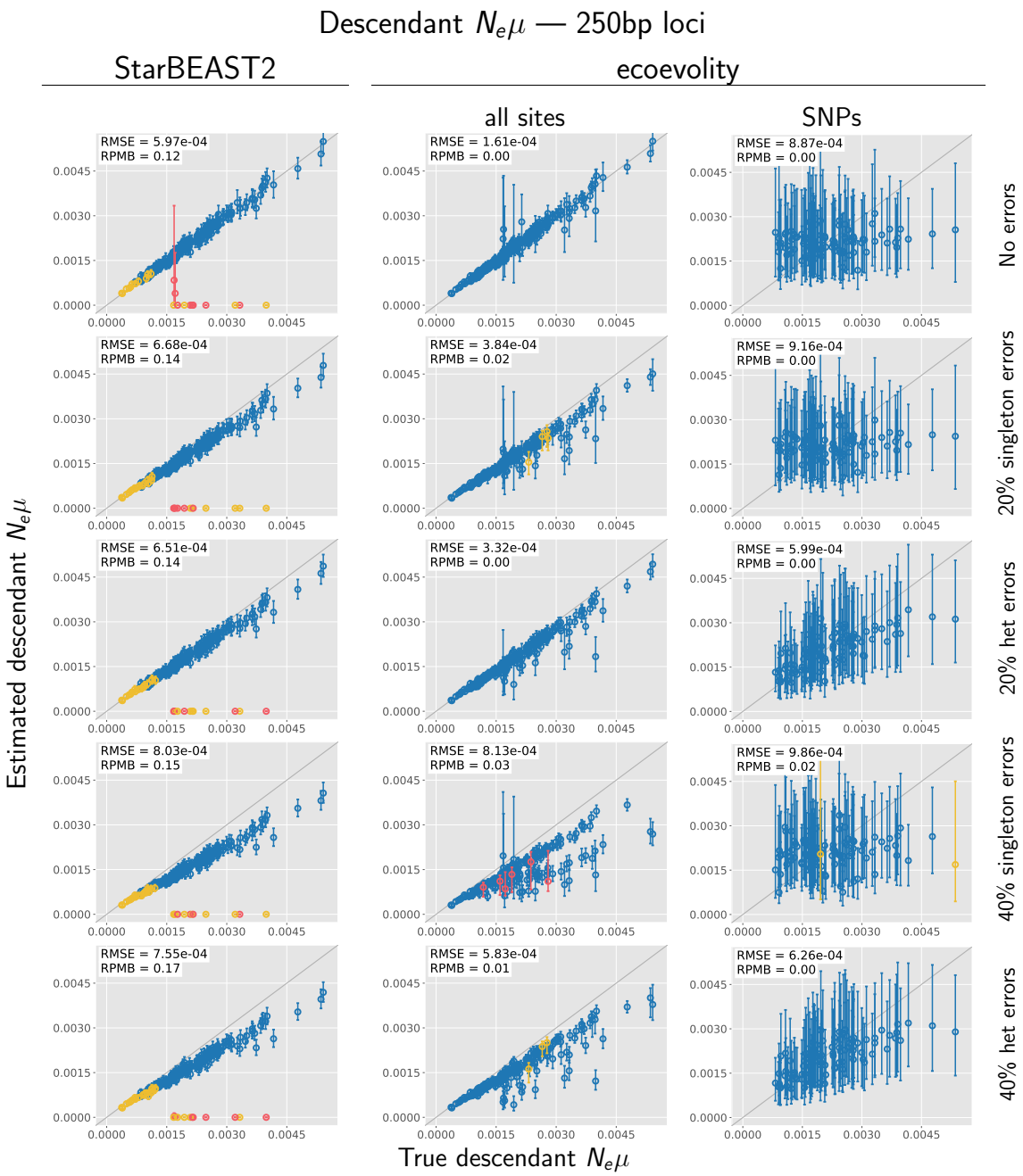

Figure 3.10. **Accuracy and precision of estimates of effective population sizes of the descendant branches of the tree scaled by the mutation rate ($N_e^D\mu$) with 250 base pair loci.** The left column shows estimates from *StarBEAST2*, and the center and right column shows estimates from *ecoevolity*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

