Theoretical Population Biology 80 (2011) 100–113

Contents lists available at ScienceDirect

Theoretical Population Biology

journal homepage: www.elsevier.com/locate/tpb

On the size distribution of private microsatellite alleles

Zachary A. Szpiech[a,*], Noah A. Rosenberg[a,b,c]

[a] *Center for Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI 48109, USA*
[b] *Department of Human Genetics, University of Michigan, Ann Arbor, MI 48109, USA*
[c] *Life Sciences Institute, University of Michigan, Ann Arbor, MI 48109, USA*

ARTICLE INFO

Article history:
Received 17 December 2010
Available online 15 April 2011

Keywords:
Coalescent
Microsatellites
Private alleles

ABSTRACT

Private microsatellite alleles tend to be found in the tails rather than in the interior of the allele size distribution. To explain this phenomenon, we have investigated the size distribution of private alleles in a coalescent model of two populations, assuming the symmetric stepwise mutation model as the mode of microsatellite mutation. For the case in which four alleles are sampled, two from each population, we condition on the configuration in which three distinct allele sizes are present, one of which is common to both populations, one of which is private to one population, and the third of which is private to the other population. Conditional on this configuration, we calculate the probability that the two private alleles occupy the two tails of the size distribution. This probability, which increases as a function of mutation rate and divergence time between the two populations, is seen to be greater than the value that would be predicted if there was no relationship between privacy and location in the allele size distribution. In accordance with the prediction of the model, we find that in pairs of human populations, the frequency with which private microsatellite alleles occur in the tails of the allele size distribution increases as a function of genetic differentiation between populations.

1. Introduction

Private alleles are alleles that are found only in a single population among a broader collection of populations. These alleles have proven to be informative for diverse types of population-genetic studies, in such areas as molecular ecology and conservation genetics (e.g. Petit et al., 1998; Parker et al., 1999; Fiumera et al., 2000; Neel and Cummings, 2003; Torres et al., 2003; Kalinowski, 2004) and human evolutionary genetics (e.g. Neel, 1973, 1978; Neel and Thompson, 1978; Calafell et al., 1998; Schroeder et al., 2007; Szpiech et al., 2008).

Some of the first investigations of private alleles trace to studies of private electrophoretic variants in native American groups from South America (Neel, 1973, 1978; Neel and Thompson, 1978). Using private alleles, Neel and colleagues obtained estimates of mutation rates in these populations. Slatkin (1985) and Barton and Slatkin (1986) showed that private alleles can contribute to indicators of gene flow, finding in theoretical models of population structure that the occurrence of private alleles was related to the mean number of migrants exchanged per generation between populations. Private alleles have also been used in empirical studies of human migrations. Calafell et al. (1998) noted that in human populations, the mean number of private alleles is greater in Africa, providing support to models of human migration out of Africa. Schroeder et al. (2007) argued on the basis of a private allele ubiquitous in the Americas that all modern native American populations are descended from the same founding population.

One recent study, which investigated 678 microsatellite markers in 29 native American populations from North, Central, and South America (Wang et al., 2007), has identified a peculiar property of private alleles. Wang et al. (2007) characterized the distribution of private alleles across four subregions in the Americas, observing that private microsatellite alleles were found in the tails rather than in the interior of the allele size distribution more often than was expected by chance. In other words, private alleles at a locus frequently had very long or very short repeat lengths with respect to the other alleles at the locus.

Here we take a modeling approach to examine the reasons underlying the frequent occurrence of private alleles on the edges of the allele size distribution. Using a simple coalescent model, we assess the properties of microsatellite private alleles, thereby helping to explain patterns that exist in the relationship between privacy and allele size across human populations.

2. Theory

Let $\{x_1 x_2 / x_3 x_4\}$ denote four sampled microsatellite alleles in two populations, where x_i indicates the allele size for sampled

[*] Corresponding author.
E-mail address: szpiechz@umich.edu (Z.A. Szpiech).

0040-5809/$ – see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.tpb.2011.03.006

allele i, and the forward slash separates alleles from different populations. We restrict our attention to cases with four alleles; a scenario with two alleles each in two populations gives the smallest sample size useful for examining the phenomenon of interest, as we will explain below. Because the four-allele case involves a tractable number of calculations, it is possible in this case to mathematically investigate the position of private alleles in the size distribution.

We map sets of four allele sizes in two populations to one of seven possible configurations of identity and nonidentity, using the letters A, B, C, and D to denote distinct allele sizes. Thus, if two sampled alleles are identical by state (IBS), we indicate this identity by assigning the alleles the same letter. For example, if all four sampled alleles are IBS, we represent the allele configuration by $\{AA/AA\}$. If one allele in population 1 is IBS to an allele in population 2 and the other allele in population 1 is IBS to the other allele in population 2 (and no alleles are IBS within populations), then we represent the allele configuration by $\{AB/AB\}$. We label the seven possible configurations by C_i for $i \in \{1, \dots, 7\}$, and we list them in Table 1.

