Population structure and coalescence in pedigrees: comparisons to the structured coalescent and a framework for inference

Peter R. Wilton[a,*], Pierre Baduel[a,c], Matthieu M. Landon[b,c], John Wakeley[a]

[a] *Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA, 02138, USA*
[b] *Department of Systems Biology, Harvard University, Cambridge, MA, 02138, USA*
[c] *École des Mines de Paris, Mines ParisTech, Paris 75272, France*

Abstract

Contrary to what is often assumed in population genetics, independently segregating loci do not have completely independent ancestries, since all loci are inherited through a single, shared population pedigree. Previous work has shown that the non-independence between gene genealogies of independently segregating loci created by the population pedigree is weak in panmictic populations, and predictions made from standard coalescent theory are accurate for populations that are at least moderately sized. Here, we investigate patterns of coalescence in pedigrees of structured populations. We find that population structure creates deviations away from the predictions of standard theory that persist on a longer timescale than in panmictic populations. Nevertheless, we find that the structured coalescent provides a reasonable approximation for the coalescent process in structured population pedigrees so long as migration events are moderately frequent and there is no admixture in the recent pedigree of the sample. When the sampled individuals have admixed ancestry, we find that distributions of coalescence in the sample can be modeled as a mixture of distributions from different sample configurations. We use this observation to motivate a maximum-likelihood approach for inferring migration rates and population sizes jointly with features of the recent pedigree such as admixture and relatedness. Using simulation, we demonstrate that our inference framework accurately recovers long-term migration rates in the presence of recent admixture in the sample pedigree.

Keywords: pedigree, coalescent theory, mixture distribution, identity-by-descent, demographic inference

1. Introduction

The coalescent is a stochastic process that describes how to construct gene genealogies, the tree-like structures that describe relationships among the sampled copies of a gene. Since its introduction (Kingman, 1982a,b; Hudson, 1983; Tajima, 1983), the coalescent has been extended and applied to numerous areas in population genetics and is now one of the foremost mathematical tools for modeling genetic variation in samples (Hein *et al.*, 2005; Wakeley, 2009).

In a typical application to data from diploid sexual organisms, the coalescent is applied to multiple loci that are assumed to have independent ancestries because they are found on different chromosomes and thus segregate independently, or are far enough apart along a single chromosome that they effectively segregate independently. Even chromosome-scale coalescent-based inference methods that account for linkage and recombination (e.g., Li and Durbin, 2011; Sheehan *et al.*, 2013; Schiffels and Durbin, 2014) multiply probabilities across distinct chromosomes that are assumed to have completely independent histories due to their independent segregation.

In reality, the ancestries of independently segregating loci are independent only after conditioning on the population pedigree, which is the set of relationships between all individuals in the population throughout all time. The predictions of standard coalescent theory implicitly average over the outcome of reproduction and thus over pedigrees. Since all genetic material is inherited through a single, shared population pedigree, applying these probabilities to multiple loci sampled from a single population ignores the non-independence between unlinked loci induced by the population pedigree.

This non-independence was examined by Wakeley *et al.* (2012), who studied gene genealogies of loci segregating independently through pedigrees of diploid populations generated under basic Wright-Fisher-like reproductive dynamics, *i.e.*, populations with constant population size, random mating, non-overlapping generations, and lacking population structure. In this context, it was found that the shape of the population pedigree affected coalescence probabilities mostly in the first $\sim \log_2(N)$ generations back in time, where N is the population size, and that in general it was difficult to distinguish distributions of coalescence times that were generated by segregating independent chromosomes back in time through a fixed, randomly-generated pedigree from the predictions of the

[*] Corresponding author
Email address: pwilton@fas.harvard.edu (Peter R. Wilton)

standard coalescent.

That the pedigree should have substantial effects on coalescence probabilities only during the most recent $\sim \log_2(N)$ generations is in agreement with other theoretical studies of population pedigrees. Chang (1999) found that the number of generations until two individuals share an ancestor in the biparental, pedigree sense converges to $\log_2(N)$ as the population size grows. Likewise, Derrida *et al.* (2000) showed that the distribution of the number of repetitions in an individual's pedigree ancestry becomes stationary around $\log_2(N)$ generations in the past. This $\log_2(N)$-generation timescale is the natural timescale of convergence in pedigrees due to the approximate doubling of the number of possible ancestors each generation back in time.

In each of these studies it is assumed that the population is panmictic, *i.e.*, that individuals mate with each other uniformly at random. One phenomenon that may alter this convergence in pedigrees is population structure with migration between subpopulations or demes. In a subdivided population, the exchange of ancestry between demes depends on the particular history of migration events embedded in the population pedigree. These past migration events may be infrequent or irregular enough that the convergence in the pedigree depends on the details of the migration history rather than on the reproductive dynamics underlying convergence in panmictic populations.

Rohde *et al.* (2004) studied the sharing of pedigree ancestry in structured populations and found that population structure did not change the $\log_2(N)$-scaling of the number of generations until a common ancestor of everyone in the population is reached. Barton and Etheridge (2011) studied the expected number of descendants of an ancestral individual, a quantity termed the reproductive value, and similarly found that population subdivision did not much slow the convergence of this quantity over the course of generations.

While these results give a general characterization of how pedigrees are affected by population structure, a direct examination of the coalescent process for loci segregating through a fixed pedigree of a structured population is still needed. Fixing the migration events in the pedigree may produce long-term fluctuations in coalescent probabilities that make the predictions of the structured coalescent break down. The pedigree may also bias inference of demographic history. Each sampled locus will have a relatively great probability of being affected by any migration events or overlap in ancestry contained in the most recent generations of the sample pedigree, and thus demographic inference methods that do not take into account the pedigree of the sample may be biased by how these events shape genetic variation in the sample.

Here, we explore how population structure affects coalescence through a fixed population pedigree. Using simulations, we investigate how variation in the migration history embedded in the pedigree affects coalescence probabilities, and we determine how these pedigree effects depend on population size and migration rate. We also study the effects of recent admixture on coalescence-time distributions and use our findings to develop a simple framework for modeling the sample as a probabilistic mixture of multiple non-admixed ancestries. We demonstrate how this framework can be incorporated into demographic inference by developing a maximum-likelihood method of inferring scaled mutation and migration rates jointly with the recent pedigree of the sample. We test this inference approach with simulations, showing that including the pedigree in inference corrects a bias that is present when there is unaccounted-for admixture in the ancestry of the sample.

2. Theory and Results

2.1. Pedigree simulation

Except where otherwise stated, each population we model has two demes of constant size, exchanging migrants symmetrically at a constant rate. This model demonstrates the effects of population structure in one of the the simplest ways possible and has a relatively simple mathematical theory (Wakeley, 2009).

We assume that generations are non-overlapping and that the population has individuals of two sexes in equal number. In each generation, each individual chooses a mother uniformly at random from the females of the same deme with probability $1 - m$ and from the females of the other deme with probability m. Likewise a father is chosen uniformly at random from the males of the same deme with probability $1 - m$ and from the males of the other deme with probability m.

All simulations were carried out with `coalseam`, a program for simulation of coalescence through randomly-generated population pedigrees. The user provides parameters such as population size, number of demes, mutation rate, and migration rate, and `coalseam` simulates a population pedigree under a Wright-Fisher-like model meeting the specified conditions. Gene genealogies are constructed by simulating segregation back in time through the pedigree, and the resulting genealogies are used to produce simulated genetic loci. Output is in a format similar to that of the program `ms` (Hudson, 2002), and various options allow the user to simulate and analyze pedigrees featuring, for example, recent selective sweeps or fixed amounts of identity by descent and admixture.

The program `coalseam` is written in C and released under a permissive license. It is available online at `https://github.com/ammodramus/coalseam`.

2.2. Structured population pedigrees and probabilities of coalescence

In a well-mixed population of size N, the distribution of coalescence times for independently segregating loci sampled from two individuals shows large fluctuations over the first $\sim \log_2(N)$ generations depending on the degree of

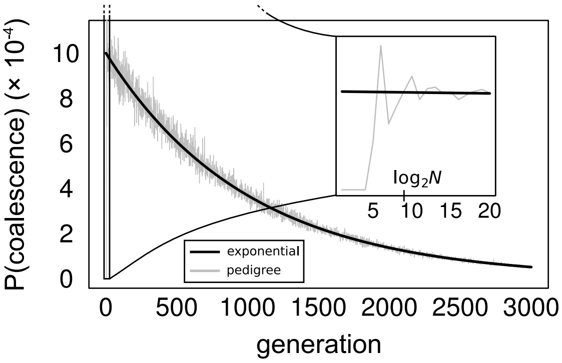

Figure 1: Coalescence time distribution for independently segregating loci sampled from two individuals in a panmictic population. The gray line shows the distribution from the pedigree, and the black line shows the exponential prediction of the standard coalescent. The population size is $N = 500$ diploid individuals.

overlap in the pedigree of the two individuals. After this initial period, the coalescence probabilities quickly converge to the expectation under standard coalescent theory, with small fluctuations around that expectation. (Wakeley *et al.*, 2012, Fig. 1). The magnitude of these fluctuations depends on the population size, but even for small to moderately sized populations (e.g., $N = 500$), the exponential prediction of the standard coalescent is a good approximation to the true distribution after the first $\log_2(N)$ generations.

When the population is divided into multiple demes, deviations from the coalescent probabilities predicted by the structured coalescent depend on the particular history of migration in the population pedigree. The effect of the migration history is especially pronounced when the average number of migration events per generation is of the same order as the per-generation pairwise coalescent probability (Fig. 2A). In this migration-limited regime, two lineages in different demes have zero probability of coalescing before a migration event in the pedigree can bring them together into the same deme. This creates large peaks in the coalescence time distributions for loci segregating independently through the same pedigree, with each peak corresponding to a migration event (Fig. 2A).

Even when the migration rate is greater and there are many migration events per coalescent event, the pedigree can still cause coalescence probabilities to differ from the predictions of the structured coalescent. Under these conditions, coalescence is not constrained by individual migration events, but there may be stochastic fluctuations in the realized migration rate, with some periods experiencing more migration and others less. These fluctuations can cause deviations in the predicted coalescence probabilities long past the $\log_2(N)$-generation timescale found in well-mixed populations (Figs. 2B, S1, S2). The degree of these deviations depends on the rate of migration and the population size, with smaller populations and lower migration rates causing greater deviations, and deviations from predictions are generally larger for samples taken between demes than samples within demes. When there are many migration events per coalescent event (i.e., when $Nm >> 1/N$), the predictions of the structured coalescent fit the observed distributions in pedigrees reasonably well (Figs. S1–S2).

