American Genetic Association

Journal of Heredity, 2020, 573–582
doi:10.1093/jhered/esaa042
Brief Communication
Advance Access publication October 8, 2020

OXFORD

Brief Communication

Ghosts of a Structured Past: Impacts of Ancestral Patterns of Isolation-by-Distance on Divergence-Time Estimation

Zachary B. Hancock and Heath Blackmon

From the Department of Biology at Texas A&M University, Butler Hall, 3528, 525 Lubbock St, College Station, TX 77845-3424 (Hancock and Blackmon) and Ecology & Evolutionary Biology Interdisciplinary Program at Texas A&M University, Butler Hall, 3528, 525 Lubbock St, College Station, TX 77845-3424 (Hancock and Blackmon).

Address correspondence to Z. B. Hancock at the address above, or e-mail: zhancock@bio.tamu.edu.

Received June 16, 2020; First decision July 30, 2020; Accepted October 1, 2020.

Corresponding Editor: Mark Springer

Abstract

Isolation-by-distance is a widespread pattern in nature that describes the reduction of genetic correlation between subpopulations with increased geographic distance. In the population ancestral to modern sister species, this pattern may hypothetically inflate population divergence time estimation due to allele frequency differences in subpopulations at the ends of the ancestral population. In this study, we analyze the relationship between the time to the most recent common ancestor and the population divergence time when the ancestral population model is a linear stepping-stone. Using coalescent simulations, we compare the coalescent time to the population divergence time for various ratios of the divergence time over the population size. Next, we simulate whole genomes to obtain single nucleotide polymorphisms (SNPs), and use the Bayesian coalescent program SNAPP to estimate divergence times. We find that as the rate of migration between neighboring demes decreases, the coalescent time becomes significantly greater than the population divergence time when sampled from end demes. Divergence-time overestimation in SNAPP becomes severe when the divergence-to-population size ratio < 10 and migration is low. Finally, we demonstrate the impact of ancestral isolation-by-distance on divergence-time estimation using an empirical dataset of squamates (*Tropidurus*) endemic to Brazil. We conclude that studies estimating divergence times should be cognizant of the potential ancestral population structure in an explicitly spatial context or risk dramatically overestimating the timing of population splits.

Subject area: Tree of Life: Population structure, phylogeography and phylogenomics
Keywords: phylogenetics, multispecies coalescent, pairwise divergence, *Tropidurus*

A major goal in phylogenetic and phylogeographic studies is the estimation of species divergence times. The topic has a long and contentious history largely centered around questions of how to appropriately apply fossil calibrations (e.g., Heath et al. 2014; Brown and Smith 2018), rate heterogeneity (Pond and Muse 2005), rate of morphological evolution (Lynch 1990), and selecting an adequate clock model (Douzery et al. 2004; Lepage et al. 2007).

Beyond methodological concerns are those that emerge from the nature of the data itself. Most phylogenetic models assume that fixed differences between species are the result of genetic drift, and under the neutral theory of molecular evolution (Kimura 1968; King and Jukes 1969) the rate of evolution (or substitution rate) is equal to the per generation neutral mutation rate, μ (Kimura 1983). For well-calibrated molecular clocks (e.g., Knowlton and Weigt 1998; Weir

and Schluter 2008; Herman et al. 2018), we can estimate the time of divergence (usually in years) as $\pi_{12} / 2\mu$, where π_{12} is the pairwise sequence divergence between species 1 and 2. However, in general we are not interested in estimating the divergence time of specific genetic variants, but rather the time of population divergence (T_D). For example, we might be interested in estimating the timing of a vicariant event that we suspect corresponds to a past geological upheaval.

There is a known discrepancy between the coalescent time of neutral genetic variants (T_{MRCA}) and T_D (Nei and Li 1979; Nei and Takahata 1993). The degree of this discrepancy is determined by the ratio of T_D / N_e, where N_e is the effective population size (Edwards and Beerli 2000; Rosenberg and Feldman 2002). This is because lineages must first be within the same population, which occurs T_D generations in the past, followed by coalescence, which on average requires $2N_e$ generations. Therefore, for a completely panmictic population: $T_{MRCA} = T_D + 2N_e$. The expected amount of pairwise sequence divergence is

$$E(\pi_{12}) = 2\mu \ [T_D + 2N_e] \tag{1}$$

(Wakeley 2000). When the ratio of T_D / N_e is large, the bias in coalescent time in the ancestral population is minimal compared to T_D (Edwards and Beerli 2000; Arbogast et al. 2002). However, as T_D / N_e becomes small, $2N_e$ plays a major role in the overall sequence divergence between species. Rosenberg and Feldman (2002) evaluated the relationship between T_{MRCA} and T_D in a simple 2 population split model using coalescent simulations. They found that T_{MRCA} converged on T_D when the ratio of $T_D / N_e \approx 5$. Importantly, the N_e in these models is that of the ancestral population; therefore, the extent of overestimation is the result of demographic conditions present in the ancestor. Demographic conditions that inflate N_e, such as ancestral population structure or a bottleneck following the split, are expected to have a major impact on divergence-time estimation (Gaggiotti and Excoffier 2000; Arbogast et al. 2002; Angelis and Dos Reis 2015).

Wakeley (2000) demonstrated that in descendant species who share an ancestor whose population dynamics are characterized by an island model (Wright 1931) with free migration between demes, overestimation of divergence-times are on the order of $2N_eD[1 + 1/(2M)]$ where $M = 2N_emD/(D - 1)$, m is the migration rate and D is the number of demes. The expected amount of pairwise sequence divergence is therefore

$$E(\pi_{12}) = 2\mu \ \left[T_D + 2N_eD\left(1 + \frac{1}{2M}\right)\right]. \tag{2}$$

Population subdivision initially leads to shallow coalescent times where individuals within a shared deme rapidly find ancestors (the "scattering phase"; Wakeley 1998). However, since ancestral lineages must be in the same deme to coalesce, the rate in the "collecting phase" is characterized by the migration rate that shuffles ancestors around the range, reducing the probability that lineages coalesce (Wakeley 1998, 1999).

In the context of real populations, the island model of migration rarely applies (Whitlock and McCauley 1999; Meirmans 2012). Instead, population structure is the product of the spatial distribution and dispersal potential of the organism in question. Often this structure is in the form of isolation-by-distance (IBD). IBD is a widespread pattern in natural systems, characterized by a reduction in the probability of identity by descent (Wright 1943) or genetic correlation (Malécot 1968) with geographic distance. Patterns of IBD are most pronounced in stepping-stone models (Kimura 1953; Kimura and Weiss 1964) in which migration is restricted to neighboring demes. In this way, demes in close proximity share a greater proportion of migrants than they do with more distant demes. Distributions of coalescent times in stepping-stone models have been studied both in the context of 1-dimensional and 2-dimensional models that are circular or toroidal (Maruyama 1970a, 1970b; Slatkin 1991), and in continuous models with joined ends (Maruyama 1971) or with discrete edges (Wilkins and Wakeley 2002). Slatkin (1991), using a circular stepping-stone model, showed that the probability for 2 genes sampled i steps apart have an average coalescent time:

$$T_{MRCA} = 2N_eD + \frac{(D - i)i}{2m} \tag{3}$$

Therefore, the amount of expected pairwise sequence divergence is:

$$E(\pi_{12}) = 2\mu \ \left[T_D + 2N_eD + \frac{(D - i)i}{2m}\right] \tag{4}$$

The circular stepping-stone model should overestimate T_D more dramatically as the number of demes becomes large and the distance between them increases. However, like the island model of free migration, circular ranges are likely rare in nature. Instead, natural populations are characterized by discrete range edges where end demes may only receive migrants from one direction (e.g., Peterson and Denno 1998; Broquet et al. 2006; Aguillon et al. 2017). Hey (1991) showed analytically in the case of a linear stepping-stone model that the distribution of coalescent times of 2 alleles from demes at the extremes of the range should coalesce much deeper than any 2 alleles chosen randomly from the population.