We are interested in comparing private and shared alleles on the basis of size. In particular, we wish to examine whether alleles lie on the edges of the size distribution, that is, whether they have the longest or shortest lengths. To have a sensible definition of the "edges" and interior of the allele size distribution, we must have at least three distinct alleles among the four sampled alleles that we consider. Furthermore, because we are concerned with the location of private alleles with respect to shared alleles, we must have at least one shared allele and one private allele. The only one of the seven configurations of four alleles that satisfies both of these requirements—and that therefore enables a computation of the probability that private alleles lie on the edges of the allele size distribution—is C_6 (configuration $\{AB/AC\}$). This configuration, with sample size four, provides the smallest scenario that contains both private alleles and shared alleles and that contains both edges and an interior of the allele size distribution. We aim to compute the probability that B and C, the two private alleles in configuration $\{AB/AC\}$, both lie on the edges of the size distribution, conditional on this configuration being produced.

2.1. A naïve argument

If we disregard the genealogical relatedness of the alleles in our two-population four-allele model, what do we expect for the probability that the private alleles lie on the edges? There are six possible orderings of the three allele sizes A, B, and C ($A < B < C$, $A < C < B$, etc.), and, if no relationship exists between the size of an allele and its status as shared or private, we expect the six orderings to be equiprobable. Two of the six orderings place the private alleles B and C on the edges of the size distribution. Under this simple argument, we would expect the probability that both private alleles lie on the edges to be $1/3$.

This argument gives an initial sense of what might be predicted for the probability that the private alleles lie on the edges of the size distribution. However, it disregards the fact that the alleles are related through a common ancestor. We now turn to a genealogical argument that more directly models this relationship.

2.2. The probability of microsatellite configurations

To account for the genealogical relatedness of the four alleles in obtaining a prediction of the probability that private alleles lie on the edges of the allele size distribution, we use the coalescent with symmetric stepwise mutation. Initially, we consider the two populations to have instantaneously diverged zero coalescent time units in the past ($t_d = 0$). Later, we will consider arbitrary values of the divergence time t_d.

Table 1
The seven possible configurations of four alleles in two populations and the counts of shared, private, and total distinct alleles for each configuration.

Event	Configuration	Number of shared alleles	Number of private alleles	Total number of distinct alleles
C_1	$\{AA/AA\}$	1	0	1
C_2	$\{AA/AB\}$	1	1	2
C_3	$\{AA/BB\}$	0	2	2
C_4	$\{AB/AB\}$	2	0	2
C_5	$\{AA/BC\}$	0	3	3
C_6	$\{AB/AC\}$	1	2	3
C_7	$\{AB/CD\}$	0	4	4

To calculate the desired probability, we first condition on the $\{AB/AC\}$ allele configuration (configuration C_6), the mutation rate, and the coalescence times of the genealogy. By considering the probability of a net change by d mutational steps along a genealogical branch, we construct the joint probability of an allele configuration and a particular labeled history for the four alleles, where the allele configuration refers to one of the seven scenarios in Table 1 and the labeled history refers to the sequence of coalescences (Fig. 1). We then calculate the total probability that the private alleles lie on the edges of the allele size distribution, summing across all labeled histories, and integrating over coalescence times to arrive at the desired probability, conditional only on a mutation rate θ.

Consider the events E_1: size$(B) <$ size$(A) <$ size(C), and E_2: size$(C) <$ size$(A) <$ size(B). These events are equiprobable, and we aim to calculate the probability

$$\mathbb{P}[E_1 \cup E_2 | C_6, \theta] = \frac{2\mathbb{P}[E_1, C_6|\theta]}{\mathbb{P}[C_6|\theta]}. \tag{1}$$

Under the symmetric single stepwise mutation model, a microsatellite allele can mutate by only one step at a time in either a positive or negative direction, and the probability of mutating $+1$ step is equal to the probability of mutating -1 step, independent of the size of the allele. We work with coalescent time units (units of $2N_e$ generations, where N_e is the effective size of each population, treated as containing diploid individuals) and with the population-scaled mutation rate $\theta = 4N_e\mu$, where μ is the per-locus per-generation mutation rate.

2.3. Mutations on a genealogical branch

The probability that a marker evolving according to the symmetric stepwise mutation model with population-scaled mutation rate θ has net change d units along a branch of length t coalescent time units is (Wehrhahn, 1975; Wilson and Balding, 1998)

$$f(|d|; t, \theta) = e^{-t\theta/2} I_{|d|}(t\theta/2), \tag{2}$$

where

$$I_{|d|}(x) = \sum_{k=0}^{\infty} \frac{(x/2)^{(2k+|d|)}}{k!(k+|d|)!}$$

is the modified Bessel function of the first kind (Gradshteyn and Ryzhik, 2000). Because positive and negative mutations are equally likely, we write f as a function of $|d|$ rather than d, which can be positive, negative or zero.