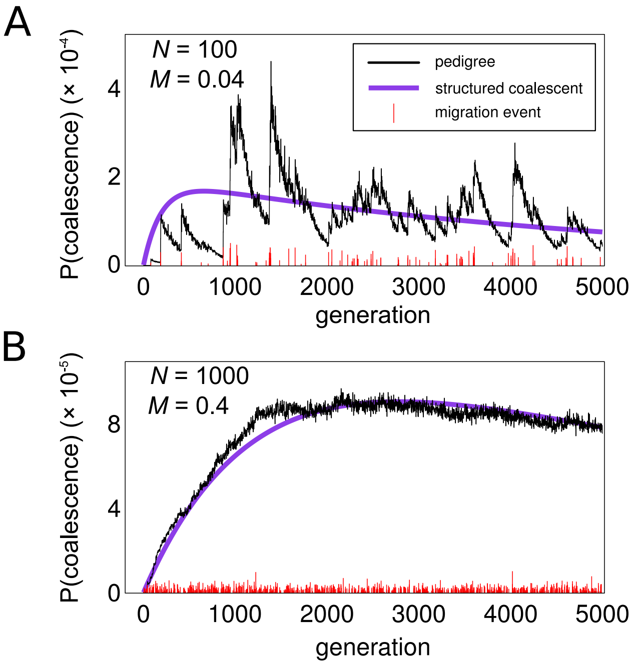

Figure 2: Distribution of coalescence times for two individuals sampled from different demes. In both panels, the black line shows the distribution calculated from a simulated pedigree, and the purple line shows the prediction from the structured coalescent. Red vertical lines along the horizontal axis show the occurrence of migration events in the population, with the relative height representing the total reproductive weight (see Barton and Etheridge, 2011) of the migrant individual(s) in that generation. **A)** Low migration pedigree, with $M = 4Nm = 0.04$ and $N = 100$. Under these conditions, coalescence is limited by migration events, so there are distinct peaks in the coalescence time distribution corresponding to individual migration events. **B)** Higher migration pedigree, with $M = 0.4$ and $N = 1000$. With the greater migration rate, coalescence is no longer limited by migration, but stochastic fluctuations in the migration process over time cause deviations away from the standard-coalescent prediction on a timescale longer than $\log_2(N)$ generations.

To investigate the dependence of the coalescence time distribution on the pedigree more systematically, we simulated 20 replicate population pedigrees for a range of populations sizes and migration rates. From each pedigree, we sampled two individuals in different demes and calculated the distribution of pairwise coalescence times for independently segregating loci sampled from those two individuals. We measured the total variation distance from the distribution predicted under a discrete-time model of coalescence and migration analogous to the continuous-time structured coalescent. The total variation distance of two discrete distributions P and Q is defined as

$$D_{TV}(P, Q) = \frac{1}{2} \sum_i |P(i) - Q(i)| \,. \tag{1}$$

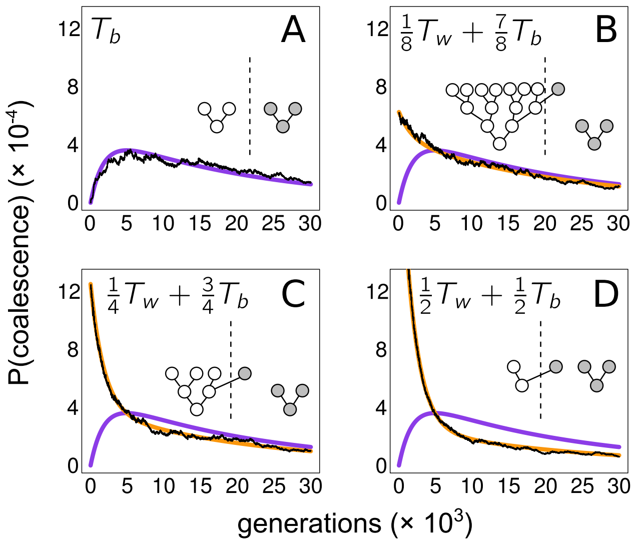

Figure 3: Pairwise coalescence time distributions for samples with admixed ancestry. Each panel shows the distribution of pairwise coalescence times for a sample whose pedigree contains some amount of admixture. In each simulation, there are two demes of size $N = 1000$, and the scaled migration rate is $4Nm = 0.1$. In each panel, the population pedigree was simulated conditional on the sample having the pedigree shown in the panel. The purple line shows the between-deme coalescence time distribution that would be expected in the absence of admixture, and the gold line shows the the mixture of the within- and between-deme coalescence time distributions that corresponds to the degree of admixture. Black lines are numerically calculated coalescence time distributions for the simulated example pedigrees.

We found that the total variation distance between the distributions of pairwise coalescence times from pedigrees and the distributions from standard theory decreases as both the population size and migration rate increase and that, in general, the total variation distance is more sensitive to the migration rate than to population size (Fig. S3).

2.3. Admixture and coalescence distributions in pedigrees

As is the case for panmictic populations, the details of the *recent* sample pedigree are most important in determining the patterns of genetic variation in the sample. In panmictic populations, overlap in ancestry in the recent past creates identity-by-descent. In structured populations, an individual may also have recent relatives from another deme, resulting in admixed ancestry. When this occurs, the distribution of pairwise coalescence times is potentially very different from the prediction in the absence of admixture due to the admixture paths in the pedigree that lead to a recent change in demes. The degree of the difference in distributions is directly related to the degree of admixture, with more recent admixture causing greater changes in the coalescence time distribution (Fig. 3).

If an individual in the sample has a migration event in its recent pedigree, there will be some probability that a gene copy sampled from this individual was inherited via the path through the pedigree including this recent migration event. If this is the case, it will appear as if that gene copy was actually sampled from a deme other than the one it was sampled from. In this way, the recent sample pedigree "reconfigures" the sample by changing the location of sampled gene copies, with the probabilities of the different sample reconfigurations depending on the number and timing of migration events in the recent pedigree.

If many independently segregating loci are sampled from individuals having admixed ancestry, some loci will be reconfigured by recent migration events, and others will not. In this scenario, the sample itself can be modeled as a probabilistic mixture of samples taken in different configurations, and probabilities of coalescence can be calculated by considering this mixture. For example, consider a sample of independently segregating loci taken from two individuals related by the pedigree shown in Fig. 3C, where one of two individuals sampled from different demes has a grandparent from the other deme. The distribution of pairwise coalescence times for loci sampled from this pair resembles the distribution of $T_w/4 + 3T_b/4$, where T_b is the standard between-deme pairwise coalescence time for a structured-coalescent model with two demes, and T_w is the corresponding within-deme pairwise coalescence time (Fig. 3C). This particular mixture reflects the fact that a lineage sampled from the admixed individual follows the admixture path with probability 1/4.

This sample reconfiguration framework can also be used to model identity-by-descent (IBD), where overlap among branches of the recent sample pedigree causes early coalescence with unusually high probability. If the pedigree causes an IBD event to occur with probability Pr(IBD), then the pairwise coalescence time is a mixture of the standard distribution (without IBD) and instantaneous coalescence (on the coalescent timescale) with probabilities $1 - \text{Pr(IBD)}$ and Pr(IBD), respectively. If there is both IBD and admixture in the recent sample pedigree (or if there are multiple admixture or IBD events), the sample can be modeled as a mixture of several sample reconfigurations (e.g., Fig. 4).

This approach to modeling the sample implicitly assumes that there is some threshold generation separating the recent pedigree, which determines the mixture of sample reconfigurations, and the more ancient pedigree, where the standard coalescent models are assumed to hold well enough. The natural boundary between these two periods is around $\log_2(N)$ generations, since any pedigree feature more ancient than that tends to be shared by most or all of the population (making such features "population demography"), whereas any features more recent tend to be particular to the sample.

We note that there is a long history in population genetics of modeling genetic variation in pedigrees as a mixture of different sample reconfigurations. Wright (1951) wrote the probability of observing a homozygous A_1A_1 genotype as $p^2(1 - F) + pF$, where p is the frequency of A_1 in the population and F is essentially the probability of IBD calculated from the sample pedigree. This can be thought of as a probability for a mixture of two samples of

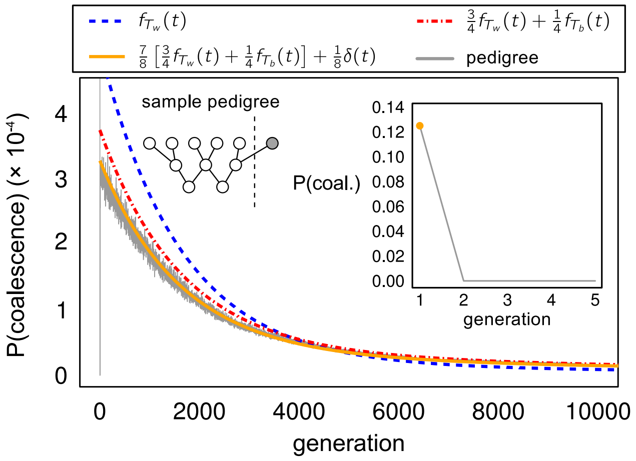

Figure 4: Distribution of pairwise coalescence times for a sample whose recent pedigree contains both admixture and IBD. The recent pedigree of the sample is shown, with the two sampled individuals located at the bottom of the pedigree. The distribution for the simulated pedigree (gray line) is based on numerical coalescence probabilities calculated in a pedigree of two demes of size $N = 1000$ each, with migration rate $M = 4Nm = 0.2$. The colored lines show mixtures of the within-deme coalescence time distribution (f_{T_w}) and the between-deme distribution (f_{T_b}). The inset shows the probability of coalescence during the first five generations; the probability mass at generation 1 is predicted by the mixture model accounting for both IBD and admixture (orange line). The red line shows the distribution if IBD is ignored, and the blue line shows the distribution if both IBD and admixture are ignored.

size $n = 2$ (with probability $1 - F$) and $n = 1$ (probability F). The popular ancestry inference program STRUCTURE (Pritchard *et al.*, 2000) and related methods similarly write the likelihood of observed genotypes as a mixture over different possible subpopulation origins of the sampled alleles. Here, motivated by our simulations of coalescence in pedigrees, we explicitly extend this kind of approach to modeling coalescence. In the next section, we propose an approach to inferring population parameters such as mutation and migration rates jointly with features of the sample pedigree such as recent IBD and admixture. In the Discussion, we further consider the similarities and differences between our inference approach and those of existing methods.