Vicariant speciation is considered one of the most common forms of allopatry (Coyne and Orr 2004), and results from the cessation of gene flow at some discrete barrier in a species range. This form of speciation has been invoked across many empirical systems (e.g., Riddle et al. 2000; Van Bocxlaer et al. 2006; Hancock et al. 2019). Vicariant speciation in organisms with low dispersal abilities may maintain strong allelic differences at the range edges ancestrally consistent with a pattern of IBD. For example, Hancock et al. (2019) found that sister species of beach amphipods in the Gulf of Mexico with large ranges showed patterns of IBD within species. In addition, they identified a distinct barrier to gene flow (the Mississippi River) and posited that this resulted in vicariant speciation. Since both IBD and vicariant speciation are presumed common in nature, biases in divergence-estimation based on π_{12} could be widespread.

Ultimately, the degree to which T_{MRCA} impacts phylogenetic inference and divergence-time estimation is dependent on its impact on π_{12}. Given that lower migration rates lead to greater T_{MRCA} (Hey 1991), we expect that differentiation (π_{12}) between end demes compared to center demes will become more pronounced at smaller m. If the difference between the T_{MRCA} of central demes and end demes is dramatic enough, we expect that divergence dating of species that arose from ancestral end demes may significantly overestimate T_D.

In this study, we estimate mean T_{MRCA} for 2 genes sampled in descendant species (either from the ends or the center of the ancestral range) in which the ancestral population is characterized by a stepping-stone model with discrete ends using a simulation

approach. In particular, we are interested in what value of T_D / *ND* we expect T_{MRCA} to converge on T_D. We use *ND* (the product of the census size and deme number) as our expected N_e under panmixia (Wakeley 2009). Next, we examine the distribution of π_{12} across the genome under different simulated migration conditions to compare with expectations under a panmictic model. We then test the performance of the phylogenetic inference program SNAPP (Bryant et al. 2012) on simulated single nucleotide polymorphism (SNP) data to evaluate how these trends may bias our inference of species divergence times. SNAPP is a BEAST (Bouckaert et al. 2014) package that operates under an explicit coalescent framework, inferring gene trees from individual SNPs. The program is ideal for phylogeographic studies that utilize RADseq and other genotype-by-sequencing (GBS) technologies to generate thousands of SNPs across the genome (e.g., Manthey et al. 2015; Dowle et al. 2017; Manthey et al. 2017; Leslie and Morin 2018), and has been used explicitly in divergence-time dating previously (Strange et al. 2018; Spalink et al. 2019; Fang et al. 2020). Finally, we illustrate how ancestral IBD can inflate divergence-time estimates on an empirical phylogenomic dataset of lizards (Domingos et al. 2017).

Methods

In the following methods, we use the term "deme" to represent a subpopulation of randomly mating individuals within a broader collection of demes that we refer to as the "population." To be a part of the population, a deme must be able to share migrants with other demes within the population. We use the term "species" to represent an isolated randomly mating unit that no longer shares migration with other populations or demes. This is not meant to reflect any species definition. Finally, we use the term "end species" and "center species" to refer to a set of sister species that either descend from demes on opposite ends of the ancestral range (end species) or descend from neighboring demes in the range center (center species). Our focus below is on evaluating the coalescent times, pairwise differences, and estimated divergence times between these sister species (i.e., the outgroup – "sp3" in Figure 1a—is only meant to root the tree).

For all simulation models, we used the product of the census population size (*N*) and the deme number (*D*) to evaluate the relationship of T_D / *N* (Rosenberg and Feldman 2002). Since sp1 and sp2 (Figure 1) transition to panmixia following the split at time T_D, *ND* should approximate N_e, though there will be a period of nonequilibrium immediately following divergence. Therefore, the contribution to π_{12} from T_D is on the order of *ND*. The ancestral contribution of π_{12} will necessarily be some value greater than *ND* due to population structure (i.e., $N_e > ND$; Wakeley 2000). Our interest here is explicitly on how much greater this contribution is relative to T_D.

Coalescent Simulations

To evaluate the relationship between T_D / *N* and (T_{MRCA} – T_D) / T_D when the ancestral population is characterized by a stepping-stone model of migration, we used *fastsimcoal2* (Excoffier et al. 2013) simulations over a wide range of T_D / *N* values. Specifically, starting at time 0 and going backwards, each simulation consisted of initially 2 species with no migration between them until time T_D in the past. At T_D, these 2 species merge into an ancestral population with 10 demes (*D*) following a linear stepping-stone migration model. For simulations of end species, the ancestral deme of each species was on opposite ends of the range (i.e., demes 1 and 10 in Figure 1a). For the center species, the ancestral demes were neighboring and in the range center. For each of these 2 models, we sampled $k = 2$ individuals to coalesce, and each simulation terminated upon coalescence.

In the ancestral population, center demes received migrants from neighboring demes at rate $2m$, whereas demes at the end of the range received migrants at rate m. This is due to the fact that end demes have only a single neighbor, whereas all center demes have 2 neighbors (Figure 1a). The ancestral population was simulated for migration rates of 0.1, 0.01, and 0.001, and a range of T_D / *ND* values from 0.01–10. In addition, we simulated an island model of migration for comparison with the stepping-stone model. In the island model, the ancestral population consisted of 10 demes with free migration between each at rate m. This resulted in a total of 84 distinct simulation scenarios, and each were replicated 1000 times. We did not explicitly model chromosomes; instead, replicates were treated as independent loci.

To statistically compare between the 3 models (end species sampled in stepping-stone, center species in stepping-stone, and the island model), we subset ratios of T_D / *ND* to values of 10, 5, 2, 1, 0.5, and 0.1. Resulting T_{MRCA} distributions for each population model were compared using a pairwise Wilcoxon test in the R platform (R Core Team 2019), as the resulting distributions were non-normal.

Genome Simulations

To evaluate how ancestral IBD impacts pairwise sequence divergence (π_{12}), genome-wide coalescent times (T_{MRCA}), and divergence-time estimation, we performed hybrid simulations that combined the

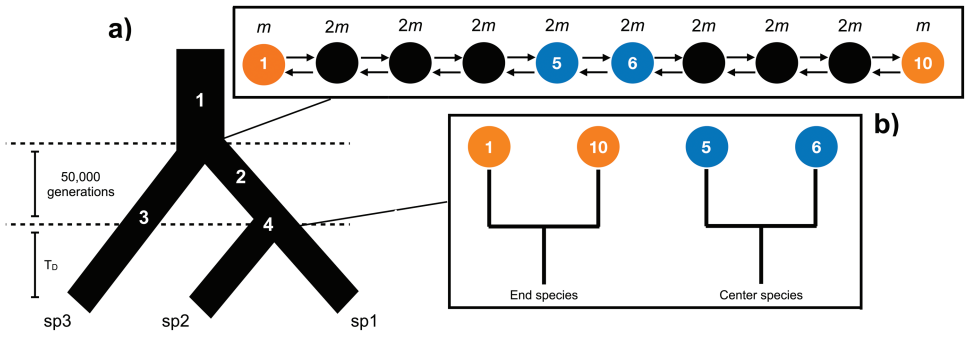

Figure 1. Population model for SLiM simulations. **(a)** Three-taxon species tree: 1) coalescent simulations in *msprime* with $N = 2000$; 2) ancestral stepping-stone conditions begin (see **b**); 3) $N = 1000$, panmictic; 4) population split, leaving end or center species surviving as sp1 and sp2. **(b)** Ancestral population dynamics. Circles designated "1" and "10" are end species; center species are "5" and "6."

