Phylogeny Estimation by Integration over Isolation with Migration Models

Jody Hey,*,[1] Yujin Chung,[†,1] Arun Sethuraman,[‡,1] Joseph Lachance,[§,2] Sarah Tishkoff,[2] Vitor C. Sousa,[¶,3] and Yong Wang[‖,3]

[1]Department of Biology, Center for Computational Genetics and Genomics, Temple University, Philadelphia, PA

[2]Department of Genetics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA

[3]Department of Genetics, Rutgers, the State University of New Jersey, Piscataway, NJ

[†]Present address: The Department of Applied Statistics, Kyonggi University, Suwon, South Korea
[‡]Present address: Department of Biological Sciences, California State University San Marcos, San Marcos, CA
[§]Present address: Georgia Institute of Technology, Atlanta, GA
[¶]Present address: University of Lisbon, Lisboa, Portugal
[‖]Present address: Ancestry, San Francisco, CA

*Corresponding author: E-mail: hey@temple.edu.

Associate editor: Yuseob Kim

Abstract

Phylogeny estimation is difficult for closely related populations and species, especially if they have been exchanging genes. We present a hierarchical Bayesian, Markov-chain Monte Carlo method with a state space that includes all possible phylogenies in a full Isolation-with-Migration model framework. The method is based on a new type of genealogy augmentation called a "hidden genealogy" that enables efficient updating of the phylogeny. This is the first likelihood-based method to fully incorporate directional gene flow and genetic drift for estimation of a species or population phylogeny. Application to human hunter-gatherer populations from Africa revealed a clear phylogenetic history, with strong support for gene exchange with an unsampled ghost population, and relatively ancient divergence between a ghost population and modern human populations, consistent with human/archaic divergence. In contrast, a study of five chimpanzee populations reveals a clear phylogeny with several pairs of populations having exchanged DNA, but does not support a history with an unsampled ghost population.

Key words: isolation with migration, speciation, population genetics, coalescent, gene flow, divergence.

Introduction

Phylogeny estimation is difficult for closely related species or populations, not simply because branching events were recent, but also because of population genetic factors of incomplete lineage sorting (ILS) and gene flow between populations. ILS is the byproduct of normal genetic drift that causes gene trees to coalesce in ways that appear inconsistent with the phylogeny; and it is a major complication when speciation events have occurred close together in time, or very recently (Neigel and Avise 1986; Pamilo and Nei 1988; Takahata 1989). The challenges of phylogenetic inference in the face of ILS can be addressed by using a multispecies coalescent framework (Liu and Pearl 2007; Kubatko et al. 2009; Liu et al. 2009; Heled and Drummond 2010; Bryant et al. 2012; Yang 2015; Rannala and Yang 2017). And recently there have appeared methods for phylogenetic inference under a multispecies coalescent that include rare migration (Jones 2018) or discrete admixture, or hybridization events between populations (Zhang et al. 2017; Wen and Nakhleh 2018). However, there do not exist model-based methods for estimating species phylogenies that include continuous gene flow, and the lack of such methods has long been recognized as a significant limitation (Degnan and Rosenberg 2009; Liu et al. 2009; Leaché et al. 2014; Xu and Yang 2016).

Gene flow can have a very large impact on speciation and divergence (Pinho and Hey 2010), and it can complicate phylogeny estimation in multiple ways. 1) Gene flow removes information about divergence and even a small amount of regular gene flow over many generations can make two populations appear closely related (Wright 1931). 2) Unidirectional gene flow can create asymmetries in patterns of variation such that a population that contributes genes to another may appear to be derived from the receiving population (DeSalle and Giddings 1986). 3) Even if gene flow has not occurred among sampled populations, phylogeny estimation for sampled populations may be disrupted if exchange has occurred with unsampled populations; for example if two phylogenetically distantly related sampled populations have each received genes from the same unsampled population, then they may appear to be phylogenetically closely related.

Notwithstanding these challenges, it is possible to model the movement of genes between populations in a probabilistic framework. By altering the sample configuration across populations in the genealogical history, gene flow alters the coalescent rates and thus changes the depth and structure of

Open Access

gene genealogies (Takahata 1988). Because we can calculate the probabilities of coalescent genealogies that fall within a phylogeny, with associated demographic parameters (Hey 2010b), it is possible to take a model-based and likelihood-based approach to demographic inference using coalescent theory that includes gene flow. Here we describe a new phylogenetic method that uses a formal parameterization of divergence in the presence of gene flow known as an Isolation-with-Migration (IM) model (Wakeley and Hey 1998). Such models have been implemented in numerous methods for studying divergence (Hey and Nielsen 2004; Becquet and Przeworski 2007; Gutenkunst et al. 2009; Lopes et al. 2009; Hey 2010b; Mailund et al. 2012; Dalquen et al. 2016; Chung and Hey 2017).

Results

Methods Overview

To connect sequence data to population genetic models, it has become common to approximate an integration over gene trees or genealogies using a Markov chain Monte-Carlo (MCMC) simulation (Kuhner et al. 1995; Wilson and Balding 1998; Beerli and Felsenstein 1999; Kuhner 2009). This approach has been adapted to include gene flow between populations in an IM model (Nielsen and Wakeley 2001), and such methods have now been extended to allow the analysis of multiple populations, albeit only for the case of a fixed phylogenetic topology (Hey 2010b; Dalquen et al. 2016; Chung and Hey 2017). Here we present the first method for phylogeny estimation that includes directional gene flow in a fully model-based Bayesian framework. Throughout this paper we use "phylogeny" to refer to the species tree (or the population tree in the case of multiple populations of a species) that is the tree representing historical relationships among species or populations, and we use "genealogy" to refer to a coalescent tree for sampled gene copies at a locus.

We developed a new augmentation of the genealogy (fig. 1) in which each locus is represented by a hidden genealogy, G_H, that consists of a bifurcating genealogy with zero or more migration events in an island model in which the sampled populations exist for infinite time (Wright 1951). A hidden genealogy is not associated with any phylogeny, but when it is overlaid by a branching phylogeny, some migration events are masked and are not relevant, given that topology, because they occur between populations that are in a single ancestral population in that topology. We call such migration events, that are obscured by ancestral populations in a phylogeny, *hidden migrations*. A given augmented genealogy, such as that shown in the upper left of figure 1, can be overlaid by any phylogeny to reveal a conventional genealogy. Hidden migrations do not represent a real evolutionary process and do not enter into likelihood calculations. What they provide is access to a much simpler Metropolis-Hastings update of the population phylogeny than would otherwise be possible. A change to the phylogeny does not require an update to the hidden genealogy, and as a result, it is possible

to have reasonable update acceptance rates and MCMC mixing even with relatively large numbers of loci.

We developed a program (IMa3) that implements an MCMC simulation for an IM model with multiple populations, with uniform prior on population topology, and with user specified priors on effective population sizes, migration rates and population splitting times. In this model the phylogenetic topology is rooted, and the sequence of internal nodes is ordered in time such that, for example, a tree in which populations 1 and 2 join more recently than do 3 and 4, is distinct from one in which populations 3 and 4 join more recently than the junction of 1 and 2. (Edwards 1970) identified such trees as "labelled histories." The method uses the new hidden genealogy approach to update phylogenetic trees, in conjunction with analytic integration over demographic parameters (Hey and Nielsen 2007). Because the number of demographic parameters becomes quite large, and to facilitate mixing of the Markov chain simulation, we implemented hierarchical prior distributions in which the user specifies the hyperprior densities for the parameters of population size and migration rate prior distributions.

Because the ancestral populations in the model change as phylogenetic topology changes over the course of an IMa3 run, it is expected that most analyses of phylogeny and demography will be done, as we did for the empirical data sets reported here, as two separate steps. First, a run is conducted to estimate the marginal posterior probability distribution of topologies, from which we obtain the phylogeny with the greatest probability that we use as our estimate, $\hat{\Phi}$; followed by a second run to estimate demography that is done with the phylogenetic topology fixed at $\hat{\Phi}$.

Simulation Results

To assess performance of the method as population splitting times converge, we simulated 50 single locus data sets for each of a series of splitting time values, in a 3-population model with gene-flow. Figure 2a shows the mean posterior probabilities for each splitting time. In each case the true topology is (2,(0, 1)) and the splitting time between 2 and the ancestor of 0 and 1, $t_{2,(0,1)}$, was set to 1.0, whereas the splitting time between populations 0 and 1, $t_{0,1}$, varied between 0 and 1 (see Materials and Methods for details on parameter units). As expected the mean posterior probabilities suggest strong support for the true tree when $t_{0,1} = 0.0$ and they suggest even support for all three trees when $t_{0,1} = t_{2,(0,1)} = 1.0$. Figure 2a (right axis) also shows a similar result for the proportion of the maximum a posteriori (MAP) trees that match the true tree.