2.4. Joint inference of the recent sample pedigree and population demography

The sample reconfiguration framework for modeling genetic variation in pedigrees, described above, can be used to calculate how admixture or IBD in the recent sample pedigree bias estimators of population-genetic parameters. In Appendix A, we calculate the bias of three estimators of the population-scaled mutation rate $\theta = 4N\mu$ in a panmictic population when the recent sample pedigree contains IBD. In Appendix B, we calculate the bias of a moments-based estimator of M due to recent admixture in the sample pedigree. We use simulations to confirm these calculations (Figs. S5,S6). In both cases, if the recent sample pedigree is known, it is straightforward to correct these estimators to eliminate bias.

It is uncommon that the recent pedigree of the sample is known, however, and if it is assumed known, it is often estimated from the same data that is used to infer demographic parameters. Ideally, one would infer long-term demographic history jointly with sample-specific features of the pedigree. In this section, we develop a maximum-likelihood approach to inferring IBD and admixture jointly with scaled mutation and migration rates. The method uses the approach proposed in the previous section: the sample pedigree defines some set of possible outcomes of Mendelian segregation in recent generations, and the resulting, reconfigured sample is modeled by the standard coalescent process.

We model a population with two demes each of size N, with each individual having probability m of migrating to the other deme in each generation. We rescale time by N so that the rate of coalescence within a deme is 1 and the rescaled rate of migration per lineage is $M/2 = 2Nm$. We assume that we have sampled two copies of each locus from each of n_1 (diploid) individuals from deme 1 and n_2 individuals from deme 2. We write the total diploid sample size as $n_1 + n_2 = n$ so that the total number of sequences sampled at each locus is $2n$. We index our sequences with $\mathcal{I}_n := \{1^{\mathrm{m}}, 1^{\mathrm{p}}, 2^{\mathrm{m}}, 2^{\mathrm{p}}, \ldots, n^{\mathrm{m}}, n^{\mathrm{p}}\}$, where i^{m} and i^{p} index the maternal and paternal sequences sampled from individual i. These are simply notational conventions; we do not assume that these maternal and paternal designations are known or observed.

Each recent pedigree $\mathcal{P}$ has some set of possible outcomes of segregation through the recent past, involving coalescence of lineages (IBD) and movement of lineages between demes (admixture). The set of these sample reconfigurations is denoted $\mathcal{R}(\mathcal{P})$, and each reconfiguration $r \in \mathcal{R}(\mathcal{P})$ is a partition of $\mathcal{I}_n$, with the groups in r representing the lineages that survive after segregation back in time through the recent pedigree. Each group in a reconfiguration is also labeled with the deme in which the corresponding lineage is found segregation back in time through the recent pedigree. Corresponding to each sample reconfiguration $r \in \mathcal{R}(\mathcal{P})$, there is a probability $\Pr(r \mid \mathcal{P})$ of that sample reconfiguration being the outcome of segregation back in time through the recent pedigree.

The data $\boldsymbol{X} = \{\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_L\}$ consist of sequence data at L loci. The data at locus i are represented by the sequences $\boldsymbol{X}_i = \{X_{i,1}^{(\mathrm{a})}, X_{i,1}^{(\mathrm{b})}, X_{i,2}^{(\mathrm{a})}, X_{i,2}^{(\mathrm{b})}, \ldots, X_{i,n}^{(\mathrm{a})}, X_{i,n}^{(\mathrm{b})}\}$, where $X_{i,j}^{(\mathrm{a})}$ and $X_{i,j}^{(\mathrm{b})}$ are the two sequences at locus i from individual j. We label them (a) and (b) because we assume that they are of unknown parental origin. In order to make calculation of sampling probabilities feasible, we assume that each sequence evolves under the infinite-sites mutation model and can thus be represented by a binary sequence. We further assume that there is free recombination between loci and no recombination within loci.

Our goal is to find the θ, M, and $\mathcal{P}$ that maximize the likelihood

$$L(\mathcal{P},\theta,M \mid \boldsymbol{X}) = \Pr(\boldsymbol{X} \mid \mathcal{P};\theta,M) = \prod_{i=1}^{L} \sum_{r\in\mathcal{R}(\mathcal{P})} \Pr(\boldsymbol{X}_i \mid r;\theta,M)\Pr(r \mid \mathcal{P}). \quad (2)$$

Probabilities are multiplied across independently segregating loci because their ancestries are independent conditional on the pedigree.

In order to calculate $\Pr(\boldsymbol{X}_i \mid r;\theta,M)$, it is necessary to consider all the ways the sequences $\boldsymbol{X}_i$ could have been inherited maternally versus paternally, since we assume that we do not know which sequence is maternal and which is paternal. For sequences $\boldsymbol{X}_i$, let $\Lambda(\boldsymbol{X}_i)$ be the set of all possible ways of labeling $\boldsymbol{X}_i$ as maternal and paternal, thus associating each sequence with an index in $\mathcal{I}_n$. The sampling probability of the data at locus i is then

$$\Pr(\boldsymbol{X}_i \mid r;\theta,M) = \frac{1}{2^n}\sum_{\lambda\in\Lambda(\boldsymbol{X}_i)} \Pr(X_i \mid \lambda,r;\theta,M), \quad (3)$$

since there are 2^n ways that the X_i could have segregated as maternal and paternal alleles, and each is equally likely to have occurred.

Together the reconfiguration $r \in \mathcal{R}(\mathcal{P})$ and the maternal-paternal labeling $\lambda \in \Lambda(\boldsymbol{X}_i)$ imply a partition $\mathbb{P}(\boldsymbol{X}_i, r, \lambda)$ of the sequences $\boldsymbol{X}_i$ corresponding to the partition of sequence indices represented by r. For each group $h \in \mathbb{P}(\boldsymbol{X}_i, r, \lambda)$, there is a group $g \in r$ that can be mapped onto h such that 1) each index $i \in g$ indexes a distinct sequence in h sampled from the individual indexed by i, and 2) the deme labeling of g matches the deme labeling of h. Denote the unique elements of a set A as $A_{\neq}$. The conditional sampling probability of the sequences $\boldsymbol{X}_i$ given maternal-paternal labeling $\lambda \in \Lambda(\boldsymbol{X}_i)$ and sample reconfiguration $r \in \mathcal{R}(\mathcal{P})$ is

$$\Pr(X_i \mid \lambda,r;\theta,M) = \Pr(\mathbb{P}(\boldsymbol{X}_i,r,\lambda);\theta,M) = \begin{cases} \phi(\{h_{\neq} : h \in \mathbb{P}(\boldsymbol{X}_i,r,\lambda)\};\theta,M) & \text{if } |h_{\neq}| = 1 \\ & \forall h \in \mathbb{P}(\boldsymbol{X}_i,r,\lambda) \\ 0 & \text{otherwise,} \end{cases} \quad (4)$$

where each $h \in \mathbb{P}(\boldsymbol{X}_i, r, \lambda)$ is one of the non-empty subsets in the partitioned sequences, $|h_{\neq}|$ is the number of unique elements in such a subset, $\{h_{\neq} : h \in \mathbb{P}(\boldsymbol{X}_i, r, \lambda)\}$ is the "reduced" set of sequences, such that each subset in the partition is replaced the unique elements in the subset, and $\phi(\{h_{\neq} : g \in \mathbb{P}(\boldsymbol{X}_i, r, \lambda)\};\theta, M)$ is the standard infinite-sites sampling probability of the sample after it has been reconfigured by the recent pedigree. This sampling probability can be calculated numerically using a dynamic programming approach (Griffiths and Tavaré, 1994; Wu, 2010, see below).

In other words, conditional on certain sequences being IBD (*i.e.*, they are in the same group in the partitioned sequences), the sampling probability is the standard infinite-sites probability of the set of sequences with duplicate IBD sequences removed and the deme labelings of the different groups made to match any admixture events that may have occurred. If any of the sequences designated as IBD are not in fact identical in sequence, the sampling probability for that reconfiguration and maternal-paternal labeling is zero. This is equivalent to assuming that no mutation occurs in the recent part of the pedigree.

Together, (2), (3), and (4) give the overall joint log-likelihood of mutation rate θ, migration rate M, and pedigree $\mathcal{P}$ given sequences $\boldsymbol{X}$:

$$LL(\theta,M,\mathcal{P} \mid \boldsymbol{X}) = \sum_{i=1}^{L}\log\left(\sum_{r\in\mathcal{R}(\mathcal{P})}\Pr(r \mid \mathcal{P})\sum_{\lambda\in\Lambda(\boldsymbol{X}_i,r)}\Pr(\mathbb{P}(\boldsymbol{X}_i,r,\lambda);\theta,M)\right) - nL\log(2) \quad (5)$$

Our goal is to maximize (5) over θ, M, and $\mathcal{P}$ in order to estimate these parameters. One naive approach would be to generate all possible recent sample pedigrees and maximize the log-likelihood conditional on each pedigree in turn. However, the number of pedigrees to consider is prohibitively large even if only the first few generations back in time are considered. Many sample pedigrees will contain many IBD or admixture events and thus be unlikely to occur in nature, and in many populations, it is more probable that the sample will have few IBD or admixture events in the very recent pedigree, if any. With this in mind, we consider only pedigrees containing no more than two events, whether they be IBD events or admixture events. We further limit the number of pedigrees by considering only the past three generations. Besides reducing the number of pedigrees requiring calculations, this also limits consideration to pedigrees having the greatest effect on genetic variation. Finally, since we assume that we do not know the parental origin of each sequence, we further reduce the number of pedigrees to consider by evaluating only pedigrees that are unique up to labeling of ancestors as maternal and paternal.