coalescent simulator *msprime* (Kelleher et al. 2016) and the forward-time simulator SLiM v3.3 (Haller and Messer 2019). Since forward-time simulators begin with individuals that are completely unrelated, often a neutral burn-in period is required to allow coalescence or mutation-drift equilibrium to occur (Haller et al. 2019). This can be computationally costly and time-consuming; however, using tree-sequence recording methods in SLiM (Haller et al. 2019), we can bypass the need to equilibrate during the forward-time simulation. To generate a panmictic ancestral population with a coalescent history, we simulated 2000 individuals (N_e = 4000) using *msprime* with genome sizes of 10 Mb and a recombination rate of 10^{-8} (~0.1 recombination events per individual per generation). The resulting coalescent trees were then imported into SLiM as the basis for the starting population.

In SLiM, the initial population was split into 2 populations of N = 1000: 1) an outgroup that remained panmictic ("sp3" in Figure 1a) and 2) the ancestral population of the sister species "sp1" and "sp2," which was subdivided into 10 demes (N = 100 per deme) in a linear stepping-stone model (Figure 1a). These dynamics persisted for 50 000 generations, after which the ancestral population was split into either end species or center species (Figure 1a). This was done by removing the intermediate demes and instantaneously adjusting N for the 2 species to 1000 so that the census size remained constant. The resulting 3-species were then allowed to evolve for T_D generations before the simulation was terminated. Five different T_D values were simulated, which correspond to T_D / *ND* ratios of 50, 25, 10, 5, and 1 (T_D values of 50 000, 25 000, 10 000, 5000, and 1000 generations). These values of T_D / *ND* were chosen based on the results from the coalescent simulations in *fastsimcoal2* (see Results); for values >10, T_{MRCA} is expected to converge on T_D, whereas values <10 are expected to overestimate T_D regardless of migration rates.

The resulting tree-sequences from the SLiM simulation were imported into Python 3 using *pyslim*, and we overlaid neutral mutations ($\mu = 10^{-7}$ per base per generation) onto the trees using *msprime*. Pairwise divergence (π_{12}) was then estimated across the genome in windows of 100 kb for both end demes and center demes. These values were also converted into generations using $\pi_{12} / 2\mu$, which gives a rough estimate of divergence time per window.

Equation 1 primes our expectation for the amount of sequence divergence expected given some value of T_D and ancestral N_e. By rearranging equation 1, we can naively calculate the ancestral N_e from genome-wide π_{12} as:

$$N_e = \frac{\pi_{12} - 2T_D\mu}{4\mu} \quad (5)$$

From this, we plot estimated ancestral N_e within 100 kb windows across the genome to compare with the known census population size (N_c = 1000), and to evaluate the relationship between N_e and N_c in the presence of IBD.

Next, we plotted the distribution of coalescent times (T_{MRCA}) across the genome to visualize differences between T_{MRCA} of end and center species. Median T_{MRCA} for each ratio and migration rate was compared via a Kruskal-Wallis test and a pairwise Wilcoxon rank test in R due to the data violating normality.

Each simulation produced >200 000 SNPs. For divergence-time analysis, we randomly sampled 3000 SNPs—a number found by Strange et al. (2018) to optimally perform in SNAPP. Each run consisted of 10 individuals from species sp1 and sp2, and 1 individual from the outgroup population, sp3 (Figure 1). Unlike other fully coalescent models, SNAPP does not sample from gene trees directly to estimate the species tree, but instead integrates over all possible gene trees using biallelic SNPs. The method has been found previously to perform well on both simulated and empirical data (Bryant et al. 2012; Strange et al. 2018). We designated a gamma-distributed prior on θ ($=4N_e\mu$) with a mean equal to the expected π_{12} (equation 1). Forward (u) and backward (v) mutation rates were estimated within BEAUti (Bouckaert et al. 2014) from the empirical SNP matrix using the tab *Calc_mutation_rates*, and these values were sampled during the MCMC. The rate parameter λ, which is the birth-rate on the Yule tree prior, was gamma-distributed with $\alpha = 2$ and $\beta = 200$, where the mean is α / β (Leaché and Bouckaert 2018).

SNAPP is designed to handle incomplete lineage sorting (ILS), but to minimize its effects—since we are not interested in the program's ability to estimate topology but rather branch-lengths—we applied a fixed species tree. Branch-lengths in SNAPP do not scale to time, but instead are measured in number of substitutions. Given a fixed mutation rate, we convert the number of substitutions separating sp1 and sp2 to the number of generations as $g = s / \mu$, where s is branch-lengths in units of substitutions (Bouckaert and Bryant 2015). The MCMC chain length was 10–50 million sampling every 1000 with a burn-in of 10%, ensuring that ESS values of interest were all >200. Runs were performed on the high-performance computing cluster CIPRES (www.phylo.org; Miller et al. 2010).

MCMC log files were then downloaded and analyzed in R. The performance of SNAPP was evaluated by comparing traces of end and center species across migration rates and T_D / *ND* values. Results were evaluated by first randomly sampling 1000 rows for each migration rate, and then performing a Kruskal-Wallis test. Trees from the MCMC were summarized in TreeAnnotator v.2.6.0 (Bouckaert et al. 2014) and visualized in R using the package *ggtree* (Yu et al. 2017). Branch colors were scaled by estimated median θ per branch.

To ensure the trends observed were the result of inflated π_{12} when T_D / *ND* and migration rate is low and not an issue unique to SNAPP, we performed pairwise F_{ST} tests of end and center demes that were used in the divergence-time estimation (Supplementary Material; Supplementary Figure S10). These tests were performed using *tskit*. All SLiM recipes, python and R code, and xml files can be found at https://github.com/hancockzb/ancestralIBD.

Empirical Dataset

To evaluate how ancestral patterns of IBD may impact divergence-time estimation in practice, we analyzed the phylogenomic dataset of endemic squamates from Domingos et al. (2017). The dataset consists of 12 species (including the identified cryptic lineages) sampled broadly across the geographic range of *Tropidurus itambere*, which is native to the Cerrado, a tropical savanna in Brazil that stretches across the states of Goiás, Mato Grosso do Sul, Mato Grosso, Tocantins, Minas Gerais, and the Federal District. Domingos et al. (2017), using anchored hybrid enrichment, identified 5 cryptic species they designated A–E within *T. itambere*. Each of these species were sampled across multiple localities with some locations spatially nearer to their close relatives than others (see figure 1 in Domingos et al. 2017).

This geographically broad sample scheme is ideal to test the impact of IBD on divergence-time estimation. If ancestral IBD has influenced π_{12}, we expect species localities more distant to one another to be more deeply diverged than when 2 species are sampled from locations nearby. Importantly, this pattern should only hold if the range was once continuous (as in our simulations above); otherwise, there should be no difference in π_{12} between species even if there

are current patterns of IBD within species. Ongoing gene flow between the spatially close localities could also generate this pattern, but Domingos et al. (2017) found no evidence for this.