2.4. Probability of the set of allele sizes on a genealogical tree

We can use Eq. (2) to calculate the probability that changes along a coalescent tree ultimately give rise to a specified set of allele sizes. Considering that each branch evolves independently of

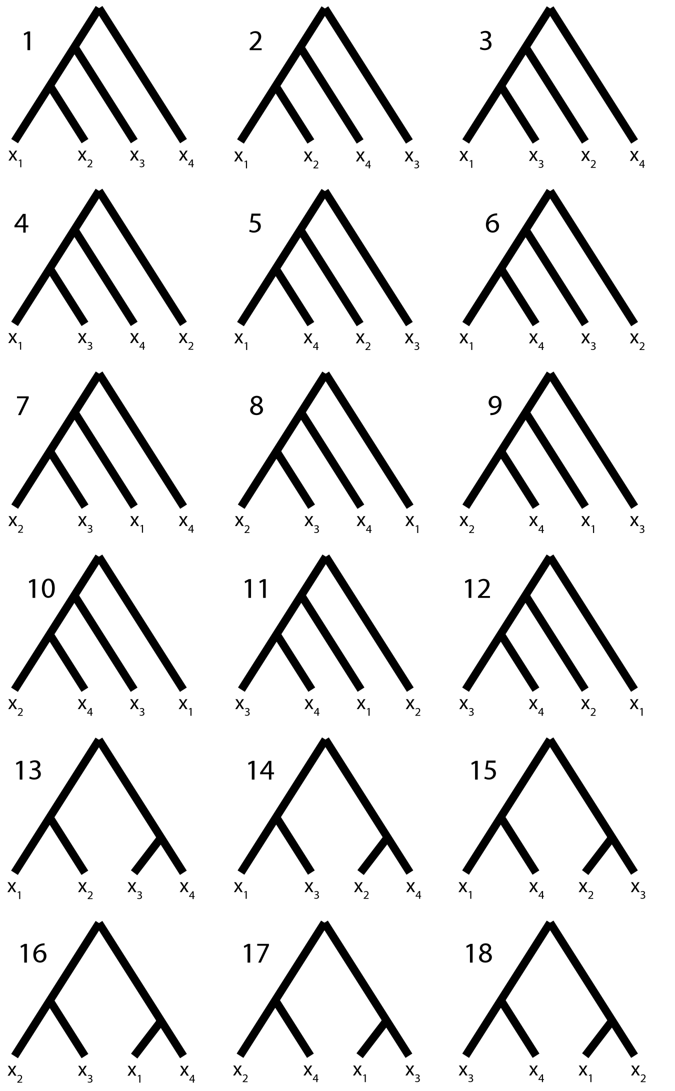


Fig. 1. An enumeration of all possible labeled histories relating four sampled alleles. Twelve histories have asymmetric topologies (1–12), and six histories have symmetric topologies (13–18).

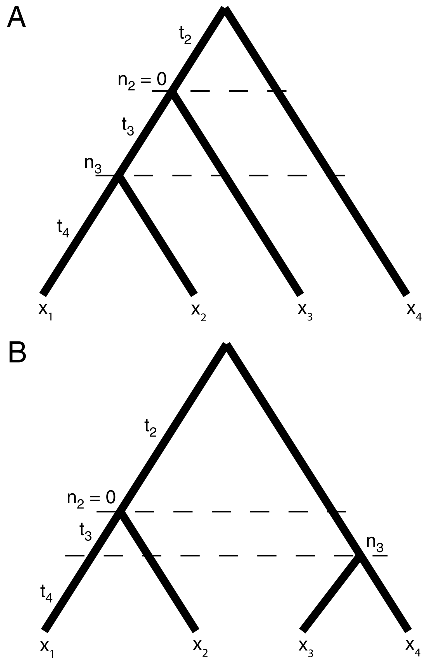


Fig. 2. Example labelings of the two possible unlabeled topologies for gene genealogies with four lineages. (A) The asymmetric topology and (B) the symmetric topology are parameterized by allele sizes at the nodes, $(x_1, x_2, x_3, x_4, n_3, n_2)$, and by coalescence times (t_2, t_3, t_4) indicating the lengths of certain segments of the branches.