In a two-deme population, each recent sample pedigree with two or fewer IBD or admixture events has the shape of one of the pedigrees shown in Figure S4. There are at most 21 distinct shapes of pedigrees with two or fewer events, and for each such pedigree shape, there exist some number of pedigrees with unique labelings of the sampled individuals and timings of the events in the pedigree. (Fewer distinct shapes are possible if the sample size is too small to permit certain shapes.) Table 1 gives the number of distinct pedigrees that must be considered for different sample sizes and numbers of past generations considered. We note that we consider only pedigrees with

Table 1: Number of distinct pedigrees with two or fewer IBD or admixture events. Pedigrees that differ only in maternal-paternal labeling of individuals are not counted as distinct.

generations	sample size			
	$n = 1$	$n = 2$	$n = 3$	$n = 4$
$g = 2$	16	123	434	1109
$g = 3$	41	328	1144	2879
$g = 4$	78	631	2190	5477

non-overlapping generations and the sampled individuals found in the current generation. This is sufficient for the Wright-Fisher-like model of pedigrees that we model here, but real pedigrees will not in general satisfy these criteria. Allowing overlapping generations will require consideration of many additional pedigrees and introduce problems of identifiability, with multiple pedigrees having the same set of reconfigurations with the same probabilities. We do not explore these issues here.

To calculate the standard sampling probabilities needed in (4), we use the dynamic-programming method described by Wu (2010). In principle, it should be possible to calculate the log-likelihood of all pedigrees simultaneously, since any reconfiguration of the sample by recent IBD or admixture must correspond to one of the ancestral configurations in the recursion solved by Wu's (2010) method (see also Griffiths and Tavaré, 1994). Thus, for particular values of θ and M, after solving the ancestral recursion only once (and storing the sampling probabilities of all ancestral configurations), the likelihood of any pedigree can be found by extracting the relevant probabilities from the recursion. However, in order to take this approach to maximize the log-likelihood, it is necessary to solve the recursion on a large grid of θ and M. In practice, we find that it is faster to maximize the log-likelihood separately for each pedigree, using standard derivative-free numerical optimization procedures to find the M and θ that maximize the log-likelihood for the pedigree. We use ordered sampling probabilities throughout, since the typical unordered probabilities would be inappropriate in this context due to the partial ordering of the sample by the pedigree.

To test our inference method we simulated datasets of 1000 independently segregating loci, generated by simulating coalescence through a randomly generated pedigree of a two-deme population with deme size $N = 1000$ and migration rate $M \in \{0.2, 2.0\}$. We sampled one individual (two sequences) from each deme. Sequence data were generated by placing mutations on the simulated gene genealogies according to the infinite-sites mutation model with rate $\theta/2 = 0.5$ (when $M = 0.2$) or $\theta/2 = 1.0$ (when $M = 2.0$). In order to investigate the effects of admixture on the estimation on θ and M, each replicate dataset was conditioned upon having one of three different sample pedigrees with differing amounts of admixture (see Fig. 5). We calculated maximum-likelihood estimates of θ and M for each of the 41 distinct pedigrees containing two or fewer IBD or admixture events occurring in the past three generations. We compared these estimates to the estimates that would be obtained from a similar maximum-likelihood procedure that ignores the pedigree (*i.e.*, assuming the null pedigree of no sample reconfiguration).

When there was recent admixture in the sample, assuming the null pedigree to be the true pedigree produced a bias towards overestimation of the migration rate (Fig. 5), since the early probability of migration via the admixture path must be accommodated by an increase in the migration rate. For this reason, the overestimation of the migration rate was greater when the degree of admixture was greater. The mutation rate was also overestimated when the admixture in the pedigree was ignored, presumably because migration via the admixture path did not decrease allelic diversity as much as the overestimated migration rate should. Including the pedigree as a free parameter in the estimation corrected this biased estimation of M in the presence of admixed ancestry in the sample. Estimates from simulations of samples lacking any features in the recent pedigree produced approximately unbiased estimates of θ and M (Fig. 5).

The pedigree was not inferred as reliably as the mutation and migration rates (Fig. 6). When the simulated pedigree contained admixture, the estimated pedigree was the correct pedigree (out of 41 possible pedigrees) roughly half of the time. For pedigrees with no admixture and no IBD, the correct pedigree was inferred about one third of the time. In addition to calculating a maximum-likelihood pedigree, it is possible to construct an approximate 95% confidence set of pedigrees using the fact that the maximum of the log-likelihood is approximately χ^2 distributed when the number of loci is large. These pedigree confidence sets contained the true pedigree $\sim 88 - 99\%$ of the time, depending on the true sample pedigree, mutation rate, and migration rate. A log-likelihood ratio test has nearly perfect power to reject the null pedigree for the simulations with the lesser migration rate; for simulations with the greater migration rate the power depended on the degree of admixture, with more recent admixture producing greater power to reject the null pedigree (Fig. 6). Type I error rates for simulations where the null pedigree is the true pedigree were close to $\alpha = 0.05$.

We also simulated datasets of 1000 loci sampled from individuals with completely random pedigrees (i.e., sampled without conditioning on any admixture in the recent sample pedigree) in a small two-deme population of size $N = 50$ per deme with one of two different migration rates ($M \in \{0.2, 2.0\}$). Whether or not the pedigree was included as a free parameter mostly had little effect on the estimates of θ and M (Fig. 7). However, in the few cases when the sample pedigree included recent admixture, the estimates were biased when the sampled pedigree was not considered as a free parameter. Inferring the pedigree together with the other parameters corrected this bias. (Fig. 7).

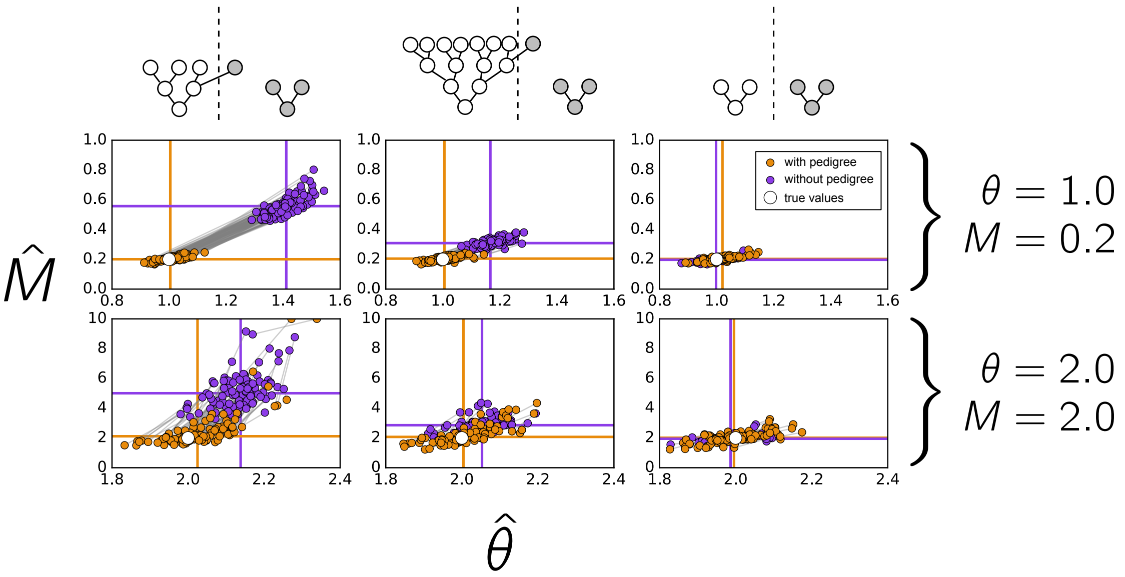

Figure 5: Maximum-likelihood mutation rates and migration rates for datasets simulated through different sample pedigrees. Each point depicts the maximum-likelihood estimates of θ and M for a particular simulation. Orange points show estimates obtained when the pedigree is included as a free parameter, and purple points show estimates obtained when the pedigree is assumed to have no effect on the data. In the first two columns, one of the sampled individuals is conditioned upon having a relative from the other deme, and in the third column the data are generated from completely random pedigrees. In each panel the true parameter values are shown with a solid white circle, and horizontal and vertical lines show means across replicates. Gray lines connect estimates calculated from the same dataset.

3. Discussion

Here we have explored the effects of fixed migration events in the population pedigree on the patterns of coalescence of independently segregating loci. In contrast to the case of panmictic populations, in structured populations the population pedigree can influence coalescence well beyond the time scale of $\sim \log_2(N)$ generations in the past. These effects are greater when the migration rate is small and are particularly pronounced when the total number of migration events occurring per generation is of the same order as the coalescence probability. When migration occurs more frequently than this (and there is no admixture in the recent sample pedigree), the particular history of migration events embedded in the population pedigree has less of an effect on coalescence, and the coalescence distributions based on the structured coalescent serve as good approximations to coalescent distributions within pedigrees.

We have also proposed a framework for inferring demographic parameters jointly with the recent pedigree of the sample. This framework considers how recent migration and shared ancestry events reconfigure the sample by moving lineages between demes and coalescing lineages. The inferred sample pedigree is the sample pedigree that has the maximum-likelihood mixture distribution of sample reconfigurations. In our implementation of this framework, we consider only sample pedigrees having two or fewer events, which must have occurred in the most recent three generations. At the expense of computational runtimes, the set of possible sample pedigrees could be expanded to include pedigrees with more events extending a greater number of generations into the past, but as these limits are extended, the number of pedigrees to consider grows rapidly. Another approach one could take would be to consider all of the reconfigurations (rather than sample pedigrees) with fewer than some maximum number of differences from the original sample configuration, and find the mixture of reconfigurations that maximizes the likelihood of the data, without any reference to the pedigree that creates the mixture. Such an approach may allow greater flexibility in modeling the effects of the recent pedigree, but searching the space of possible mixtures of reconfigurations — the k-simplex, if k is the number of possible reconfigurations — would likely be more challenging than maximizing over the finite set of sample pedigrees in the method we have implemented.