The alignments used in the coalescent species delimitation in Domingos et al. (2017) were downloaded from https://datadryad.org/stash/dataset/doi:10.5061/dryad.1hs2m. We randomly selected 10 loci conditional on them containing samples from all locations (total length = 15 303 bp). We generated 2 separate alignment files based on the sampled locality's distance from their nearest relative. Preliminary analyses, including the entire dataset, showed that species (DE) were sister to A, and B was sister to C (which was also found from the entire concatenated dataset in Domingos et al. 2017). Therefore, the first dataset we designated as "far," and it included only the location of E that was most distant from A (i.e., São João D'Aliança; roughly 1000 km), and the location of C that was the most distant to B (i.e., Ribeirão Cascalheira; ~1000 km). The second dataset was designated "near" which included localities of E geographically closest to A (Brasilia and Pirenópolis; ~160 km) and C to B (Barra do Garças; ~600 km).

Phylogenetic inference was performed using the fully coalescent software *BEAST (Heled and Drummond 2010). For each locus, we applied the HKY model (Hasegawa et al. 1985) and a strict molecular clock. We also performed runs with a relaxed lognormal clock for each locus to ensure age differences across the tree were not the result of restricting substitution rate variation. Domingos et al. (2017) only focus on topology and time is not considered; therefore, we estimate branch-lengths in units of substitutions per site. We use an arbitrary rate of mutation (10^{-8}) to convert substitutions to generations. This analysis is not meant to be a rigorous evaluation of the true time of divergence but merely a demonstration of the impacts of ancestral IBD on empirical data. Each analysis was run for 100 million generations with a burn-in of 10%. Estimated divergence times between the 2 models were compared using a 2-way ANOVA in R. Differences in the densities of estimated gene trees from the posterior were visualized using DensiTree (Bouckaert 2010).

Results

Coalescent Simulation Results

The coalescent simulations produced trends superficially similar to those found by Rosenberg and Feldman (2002). At the lowest T_D / *ND*, the proportion of deep coalescence was dramatically greater than at higher values with the curve producing a similar logarithmic relationship (Supplementary Figure S1). However, T_D and T_{MRCA} did not necessarily converge when T_D / *ND* = 5. Instead, the rate of convergence was dependent on both the deme sampled and the migration rate.

When migration was high (m = 0.1) and T_D / *ND* was less than 0.5, there was no significant difference between center or end species in the stepping-stone model or the island model. However, for values of T_D / *ND* > 0.5, the T_{MRCA} of end species became significantly different from both island (P < 0.02) and center species (P < 0.01; see Supplementary Table S1). When migration was reduced below 0.1, this pattern became more extreme. End species were significantly different in all pairwise comparisons of models (P < 0.000001), and center species differed from the island model at T_D / *ND* ratios of 0.5, 2, and 10 (P < 0.03) when m = 0.01. At the lowest migration rate simulated (m = 0.001), all pairwise model comparisons were significantly different when T_D / *ND* > 0.5 (P < 0.001; see Supplementary Table S1).

Genome Simulation Results

Results from the genome simulation approach corroborated those found with *fastsimcoal2*. Regardless of T_D / *ND*, when m = 0.1 the difference between center and end species was less severe relative to when m < 0.1 (Supplementary Table S2). Across the simulated genomes, T_{MRCA} became dramatically deeper between end than center species as migration fell below 0.01. For the genome-wide divergence estimates, the degree of overestimation depended on the ratio of T_D / *ND*. While all scenarios where m = 0.001 overestimated the true T_D, when T_D / *ND* < 10 end species were 5–60 times more diverged than expected (Supplementary Figure S3). This is a direct result of the deeper coalescent times between end species when m < 0.1, as these longer branches provide more time for mutations to occur and accumulate (Supplementary Figure S2).

Genome-wide coalescent times (T_{MRCA}) are shown in Supplementary Figure S2. When m = 0.1, only T_D / *ND* = 25 and 10 were significantly different between end and center species (P < 0.005). Regardless of T_D / *ND*, the variance in T_{MRCA} steadily increased with decreasing m. Indeed, the increase in mean T_{MRCA} when m = 0.001 appears largely driven by an increase in the variance at this lower rate. Due to this, we find that ancestral N_e dramatically exceeds N_c when m = 0.001 (Figure 3).

Despite the potential for divergence-time overestimation to be extreme, SNAPP was relatively resilient when T_D / *ND* > 10 and when m > 0.001. When T_D / *ND* = 50, SNAPP was overly conservative and underestimated the number of substitutions expected to occur (Figure 2). When T_D / *ND* = 25, the mean estimate of both center and end species when m > 0.001 either underestimated the true age or was within 5%. However, for end species where m = 0.001 the estimated divergence time exceeded the true age by ~80% (Supplementary Table S3). A similar trend occurred when T_D / *ND* = 10 and 5. Here, both center and end species overestimated the true age, but the end species did so more dramatically (138% the true age versus 81% for 10; 184% versus 67% for 5). The most dramatic overestimation occurred between end species when T_D / *ND* = 1 at ~700% the true age. Importantly, this was not merely the result of a low T_D / *ND* ratio, as the other migration regimes performed well. In fact, most were closer to the true T_D than the expected $\pi_{12,}$ accounting for 2*N* (Supplementary Table S3).

Estimated θ for each branch is shown in Figure 4 for T_D / *ND* = 10, and in Supplementary Figures S4–S7 for the remaining ratios. For all T_D / *ND* values except 1, the median ancestral θ was higher for end species than center when m = 0.001, and the estimated θ for the descendant species (sp1 and sp2 in Figure 1) was considerably lower than for the ancestor or the outgroup, sp3 (Figure 4; Supplementary Figures S4–S7). These patterns are consistent with a population bottleneck, despite *N* being maintained throughout the simulation.

Empirical Results

The estimated divergence-time for the clades ((DE)A) and (BC) were significantly older when samples were from geographically distant localities as opposed to those nearby (p < 0.00001; Figure 4c). For the ((DE)A) clade, the "far" dataset inferred an age of divergence of 270 000 generations, which was 40 000 generations (or ~14%) higher than the "near" estimate. The (BC) divergence was even more extreme, with the "far" being ~24% older than the "near" (Figure 4c). Interestingly, despite the fact that all other samples were included in both analyses, the "far" dataset estimated older ages for most of the other nodes in the tree as well (Figure 4). The total tree height of the "far" was 220 000