the others, we calculate the probabilities of changes along individual branches and then multiply probabilities across branches to get the joint probability of all changes on the tree. There are two unlabeled topologies that we need to consider: an asymmetric topology (Fig. 2(A)) and a symmetric topology (Fig. 2(B)). Each topology is parameterized by a vector of allele sizes, $(n_2, n_3, x_1, x_2, x_3, x_4)$, and a vector of coalescence times, (t_2, t_3, t_4). The x variables represent the sizes of alleles at the leaf nodes, and n_i represents the size of the allele at the interior node located at the reduction of the number of distinct lineages to i. The coalescence time t_i represents the length of time during which there exist i distinct lineages. Initially, we treat the coalescence times as fixed, and later we will integrate the probabilities against the density of coalescence times to obtain a probability unconditional on t_2, t_3, and t_4. Because we assume that mutation probabilities do not depend on allele size, we can set the allele size of the most recent common ancestor of the four-allele sample (the root node) to 0 without loss of generality. However, following a choice similar to that of Pritchard and Feldman (1996) and Zhang and Rosenberg (2007), we instead choose to set $n_2 = 0$ rather than setting the root node to 0, and we treat the two branches that descend from the root as one branch with length equal to the sum of the lengths of its two constituent branches. This choice makes it possible to consider coalescent trees with five rather than six separate branches, thereby simplifying the computation.

Considering the asymmetric caterpillar topology (Fig. 2(A)), we obtain the joint probability of $(n_2, n_3, x_1, x_2, x_3, x_4)$ given (t_2, t_3, t_4) by calculating the probability of changing from n_2 to x_4 repeats along a branch of length $2t_2 + t_3 + t_4$, from n_2 to x_3 repeats along a branch of length $t_3 + t_4$, from n_2 to n_3 repeats along a branch of length t_3, from n_3 to x_2 repeats along a branch of length t_4, and from n_3 to x_1 repeats along a branch of length t_4. Assuming $n_2 = 0$ and multiplying these five probabilities together gives

$$
\begin{aligned}
&V^{\mathrm{cat}}(x_1, x_2, x_3, x_4, n_3, \Psi)\\
&\quad = f(|x_4|; 2t_2 + t_3 + t_4, \theta) \times f(|x_3|; t_3 + t_4, \theta) \times f(|n_3|; t_3, \theta)\\
&\qquad \times f(|n_3 - x_2|; t_4, \theta) \times f(|n_3 - x_1|; t_4, \theta),
\end{aligned}
$$

where $\Psi = (\tau, \theta)$ is a vector of parameters and $\tau = (t_2, t_3, t_4)$ is the vector of coalescence times. Similarly, for the symmetric topology, we calculate the probability of the set of allele sizes in Fig. 2(B) to get

$$
\begin{aligned}
&V^{sym}(x_1, x_2, x_3, x_4, n_3, \Psi)\\
&\quad = f(|n_3|; 2t_2 + t_3, \theta) \times f(|n_3 - x_4|; t_4, \theta) \times f(|n_3 - x_3|; t_4, \theta)\\
&\qquad \times f(|x_2|; t_3 + t_4, \theta) \times f(|x_1|; t_3 + t_4, \theta).
\end{aligned}
$$

2.5. Assigning alleles the roles of A, B, and C

There are 18 labeled histories for the alleles $\{x_1, x_2, x_3, x_4\}$, which we denote by T_i for $i \in \{1, \ldots, 18\}$ (Fig. 1). We can then calculate $\mathbb{P}[C_6|T_i, \Psi]$ by considering all possible ways to get configuration C_6 with labeled history T_i. Because we have defined $\{x_1, x_2\}$ to be in population 1 and $\{x_3, x_4\}$ to be in population 2, we need to consider four cases for each history, reflecting the four possible assignments of the allele sizes x_1, x_2, x_3, and x_4 to the

Table 2
The four allele size relationships possible for the $\{AB/AC\}$ allele configuration.

Case	Allele size relationship				Allele roles in $\{AB/AC\}$			
					x_1	x_2	x_3	x_4
1	$x_1 = x_3$	$x_2 \neq x_4$	$x_1 \neq x_2$	$x_1 \neq x_4$	A	B	A	C
2	$x_1 = x_4$	$x_2 \neq x_3$	$x_1 \neq x_2$	$x_1 \neq x_3$	A	B	C	A
3	$x_2 = x_3$	$x_1 \neq x_4$	$x_2 \neq x_1$	$x_2 \neq x_4$	B	A	A	C
4	$x_2 = x_4$	$x_1 \neq x_3$	$x_2 \neq x_1$	$x_2 \neq x_3$	B	A	C	A

roles of distinct alleles A, B, and C. These four cases are shown in Table 2.