To assess the effect of migration that tends to obscure the true phylogeny, we simulated 50 single locus data sets with gene exchange between two populations that are not sisters. In figure 2b, we see that as the population migration rate, $2Nm$, goes from 0 to 10 between populations 1 and 2 (the true tree is (2,(0, 1))), support for the true tree drops and support for (0,(1, 2)) rises. As the true population migration rate exceeds the upper bound on the prior on migration, 1.0,
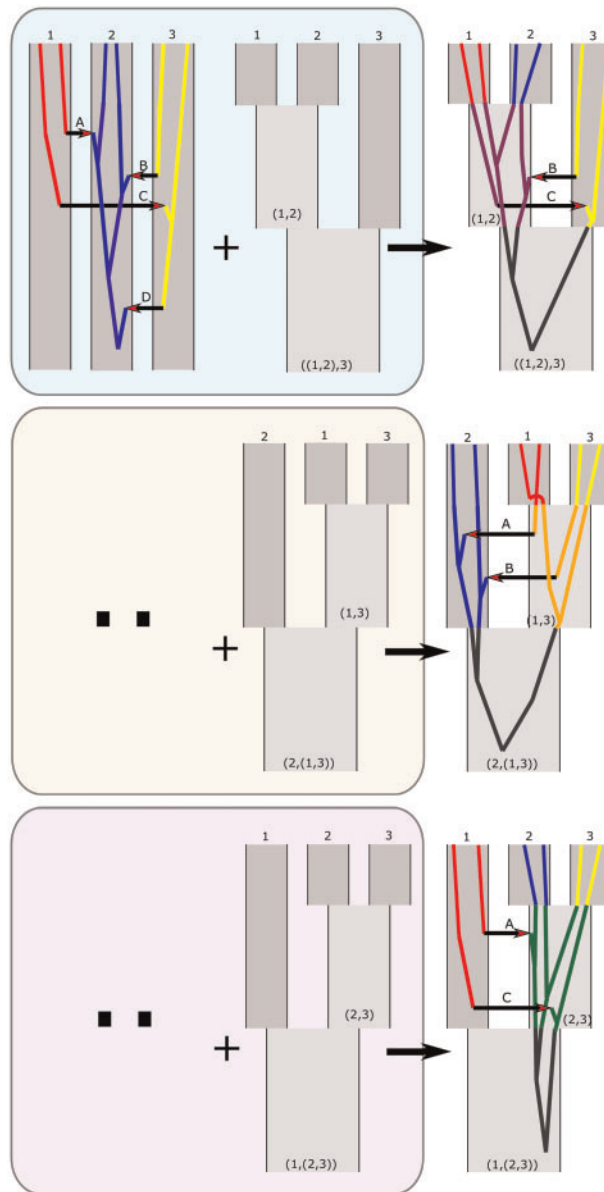
FIG. 1. Phylogenies and Hidden Genealogies. The upper left panel shows a hidden genealogy in an island model of 3 populations adjacent to a phylogeny in which species 1 and 2 are most closely related. The operation of overlaying the phylogeny on the hidden genealogy generates the genealogy shown within a phylogeny on the right side of the upper panel. This operation leave some migration events irrelevant (hidden) because they occur between two populations that are not present in the phylogeny at the time of the migration event. In the middle row the same kind of operation is shown, using the same hidden genealogy as in the top row, but with a different phylogeny that causes a different genealogy (note that the order in which the populations are listed changes in this phylogeny). A third example with the same hidden genealogy is shown in the lower panel.

support for the incorrect tree exceeds the support for the true tree.

To assess the effect of increasing the number of loci, of widening the migration prior, and of using hyperpriors, we simulated multiple data sets, each of 50 loci, for a 4-population model. With four populations, there are 18 distinct ordered topologies, so results are shown for the true tree, the two trees most similar to the true tree, and all other trees. The results for a narrow prior on migration rates are shown in figure 3A. When a wider uniform prior on migration rate is used (fig. 3B), the posterior probabilities for topologies are

considerably flatter than when a narrow prior is used, as expected for these data sets which were simulated without any migration. Figure 3B also shows a benefit, in terms of increased posterior probability of the true tree, when using hyperprior distributions for population size parameters and migration rate parameters.

We also assessed IMa3 performance on a challenging phylogenetic problem with seven populations and substantial gene flow under an IM model, and we compared performance to that found using other methods on the same data. With seven populations, IMa3 estimates an ordered
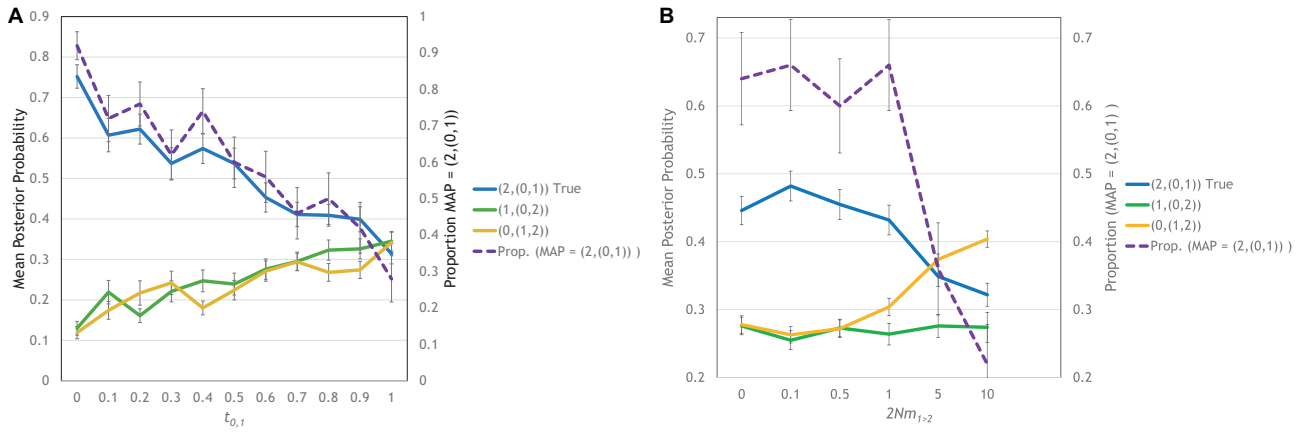
Fig. 2. Phylogeny estimation while varying splitting time and migration. (A) A 3 population model was examined as $t_{0,1}$ varies from 0 to 1, where $t_{2,(0,1)} = 1$ and $(2,(0,1))$ is the true phylogeny. Means and standard errors of estimated posterior probabilities for each phylogeny for 50 data sets simulated under each $t_{0,1}$ value are shown on the left axis. The proportion of maximum a posteriori (MAP) trees (out of 50) that matched the tree is shown on the right axis. (B) A 3 population model with migration that tends to obscure the true phylogenic topology, $(2,(0,1))$. Mean and standard error of estimated posterior probabilities (left axis) for each phylogeny for 50 data sets simulated under each of a range of $2Nm_{1>2}$ values. The proportion of MAP trees matching the true tree is on the right axis.
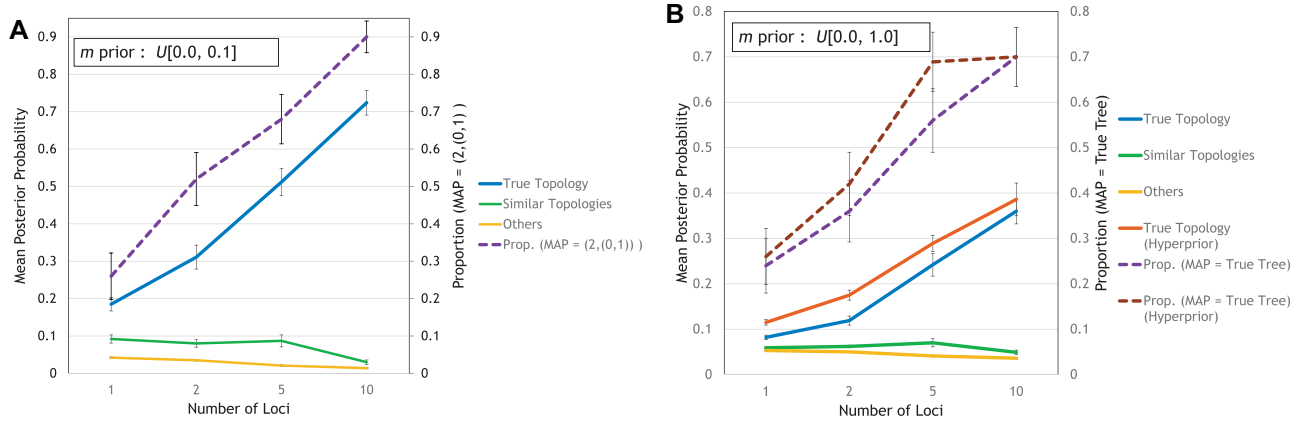

Fig. 3. Phylogeny estimation with four populations and zero migration, while varying migration priors, numbers of loci, and use of hyperprior distributions. Means and standard errors are shown for posterior probabilities (left axis) of phylogenies for 50 data sets simulated for each number of loci under a fixed phylogenetic topology, $((0, 1)4,(2, 3)5)6$ where the ancestor populations (4, 5, and 6) are ordered in time (i.e., populations 0 and 1 split most recently, followed by 2 and 3). In each panel the mean posterior probability is shown for the estimated posteriors for the true tree, the mean of the two most similar trees $((2,(3,(0, 1)4)5)6$ and $(3,(2,(0, 1)4)5)6)$, and the mean of all other trees. The proportion of MAP trees matching the true tree is on the right axis. (A) Migration rate priors have a $U[0, 0.1]$ distribution. (B) Migration rate priors have a $U[0, 1]$ distribution. Also shown for the true tree are results using a hyperprior distribution for drift hyperparameters of U[0, 20] and for migration hyperparameters U[0.0, 1.0].

topology by effectively integrating over all branch lengths, 13 population size parameters, and 72 migration rates (Hey 2010b). Twenty data sets, each of 50 loci, were generated with the ms program (Hudson 2002), with each data set having a randomly sampled topology, as well as having each population size and migration rate parameter sampled randomly from a uniform prior. Phylogeny estimates were also obtained on these same data using Neighbor-Joining (Saitou and Nei 1987) on distance matrices made using F_{st} values and net d_{xy} values (Nei 1987), and using the TreeMix program (Pickrell and Pritchard 2012). Because TreeMix is a model-based phylogeny estimation program that can handle large amounts of data, we also simulated data sets with 10,000

loci under the same IM models used for the 50-locus data sets and ran TreeMix on these. In contrast to IMa3, these other methods generate unrooted phylogeny estimates. We also compared IMa3 to other MCMC genealogy samplers, BPP (Yang 2015) and StarBEAST2 (Ogilvie et al. 2017), that implement the multi-species coalescent, but without gene flow. Like IMa3, these methods generate estimates of rooted topologies.