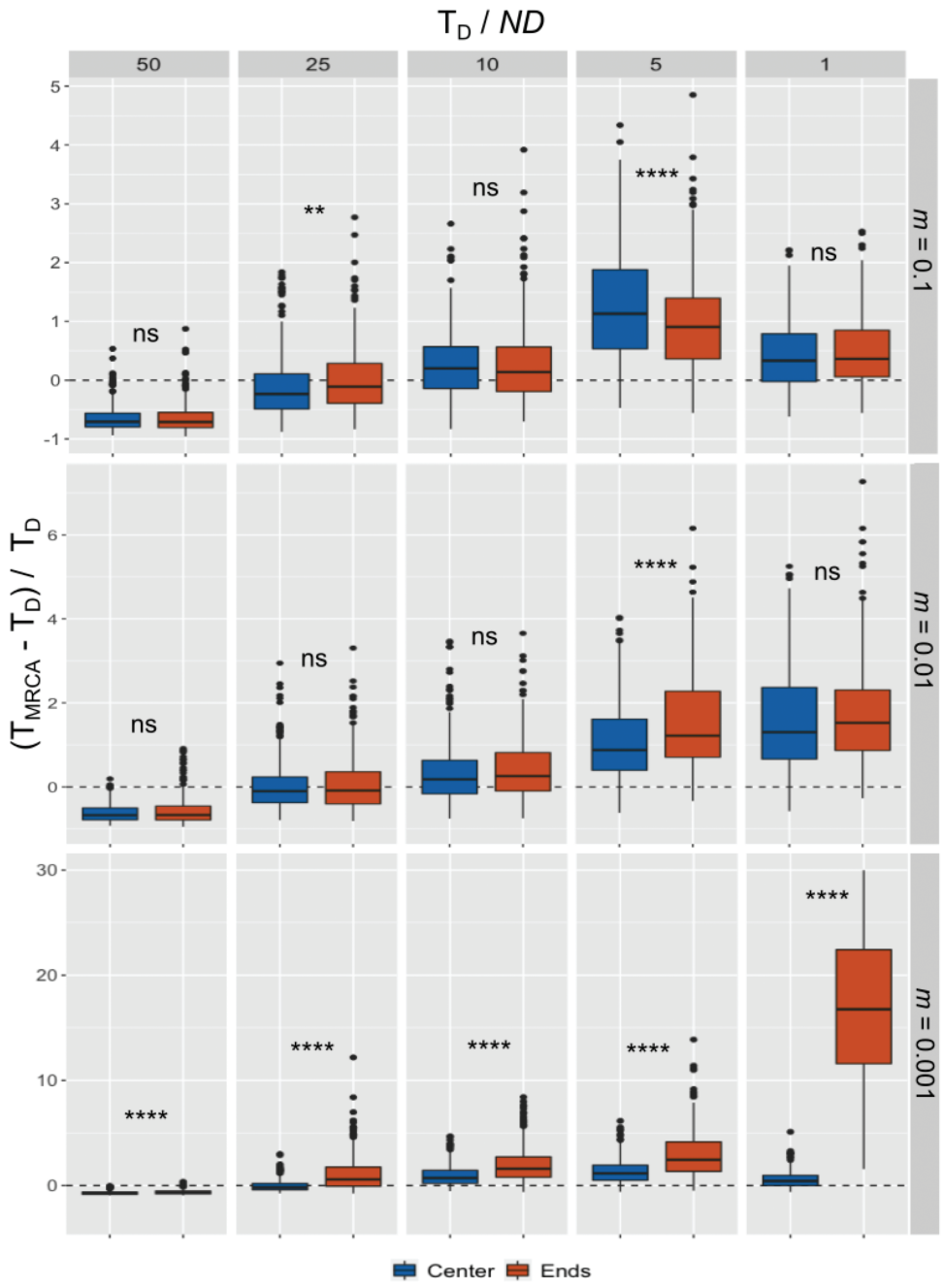

Figure 2. Box plots of the estimated T_{MRCA} by SNAPP; ns = "not significant," $P < 0.05$ (*), $P < 0.001$ (**), $P < 0.0001$ (***), $P < 0.00001$ (****). Dashed lines represent when the estimated age converges on the true age (i.e., at 0). Note that the *y* axis is different between the panels. Center species are on the left, end species on the right (see online version for full color).

generations deeper in time than the "close" (or ~ 9%). The relaxed clock estimates were more extreme, with all node heights being higher in the "far" versus "near" datasets (Supplementary Figure S11).

Discussion

Macroevolutionary patterns are ultimately governed by microevolutionary processes (Li et al. 2018), an observation Lynch (2007), extending Dobzhansky's (1973) maxim, summed up as "nothing in evolution makes sense except in light of population genetics." In this light, we have demonstrated that the population genetic environment of the ancestor shapes the genetic landscape of descendant species. This has been known to impact tree topology when ILS is common (Kubatko and Degnan 2007) and overestimate divergence times in the presence of population structure caused by an island model of migration (Edwards and Beerli 2000; Wakeley 2000). Extensive prior work has shown that the stepping-stone model of migration reduces genetic correlation between demes (Kimura and Weiss 1964; Maruyama 1970a) and that demes farther apart should

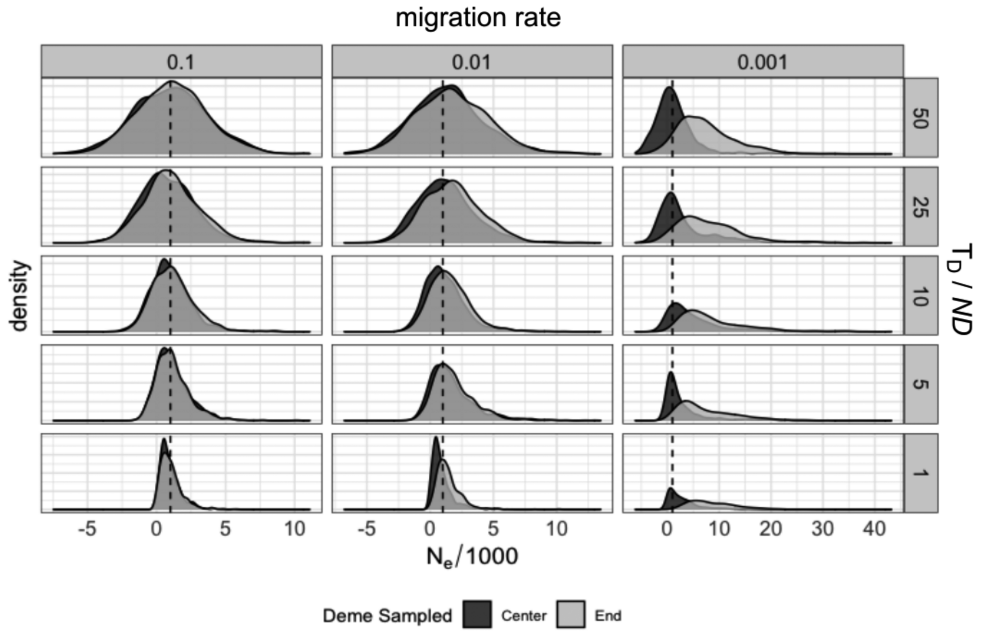

Figure 3. Density plot of scaled ancestral N_e (/1000) based on mean π_{12} across genomic windows of 100kb. Dashed line is when N_e / N_c = 1.