If we represent the size of the shared allele (allele A) by n_A and the sizes of the two private alleles (B and C) by n_B and n_C, respectively, then we can calculate $\mathbb{P}[C_6|T_i, \Psi]$ by summing the individual probabilities of each of the four cases in Table 2. For example, consider T_1:

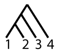
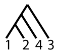
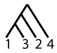
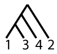
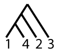
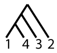
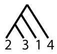
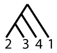
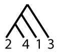
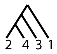
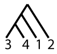
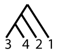
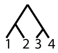
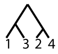
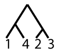
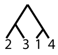
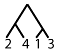
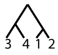
$$\begin{aligned}\mathbb{P}[C_6|T_1, \Psi]\\ &= \sum_{n_3=-\infty}^{\infty} \sum_{n_A=-\infty}^{\infty} \sum_{n_B=-\infty}^{\infty} \sum_{n_C=-\infty}^{\infty} V^{\text{cat}}(n_A, n_B, n_A, n_C, n_3, \Psi)\\ &\quad + V^{\text{cat}}(n_A, n_B, n_C, n_A, n_3, \Psi) + V^{\text{cat}}(n_B, n_A, n_A, n_C, n_3, \Psi)\\ &\quad + V^{\text{cat}}(n_B, n_A, n_C, n_A, n_3, \Psi). \qquad (3)\end{aligned}$$

Here, without loss of generality, we treat the private allele in the first population as the B allele and the private allele in the second population as the C allele. Similar calculations can be performed for the 17 remaining labeled histories (Table 3).

2.6. Summing over labeled histories

In order to calculate $\mathbb{P}[E_1, C_6|\Psi]$, we proceed exactly as in Eq. (3), conditioning on each history T_i, but we restrict the bounds

Table 3
The probability contributions for a given set of values $(n_A, n_B, n_C, n_3, \Psi)$ for each of the 18 labeled histories. These probabilities occur in the sums in Eqs. (6) and (7).