Table 1 shows results for the 7-population simulations in terms of the numbers of internal edges in estimated trees that are found in the true phylogeny. For the methods that return a rooted tree, the correct tree will match five internal edges when there are seven external nodes (populations).

Table 1. Counts of the Number of Simulated Seven-Population Data Sets that Matched Estimated Phylogenies (20 total).[a]

		Number of Correct Internal Edges					
	Method	5	4	3	2	1	0
Methods that provide a rooted tree	IMa3	19	1	0	0	0	0
	BPP	4	3	8	3	0	2
	*Beast2	2	8	7	3	0	0
Methods that provide an unrooted tree	NJ-Fst	—	6	8	5	0	0
	NJ-netDxy	—	6	9	4	0	1
	Treemix	—	5	8	7	0	0
	Treemix (10,000 loci)	—	18	2	0	0	0

[a]Counts are ordered by method, and by number of correct internal edges. Unrooted trees cannot share >4 internal edges with the true rooted tree.

IMa3 returned the true tree in 19 out of 20 data sets. The other MCMC methods, that do not include migration in their models, did not perform nearly as well. The true model for these simulations had large amounts of gene flow, which would account for the relatively poor performance of the methods that assume no migration. In other studies, analyses of simulated data have shown relatively good performance of such methods when gene flow is modest (Leaché et al. 2014). Among the methods that do not provide a rooted tree estimate, performance was not very good on these high gene-flow problems, with one important exception. When TreeMix was run on data sets with 10,000 loci simulated under the same IM models as used for the 50 locus data sets, it performed well and returned the correct unrooted tree the large majority of the time.

Human Hunter Gatherer Analyses

We studied the phylogenetic and demographic history of three African hunter gatherer populations, including Baka Pygmies from Cameroon, and Hadza and Sandawe, both click-language speakers, from Tanzania (Lachance et al. 2012). We also included a sample of Yorubans from Nigeria as representative Bantu-speaking nonhunter gatherers. As these represent only a small fraction of human populations in Africa, and given the probability that these populations have exchanged genes with others that are not in the analysis (Wall 2000; Plagnol and Wall 2006; Hammer et al. 2011; Henn et al. 2011), we were particularly interested in the comparison between analyses with and without an unsampled ghost population that was set to be a sister population to the clade of sampled populations (i.e., an outgroup population). The ability to include an unsampled population is a major advantage of some coalescent-based methods (Beerli 2004). Gene flow between sampled and unsampled populations can have a large effect on the pattern of variation within sampled populations and can in turn disrupt estimates of phylogenetic and demographic history if that gene flow is not accounted for.

Both analyses supported a common phylogeny in which the Tanzanian populations are most closely related, (Baka, [Yoruba,{Hadza, Sandawe}]), which is the tree supported in the original analyses (Lachance et al. 2012). However, support for this topology was stronger in the analysis with a ghost population, with the estimated posterior probability for the most strongly supported tree at 0.42 for the no-ghost model,

and 0.86 for the ghost model (see supplementary tables 1 and 2, Supplementary Material online). To assess whether the model with the ghost performs better, in IMa3 runs conditioned on the estimated phylogeny, we implemented a thermodynamic integration approach (Lartillot and Philippe 2006) to estimate the marginal likelihood and ran this under both the model with the ghost and without. The logs of the marginal likelihoods suggest a Bayes factor that strongly favors the ghost model (estimated marginal likelihood for the ghost model: −4,885; nonghost model: −4,905).

Figure 4 shows the estimated histories, for IMa3 runs using the estimated phylogeny, with splitting times for the sampled populations ranging from about 20 to 60 KYA. Effective population sizes for sampled populations and their ancestors ranged from 29,000 for the Yoruba to 3,000 for the Hadza (supplementary table 5, Supplementary Material online), which were known to harbor low levels of variation (Henn et al. 2011). None of the migration rate parameters shared by sampled populations were found to differ significantly from zero, however, there were several strong and significant signals of gene flow from the ghost population.

For the sampled populations, the ghost can best be considered as reflecting the existence of multiple unsampled African populations, that have themselves experienced gene exchange. The large effective population size of the ghost (33,000 fig. 4 and supplementary table 6, Supplementary Material online) is consistent with the ghost actually representing a large structured population. However for the ancient parts of this model, much older than the common ancestry of the sampled populations, the common ancestry time of the ghost and the ancestor of the sampled populations (850 KYA) suggests the presence of a sister population to the ancestor of modern humans that exchanged genes for some period of time. A plausible candidate for that sister population would be that which ultimately gave rise to Neanderthals. Our estimated time and broad confidence interval is not inconsistent with other population genetic-based estimates of human/Neanderthal divergence, which range from 550 to 800 KYA (Beerli and Edwards 2003; Prufer et al. 2014; Mendez et al. 2016).

Chimpanzee Analyses

The subspecies of the common chimpanzee (*Pan troglodytes*) share much of their genetic variation, with most estimates of divergence times falling within the past half million years
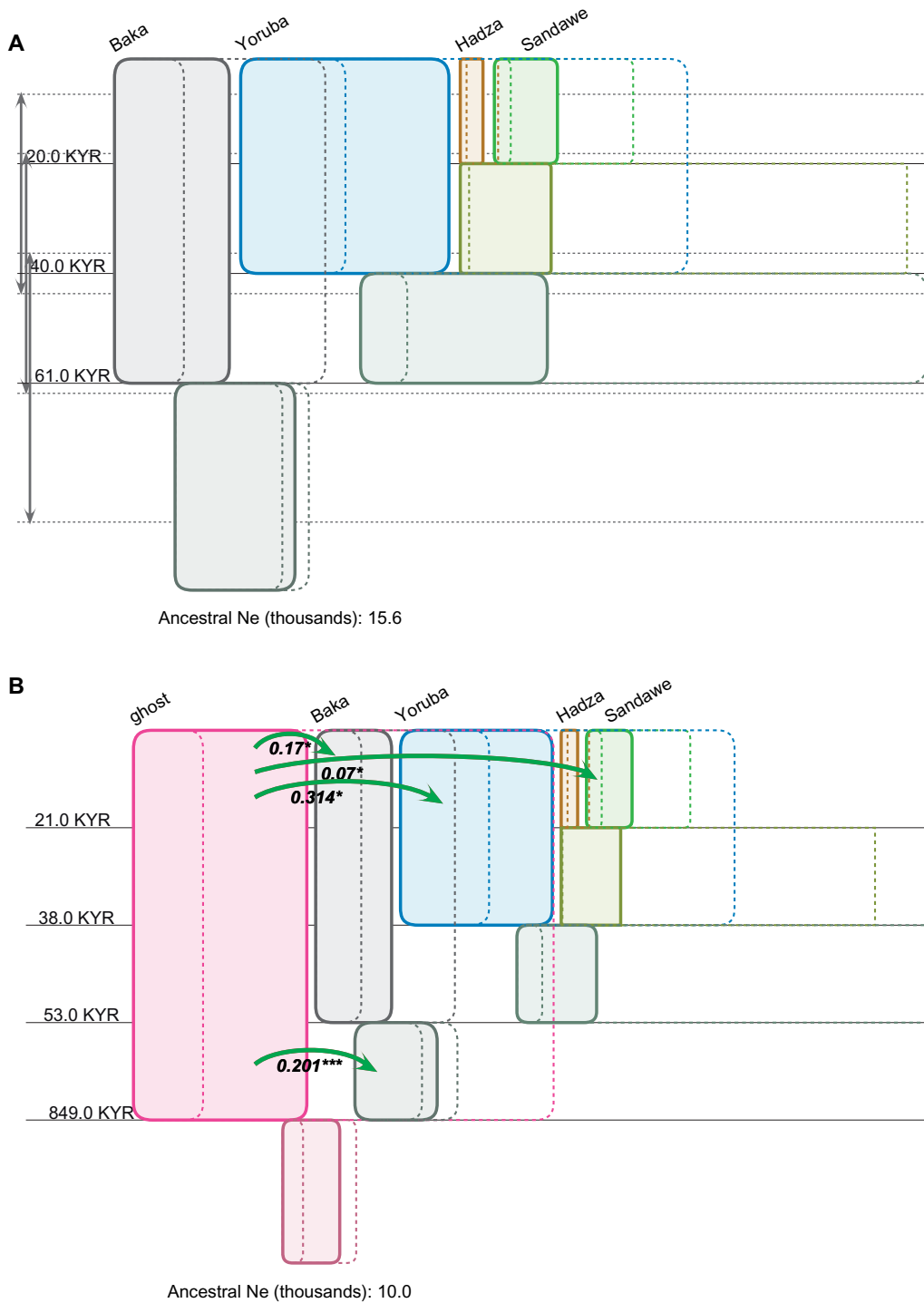
FIG. 4. Estimated histories for human populations. Boxes represent populations, with widths proportional to estimated effective population sizes (ancestral N_e is given for scale). Confidence intervals are indicated as dashed-line boxes aligned with the corresponding population's box on the left side. Estimated population migration ($2N_e m$)rates that are associated with a migration rate significantly >0 based on a marginal likelihood ratio test (Nielsen and Wakeley 2001) are shown together with their estimated $2N_e m$ values (*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$) Migration rates not significantly different from zero at $p < 0.01$ are not shown (a) Without a ghost population. Estimated splitting times are shown to scale with 95% confidence intervals. No migration rates were significantly different from zero. (b) With a ghost population. Splitting times are shown evenly distributed because of the great depth of the first split. The 95% confidence intervals for three recent splits are similar to part figure "a". Confidence interval for the oldest split was 554–1,663 KYA.