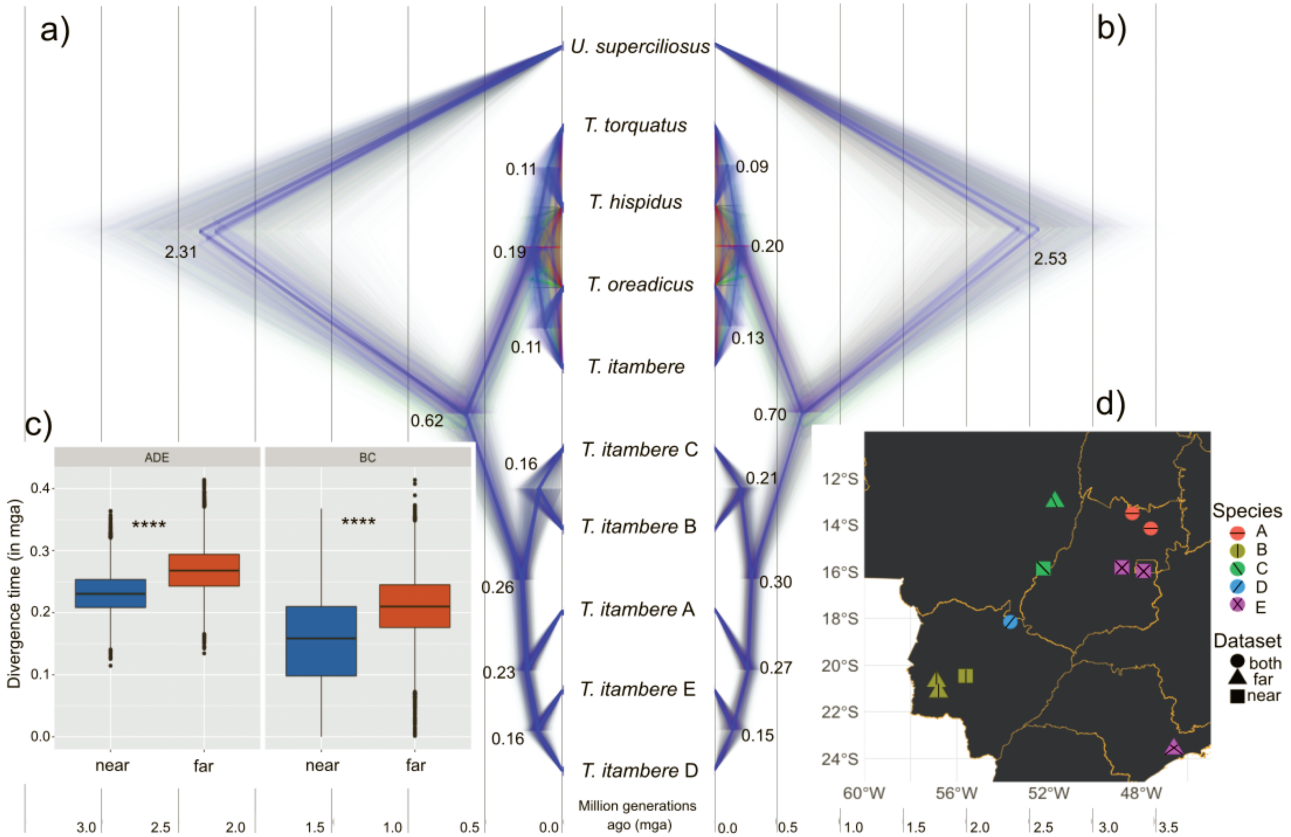

Figure 4. Estimated divergence-times for the Domingos et al. (2017) dataset. (**a**) the "near" trees from the posterior distribution; (**b**) the "far" trees; (**c**) boxplot of the 2 nodes of focus, where ADE is for clade ((DE)A) and BC is (BC); (**d**) map of included sites from Domingos et al. (2017). Significance is as in Figure 2. Numbers at nodes represent the median height in units of millions of generations ago (mga) (see online version for full color.)

coalesce deeper in time than those geographically closer (Hey 1991; Slatkin 1991). However, to our knowledge, the impact of ancestral IBD has not been evaluated in the context of divergence-time estimation previously.

Rosenberg and Feldman (2002) found previously that when $T_D / N = 5$, T_{MRCA} and T_D largely converged in a simple population split model. However, when in the presence of ancestral IBD we found that convergence was dependent on the migration rate (i.e., the strength of ancestral IBD) and whether surviving species neighbored each other or were at the range ends in the ancestral population.

When $T_D / ND > 10$, the ancestral dynamics contribute little to the divergence-time estimate differences between center and end species. However, as this ratio decreases the contribution of $2N_e$ to overall sequence divergence becomes non-trivial. The probability that genetic variants share an ancestor just prior to the population split is higher between species that were geographically closer than those more distant in the ancestral population. This is mediated by the migration rate, which, when high enough, can largely erase the differences between center and end demes. When migration is high (10%, or $m = 0.1$), individuals move well between demes and the coalescent times largely converge (though deeper in time depending on the ratio of T_D / ND). However, as m falls below 1% ($m = 0.01$), or less than one migrant per generation being shared between demes, dispersal cannot keep up with genetic differentiation. Despite all migration regimes producing similar patterns of IBD (Supplementary Figure S9), F_{ST} becomes dramatically higher as migration drops below 1%. This differentiation in the ancestor contributes to the overall sequence divergence (π_{12}) between species, which drives an overestimation of the time of the population split (T_D) when surviving lineages descend from demes on the opposite ends of the range.

As expected, ancestral IBD skews π_{12} and T_{MRCA} away from expected values in a panmictic population, and this caused an inflation in N_e relative to N_c. For $T_D / ND = 50$ and $m = 0.1$, the mean π_{12} for end species was 0.010459 and 0.010419 for center species. Using equation 5, ancestral N_e was 1147.5 for end species and 1047.5 for central. However, when $m = 0.001$, π_{12} for end species was 0.012948, an ancestral $N_e = 7370$. Center species, on the other hand, only increased to $N_e = 1255$. As with the coalescent times, at lower migration rates the variance in N_e becomes exceedingly large, driving up the mean. Importantly, mean genome-wide N_e always exceeds N_c in the presence of ancestral IBD at a level dictated by the migration rate.

This feature of ancestral IBD has important consequences for conservation genetics. Many studies use N_e as a rough biological measure of population size (Turner et al. 2002; Hare et al. 2011; Rieman and Allendorf 2011), and therefore a metric of the health of a population. However, a common phenomenon in range contractions is fragmentation and isolation (Ceballos et al. 2017), which may result in IBD. If many of the demes once contributing to the connectivity of the population have become extinct, and N_e is estimated based on the surviving demes, it will overestimate the actual number of individuals within the population (i.e., the census size, N_c). Thus, we might incorrectly conclude that a species has a larger population size than it actually does, which may lead to mismanagement.

Since N_e is inflated in the ancestral lineage, the descendant species appear to pass through a bottleneck despite N remaining constant (Supplementary Figures S4–S8). Estimated θ in SNAPP captured this dynamic with more extreme differences in θ (i.e., more dramatic bottlenecks) being inferred between end species and when $m = 0.001$. Population bottlenecks have been found to cause divergence-time overestimation due to random differential survival of ancestral alleles into the descendant species (Gaggiotti and Excoffier 2000). In the presence of IBD, this differential allelic persistence between species is mimicking a bottleneck—when demes were far apart, this pattern is more extreme as they already maintain different allelic patterns ancestrally. However, because this pattern is recognizable (Supplementary Figures S4–S8), it can be used to signal when ancestral IBD may be impacting our divergence-time estimation. Unfortunately, without prior range-size knowledge, it may be impossible to differentiate between ancestral IBD and a bottleneck since these produce virtually identical genetic patterns. However, it may not be necessary to do so for simple divergence estimates.

Demes need not necessarily go extinct for ancestral IBD to still influence π_{12}, as seen from the results of the empirical dataset; however, the persistence of demes into the present allows for a geographically aware sampling scheme. Since Domingos et al. (2017) sampled broadly across the range of *T. itambere*, they would be well-positioned to identify inflated π_{12} resulting from ancestral structure. For example, as we have done here, by subsetting the dataset by geographic proximity one can explicitly test for ancestral IBD. For ancestrally structured populations, geography should dictate the degree of π_{12}. Importantly, this also requires that species diverged via vicariance (Coyne and Orr 2004)—the splitting of a once larger range by a discrete barrier—and not some other means, such as population expansions following divergence from a more restricted habitat.

The broader impact of ancestral IBD on divergence-time estimation when in the context of large phylogenies is beyond the scope of this work, but it is conceivable that the longer than expected branches between sister species might bias rate estimation (Arisbrosou and Excoffier 1996; Magallón 2010). In the case of ancestral IBD, the inflated N_e is mimicking a pattern of substitution rate increase. Under neutrality, the rate of substitution is equal to the per generation mutation rate, μ (Kimura 1983); however, in the presence of population structure, substitutions may occur in the ancestral lineages between demes separated by large geographic distances. If the true age of the sister taxa is known but ancestral structure is not accounted for, the substitution rate will be upwardly biased. We found some evidence for this in the *T. itambere* dataset, in which we found higher estimates of π_{12} in the "far" versus the "close" dataset even for nodes more distantly related to the focal clades (Figure 4).