History number	History	Contribution
1		$V^{\text{cat}}(n_A, n_B, n_A, n_C, n_3, \Psi) + V^{\text{cat}}(n_A, n_B, n_C, n_A, n_3, \Psi) + V^{\text{cat}}(n_B, n_A, n_A, n_C, n_3, \Psi) + V^{\text{cat}}(n_B, n_A, n_C, n_A, n_3, \Psi)$
2		$V^{\text{cat}}(n_A, n_B, n_C, n_A, n_3, \Psi) + V^{\text{cat}}(n_A, n_B, n_A, n_C, n_3, \Psi) + V^{\text{cat}}(n_B, n_A, n_C, n_A, n_3, \Psi) + V^{\text{cat}}(n_B, n_A, n_A, n_C, n_3, \Psi)$
3		$V^{\text{cat}}(n_A, n_A, n_B, n_C, n_3, \Psi) + V^{\text{cat}}(n_A, n_C, n_B, n_A, n_3, \Psi) + V^{\text{cat}}(n_B, n_A, n_A, n_C, n_3, \Psi) + V^{\text{cat}}(n_B, n_C, n_A, n_A, n_3, \Psi)$
4		$V^{\text{cat}}(n_A, n_A, n_C, n_B, n_3, \Psi) + V^{\text{cat}}(n_A, n_C, n_A, n_B, n_3, \Psi) + V^{\text{cat}}(n_B, n_A, n_C, n_A, n_3, \Psi) + V^{\text{cat}}(n_B, n_C, n_A, n_A, n_3, \Psi)$
5		$V^{\text{cat}}(n_A, n_C, n_B, n_A, n_3, \Psi) + V^{\text{cat}}(n_A, n_A, n_B, n_C, n_3, \Psi) + V^{\text{cat}}(n_B, n_C, n_A, n_A, n_3, \Psi) + V^{\text{cat}}(n_B, n_A, n_A, n_C, n_3, \Psi)$
6		$V^{\text{cat}}(n_A, n_C, n_A, n_B, n_3, \Psi) + V^{\text{cat}}(n_A, n_A, n_C, n_B, n_3, \Psi) + V^{\text{cat}}(n_B, n_C, n_A, n_A, n_3, \Psi) + V^{\text{cat}}(n_B, n_A, n_C, n_A, n_3, \Psi)$
7		$V^{\text{cat}}(n_B, n_A, n_A, n_C, n_3, \Psi) + V^{\text{cat}}(n_B, n_C, n_A, n_A, n_3, \Psi) + V^{\text{cat}}(n_A, n_A, n_B, n_C, n_3, \Psi) + V^{\text{cat}}(n_A, n_C, n_B, n_A, n_3, \Psi)$
8		$V^{\text{cat}}(n_B, n_A, n_C, n_A, n_3, \Psi) + V^{\text{cat}}(n_B, n_C, n_A, n_A, n_3, \Psi) + V^{\text{cat}}(n_A, n_A, n_C, n_B, n_3, \Psi) + V^{\text{cat}}(n_A, n_C, n_A, n_B, n_3, \Psi)$
9		$V^{\text{cat}}(n_B, n_C, n_A, n_A, n_3, \Psi) + V^{\text{cat}}(n_B, n_A, n_A, n_C, n_3, \Psi) + V^{\text{cat}}(n_A, n_C, n_B, n_A, n_3, \Psi) + V^{\text{cat}}(n_A, n_A, n_B, n_C, n_3, \Psi)$
10		$V^{\text{cat}}(n_B, n_C, n_A, n_A, n_3, \Psi) + V^{\text{cat}}(n_B, n_A, n_C, n_A, n_3, \Psi) + V^{\text{cat}}(n_A, n_C, n_A, n_B, n_3, \Psi) + V^{\text{cat}}(n_A, n_A, n_C, n_B, n_3, \Psi)$
11		$V^{\text{cat}}(n_A, n_C, n_A, n_B, n_3, \Psi) + V^{\text{cat}}(n_C, n_A, n_A, n_B, n_3, \Psi) + V^{\text{cat}}(n_A, n_C, n_B, n_A, n_3, \Psi) + V^{\text{cat}}(n_C, n_A, n_B, n_A, n_3, \Psi)$
12		$V^{\text{cat}}(n_A, n_C, n_B, n_A, n_3, \Psi) + V^{\text{cat}}(n_C, n_A, n_B, n_A, n_3, \Psi) + V^{\text{cat}}(n_A, n_C, n_A, n_B, n_3, \Psi) + V^{\text{cat}}(n_C, n_A, n_A, n_B, n_3, \Psi)$
13		$V^{\text{sym}}(n_A, n_B, n_A, n_C, n_3, \Psi) + V^{\text{sym}}(n_A, n_B, n_C, n_A, n_3, \Psi) + V^{\text{sym}}(n_B, n_A, n_A, n_C, n_3, \Psi) + V^{\text{sym}}(n_B, n_A, n_C, n_A, n_3, \Psi)$
14		$V^{\text{sym}}(n_A, n_A, n_B, n_C, n_3, \Psi) + V^{\text{sym}}(n_A, n_C, n_B, n_A, n_3, \Psi) + V^{\text{sym}}(n_B, n_A, n_A, n_C, n_3, \Psi) + V^{\text{sym}}(n_B, n_C, n_A, n_A, n_3, \Psi)$
15		$V^{\text{sym}}(n_A, n_C, n_B, n_A, n_3, \Psi) + V^{\text{sym}}(n_A, n_A, n_B, n_C, n_3, \Psi) + V^{\text{sym}}(n_B, n_C, n_A, n_A, n_3, \Psi) + V^{\text{sym}}(n_B, n_A, n_A, n_C, n_3, \Psi)$
16		$V^{\text{sym}}(n_B, n_A, n_A, n_C, n_3, \Psi) + V^{\text{sym}}(n_B, n_C, n_A, n_A, n_3, \Psi) + V^{\text{sym}}(n_A, n_A, n_B, n_C, n_3, \Psi) + V^{\text{sym}}(n_A, n_C, n_B, n_A, n_3, \Psi)$
17		$V^{\text{sym}}(n_B, n_C, n_A, n_A, n_3, \Psi) + V^{\text{sym}}(n_B, n_A, n_A, n_C, n_3, \Psi) + V^{\text{sym}}(n_A, n_C, n_B, n_A, n_3, \Psi) + V^{\text{sym}}(n_A, n_A, n_B, n_C, n_3, \Psi)$
18		$V^{\text{sym}}(n_A, n_C, n_A, n_B, n_3, \Psi) + V^{\text{sym}}(n_C, n_A, n_A, n_B, n_3, \Psi) + V^{\text{sym}}(n_A, n_C, n_B, n_A, n_3, \Psi) + V^{\text{sym}}(n_C, n_A, n_B, n_A, n_3, \Psi)$

of summation on n_B and n_C to $-\infty < n_B < n_A$ and $n_A < n_C < \infty$, respectively.

We now have

$$\mathbb{P}[C_6|\Psi] = \sum_{i=1}^{18} \mathbb{P}[C_6|T_i, \Psi]\mathbb{P}[T_i|\Psi] \tag{4}$$

and

$$\mathbb{P}[E_1, C_6|\Psi] = \sum_{i=1}^{18} \mathbb{P}[E_1, C_6|T_i, \Psi]\mathbb{P}[T_i|\Psi]. \tag{5}$$

Here, $\mathbb{P}[T_i|\Psi] = 1/18$ for all i because each labeled history of four lineages is equally likely under the assumption of the coalescent process that lineages join randomly going back in time. Note that symmetries exist in V^{cat} and V^{sym} as a result of exchangeability of certain nodes in the topologies that they consider. For asymmetric topologies,

$$V^{\text{cat}}(W, X, Y, Z, n_3, \Psi) = V^{\text{cat}}(X, W, Y, Z, n_3, \Psi).$$