(Caswell et al. 2008; Hey 2010a; Prado-Martinez et al. 2013; de Manuel et al. 2016). Several studies have reported evidence of gene flow among subspecies based upon population genetic analyses (Won and Hey 2005; Becquet et al. 2007;

Prado-Martinez et al. 2013; de Manuel et al. 2016). The bonobo (*Pan paniscus*) is thought to have diverged from the ancestor of the common chimpanzee over 1 Ma, and they show some evidence of gene flow in the past with the

common chimpanzee (de Manuel et al. 2016). We used whole genome assemblies for 3–5 individuals per species/subspecies from the Great Ape Genome Project (Prado-Martinez et al. 2013).

For a five population IM model, the phylogenetic component has four population splitting times and 180 distinct ordered topologies, whereas the demographic component has 9 population size parameters and 32 migration rate parameters. We also ran the data under a 6 population model, with one unsampled "ghost" population (Beerli 2004) included as an outgroup to the sampled populations. The estimated histories are shown in figure 5 with the topology having the highest estimated posterior probability (0.956). The second most probable tree had the same overall topology but with the order of the first two splitting events reversed in time (see supplementary table 1, Supplementary Material online). This is the same tree estimated with various distance metrics for a microsatellite data set (Gonder et al. 2011) and later with full genome data (Prado-Martinez et al. 2013). Overall, the splitting time and effective population size estimates are quite consistent with previous estimates (Hey 2010a), with the largest exception being the recent divergence time of the eastern and central subspecies (Pan t. schweinfurthii and Pan t. troglodytes 13,000 years, with 95% CI range: 4,200–51,600 years). Previous estimates for this pair of subspecies, using an IM framework, ranged from 32,000 to 46,000 (Hey 2010a).

Figure 5 also shows the 2Nm values and the directions of migration parameters that are statistically significantly different from zero (at the $p < 0.01$ level) using a likelihood ratio test (Nielsen and Wakeley 2001). Five of the 32 migration parameters were significant based on this test. Three of these involve Pan t. ellioti and its geographical neighbors, Pan t. verus to the west and the ancestor of Pan t. troglodytes and Pan t. schweinfurthii to the east. The signal of low migration between the common ancestors of the common chimpanzee and Pan paniscus is also consistent with suggestions of gene flow early in the divergence of the two chimpanzee subspecies, however, we do not see evidence of more recent exchange between the species that has also been suggested (de Manuel et al. 2016). Figure 5 also shows low significant gene flow from Pan t. ellioti into Pan t. schweinfurthii, which has been previously reported based on an analyses of whole genome data using the TreeMix program on the same genomes used here (Prado-Martinez et al. 2013). If the ranges of common chimpanzee subspecies are relatively stable, as might be the case if they are primarily proscribed by large rivers (Goldberg 1998; Mitchell et al. 2015), then this signal of migration has probably been caused by some departure of the data from the models being used.

When the data were analyzed with a ghost population as an outgroup to the five sampled populations, IMa3 returned the same phylogeny as it did when run without a ghost. However, the posterior probability distribution of topologies was much flatter, with the expected tree having Pan t. schweiinfurthii and Pan t. troglodytes as most closely related, followed in time by the pair of Pan t. ellioti and Pan t. verus (posterior probability 0.303) followed very closely by a tree

with the order of these pairings reversed (supplementary table 2, Supplementary Material online). Results for the ghost model for splitting times, migration rates and effective population size estimates were all quite similar to the nonghost model, suggesting that there has not been a large impact on chimpanzee phylogenetic and demographic history from other unsampled populations. Where the ghost and nonghost histories differ the most is in the size of the chimpanzee common ancestor and the presence in the nonghost model of significant gene flow into the Bonobo. Although these differences are suggestive of the presence of one or more nonsampled populations early in the divergence of the Pan genus, we did not find that the ghost population model fit the data substantially better than the model without a ghost population. Overall support for the nonghost model was higher than for the ghost model (marginal likelihood without a ghost: −5291.8, with a ghost: −5307.9), notwithstanding the fact that the nonghost model has 21 fewer parameters than the ghost model (245 vs. 266).

Discussion

Evolutionary biologists have long recognized that one cannot well address the phylogenetic history of closely related species without simultaneously considering their population genetic history (Gillespie and Langley 1979; Tajima 1983; Felsenstein 1988; Avise 1994; Maddison 1997; Arbogast et al. 2002; Maddison and Knowles 2006). The new method presented here makes possible the simultaneous study of population genetics and phylogenetics. When run as a phylogeny estimator, the method integrates over all IM models within the bounds of the user-specified prior, and returns a posterior probability density over all ordered, rooted topologies. IMa3 overcomes much of the difficulty associated with MCMC-based genealogy samplers by implementing a new kind of genealogy augmentation method (the hidden genealogy) that minimizes the changes being made to genealogies. The method cannot overcome some challenges that inherently arise when gene exchange is present. Gene flow will contribute to populations being more similar and to an overall flattening of the posterior probabilities for both population genetic and phylogenetic components of the model. Gene flow can also generate a clear signal of the incorrect phylogeny if it exceeds the prior distribution (fig. 2), and it can cause some aspects of population genetic history to not be identifiable using genealogical models (Than et al. 2006; Sousa et al. 2011). Like many other genealogy samplers, IMa3 is limited by the assumptions that loci are separated by high recombination, whereas recombination within loci is absent.

Importantly, IMa3 makes full use of the data, and it fits a very general and widely used phylogenetic and demographic model to the data without approximation. For divergence problems where gene flow is not an issue, investigators have many tools upon which to draw; but until now this has not been the case for the many contexts where high levels of gene flow are known or likely and in which the divergence history may be refractory to conventional analyses (Nosil 2008; Pinho and Hey 2010; Leaché et al. 2014).
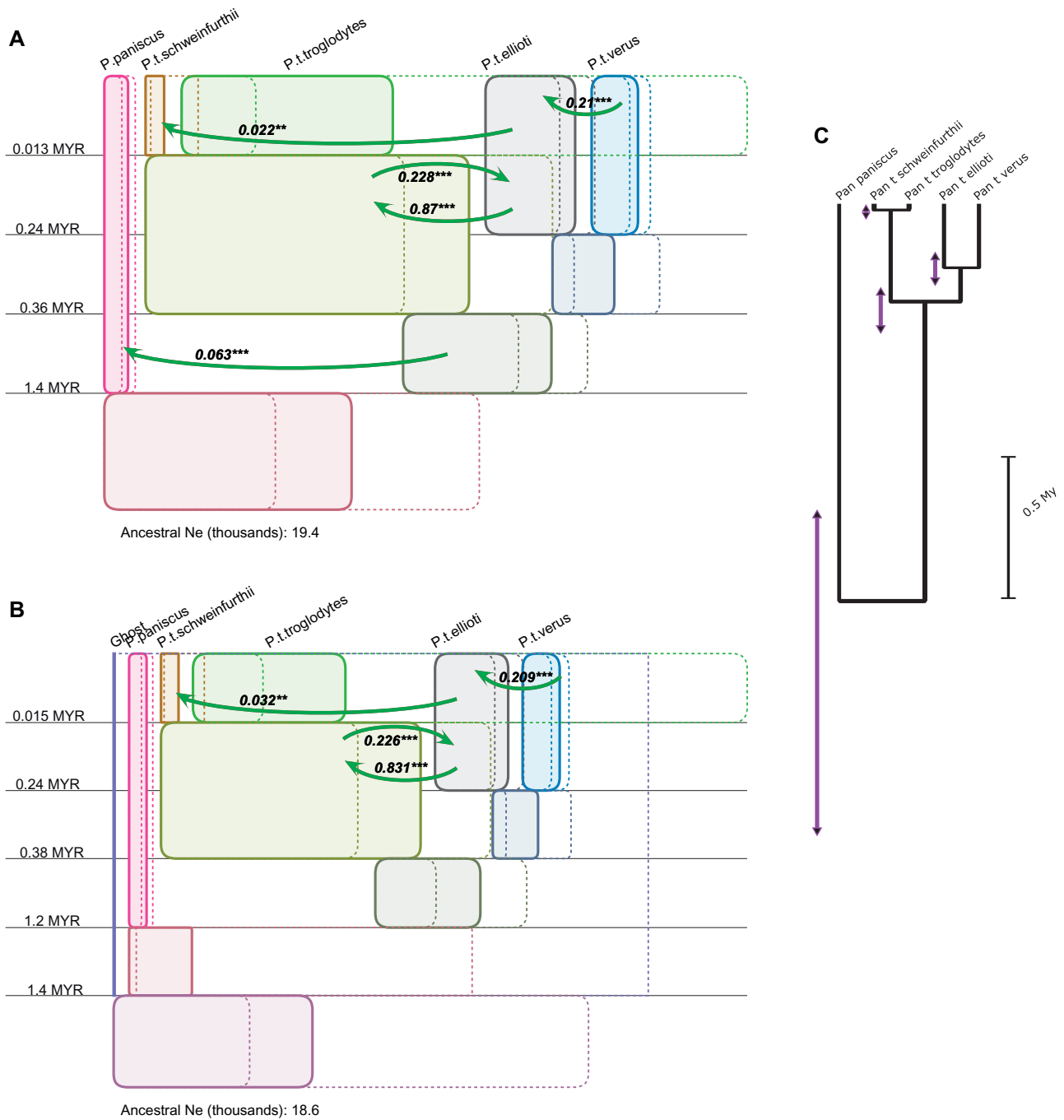
Fig. 5. Estimated phylogenetic and demographic history for bonobo and common chimpanzee subspecies. Estimated population migration ($2N_em$) rates that are associated with a migration rate significantly >0 based on a likelihood ratio test (Nielsen and Wakeley 2001) are shown together with their estimated $2N_em$ values (**$p < 0.01$; ***$p < 0.001$). Migration rates not significantly different from zero at $p < 0.01$ are not shown. Population topology is shown by the position of ancestral population boxes with respect to pairs of descendant population boxes. Estimating splitting times are given on the left, with distances not proportional to the values, but evenly distributed for clarity. (A) Without a ghost population. (B) With a ghost population. (C) The phylogeny from A drawn to scale and showing confidence intervals of splitting times.