Ancestral structured populations leave their imprint on descendent species in the form of greater coalescent times, and therefore larger than expected pairwise divergences between species. Further, these patterns cause inflated N_e relative to census sizes. Since ancestral IBD mimics the signature of a population bottleneck, coalescent methods that co-estimate θ along with the topology and π_{12}, such as SNAPP and *BEAST, may be the best suited to reveal this potential source of bias. However, fully coalescent models such as these are infamously computationally costly and not presently used for whole-genome sequence data or for phylogenies with large numbers of tips. Indeed, SNAPP becomes prohibitively slow when the number of tips is ~30 (Leaché and Bouckaert 2018).

In the context of larger phylogenies or organisms in which little is known about their ancestral range, it may be impossible to know if extant species descend from range centers or ends, or the level of IBD present in the ancestor. The genetic consequences of ancestral structure therefore behave much like "ghost" populations (Slatkin 2005); despite being extinct, their influence haunts our ability to adequately assess the phylogenetic history of their descendants.

Supplementary Material

Supplementary material is available at *Journal of Heredity* online.

Funding

National Institute of General Medical Sciences at the National Institutes of Health (R35GM138098).

Acknowledgments

Thanks to Ben Haller and Wesley Brashear for coding help.

Data Availability

SLiM recipes, R and python code, and .XML files have been uploaded to https://github.com/hancockzb/ancestralIBD.

References

Aguillon SM, Fitzpatrick JW, Bowman R, Schoech SJ, Clark AG, Coop G, Chen N. 2017. Deconstructing isolation-by-distance: the genomic consequences of limited dispersal. *PLoS Genet*. 13:e1006911.

Angelis K, Dos Reis M. 2015. The impact of ancestral population size and incomplete lineage sorting on Bayesian estimation of species divergence times. *Curr Zool*. 61:874–885.

Arbogast BS, Edwards SV, Wakeley J, Beerli P, Slowinski JB. 2002. Estimating divergence times from molecular data on phylogenetic and population genetic timescales. *Ann. Rev. Ecol. Syst*. 33:707–740.

Aris-Brosou S, Excoffier L. 1996. The impact of population expansion and mutation rate heterogeneity on DNA sequence polymorphism. *Mol Biol Evol*. 13:494–504.

Bouckaert RR. 2010. DensiTree: making sense of sets of phylogenetic trees. *Bioinformatics*. 26:1372–1373.

Bouckaert R, Bryant D. 2015. A rough guide to SNAPP. Available from: https://www.beast2.org/snapp/

Bouckaert R, Heled J, Kühnert D, Vaughan T, Wu CH, Xie D, Suchard MA, Rambaut A, Drummond AJ. 2014. BEAST 2: a software platform for Bayesian evolutionary analysis. *PLoS Comput Biol*. 10:e1003537.

Broquet T, Ray N, Petit E, Fryxell JM, Burel F. 2006. Genetic isolation by distance and landscape connectivity in the American marten (*Martes americana*). *Land. Eco*. 21:877–889.

Brown JW, Smith SA. 2018. The past sure is tense: on interpreting phylogenetic divergence time estimates. *Syst Biol*. 67:340–353.

Bryant D, Bouckaert R, Felsenstein J, Rosenberg NA, RoyChoudhury A. 2012. Inferring species trees directly from biallelic genetic markers: bypassing gene trees in a full coalescent analysis. *Mol Biol Evol*. 29:1917–1932.

Ceballos G, Ehrlich PR, Dirzo R. 2017. Biological annihilation via the ongoing sixth mass extinction signaled by vertebrate population losses and declines. *Proc Natl Acad Sci U S A*. 114:E6089–E6096.

Coyne J, Orr HA. 2004. *Speciation*. Sunderland (MA): Sinauer Associates, Inc. p. 545.

Dobzhansky T. 1973. Nothing in biology makes sense except in the light of evolution. *American Biology Teacher*. 35:125–129.

Domingos FMCB, Colli GR, Lemmon A, Lemmon EM, Beheregaray LB. 2017. In the shadows: phylogenomics and coalescent species delimitation unveil cryptic diversity in a Cerrado endemic lizard (Squamata: Tropidurus). *Mol Phylogenet Evol*. 107:455–465.

Douzery EJ, Snell EA, Bapteste E, Delsuc F, Philippe H. 2004. The timing of eukaryotic evolution: does a relaxed molecular clock reconcile proteins and fossils? *Proc Natl Acad Sci U S A*. 101:15386–15391.

Dowle EJ, Bracewell RR, Pfrender ME, Mock KE, Bentz BJ, Ragland GJ. 2017. Reproductive isolation and environmental adaptation shape the phylogeography of mountain pine beetle (*Dendroctonus ponderosae*). *Mol Ecol*. 26:6071–6084.

Edwards SV, Beerli P. 2000. Gene divergence, population divergence, and the variance in coalescent time in phylogeographic studies. *Evolution*. 54:1839–1854.

Excoffier L, Dupanloup I, Huerta-Sánchez E, Sousa VC, Foll M. 2013. Robust demographic inference from genomic and SNP data. *PLoS Genet*. 9:e1003905.

Fang B, Merilä J, Matschiner M, Momigliano P. 2020. Estimating uncertainty in divergence times among three-spined stickleback clades using the multispecies coalescent. *Mol Phylogenet Evol*. 142:106646.

Gaggiotti OE, Excoffier L. 2000. A simple method for removing the effect of a bottleneck and unequal population sizes on pairwise genetic distances. *Proc. B*. 267:81–87.

Haller BC, Galloway J, Kelleher J, Messer PW, Ralph PL. 2019. Tree-sequence recording in SLiM opens new horizons for forward-time simulation of whole genomes. *Mol Ecol Resour*. 19:552–566.

Haller BC, Messer PW. 2019. SLiM 3: forward genetic simulations beyond the wright-fisher model. *Mol Biol Evol*. 36:632–637.

Hancock ZB, Hardin FO, Light JE. 2019. Phylogeography of sand-burrowing amphipods (Haustoriidae) supports an ancient suture zone in the Gulf of Mexico. *J. Biogeo*. 46:2532–2547.

Hare MP, Nunney L, Schwartz MK, Ruzzante DE, Burford M, Waples RS, Ruegg K, Palstra F. 2011. Understanding and estimating effective population size for practical application in marine species management. *Conserv Biol*. 25:438–449.

Hasegawa M, Kishino H, Yano T. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J Mol Evol*. 22:160–174.

Heath TA, Huelsenbeck JP, Stadler T. 2014. The fossilized birth-death process for coherent calibration of divergence-time estimates. *Proc Natl Acad Sci U S A*. 111:E2957–E2966.

Heled J, Drummond AJ. 2010. Bayesian inference of species trees from multilocus data. *Mol Biol Evol*. 27:570–580.

Herman A, Brandvain Y, Weagley J, Jeffery WR, Keene AC, Kono TJY, Bilandžija H, Borowsky R, Espinasa L, O'Quin K, *et al.* 2018. The role of gene flow in rapid and repeated evolution of cave-related traits in Mexican tetra, Astyanax mexicanus. *Mol Ecol*. 27:4397–4416.

Hey J. 1991. A multi-dimensional coalescent process applied to multi-allelic selection models and migration models. *Theor Popul Biol*. 39:30–48.

Kelleher J, Etheridge AM, McVean G. 2016. Efficient coalescent simulation and genealogical analysis for large sample sizes. *PLoS Comput Biol*. 12:e1004842.