The sample reconfiguration inference framework is complementary to existing procedures for inferring recent admixture and relatedness in structured populations. The popular program STRUCTURE (Pritchard *et al.*, 2000) and related methods (Raj *et al.*, 2014; Alexander *et al.*, 2009; Tang *et al.*, 2005) are powerful and flexible tools for inferring admixture and population structure. Likewise, the inference tools RelateAdmix (Moltke and Albrechtsen, 2014), REAP (Thornton *et al.*, 2012), and KING-robust (Manichaikul *et al.*, 2010) all offer solutions to the problem of inferring relatedness in the presence of population structure and admixture. Perhaps the most similar in scope is the method of Wilson and Rannala (2003), which uses inferred ancestry proportions to estimate migration rates in the most recent generations. For input, each of these methods take genotypes at poly-

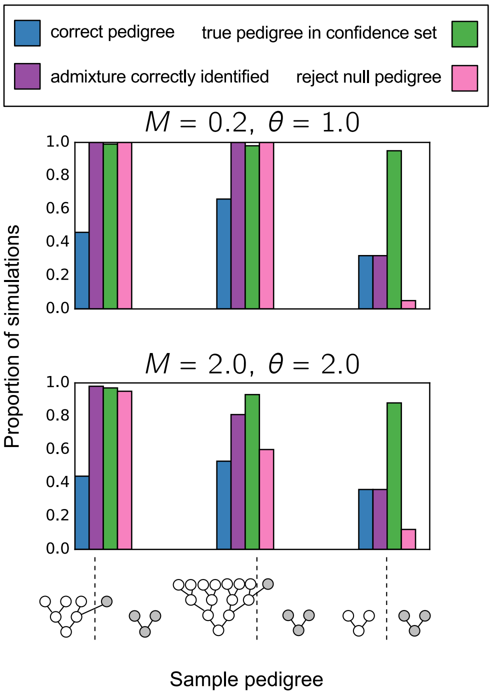

Figure 6: Inference of sample pedigrees. For simulations of 1000 infinite-sites loci, with $\theta = 0.5$ and $M = 0.2$ (**A**) or $\theta = 1.0$ and $M = 2.0$ (**B**), different measurements of the accuracy and power of pedigree inference are shown. The conditioned-upon sample pedigrees are shown at the bottom of the figure. Blue bars show the proportion of simulations in which the maximum-likelihood pedigree was the true pedigree. Purple bars show the proportion of simulations where it was inferred that sampled individuals had admixed ancestry. Green bars show the proportion of simulations in which the true pedigree was found within the approximate 95% confidence set of pedigrees, and pink bars show the proportion of simulations in which the null pedigree is rejected by a log-likelihood ratio test.

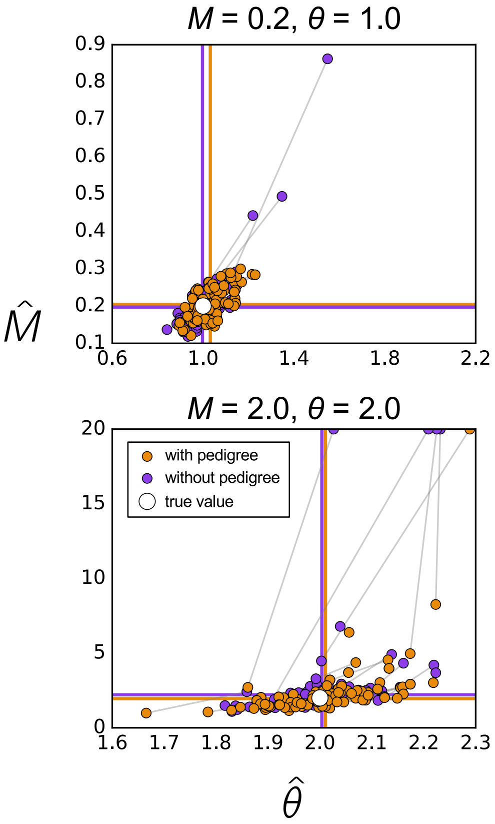

Figure 7: Maximum-likelihood mutation rates and migration rates for datasets of 1000 loci segregated through random pedigrees simulated in a two-deme population with deme size $N = 50$. Each point depicts the maximum-likelihood estimates of θ and M for a particular simulation. Orange points show estimates obtained when the pedigree is included as a free parameter, and purple points show estimates obtained when the pedigree is assumed to have no effect on the data. True parameter values are shown with white circles, and horizontal and vertical lines show means across replicates. Gray lines connect estimates calculated from the same dataset.

morphic sites, often biallelic SNPs, that are assumed to segregate independently. Likelihoods are calculated from the probabilities of observing the observed genotypes under the rules of Hardy-Weinberg equilibrium. These methods are well suited for samples of a large number of SNP loci sampled from a large number of individuals. The inference procedure we have implemented, on the other hand, is capable of handling a sample of only a few individuals ($n \approx 4$), and likelihood calculations in our method are based on the coalescent in an explicit population genetic model, the parameters of which being the primary objects of inference. The pedigree is also explicitly modeled. While we have shown that our approach works well when its assumptions hold, if the narrow assumptions of our model do not hold, or if the primary goal of inference is to infer recent features of the sample pedigree *per se*, other, more flexible methods are likely to perform better.

Underlying our inference method is a hybrid approach to modeling the coalescent. Probabilities of coalescence are determined by the sample pedigree in the recent past, and then the standard coalescent is used to model the more distant past. This is similar to how Bhaskar *et al.* (2014) modeled coalescence when the sample size approaches the population size. In such a scenario, they suggest using a discrete-time Wright-Fisher model to model coalescence for the first few generations back in time and then use the standard coalescent model after the number of surviving ancestral lineages becomes much less than the population size. We note that in a situation where the sample size nears the population size in a diploid population, there will be numerous common ancestor events and admixture events in the recent sample pedigree, so it may be important to consider the pedigree when genetic variation is sampled from a fixed set of individuals at independently segregating loci.

The sample reconfiguration framework could be extended to models that allow the demography of the population to vary over time (e.g., Gutenkunst *et al.*, 2009; Kamm *et al.*, 2015). In such an application, if only a few individuals are sampled, it would be important to distinguish between the effects of very recent events that are likely particular to the sample and the effects of events that are shared by all individuals in the population. The latter category of events are more naturally considered demographic history. On the other hand, if a sizable fraction of the population is sampled, inferred pedigree features may be used to learn more directly about the demography of the population in the last few generations. As sample sizes increase from the tens of thousands into the hundreds of thousands and millions (Stephens *et al.*, 2015), it will become more and more possible to reconstruct large (but sparse) pedigrees that are directly informative about recent demographic processes.

Unexpected close relatedness is frequently found in large genomic datasets (e.g. Gazal *et al.*, 2015; Pemberton *et al.*, 2010; Rosenberg, 2006). It is common practice to remove closely related individuals (and in some cases, individuals with admixed ancestry) from the sample prior to analysis, but this unnecessarily reduces the amount of information that is available for inference. What is needed is a fully integrative method of making inferences from pedigrees and genetic variation, properly incorporating information about both the recent past contained in the sample pedigree and the more distant past that is the more typical domain of population genetic demographic inference. Here, by performing simulations of coalescence through pedigrees, we have justified a sample reconfiguration framework for modeling coalescence in pedigrees, and we have demonstrated how this can be incorporated into coalescent-based demographic inference in order to produce unbiased estimates of demographic parameters even when there is recent relatedness or admixed ancestry amongst the sampled individuals.

4. Acknowledgments

We thank Seungsoo Kim, Mark Martinez, Janet Song, and Katherine Xue for their their assistance during the early phases of this project. We also thank Léandra King and Julia Palacios for helpful discussion and the Research Computing Group at Harvard University for computing resources.

5. References

Alexander, D. H., J. Novembre, and K. Lange, 2009 Fast model-based estimation of ancestry in unrelated individuals. Genome Research **19**: 1655–1664.

Barton, N. H., and A. M. Etheridge, 2011 The relation between reproductive value and genetic contribution. Genetics **188**: 953–973.

Bhaskar, A., A. G. Clark, and Y. S. Song, 2014 Distortion of genealogical properties when the sample is very large. Proceedings of the National Academy of Sciences **111**: 2385–2390.

Chang, J. T., 1999 Recent common ancestors of all present-day individuals. Advances in Applied Probability **31**: 1002–1026.

Derrida, B., S. C. Manrubia, and D. H. Zenette, 2000 On the genealogy of a population of biparental individuals. Journal of Theoretical Biology **203**: 303–315.

Gazal, S., M. Sahbatou, M.-C. Babron, E. Génin, and A.-L. Leutenegger, 2015 High level of inbreeding in final phase of 1000 Genomes Project. Scientific Reports **5**.

Griffiths, R. C., and S. Tavaré, 1994 Ancestral Inference in Population Genetics. Statistical Science **9**: 307–319.

Gutenkunst, R. N., R. D. Hernandez, S. H. Williamson, and C. D. Bustamante, 2009 Inferring the Joint Demographic History of Multiple Populations from Multidimensional SNP Frequency Data. PLoS Genetics **5**: e1000695.

Hein, J., M. H. Schierup, and C. Wiuf, 2005 *Gene Genealogies, Variation and Evolution: A Primer in Coalescent Theory*. Oxford University Press, Oxford, 1 edition.

Hudson, R. R., 1983 Testing the constant-rate neutral allele model with protein sequence data. Evolution **37**: 203–217.

Hudson, R. R., 2002 Generating samples under a Wright–Fisher neutral model of genetic variation. Bioinformatics **18**: 337–338.

Kamm, J. A., J. Terhorst, and Y. S. Song, 2015 Efficient computation of the joint sample frequency spectra for multiple populations. arXiv:1503.01133 .

Kingman, J. F. C., 1982a The coalescent. Stochastic Processes and their Applications **13**: 235–248.

Kingman, J. F. C., 1982b On the genealogy of large populations. Journal of Applied Probability **19**: 27–43.

Li, H., and R. Durbin, 2011 Inference of human population history from individual whole-genome sequences. Nature **475**: 493–496.

Manichaikul, A., J. C. Mychaleckyj, S. S. Rich, K. Daly, M. Sale, *et al.*, 2010 Robust relationship inference in genome-wide association studies. Bioinformatics **26**: 2867–2873.

Moltke, I., and A. Albrechtsen, 2014 RelateAdmix: a software tool for estimating relatedness between admixed individuals. Bioinformatics **30**: 1027–1028.

Pemberton, T. J., C. Wang, J. Z. Li, and N. A. Rosenberg, 2010 Inference of unexpected genetic relatedness among individuals in HapMap Phase III. The American Journal of Human Genetics **87**: 457–464.

Pritchard, J. K., M. Stephens, and P. Donnelly, 2000 Inference of population structure using multilocus genotype data. Genetics **155**: 945–959.

Raj, A., M. Stephens, and J. K. Pritchard, 2014 fastSTRUCTURE: Variational inference of population structure in large SNP data sets. Genetics **197**: 573–589.

Rohde, D. L. T., S. Olson, and J. T. Chang, 2004 Modelling the recent common ancestry of all living humans. Nature **431**: 562–566.

Rosenberg, N. A., 2006 Standardized subsets of the HGDP-CEPH Human Genome Diversity Cell Line Panel, accounting for atypical and duplicated samples and pairs of close relatives. Annals of Human Genetics **70**: 841–847.

Schiffels, S., and R. Durbin, 2014 Inferring human population size and separation history from multiple genome sequences. Nature Genetics **46**: 919–925.