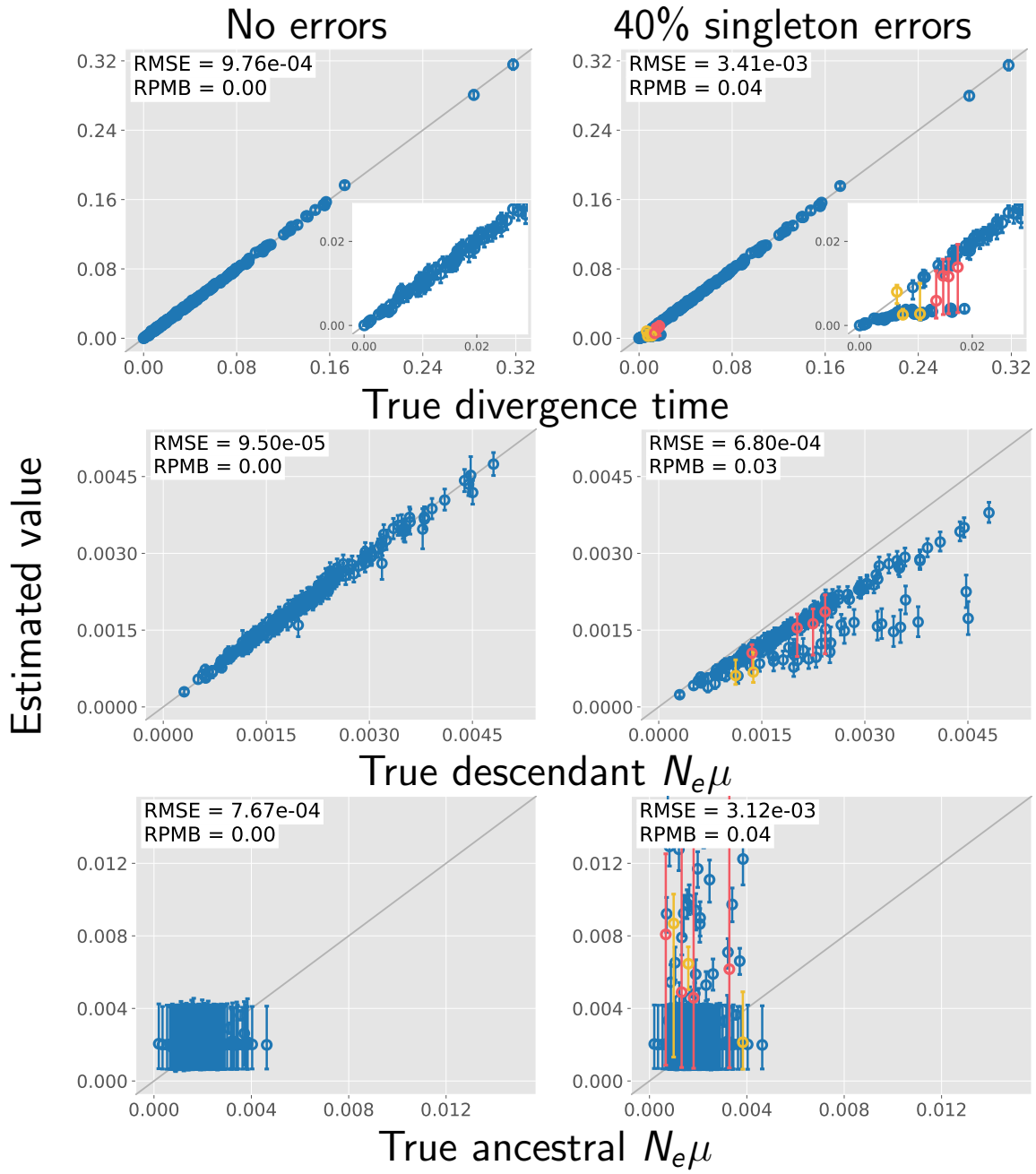

Figure 3.11. The performance of *ecoevolity* with data sets simulated with unlinked characters. Each plotted circle and associated error bars represent the posterior mean and 95% credible interval. Circles and error bars are colored yellow if the effective sample size (ESS) of the estimate was less than 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).