Kimura M. 1953. "Stepping Stone" model of population. *Ann. Rept. Nat. Inst. Genetics, Japan*. 3:62–63.

Kimura M. 1968. Evolutionary rate at the molecular level. *Nature*. 217:624–626.

Kimura M. 1983. *The neutral theory of molecular evolution*. Cambridge (UK): Cambridge University Press.

Kimura M, Weiss G. 1964. The stepping stone model of population structure and the decrease of genetic correlation with distance. *Genetics*. 49:561–576.

King JL, Jukes TH. 1969. Non-Darwinian evolution. *Science*. 164:788–798.

Knowlton N, Weigt LA. 1998. New dates and new rates for divergence across the Isthmus of Panama. *Proc. B*. 265:2257–2263.

Kubatko LS, Degnan JH. 2007. Inconsistency of phylogenetic estimates from concatenated data under coalescence. *Syst Biol*. 56:17–24.

Leaché A, Bouckaert R. 2018. Species trees estimation with SNAPP: a tutorial and example. Workshop on Population and Speciation Genomics; Český Krumlov, Czech Republic. Available from: http://evomicsorg.wpengine.netdna-cdn.com/wp-content/uploads/2018/01/BFD-tutorial-1.pdf.

Lepage T, Bryant D, Philippe H, Lartillot N. 2007. A general comparison of relaxed molecular clock models. *Mol Biol Evol*. 24:2669–2680.

Leslie MS, Morin PA. 2018. Structure and phylogeography of two tropical predators, spinner (*Stenella longirostris*) and pantropical spotted (*S. attenuata*) dolphins, from SNP data. *R Soc Open Sci*. 5:171615.

Li J, Huang JP, Sukumaran J, Knowles LL. 2018. Microevolutionary processes impact macroevolutionary patterns. *BMC Evol Biol*. 18:123.

Lynch M. 1990. The rate of morphological evolution in mammals from the standpoint of the neutral expectation. *Am. Nat*. 136:727–741.

Lynch M. 2007. *The origins of genome architecture*. Sunderland (MA): Sinauer Associates, Inc.

Magallón S. 2010. Using fossils to break long branches in molecular dating: a comparison of relaxed clocks applied to the origin of angiosperms. *Syst Biol*. 59:384–399.

Malécot G. 1968. The mathematics of heredity. Translated from the French edition (Paris, 1948), Yermanos DM, editor. San Francisco (CA): Freeman.

Manthey JD, Klicka J, Spellman GM. 2015. Chromosomal patterns of diversity and differentiation in creepers: a next-gen phylogeographic investigation of Certhia americana. *Heredity (Edinb)*. 115:165–172.

Manthey JD, Moyle RG, Gawin DF, Rahman MA, Ramji MFS, Sheldon FH. 2017. Genomic phylogeography of the endemic Mountain Black-eye of Borneo (*Chlorocharis emiliae*): montane and lowland populations differ in patterns of Pleistocene diversification. *J. Bio*. 44:2272–2283.

Maruyama T. 1970a. Effective number of alleles in a subdivided population. *Theor Popul Biol*. 1:273–306.

Maruyama T. 1970b. The rate of decrease of heterozygosity in a population occupying a circular or linear habitat. *Genetics*. 67:437–454.

Maruyama T. 1971. Analysis of population structure: II. Two-dimensional stepping stone models of finite length and other geographically structured populations. *Ann. Hum. Gen., Lon*. 35:179–196.

Meirmans PG. 2012. The trouble with isolation by distance. *Mol Ecol*. 21:2839–2846.

Miller MA, Pfeiffer W, Schwartz T. 2010. Creating the CIPRES science gateway for inference of large phylogenetic trees. Proceedings of the Gateway Computing Environments Workshop (GCE); 2010 November 14; New Orleans, LA. New Orleans, (LA): Proceeding of the Gateway Computing Environments (GCE).

Nei M, Li WH. 1979. Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proc Natl Acad Sci U S A*. 76:5269–5273.

Nei M, Takahata N. 1993. Effective population size, genetic diversity, and coalescence time in subdivided populations. *J Mol Evol*. 37:240–244.

Peterson MA, Denno RF. 1998. The influence of dispersal and diet breadth on patterns of genetic isolation by distance in phytophagous insects. *Am Nat*. 152:428–446.

Pond SK, Muse SV. 2005. Site-to-site variation of synonymous substitution rates. *Mol Biol Evol*. 22:2375–2385.

R Core Team. 2019. *R: a language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Riddle BR, Hafner DJ, Alexander LF, Jaeger JR. 2000. Cryptic vicariance in the historical assembly of a Baja California peninsular desert biota. *Proc Natl Acad Sci U S A*. 97:14438–14443.

Rieman BE, Allendorf FW. 2011. Effective population size and genetic conservation criteria for bull trout. *N Am J Fish Manag*. 21:756–764.

Rosenberg NA, Feldman MW. 2002. The relationship between coalescent times and population divergence times. In: Slatkin M and Veuille M, editors. *Modern developments in theoretical population genetics*. New York: Oxford University Press.

Slatkin M. 1991. Inbreeding coefficients and coalescence times. *Genet Res*. 58:167–175.

Slatkin M. 2005. Seeing ghosts: the effect of unsampled populations on migration rates estimated for sampled populations. *Mol Ecol*. 14:67–73.

Spalink D, MacKay R, Sytsma KJ. 2019. Phylogeography, population genetics and distribution modelling reveal vulnerability of *Scirpus longii* (Cyperaceae) and the Atlantic Coastal Plain Flora to climate change. *Mol Ecol*. 28:2046–2061.

Strange M, Sánchez-Villagra MR, Salzburger W, Matschiner M. 2018. Bayesian divergence-time estimation with genome-wide single-nucleotide polymorphism data of sea catfish (Ariidae) supports Miocene closure of the Panamanian Isthmus. *Sys. Bio*. 67:681–699.

Turner TF, Wares JP, Gold JR. 2002. Genetic effective size is three orders of magnitude smaller than adult census size in an abundant, Estuarine-dependent marine fish (Sciaenops ocellatus). *Genetics*. 162:1329–1339.

Van Bocxlaer I, Roelants K, Biju SD, Nagaraju J, Bossuyt F. 2006. Late Cretaceous vicariance in Gondwanan amphibians. *PLoS One*. 1:e74.

Wakeley J. 1998. Segregating sites in Wright's island model. *Theor Popul Biol*. 53:166–175.

Wakeley J. 1999. Nonequilibrium migration in human history. *Genetics*. 153:1863–1871.

Wakeley J. 2009. Coalescent theory. Greenwood Village (CO): Roberts and Company Publishers.

Wakeley J. 2000. The effects of subdivision on the genetic divergence of populations and species. *Evolution*. 54:1092–1101.

Weir JT, Schluter D. 2008. Calibrating the avian molecular clock. *Mol Ecol*. 17:2321–2328.

Whitlock MC, McCauley DE. 1999. Indirect measures of gene flow and migration: FST not equal to 1/(4Nm + 1). *Heredity (Edinb)*. 82 (Pt 2):117–125.

Wilkins JF, Wakeley J. 2002. The coalescent in a continuous, finite, linear population. *Genetics*. 161:873–888.

Wright S. 1931. Evolution in Mendelian populations. *Genetics*. 16:97–159.

Wright S. 1943. Isolation by distance. *Genetics*. 28:114–138.

Yu G, Smith DK, Zhu H, Guan Y, Lam TTY. 2017. ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Meth. Eco. Evo*. 8:28–36.