For symmetric topologies,

$$\begin{aligned} V^{sym}(W, X, Y, Z, n_3, \Psi) &= V^{sym}(X, W, Y, Z, n_3, \Psi) \\ &= V^{sym}(W, X, Z, Y, n_3, \Psi) \\ &= V^{sym}(X, W, Z, Y, n_3, \Psi). \end{aligned}$$

Using the list of probability contributions for each labeled history, as given in Table 3, we can exploit these symmetries and collect like terms across labeled histories to write Eq. (4) as

$$\begin{aligned} &\mathbb{P}[C_6|\Psi] \\ &= \sum_{n_3=-\infty}^{\infty} \sum_{n_A=-\infty}^{\infty} \sum_{n_B=-\infty}^{\infty} \sum_{n_C=-\infty}^{\infty} (4V^{\text{cat}}(n_A, n_A, n_B, n_C, n_3, \Psi) \\ &\quad + 4V^{\text{cat}}(n_A, n_A, n_C, n_B, n_3, \Psi) + 8V^{\text{cat}}(n_A, n_B, n_A, n_C, n_3, \Psi) \\ &\quad + 8V^{\text{cat}}(n_A, n_B, n_C, n_A, n_3, \Psi) + 8V^{\text{cat}}(n_A, n_C, n_A, n_B, n_3, \Psi) \\ &\quad + 8V^{\text{cat}}(n_A, n_C, n_B, n_A, n_3, \Psi) + 8V^{\text{cat}}(n_B, n_C, n_A, n_A, n_3, \Psi) \\ &\quad + 8V^{sym}(n_A, n_B, n_A, n_C, n_3, \Psi) \\ &\quad + 8V^{sym}(n_A, n_C, n_A, n_B, n_3, \Psi) \\ &\quad + 4V^{sym}(n_A, n_A, n_B, n_C, n_3, \Psi) \\ &\quad + 4V^{sym}(n_B, n_C, n_A, n_A, n_3, \Psi))\mathbb{P}[T_i|\Psi] \end{aligned} \tag{6}$$

and Eq. (5) as

$$\begin{aligned} &\mathbb{P}[E_1, C_6|\Psi] \\ &= \sum_{n_3=-\infty}^{\infty} \sum_{n_A=-\infty}^{\infty} \sum_{n_B=-\infty}^{n_A-1} \sum_{n_C=n_A+1}^{\infty} (4V^{\text{cat}}(n_A, n_A, n_B, n_C, n_3, \Psi) \\ &\quad + 4V^{\text{cat}}(n_A, n_A, n_C, n_B, n_3, \Psi) + 8V^{\text{cat}}(n_A, n_B, n_A, n_C, n_3, \Psi) \\ &\quad + 8V^{\text{cat}}(n_A, n_B, n_C, n_A, n_3, \Psi) + 8V^{\text{cat}}(n_A, n_C, n_A, n_B, n_3, \Psi) \\ &\quad + 8V^{\text{cat}}(n_A, n_C, n_B, n_A, n_3, \Psi) + 8V^{\text{cat}}(n_B, n_C, n_A, n_A, n_3, \Psi) \\ &\quad + 8V^{sym}(n_A, n_B, n_A, n_C, n_3, \Psi) \\ &\quad + 8V^{sym}(n_A, n_C, n_A, n_B, n_3, \Psi) \\ &\quad + 4V^{sym}(n_A, n_A, n_B, n_C, n_3, \Psi) \\ &\quad + 4V^{sym}(n_B, n_C, n_A, n_A, n_3, \Psi))\mathbb{P}[T_i|\Psi]. \end{aligned} \tag{7}$$

2.7. *Integrating out the coalescence times*

Finally, we integrate over the density of coalescence times under the standard coalescent model. Under this model, the time in coalescent time units (units of $2N_e$ generations) for i lineages to coalesce to $i - 1$ lineages is exponentially distributed with rate $\binom{i}{2}$ (Wakeley, 2009). Separate coalescence times are independent, and we can write their joint distribution in the four-taxon case as $\rho(t_2, t_3, t_4) = \binom{2}{2} e^{-\binom{2}{2}t_2} \binom{3}{2} e^{-\binom{3}{2}t_3} \binom{4}{2} e^{-\binom{4}{2}t_4} = 18e^{-t_2-3t_3-6t_4}$. Using this density, we integrate to get

$$\mathbb{P}[C_6|\theta] = \int_0^{\infty}\int_0^{\infty}\int_0^{\infty} \mathbb{P}[C_6|\Psi]\rho(t_2, t_3, t_4)\, dt_2\, dt_3\, dt_4 \tag{8}$$

and

$$\mathbb{P}[E_1, C_6|\theta] = \int_0^{\infty}\int_0^{\infty}\int_0^{\infty} \mathbb{P}[E_1, C_6|\Psi]\rho(t_2, t_3, t_4)\, dt_2\, dt_3\, dt_4. \tag{9}$$

2.8. *Implementing the computation*

To calculate $\mathbb{P}[E_1 \cup E_2|C_6, \theta]$ (Eq. (1)) in practice, we use two approaches, a numerical method and a simulation-based method.