Materials and Methods

Target Density

We use an Isolation-with-Migration (IM) model for multiple populations (Gutenkunst et al. 2009; Hey 2010b), in which sampled populations have a bifurcating phylogenetic history that includes a specified topology and periods of duration of sampled and ancestral populations, that is, branch lengths.

A model with n sampled populations will have $n - 1$ ancestral populations, and every population has an associated effective population size, N_e, as well as migration rates to and from every other population with which it coexists over a period of time, given the phylogeny. The unknowns to be estimated are partitioned into three sets, including: 1) the effective population size and migration rate parameters, Θ;

2) the rooted ultrametric population phylogeny, $\{\Phi, \tau\}$, which includes the topology Φ and the times of ancestral population splitting τ; and 3) when there are $L > 1$ loci, a set of L mutation rate scalars, u (Hey and Nielsen 2004).

The new method builds upon a multi-population version of a Bayesian method that provides a joint estimate of the elements of Θ, along with marginal estimates of the elements of u and τ, all for a fixed Φ (Hey 2010b). The method works by running a Markov chain Monte Carlo (MCMC) simulation that generates samples of genealogies, splitting times, and mutation rate scalars $\{G, \tau, u\} \sim p(G, \tau, u|X)$, (in the case when X includes data from multiple loci, G is a set of genealogies, one from each locus). Then, for a fixed Φ, Θ can be estimated by optimizing

$$p(\Theta|X) = \int\int\int p(\Theta|G, \tau, u)p(G, \tau, u|X)\ dG\ d\tau\ du$$
$$\approx \frac{1}{k}\sum_{i=1}^{k} p(\Theta|G_i, \tau_i, u_i),$$

for a sample of k values of $\{G_i, \tau_i, u_i\} \sim P(G, \tau, u|X)$.

To include Φ, not as a fixed value but as a variable in the model, we developed a new MCMC simulator to provide samples of $\{G, \tau, \Phi, u\} \sim p(G, \tau, \Phi, u|X)$. The challenge in designing a Metropolis-Hastings update of Φ is that every change in the phylogeny (either topology or branch lengths) also requires updates to the full set of genealogies (Rannala and Yang 2003; Hey and Nielsen 2004). Each instance of G includes, for all of the loci, an ultrametric branching graph that specifies the population states of all edges at all times, including the time and direction of migration events. Because Φ and τ specify which populations exist at which times, they necessarily also constrain G in terms of the kinds of migration events that occur at which times and which genealogy edges can coalesce at which times. The entanglement of phylogeny with the genealogies of all the loci leads to increasingly low acceptance rates of proposed updates to the phylogeny with larger data sets. To overcome this difficulty we developed a new kind of genealogy augmentation that minimizes those changes in G required when proposing a change in Φ.

As shown in figure 1, the new method uses a hidden genealogy, G_H, that exists in an island model in which the sampled populations persist for infinite time (Wright 1951). When the hidden genealogy is overlaid by a population phylogeny, some migration events remain relevant whereas others (hidden migrations, m_h) do not because they occur between two populations that are masked by an ancestral population in the phylogeny. Figure 1 shows how a single hidden genealogy can be masked by any phylogeny to reveal a conventional genealogy under the phylogeny.

Given a phylogeny, a hidden genealogy includes both a conventional genealogy and hidden migrations, that is, $G_H = \{G, m_H\}$. Because the probability of the data on a genealogy depends only on genealogy branch lengths and topology, and not on migration events (hidden or otherwise), $p(X|G, u) = p(X|G_H, u)$. Then the target density of the MCMC simulation is

$$p(G_H, \tau, \Phi, u|X) \propto p(X|G_H, u)\pi(G_H|\tau, \Phi)\pi(\tau, \Phi)\pi(u) =$$
$$p(X|G, u)\pi(G|\tau, \Phi)\pi(m_H|G, \tau, \Phi)\pi(\tau, \Phi)\pi(u),$$
$$(1)$$

where $\pi(G|\tau, \Phi)$ is the usual prior on genealogies found by integrating over Θ (Hey and Nielsen 2007), $\pi(m_H|G, \tau, \Phi)$ is the prior on hidden migrations (see supplementary information, Supplementary Material online), $\pi(\tau, \Phi)$ is a uniform prior on splitting time intervals (Hey 2010b) and topology, and $\pi(u)$ is the uniform prior (log scale, geometric mean of 1) of the mutation rate scalars (Hey and Nielsen 2004).

To ensure that the use of hidden genealogies does not affect parameter estimates we designed a prior density for m_H that is a proper probability distribution, that is, $\int \pi(m_H|G, \tau, \Phi, u)\ dm_H = 1$ for all $\{G, \tau, \Phi, u\}$ (see supplementary information, Supplementary Material online). By recording $\{G, \tau, \Phi, u\}$ from samples of $\{G_H, \tau, \Phi, u\}$, we can estimate the target density, $p(G, \tau, \Phi, u|X)$, with the standard likelihood, the conventional coalescent prior and the uniform priors on τ, Φ and u. A further check can be made by fixing the topology while using the new sampler with hidden genealogies, in which case the target density is the same as for a multi-population sampler that does not use hidden genealogies and that uses a fixed topology. We confirmed this by comparing results of the IMa3 program, run while using hidden genealogies to update branch lengths but not topology, with those of the IMa2 program (Hey 2010b) (results not shown).

Apart from including Φ as an unknown, all of the assumptions and parameterizations are unchanged from the first multi-locus method for demographic inference under an IM model (Hey and Nielsen 2004). Under this framework, with multiple loci, elements of Θ and τ are scaled by the geometric mean of the mutation rates, $u = \left(\prod_{i=1}^{L} u_i\right)^{1/L}$, where u_i is the mutation rate for locus i per generation (not per base pair, but for the full length of locus i). Then individual population size, migration rate, and splitting time parameters are scaled as, $4N_e u$, m/u, and tu, respectively (m is a migration rate per generation per gene copy). The mutation rate scalar for locus i is defined as $u_i = u_i/u$, and thus the geometric mean of all the mutation rate scalars is always 1 throughout the MCMC simulation. Following an analysis, estimates of parameters on demographic scales (e.g., $4N_e$ in individuals and t in generations) can be obtained using an estimate of the geometric mean of mutation rates for the loci used in the study. Population migration rate (i.e., $2N_e m$) estimates are obtained by optimization of a posterior density for $2N_e m$ that is obtained by integration over the margins of the estimated posterior density for Θ (Hey 2010b).

Hierarchical Priors

One of the challenges of Bayesian estimation of IM models is assigning prior distributions for a large number of parameters. To simplify this process and to improve parameter estimation we adopted a hierarchical framework in which the prior

distributions for the demographic parameters in Θ are sampled from a hyperprior distribution specified by the user. In the conventional (nonhierarchical) framework, a demographic parameter $\theta \in \Theta$ follows a uniform prior density π_q with lower bound 0 and upper bound q, as specified by the user. In the hierarchical framework, the uniform prior distributions are specified by hyperparameters that are drawn from the hyperprior density, $q \sim \pi_Q$, where π_Q follows a uniform density with lower bound 0 and upper bound Q, as specified by the user. We use two hyperprior distributions, one for the population size elements of Θ and one for the migration rate elements of Θ.

Use of hyperprior distributions means that at any point in time the state space of the Markov chain simulation includes the hyperparameters, one for each element of Θ. Additionally, because Φ is changing, only some of the possible ancestral populations are included in the model at any point in time. We allow for changing of hyperparameters for the elements of Θ when Φ changes by including in the Markov chain state space sampled hyperparameter values for all possible populations and all possible pairs of populations. These hyperparameters (i.e., the upper bounds of the uniform prior distributions) are subject to Metropolis-Hastings updates at intervals by proposing new values from the hyperprior distributions.

Any ancestral population can be identified in terms of the subset of sampled populations it is ancestral to, so the total number of possible populations (sampled and ancestral) in an n population model is the number of possible distinct subsets, or $2^n - 1$ (excluding the empty set). This is the number of population size hyperparameters that need to be included in the state space. For migration hyperparameters, the total number of possible pairs of populations can be determined by considering that for any one population pair one of the two will be ancestral to k sampled populations and the other will be ancestral to one or more populations that are not in the set of size k to which the first population is ancestral. There are $\binom{n}{k}$ possible subsets of size k, each of which will be paired with a population that could be ancestral to any subset of the complementary set of $n - k$ sampled populations. There are $2^{n-k} - 1$ of these, so the number of possible pairs of populations when either population is ancestral to k sampled populations is $\binom{n}{k}(2^{n-k} - 1)/2$. Summing over k from 1 to $n - 1$, we get a total of $(3^n - 2^{n+1} + 1)/2$ possible pairs of coexisting populations, each of which will have two hyperparameters, one for each direction, in the state space of the MCMC simulation. For updating the hyperparameters of one or more of the current elements of Θ, the Metropolis-Hastings ratio is simply $\frac{\pi(\acute{G}|\tau,\Phi)}{\pi(G|\tau,\Phi)}$, where the prior distributions for those elements of Θ, when integrating over Θ to obtain $\pi(\acute{G}|\tau,\Phi)$, are given by the newly proposed hyperparameters, and the integration for $\pi(G|\tau,\Phi)$ is done using the current hyperparameters.