Sheehan, S., K. Harris, and Y. S. Song, 2013 Estimating variable effective population sizes from multiple genomes: a sequentially Markov conditional sampling distribution approach. Genetics **194**: 647–662.

Stephens, Z. D., S. Y. Lee, F. Faghri, R. H. Campbell, C. Zhai, *et al.*, 2015 Big Data: Astronomical or Genomical? PLOS Biol **13**: e1002195.

Tajima, F., 1983 Evolutionary relationship of DNA sequences in finite populations. Genetics **105**: 437–460.

Tang, H., J. Peng, P. Wang, and N. J. Risch, 2005 Estimation of individual admixture: Analytical and study design considerations. Genetic Epidemiology **28**: 289–301.

Thornton, T., H. Tang, T. Hoffmann, H. Ochs-Balcom, B. Caan, *et al.*, 2012 Estimating kinship in admixed populations. The American Journal of Human Genetics **91**: 122–138.

Wakeley, J., 2009 *Coalescent Theory: An Introduction*. Roberts and Co., Greenwood Village, CO.

Wakeley, J., L. King, B. S. Low, and S. Ramachandran, 2012 Gene Genealogies Within a Fixed Pedigree, and the Robustness of Kingmans Coalescent. Genetics **190**: 1433–1445.

Watterson, G., 1975 On the number of segregating sites in genetical models without recombination. Theoretical Population Biology **7**: 256–276.

Wilson, G. A., and B. Rannala, 2003 Bayesian Inference of Recent Migration Rates Using Multilocus Genotypes. Genetics **163**: 1177–1191.

Wright, S., 1951 The genetical structure of populations. Annals of Eugenics **15**: 323–354.

Wu, Y., 2010 Exact computation of coalescent likelihood for panmictic and subdivided populations under the infinite sites model. IEEE Transactions on Computational Biology and Bioinformatics **7**: 611–618.

6. Appendix A: Pedigrees and biased estimators of $\boldsymbol{\theta}$

Estimates of the population-scaled mutation rate $\theta = 4N\mu$ will be downwardly biased if there is recent IBD in the sample, since sequences will be identical with an artificially inflated probability, and this resembles coalescence prior to any mutation between the two identical sequences.

Suppose that we sample two copies of a DNA sequence from each of n diploid individuals from a panmictic population. As in the main text, we index these sequences with $\mathcal{I}_n := \{1^{\mathrm{m}}, 1^{\mathrm{p}}, 2^{\mathrm{m}}, 2^{\mathrm{p}}, \ldots, n^{\mathrm{m}}, n^{\mathrm{p}}\}$. Let $\mathbb{P}_{\mathcal{I}_n}$ be the set of partitions of $\mathcal{I}_n$. Each pedigree $\mathcal{P}$ induces a set of sample reconfigurations $\mathcal{R}(\mathcal{P}) \subseteq \mathbb{P}_{\mathcal{I}_n}$, where each partition $r \in \mathcal{R}(\mathcal{P})$ represents a possible outcome of segregation through the recent sample pedigree. Each reconfiguration $r \in \mathcal{R}(\mathcal{P})$ contains $|r|$ non-empty, disjoint subsets, each representing a distinct lineage that survives after segregation through the recent pedigree. Associated with each pedigree is also a probability distribution $\Pr(r \mid \mathcal{P}), r \in \mathcal{R}(\mathcal{P})$ representing the Mendelian probabilities of the different sample reconfigurations.

We consider the bias of three estimators of θ that are unbiased in the absence of recent IBD. One estimator of θ we consider is Watterson's (1975) estimator

$$\hat{\theta}_S = \frac{\sum_{i=1}^{L} S^{(i)}}{a_n L}, \tag{6}$$

where n is the (haploid) sample size, $S^{(i)}$ is the number of segregating sites at locus i, L is the number of loci, and $a_n = \sum_{i=1}^{n-1} 1/i$. The expected value of $\hat{\theta}_S$ given pedigree $\mathcal{P}$ is

$$\begin{aligned} \mathrm{E}\left[\hat{\theta}_S \mid \mathcal{P}\right] &= \sum_{r\in\mathcal{R}(\mathcal{P})} \mathrm{E}\left[\hat{\theta}_S \mid r\right] \Pr(r \mid \mathcal{P}) \\ &= \sum_{r\in\mathcal{R}(\mathcal{P})} \mathrm{E}\left[\frac{S^{(1)}}{a_n} \mid r\right] \Pr(r \mid \mathcal{P}) \\ &= \theta \sum_{r\in\mathcal{R}(\mathcal{P})} \frac{a_{|r|}}{a_n} \Pr(r \mid \mathcal{R}(\mathcal{P})). \end{aligned} \tag{7}$$

This follows from the fact that when there are $|r|$ lineages surviving the recent pedigree, the expected number of segregating sites for that sample is $\theta \sum_{i=1}^{|r|-1} \frac{\theta}{i} = \theta a_{|r|}$.

A second estimator of θ is $\hat{\pi}$, the mean number of differences between all pairs of sequences in a sample, which can be written in terms of the site-frequency spectrum:

$$\hat{\pi} = \frac{1}{\binom{n}{2}} \sum_{i=1}^{n-1} i(n-i)\hat{\xi}_i, \tag{8}$$

where $\hat{\xi}_i$ is the number of segregating sites present in i sequences in the sample.

To calculate the expected value of $\hat{\pi}$ given $\mathcal{P}$, it is necessary to consider how IBD in the recent pedigree changes the site frequency spectrum. Define $\mathcal{S}^{(\mathrm{n})} := \{A \subseteq \Omega : \Omega \in \mathbb{P}_{\mathcal{I}_n}\}$ and $\psi : \mathbb{P}_{\mathcal{I}_n} \times \mathbb{N} \to \mathcal{S}^{(\mathrm{n})}$ as

$$\psi(\Omega, i) := \{\omega \subseteq \Omega : \sum_{g\in\omega} |g| = i\}. \tag{9}$$

That is, $\psi(\Omega, i)$ is the set of all subsets of the partition Ω such that the total size of all the groups in each subset is i. For example, for $n = 3$ and

$$\Omega = \left\{\left\{1^{\mathrm{m}}, 1^{\mathrm{p}}, 2^{\mathrm{m}}\right\}, \left\{2^{\mathrm{p}}, 3^{\mathrm{m}}\right\}, \left\{3^{\mathrm{p}}\right\}\right\},$$

$$\begin{aligned}
\psi(\Omega, 1) &= \{\{\{3^{\mathrm{p}}\}\}\} \\
\psi(\Omega, 2) &= \{\{\{2^{\mathrm{p}}, 3^{\mathrm{m}}\}\}\} \\
\psi(\Omega, 3) &= \{\{\{1^{\mathrm{m}}, 1^{\mathrm{p}}, 2^{\mathrm{m}}\}\}, \{\{2^{\mathrm{p}}, 3^{\mathrm{m}}\}, \{3^{\mathrm{p}}\}\}\} \\
\psi(\Omega, 4) &= \{\{\{1^{\mathrm{m}}, 1^{\mathrm{p}}, 2^{\mathrm{m}}\}, \{3^{\mathrm{p}}\}\}\} \\
\psi(\Omega, 5) &= \{\{\{1^{\mathrm{m}}, 1^{\mathrm{p}}, 2^{\mathrm{m}}\}, \{2^{\mathrm{p}}, 3^{\mathrm{m}}\}\}\}.
\end{aligned}$$

Then the expectation of $\hat{\xi}_i$ given reconfiguration $r \in \mathcal{R}(\mathcal{P})$ is

$$\mathrm{E}\left[\hat{\xi}_i \mid r\right] = \sum_{\omega \in \psi(r,i)} \frac{\theta}{|\omega| \binom{|r|}{|\omega|}}, \tag{10}$$

since each segregating site that is present in i present-day lineages must have occurred on the branch ancestral to the post-IBD lineages represented by some $\omega \in \psi(r, i)$. The expected number of mutations occurring on a branch subtending $|\omega|$ lineages is $\theta/|\omega|$, and the expected fraction of such mutations that occur on the branch ancestral to the lineages in ω is $1/\binom{|r|}{|\omega|}$, by exchangeability. This gives the expectation of $\hat{\pi}$ conditional on the pedigree:

$$\begin{aligned}
\mathrm{E}\left[\hat{\pi} \mid \mathcal{P}\right] &= \sum_{r \in \mathcal{R}(\mathcal{P})} \mathrm{E}\left[\hat{\pi} \mid r\right] \Pr(r \mid \mathcal{P}) \\
&= \frac{1}{\binom{n}{2}} \sum_{\mathcal{R} \in \mathcal{P}} \Pr(\mathcal{R} \mid \mathcal{P}) \sum_{i=1}^{n-1} i(n-i)\, \mathrm{E}[\hat{\xi}_i \mid r] \\
&= \frac{\theta}{\binom{n}{2}} \sum_{r \in \mathcal{R}(\mathcal{P})} \Pr(r \mid \mathcal{P}) \sum_{i=1}^{n-1} i(n-i) \sum_{\omega \in \psi(r,i)} \frac{1}{|\omega| \binom{|\mathcal{R}|}{|w|}}
\end{aligned} \tag{11}$$

A third estimator of θ is $\hat{\xi}_1$, the number of singletons in the sample. The conditional expectation of $\hat{\xi}_1$ given a pedigree $\mathcal{P}$ is

$$\mathrm{E}\left[\hat{\xi}_1 \mid \mathcal{P}\right] = \theta \sum_{r \in \mathcal{R}(\mathcal{P})} \frac{|\psi(r,1)|}{|\mathcal{R}(\mathcal{P})|} \Pr(r \mid \mathcal{P}), \tag{12}$$

since only those mutations that occur on lineages that have not coalesced with any other lineages in the early pedigree can produce singletons.

To validate these calculations, we performed simulations of 200 loci sampled from individuals whose pedigree includes some degree of IBD or inbreeding. The simulations confirm the calculated biases for the different estimators of θ (Fig. S5).