Estimation
At any point in time the Markov chain simulation will be on a single phylogenetic tree topology, and this will define the populations to which the elements of Θ pertain. In other words the ancestral populations, and the migration events that involve ancestral populations, will change as Φ changes, and therefore sampled values of Θ and τ must be partitioned with respect to the Φ value with which they were sampled. To estimate both phylogeny and demography, the simplest approach is to proceed in two steps. First, estimate the marginal posterior density of topology, $p(\Phi|X)$, by conducting an MCMC run over $\{G, \tau, \Phi, u\}$ and by sampling only values of Φ. Second, given an estimate, $\hat{\Phi}$, from the estimated marginal density for the topology $\hat{p}(\Phi|X)$, $\hat{\Phi} = argmax \ \hat{p}(\Phi|X)$, that was recorded in step one, run a conventional MCMC simulation with phylogeny fixed at $\hat{\Phi}$ to obtain estimates of splitting times, population sizes and migration rates.

Program Development
We developed a new program, IMa3, that is based on the parallel version of IMa2 (Sethuraman and Hey 2016) and that implements the new methods described here. Testing of the method presents a challenge because, with one exception, we do not know the true posterior density for this model, even for the smallest of data sets with three or more populations under an IM model. The exception is a null data set, in which the probability of the data is constant across all genealogies. In this case we expect and confirmed that the program returns a posterior distribution for phylogenetic topology that is indistinguishable from the prior distribution. We are also able to confirm an important prediction of the criterion that $\int \pi(m_H|G, \tau, \Phi)dm_H = 1$ by running the program with hidden genealogies in the state space, but for a fixed phylogenetic topology. In this case the target density is the same as that for the IMa2 program (i.e., fixed topology and no use of hidden genealogies), and results should be (and were confirmed to be) indistinguishable from those found using the IMa2 program.

The IMa3 program is flexible in a number of ways: 1) it allows for nonuniform priors on phylogenetic topology; 2) it implements three widely used mutation models (infinite sites [Kimura 1969], HKY [Hasegawa et al. 1985], and stepwise [Kimura and Ohta 1978]); 3) it implements both uniform and exponential prior distributions for migration rate priors and for migration rate hyperpriors; and 4) it retains the functionality of IMa2 for conducting likelihood ratio tests of nested demographic models. IMa3 is written in C++ and can be run on multiple processors. For most data sets, the mixing of the Markov chain simulation is greatly improved by inclusion of a sequence of heated chains, with Metropolis swaps of state spaces between chains (Geyer 1991). Parallelization of IMa3 is implemented by having two or more chains per CPU, with swapping between chains coded using MPI (Altekar et al. 2004; Sethuraman and Hey 2016).

Assessment of the quality of a set of sampled phylogeny values, in terms of convergence and mixing of the MCMC simulation, presents conventional and significant challenges (Gilks and Roberts 1996). We assessed sample quality by estimating effective sample sizes (ESSs) during a run, by comparing estimated posterior distributions for samples collected

in the first and second half of a run, and by repeating runs with different random start points.

Because of parallelization it is possible to get a substantial reduction in run times by using multiple processors. However, the innovations introduced here, while making it possible to estimate phylogeny in an Isolation-with-Migration context, do not alter the underlying dynamics that caused the predecessors of the IMa3 program (i.e., IMa2 etc.) to be notoriously slow (e.g., McCormack et al. 2008; Suárez et al. 2014; Palstra et al. 2015). Nevertheless, we used relatively large numbers of loci (for a genealogy-sampling MCMC-based approach) for our analyses. The chimpanzee data took about 2 weeks on 40 CPUs to return phylogeny estimates, and about half that time to return parameter estimates when run on the estimated phylogeny. The human data set took about 4 weeks and 1 week for these runs, respectively. Each of the 50 locus, seven population data sets for simulation set 4 took about 2 days with 40 CPUs for phylogeny estimation. These times do not include the time needed for replicate runs for ensuring convergence.

Human Hunter Gatherer Data

We used the data reported by (Lachance et al. 2012) for the Baka, Hadza and Sandawe populations (3, 5, and 5 genomes, respectively) in addition to 7 Yoruban genomes (Drmanac et al. 2010). All had been sequenced to an average depth of 60X coverage and aligned to the Hg19 human reference genome using the same protocol (Drmanac et al. 2010). Data were filtered to avoid regions: with less than 5X coverage for all individuals, regions within 10,000 base pairs of RefSeq loci, regions not showing conserved human/chimpanzee synteny, recent segmental duplications, CpGs, or conserved noncoding elements (Gronau et al. 2011), and regions most subject to GC-biased gene conversion (Capra et al. 2013). To generate sampled regions that do not show evidence of recombination, sequences were phased (Stephens et al. 2001) and subsampled using the 4-gamete criterion (Hudson and Kaplan 1985) as previously described (Hey 2010a). Two hundred randomly selected autosomal regions, with a mean length of 1490 base pairs, were used for estimating the phylogeny topology. We used a mutation rate of 1.25×10^{-8} per base per generation (Scally and Durbin 2012), and a generation time of 29 years as estimated from human hunter gatherer populations. IMa3 runs were conducted for four- and five-population models (without and with a ghost, respectively). Initial runs to estimate the topology of the species tree were conducted with hidden genealogies and topology updating. Hyperpriors were set to $U[0, 5]$ for the population size (genetic drift) parameters, $U[0, 0.2]$ or $U[0, 0.5]$ for migration rates, and $U[0, 1]$ for population splitting times. These runs were done with 400 chains on 20 or 40 processors, with 24 h burnin prior to sampling, and at least 500,000 topologies were sampled. Mixing was assessed by restarting runs, and comparing estimated posterior densities for topologies on runs observed at different times. To estimate the demographic model, given a phylogeny, IMa3 was run with a fixed topology using prior distributions with widths identical to the hyperpriors. These runs were done with 400 chains on 20 processors,

with a 15-h burnin prior to sampling, and at least 15,000 sampled genealogies.

Chimpanzee Data

To study the divergence of chimpanzees, including the bonobo and four subspecies of common chimpanzee, we used whole genome assemblies with 20x coverage on average and sample sizes ranging from 3 to 5 individuals per species/subspecies from the Great Ape Genome Project (GAPA supplementary table 12.4.1, Supplementary Material online) (Prado-Martinez et al. 2013). Data were filtered in the same way as for the human hunter-gatherer data. One hundred randomly selected autosomal regions were used for estimating the phylogeny topology, and then these regions and an additional 100 loci were used for estimating the demographic history conditional on the phylogeny estimate. The mean length of the loci used was 2,030 base pairs. We used a generation time of 25 years (Langergraber et al. 2012), and the same mutation rate of 1.25×10^{-8} as for the human data set.

IMa3 runs were conducted for five- and six-population models (without and with a ghost, respectively). Initial runs to estimate the topology of the species tree were conducted with hidden genealogies and topology updating. Hyperpriors were set to $U[0, 4]$ for the population size (genetic drift) parameters, $U[0, 1]$ for migration rates, and $U[0, 2]$ for population splitting times. These runs were done with 500 chains on 40 processors, with 24 h burnin prior to sampling, and at least 50,000 trees were sampled. Mixing was assessed by restarting runs, and comparing estimated posterior densities for topologies on runs observed at different times. To estimate the demographic model, given a phylogeny, IMa3 was run with a fixed topology using and prior distributions with widths identical to the hyperpriors. These runs were done with 400 chains on 20 processors, with a 15-h burnin prior to sampling, and at least 15,000 sampled genealogies.

Simulated Data

Python scripts were written to generate command lines for the ms program (Hudson 2002) for an IM model using randomly sampled population sizes, migration rates, and phylogeny branch lengths from specified prior distributions for population trees with fixed or random topologies for multiple populations. The ms program generates data under the infinite sites model of mutation (Kimura 1969). All data were simulated using demographic parameters scaled by the mutation rate per locus per generation, u (Hey and Nielsen 2004), including: the population mutation rate for each population, $4N_e u$ (N_e is the effective population size); the migration rate per mutation, m/u (m is the rate of migration per gene copy per generation); and population splitting time, $t\ u$ (t is the time since population splitting in units of generations). Four sets of simulations were generated to address different aspects of performance, with all simulations sampling 5 gene copies per population per locus.

Set 1. To assess performance as population splitting times converge, 50 single locus data sets for three populations were simulated for each of a series of splitting time values (fig. 2). Data were simulated using drift parameters, $4N_e u$, drawn at

random for each population from a uniform prior distribution $U[0, 20]$ and migration rate parameters, m/u, in both directions between all pairs of coexisting populations, drawn at random from $U[0, 0.1]$. For the IMa3 runs, these same uniform distributions were used for priors for drift and migration terms, whereas the priors for splitting times for the IMa3 runs were set to $U[0, 2]$. For each data set IMa3 was run with a burnin of 30,000 steps, after which 50,000 topologies were sampled.

Set 2. To assess performance when migration occurs between distantly related populations, 50 single locus data sets for three populations were simulated for each of a series of migration rates from population 1 to 2, whereas the true tree, $((0, 1), 2)$, has populations 0 and 1 as most closely related (fig. 3). The migration rate, m/u, for each simulation was based on the simulated drift parameter for population 1, $4N_1u$, so as to match one of a series of specified population migration rates ($2Nm$) ranging from 0 to 10. Splitting times were fixed at 0.5 and 1.0; and drift parameters were sampled from $U[0, 20]$. IMa3 was run using the same priors, burnin and sampling rates as for the Set 1 simulations.

Set 3. To assess performance over a range of data set sizes, and varying priors, 50 data sets for either 1, 2, 5, or 10 loci were simulated under a four population model for a fixed topology of $((0, 1)4,(2, 3)5)6$, where the ancestor populations (4, 5, and 6) are ordered in time (i.e., populations 0 and 1 split most recently, followed by 2 and 3). Data were simulated using migration rates fixed at 0; with drift parameters, $4N_eu$, drawn from $U[0, 20]$; and with the three splitting times fixed at 0.5, 1.0, and 1.5, respectively (fig. 4). For the IMa3 runs without hyperpriors, the priors for drift parameters and splitting times were sampled from uniform distributions the same as for Sets 1 and 2 ($U[0, 20]$ and $U[0, 2]$, respectively). Migration rate priors were set either to a narrow ($U[0, 0.1]$) or a wide ($U[0, 1]$) distribution. For runs using hyperprior distributions the hyperparameters for the genetic drift terms were drawn from $U[0, 20]$ while those for the migration terms were drawn from $U[0, 1]$. For each data set IMa3 was run with a burnin of 60,000 steps, after which 50,000 topologies were sampled and the maximum a posteriori tree taken as the estimated topology. For the 5 and 10 locus data sets, multiple metropolis-coupled chains (Geyer 1991) were used to ensure mixing.

Set 4. To assess the accuracy of estimated topologies under a large and complex model, twenty 50 locus data sets for seven populations were simulated. For each data set, a phylogenetic topology was drawn at random from a uniform distribution over all possible rooted ordered topologies. Data were simulated under this topology with drift parameters sampled at random from $U[0, 3]$. Each model has 72 migration rate parameters, with half (36) randomly picked to have zero gene flow and half having a migration parameter drawn at random from $U[0, 0.2]$. The splitting times were fixed to be at even intervals $\{0.25, 0.5, 0.75, 1.0, 1.25, 1.5\}$ and mutation rate scalars were sampled from $U[0.25, 4.0]$ (log scale) subject to the constraint that their geometric mean is 1. IMa3 runs were done using 200–400 heated chains, on 20 to 40 processors. For the population size parameters, the

hyperprior distribution was $U[0, 4]$, while for migration rate parameters the hyperprior distribution was $U[0, 0.3]$. Mixing was assessed by restarting runs, and comparing estimated posterior densities for topologies on runs observed at different times.

The 20 seven population data sets were also examined using several other methods for estimating phylogenetic trees (table 1). For BPP (Yang 2015) the prior distribution of population sizes was a fairly flat gamma distribution, $G(10, 4000)$, with a mean per base pair (0.0025), similar to that used for IMa3, and divergence time priors having gamma distribution $G(10, 8000)$. For StarBeast2 (Ogilvie et al. 2017), which is implemented in the Beast2 program (Bouckaert et al. 2014), the prior for the phylogenetic tree was a Yule model, where speciation rate were sampled from a log-normal distribution with mean of 1 and standard deviation of 1.25. The constant per-branch population sizes were assumed to follow an inverse gamma prior distribution, $\text{InvG}(\alpha, \beta)$, where the shape parameter α is set to 3 and beta is the scale parameter. The mean and variance of this distribution are both equal to $\mu = \beta/2$, where μ is sampled from a hyperprior:: $\mu \sim p(\mu) \propto 1/\mu$. That is, $\log(\mu) \sim p(\log(\mu)) \propto 1$. Sampling details are provided in supplementary information, Supplementary Material online.

URLS

Source code for IMa3 and IMfig is available at https://github.com/jodyhey/IMa3. All simulated data and associated scripts and results are available at https://bio.cst.temple.edu/~hey/nolinks/Hey_2018_IMa3paper_archive_results_simualtions_data.zip.

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

Author Contributions

Y.W. conceived of the hidden genealogy updating method. A.S. helped develop the parallel code. Y.C. and V.C.S. helped develop the methods and the code. J.L. and S.T. provided the human hunter-gatherer data and assisted with their analysis. J.H. helped develop the methods, designed the implementation, wrote most of the code, designed and conducted the analyses, and wrote most of the paper.

References

Altekar G, Dwarkadas S, Huelsenbeck JP, Ronquist F. 2004. Parallel metropolis coupled Markov chain Monte Carlo for Bayesian phylogenetic inference. *Bioinformatics* 20(3):407–415.

Arbogast BS, Edwards SV, Wakeley J, Beerli P, Slowinski JB. 2002. Estimating divergence times from molecular data on phylogenetic and population genetic timescales. *Annu Rev Ecol Syst.* 33(1):707–740.

Avise JC. 1994. Molecular markers, natural history and evolution. London: Chapman & Hall.

Becquet C, Patterson N, Stone AC, Przeworski M, Reich D. 2007. Genetic structure of chimpanzee populations. *PLoS Genet.* 3(4):e66.

Becquet C, Przeworski M. 2007. A new approach to estimate parameters of speciation models with application to apes. *Genome Res.* 17(10):1505–1519.

Beerli P. 2004. Effect of unsampled populations on the estimation of population sizes and migration rates between sampled populations. *Mol Ecol.* 13(4):827–836.

Beerli P, Edwards SV. 2003. When did Neanderthals and modern humans diverge?. *Evol Anthropol.* 11(S1):60–63.

Beerli P, Felsenstein J. 1999. Maximum-likelihood estimation of migration rates and effective population numbers in two populations using a coalescent approach. *Genetics* 152:763–773.

Bouckaert R, Heled J, Kühnert D, Vaughan T, Wu C-H, Xie D, Suchard MA, Rambaut A, Drummond AJ. 2014. BEAST 2: a software platform for Bayesian evolutionary analysis. *PLoS Comput Biol.* 10(4):e1003537.

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(8):1917–1932.

Capra JA, Hubisz MJ, Kostka D, Pollard KS, Siepel A. 2013. A model-based analysis of GC-biased gene conversion in the human and chimpanzee genomes. *PLoS Genet.* 9(8):e1003684.

Caswell JL, Mallick S, Richter DJ, Neubauer J, Schirmer C, Gnerre S, Reich D. 2008. Analysis of chimpanzee history based on genome sequence alignments. *PLoS Genet.* 4(4):e1000057.

Chung Y, Hey J. 2017. Bayesian analysis of evolutionary divergence with genomic data under diverse demographic models. *Mol Biol Evol.* 34(6):1517–1528.

Dalquen DA, Zhu T, Yang Z. 2016. Maximum likelihood implementation of an isolation-with-migration model for three species. *Syst Biol.* 29:3131–3142.

de Manuel M, Kuhlwilm M, Frandsen P, Sousa VC, Desai T, Prado-Martinez J, Hernandez-Rodriguez J, Dupanloup I, Lao O, Hallast P, et al. 2016. Chimpanzee genomic diversity reveals ancient admixture with bonobos. *Science* 354(6311):477–481.

Degnan JH, Rosenberg NA. 2009. Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends Ecol Evol.* 24(6):332–340.

DeSalle R, Giddings LV. 1986. Discordance of nuclear and mitochondrial DNA phylogenies in Hawaiian Drosophila. *Proc Natl Acad Sci U S A.* 83(18):6902–6906.

Drmanac R, Sparks AB, Callow MJ, Halpern AL, Burns NL, Kermani BG, Carnevali P, Nazarenko I, Nilsen GB, Yeung G, et al. 2010. Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science* 327(5961):78–81.

Edwards AW. 1970. Estimation of the branch points of a branching diffusion process. *J Roy Stat Soc B.* 32:155–174.

Felsenstein J. 1988. Phylogenies from molecular sequences: inference and reliability. *Annu Rev Genet.* 22:521–565.

Geyer CJ. 1991. Markov chain Monte Carlo maximum likelihood. Computing Science and Statistics, Proceedings of the 23rd Symposium on the Interface. p. 156–163.

Gilks WR, Roberts GO. 1996. Strategies for improving MCMC. In: Gilks WR, Richardson S, Spiegelhalter DJ, editors. Markov-Chain Monte Carlo in practice. Boca Raton (FL): Chapman and Hall. p. 89–114.

Gillespie JH, Langley CH. 1979. Are evolutionary rates really variable. *J Mol Evol.* 13(1): 27–34.

Goldberg TL. 1998. Biogeographic predictors of genetic diversity in populations of eastern African chimpanzees (Pan troglodytes schweinfurthi). *Int J Primatol.* 19(2):237–254.

Gonder MK, Locatelli S, Ghobrial L, Mitchell MW, Kujawski JT, Lankester FJ, Stewart C-B, Tishkoff SA. 2011. Evidence from Cameroon reveals differences in the genetic structure and histories of chimpanzee populations. *Proc Natl Acad Sci U S A.* 108:4766–4771.

Gronau I, Hubisz MJ, Gulko B, Danko CG, Siepel A. 2011. Bayesian inference of ancient human demography from individual genome sequences. *Nat Genet.* 43(10):1031–1035.

Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD. 2009. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genet.* 5(10):e1000695.

Hammer MF, Woerner AE, Mendez FL, Watkins JC, Wall JD. 2011. Genetic evidence for archaic admixture in Africa. *Proc Natl Acad Sci.* 108(37):15123–15128.

Hasegawa M, Kishino H, Yano T. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J Mol Evol.* 22(2):160–174.

Heled J, Drummond AJ. 2010. Bayesian inference of species trees from multilocus data. *Mol Biol Evol.* 27(3):570.

Henn BM, Gignoux CR, Jobin M, Granka JM, Macpherson JM, Kidd JM, Rodriguez-Botigue L, Ramachandran S, Hon L, Brisbin A, et al. 2011. Hunter-gatherer genomic diversity suggests a southern African origin for modern humans. *Proc Natl Acad Sci U S A.* 108(13):5154–5162.,

Hey J. 2010. The divergence of chimpanzee species and subspecies as revealed in multipopulation isolation-with-migration analyses. *Mol Biol Evol.* 27(4):921–933.