7. Appendix B: Pedigrees and biased estimators of M

In a constant-sized structured population with two demes and a constant rate of migration $M = 4Nm$ between demes, the expected within-deme and between-deme pairwise coalescence times are

$$\mathrm{E}[T_w] = 2 \tag{13}$$

$$\mathrm{E}[T_b] = 2 + 1/M, \tag{14}$$

Let π_w and π_b be the within-deme and between-deme mean pairwise diversity, respectively. Since $\mathrm{E}[\pi_w] = \theta\, \mathrm{E}[T_w]$ and $\mathrm{E}[\pi_b] = \theta\, \mathrm{E}[T_b]$, one estimator of M is

$$\hat{M} = \frac{\hat{\pi}_w}{2(\hat{\pi}_b - \hat{\pi}_w)}. \tag{15}$$

If some individuals in the sample are recently admixed, $\hat{M}$ will be biased. In general it is not possible to calculate $\mathrm{E}[\hat{M}]$, but it can be approximated by

$$\mathrm{E}[\hat{M}] \approx \frac{\mathrm{E}[\hat{\pi}_w]}{2(\mathrm{E}[\hat{\pi}_b] - \mathrm{E}[\hat{\pi}_w])}. \tag{16}$$

We sample two sequences from each of n_1 individuals from deme 1 and n_2 individuals from deme 2, defining $n = n_1 + n_2$ as the total (diploid) sample size. The sample is again indexed by $\mathcal{I}_n = \{1^{\mathrm{m}}, 1^{\mathrm{p}}, \ldots, n^{\mathrm{m}}, n^{\mathrm{p}}\}$, and we assume that the first $2n_1$ of these indices correspond to sequences sampled from deme 1 and the last $2n_2$ from deme 2.

In the context of a two-deme population, each group in the partitioned sample $r \in \mathcal{R}(\mathcal{P})$ is labeled 1 or 2 to indicate which deme the lineage is found in after segregation back in time through the recent sample pedigree. For two-deme reconfiguration r, let $d(r, i, j)$, $i, j \in \{1, 2\}$, be a function that gives the number of lineages originally sampled from deme i that are found in deme j after segregation through the recent sample pedigree.

Assume that the recent sample pedigree contains admixture but no IBD. In this case, we can write the expectations of $\hat{\pi}_w$ and $\hat{\pi}_b$ conditional on reconfiguration r as

$$\begin{aligned}
\mathrm{E}\left[\hat{\pi}_w \mid r\right] = {}& \frac{1}{\binom{2n_1}{2} + \binom{2n_2}{2}} \times \\
& \left\{ \theta\, \mathrm{E}[T_w] \sum_{i=1}^{2} \binom{d(r,i,i)}{2} + \right. \\
& \left. \theta\, \mathrm{E}[T_b]\left(d(r,1,1)d(r,1,2) + d(r,2,2)d(r,2,1)\right) \right\}
\end{aligned} \tag{17}$$

and

$$\begin{aligned}\mathrm{E}\left[\hat{\pi}_b \mid r\right] = \frac{1}{4n_1n_2}\times \\ \Big\{\theta\,\mathrm{E}[T_b]\big(d(r,1,1)d(r,2,2)+d(r,1,2)d(r,2,1)\big)+ \\ \theta\,\mathrm{E}[T_w]\left(d(r,1,2)d(r,2,2)+d(r,2,1)d(r,1,1)\right)\Big\}. \end{aligned} \tag{18}$$

The approximate expectation of $\hat{M}$ conditional on a pedigree $\mathcal{P}$ can be calculated using (17) and (18) together with

$$\mathrm{E}[\hat{\pi}_w \mid \mathcal{P}] = \sum_{r\in\mathcal{R}(\mathcal{P})} \mathrm{E}[\hat{\pi}_w \mid r]\,\mathrm{Pr}(r \mid \mathcal{P})$$

and

$$\mathrm{E}[\hat{\pi}_b \mid \mathcal{P}] = \sum_{r\in\mathcal{R}(\mathcal{P})} \mathrm{E}[\hat{\pi}_b \mid r]\,\mathrm{Pr}(r \mid \mathcal{P}).$$

Simulations of infinite-sites loci taken from samples with a single admixed ancestor confirm these calculations (Fig. S6). This method of approximating $\mathrm{E}[\hat{M} \mid \mathcal{P}]$ could be extended to accommodate recent sample pedigrees that contain both IBD and admixture, but we do not pursue this here.

8. Supplementary Material

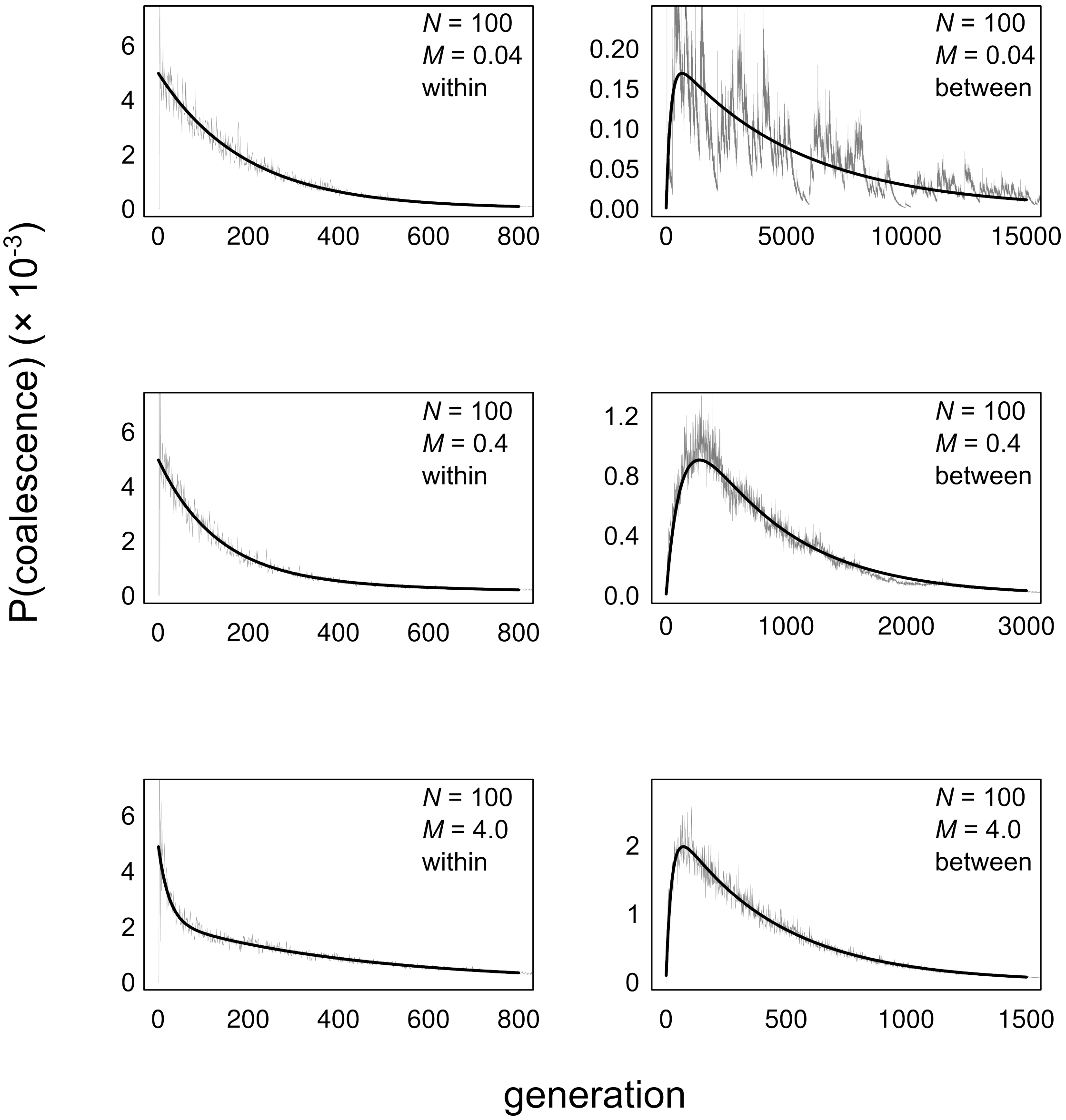

Figure S1: Distributions of coalescence times in two-deme populations with $N = 100$ individuals in each deme. Each panel shows a distribution of coalescence times for a particular value of the migration rate $M = 4Nm$. For panels in the left column, two individuals were sampled from the same deme ("within-deme" sampling), and in the right column two individuals were sampled from different demes.

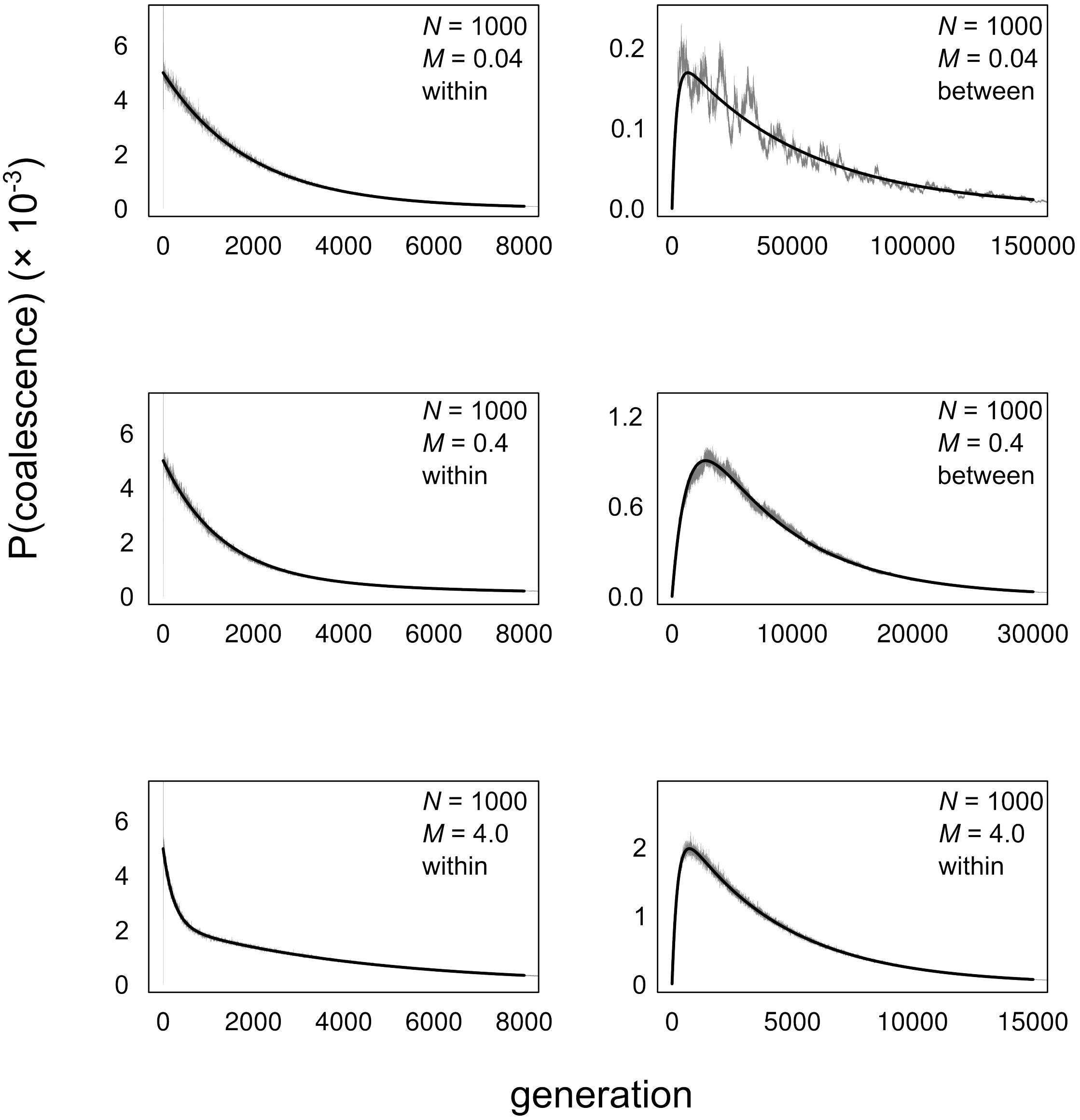

Figure S2: Distributions of coalescence times in two-deme populations with $N = 1000$ individuals in each deme. Each panel shows a distribution of coalescence times for a particular value of the migration rate $M = 4Nm$. For panels in the left column, two individuals were sampled from the same deme ("within-deme" sampling), and in the right column two individuals were sampled from different demes.

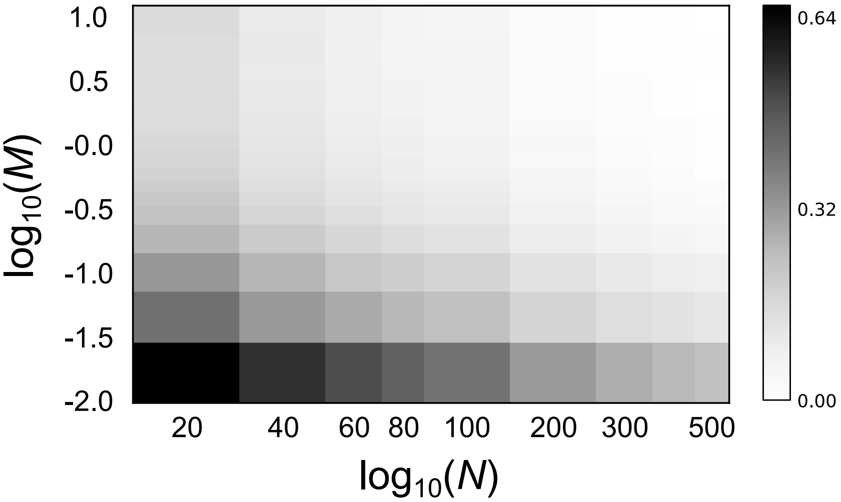

Figure S3: Total variation distance between pairwise coalescence time distributions in pedigrees versus standard theory. For each point on the grid, the total variation distance from the prediction of standard theory was averaged over 20 pedigrees with the corresponding N and M.

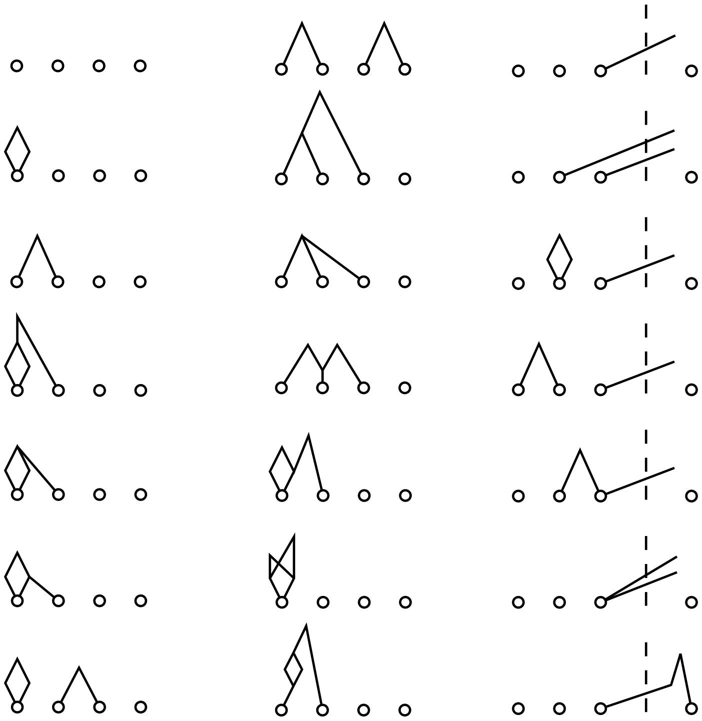

Figure S4: Distinct Wright-Fisher pedigree shapes with two or fewer IBD or admixture events in a two-deme population. Only relationships involved in IBD or admixture events are shown. All pedigrees have non-overlapping generations, and sampled individuals (white circles) are living in the present generation. To produce all distinct pedigrees, it is necessary to consider all the ways of indexing the individuals involved in the IBD and admixture events, as well as the all of the possible timings of these events.

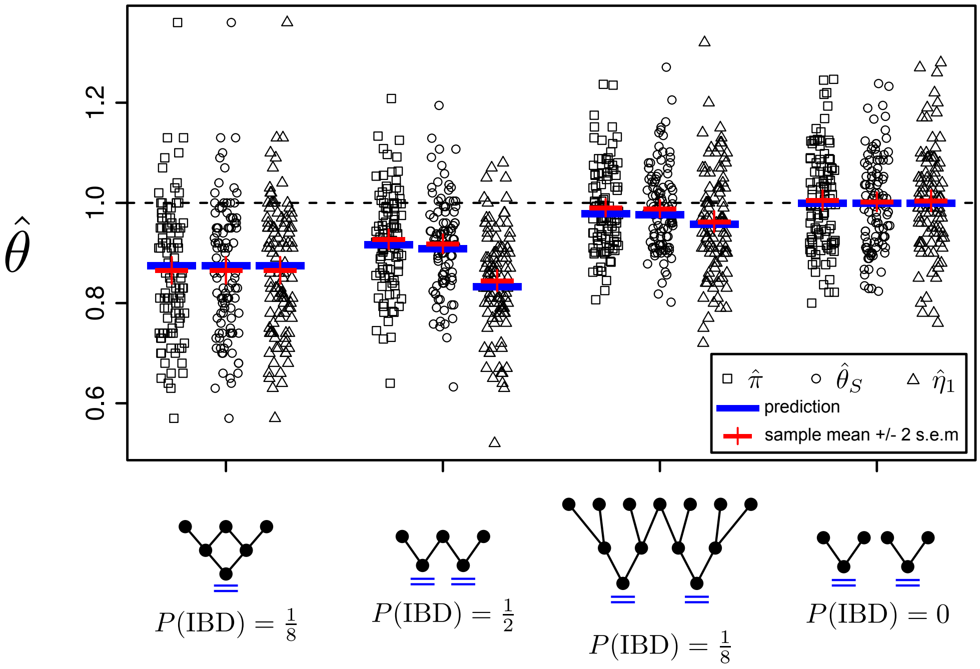

Figure S5: Estimates of $\theta = 4N\mu$ based on 100 replicate simulations of 200 loci segregating through sample pedigrees featuring identity by descent. Blue lines indicate theoretical predictions for individual pedigrees and estimators. Red horizontal lines indicate sample means, and red vertical lines indicate twice the standard error of the mean. The true value of $\theta = 1.0$ is indicated by the dashed line. Note that for the first pedigree, with $n = 2$, the three estimators are equivalent.

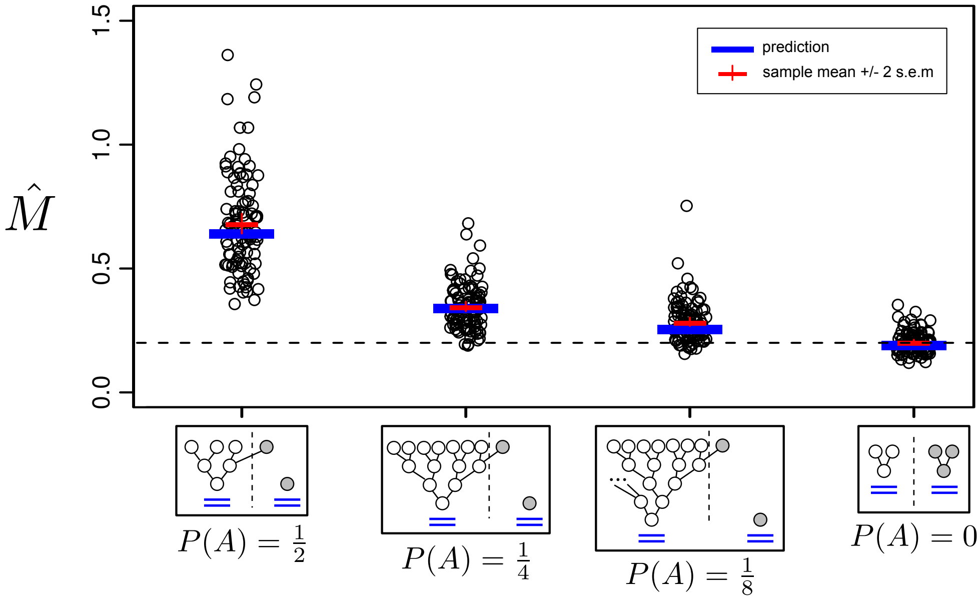

Figure S6: Estimates of M from the estimator given by (15) in replicate simulated genetic datasets of 100 loci segregating through sample pedigrees containing recent admixture. Blue lines indicate theoretical predictions for individual pedigrees. Red horizontal lines indicate sample means, and red vertical lines indicate twice the standard error of the mean. The true value of $M = 0.2$ is indicated by the dashed line. $P(A)$ indicates the probability of admixture in the indicated pedigrees.