Hey J. 2010. Isolation with migration models for more than two populations. *Mol Biol Evol.* 27(4):905–920.

Hey J, Nielsen R. 2007. Integration within the Felsenstein equation for improved Markov chain Monte Carlo methods in population genetics. *Proc Natl Acad Sci U S A.* 104(8):2785–2790.

Hey J, Nielsen R. 2004. Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of *Drosophila pseudoobscura* and *D. persimilis*. *Genetics* 167(2):747–760.

Hudson RR. 2002. Generating samples under a Wright–Fisher neutral model of genetic variation. *Bioinformatics* 18(2):337–338.

Hudson RR, Kaplan NL. 1985. Statistical properties of the number of recombination events in the history of a sample of DNA sequences. *Genetics* 111(1):147–164.

Jones GR. 2018. Divergence estimation in the presence of incomplete lineage sorting and migration. *Syst Biol.* doi: 10.1093/sysbio/syy041.

Kimura M. 1969. The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. *Genetics* 61:893–903.

Kimura M, Ohta T. 1978. Stepwise mutation model and distribution of allelic frequencies in a finite population. *Proc Natl Acad Sci U S A.* 75(6):2868–2872.

Kubatko LS, Carstens BC, Knowles LL. 2009. STEM: species tree estimation using maximum likelihood for gene trees under coalescence. *Bioinformatics* 25(7):971–973.

Kuhner MK. 2009. Coalescent genealogy samplers: windows into population history. *Trends Ecol Evol.* 24(2):86–93.

Kuhner MK, Yamato J, Felsenstein J. 1995. Estimating effective population size and mutation rate from sequence data using Metropolis-Hastings sampling. *Genetics* 140:1421–1430.

Lachance J, Vernot B, Elbers Clara C, Ferwerda B, Froment A, Bodo J-M, Lema G, Fu W, Nyambo Thomas B, Rebbeck Timothy R, et al. 2012. Evolutionary history and adaptation from high-coverage whole-genome sequences of diverse African hunter-gatherers. *Cell* 150(3):457–469.

Langergraber KE, Prufer K, Rowney C, Boesch C, Crockford C, Fawcett K, Inoue E, Inoue-Muruyama M, Mitani JC, Muller MN, et al. 2012. Generation times in wild chimpanzees and gorillas suggest earlier divergence times in great ape and human evolution. *Proc Natl Acad Sci U S A.* 109(39):15716–15721.,

Lartillot N, Philippe H. 2006. Computing Bayes factors using thermodynamic integration. *Syst Biol.* 55(2):195–207.

Leaché AD, Harris RB, Rannala B, Yang Z. 2014. The influence of gene flow on species tree estimation: a simulation study. *Syst Biol.* 63(1):17–30.

Liu L, Pearl DK. 2007. Species trees from gene trees: reconstructing Bayesian posterior distributions of a species phylogeny using estimated gene tree distributions. *Syst Biol.* 56(3):504–514.

Liu L, Yu L, Kubatko L, Pearl DK, Edwards SV. 2009. Coalescent methods for estimating phylogenetic trees. *Mol Phylogenet Evol.* 53(1):320–328.

Lopes JS, Balding D, Beaumont MA. 2009. PopABC: a program to infer historical demographic parameters. *Bioinformatics* 25(20):2747–2749.

Maddison W, Knowles L. 2006. Inferring Phylogeny Despite Incomplete Lineage Sorting. *Syst. Biol.* 55(1):21–30.

Maddison WP. 1997. Gene trees in species trees. *Syst Biol.* 46(3):523–536.

Mailund T, Halager AE, Westergaard M, Dutheil JY, Munch K, Andersen LN, Lunter G, Prüfer K, Scally A, Hobolth A, et al. 2012. A new isolation with migration model along complete genomes infers very different divergence processes among closely related great Ape species. *PLoS Genet.* 8(12):e1003125.

McCormack JE, Bowen BS, Smith TB. 2008. Integrating paleoecology and genetics of bird populations in two sky island archipelagos. *BMC Biol.* 6:28.

Mendez FL, Poznik GD, Castellano S, Bustamante CD. 2016. The divergence of Neandertal and modern human Y chromosomes. *Am J Hum Genet.* 98(4):728–734.

Mitchell MW, Locatelli S, Sesink Clee PR, Thomassen HA, Gonder MK. 2015. Environmental variation and rivers govern the structure of chimpanzee genetic diversity in a biodiversity hotspot. *BMC Evol. Biol.* 15(1):1.

Nei M. 1987. Molecular evolutionary genetics. New York: Columbia University Press.

Neigel J, Avise JC. 1986. Phylogenetic relationships of mitochondrial DNA under various demographic models of speciation. In: Nevo E, Karlin S, editors. Evolutionary processes and theory. London: Academic Press. p. 515–534.

Nielsen R, Wakeley J. 2001. Distinguishing migration from isolation. A Markov-chain Monte Carlo approach. *Genetics* 158(2): 885–896.

Nosil P. 2008. Speciation with gene flow could be common. *Mol Ecol.* 17(9):2103–2106.

Ogilvie HA, Bouckaert RR, Drummond AJ. 2017. StarBEAST2 brings faster species tree inference and accurate estimates of substitution rates. *Mol Biol. Evol.* 34(8):2101–2114.

Palstra FP, Heyer E, Austerlitz F. 2015. Statistical inference on genetic data reveals the complex demographic history of human populations in Central Asia. *Mol Biol Evol.* 32(6):1411–1424.

Pamilo P, Nei M. 1988. Relationships between gene trees and species trees. *Mol Biol Evol.* 5(5):568–583.

Pickrell JK, Pritchard JK. 2012. Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genet.* 8(11):e1002967.

Pinho C, Hey J. 2010. Divergence with gene flow: models and data. *Annu Rev Ecol. Evol. Syst.* 41(1):215–230.

Plagnol V, Wall JD. 2006. Possible ancestral structure in human populations. *PLoS Genet.* 2(7):e105.

Prado-Martinez J, Sudmant PH, Kidd JM, Li H, Kelley JL, Lorente-Galdos B, Veeramah KR, Woerner AE, O'Connor TD, Santpere G, et al. 2013. Great ape genetic diversity and population history. *Nature* 499(7459):471–475.,

Prufer K, Racimo F, Patterson N, Jay F, Sankararaman S, Sawyer S, Heinze A, Renaud G, Sudmant PH, de Filippo C, et al. 2014. The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature* 505(7481):43–49.

Rannala B, Yang Z. 2003. Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple Loci. *Genetics* 164(4):1645–1656.

Rannala B, Yang Z. 2017. Efficient Bayesian species tree inference under the multispecies coalescent. *Syst Biol.* 66:823–842.

Saitou N, Nei M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol.* 4:406–425.

Scally A, Durbin R. 2012. Revising the human mutation rate: implications for understanding human evolution. *Nat Rev Genet.* 13(10):745–753.

Sethuraman A, Hey J. 2016. IMa2p—parallel MCMC and inference of ancient demography under the Isolation with migration (IM) model. *Mol Ecol Resour.* 16(1):206–215.

Sousa VC, Grelaud A, Hey J. 2011. On the nonidentifiability of migration time estimates in isolation with migration models. *Mol Ecol.* 20(19):3956–3962.

Stephens M, Smith NJ, Donnelly P. 2001. A new statistical method for haplotype reconstruction from population data. *Am J Hum Genet.* 68(4):978–989.

Suárez N, Pestano J, Brown R. 2014. Ecological divergence combined with ancient allopatry in lizard populations from a small volcanic island. *Mol Ecol.* 23(19):4799–4812.

Tajima F. 1983. Evolutionary relationships of DNA sequences in finite populations. *Genetics* 105(2):437–460.

Takahata N. 1988. The coalescent in two partially isolated diffusion populations. *Genet Res.* 52(03):213–222.

Takahata N. 1989. Gene genealogy in three related populations: consistency probability between gene and population trees. *Genetics* 122(4):957–966.

Than C, Ruths D, Innan H, Nakhleh L. 2006. Identifiability issues in phylogeny-based detection of horizontal gene transfer. In: Bourque G, El-Mabrouk N, editors. RECOMB 2006 Comparative Genomics. Heidelberg, Springer Berlin. p. 215–229.

Wakeley J, Hey J. 1998. Testing speciation models with DNA sequence data. In: DeSalle R, Schierwater B, editors. Molecular approaches to ecology and evolution. Basel: Birkhäuser Verlag. p. 157–175.

Wall JD. 2000. Detecting ancient admixture in humans using sequence polymorphism data. *Genetics* 154:1271–1279.

Wen D, Nakhleh L. 2018. Coestimating reticulate phylogenies and gene trees from multilocus sequence data. *Syst Biol.* 67(3):439–457.

Wilson IJ, Balding DJ. 1998. Genealogical inference from microsatellite data. *Genetics* 150(1):499–510.

Won YJ, Hey J. 2005. Divergence population genetics of chimpanzees. *Mol Biol Evol.* 22(2):297–307.

Wright S. 1931. Evolution in Mendelian populations. *Genetics* 16(2):97–159.

Wright S. 1951. The genetical structure of populations. *Ann Eugenics* 15(4):323–354.

Xu B, Yang Z. 2016. Challenges in species tree estimation under the multispecies coalescent model. *Genetics* 204(4):1353–1368.

Yang Z. 2015. The BPP program for species tree estimation and species delimitation. *Curr Zool.* 61(5):854–865.

Zhang C, Ogilvie HA, Drummond AJ, Stadler T. 2017. Bayesian inference of species networks from multilocus sequence data. *Mol Biol Evol.* 35:504–517.