2.8.1. *Numerical computation*

First, we employ Gaussian quadrature to numerically estimate the numerator ($2\mathbb{P}[E_1, C_6|\theta]$, Eq. (9)) and denominator ($\mathbb{P}[C_6|\theta]$, Eq. (8)) of $\mathbb{P}[E_1 \cup E_2|C_6, \theta]$. In order to compute the integrals in finite time, we estimate the expression $e^{-t\theta/2}I_{|d|}(t\theta/2)$ using the GNU Scientific Library (GSL) function `gsl_sf_bessel_In_scaled`$(|d|, t\theta/2)$. Additionally, we truncate the bounds of the infinite sums embedded in $2\mathbb{P}[E_1, C_6|\theta]$ and $\mathbb{P}[C_6|\theta]$ to ± 10 instead of $\pm\infty$. These limits provide bounds on the size that an allele can have at any particular node. We additionally integrate all time parameters from 0 to 10 rather than from 0 to ∞. For small values of θ, these approximations are very accurate, as it is unlikely that an allele will mutate more than a few steps away from its initial number of repeats. However, for large θ, the approximation will become less accurate, as large numbers of mutations are likely to occur. These mutations ultimately cause alleles to shift further from the initial base size and beyond the arbitrary truncation in our approximation, so that the calculation fails to account for a non-trivial portion of probability mass.

2.8.2. *Simulation-based computation*

In order to calculate $\mathbb{P}[E_1 \cup E_2|C_6, \theta]$ accurately for large θ, we obtain the ratio in Eq. (1) directly by simulating the coalescent and mutation processes and tabulating the outcomes of interest. The simulation proceeds as follows.

1. Beginning with $k = 4$ alleles, arbitrarily define two alleles to be in one population and the other two alleles to be in the other population.
2. Generate a random time to coalescence from an exponential $\left(\binom{k}{2}\right)$ distribution.
3. Randomly choose two alleles to coalesce; set $k = k - 1$.
4. If $k \neq 1$, go to 2.
5. For each branch of the genealogy, generate a random number of mutation events, x, from a Poisson distribution with rate $\theta t/2$, where t is the branch length.
6. Assign each mutation a value of $+1$ or -1 by sampling the number of $+1$ mutations from a binomial$(x, 1/2)$ distribution. Those mutations not chosen to be $+1$ are assigned a value of -1.
7. Determine the allele size of each of the four sampled alleles by summing the net value of mutations from the root (allele size 0) down to the leaves.

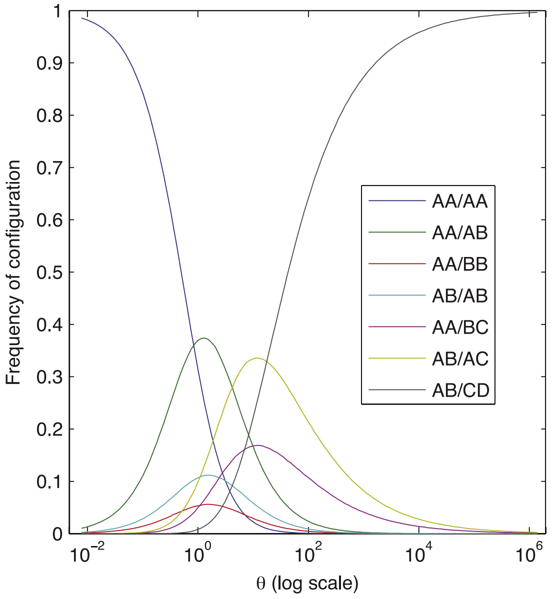


Fig. 3. The simulated frequency of occurrence of seven possible allele configurations as a function of scaled mutation rate (θ) on a log scale. 10^6 trees are simulated per θ step. These simulations utilize four alleles, two in each of two populations. Alleles are related by the coalescent, and they mutate according to the symmetric stepwise mutation model.

8. Classify the collection of four alleles into one of the seven allele configurations (Table 1).
9. If the alleles are in the C_6 configuration, accept the simulation and determine if the sizes of the private alleles (B and C) are on the ends of the distribution ($n_B < n_A < n_C$ or $n_C < n_A < n_B$). If yes, count a success.

By repeating this algorithm until the number of accepted simulations reaches some pre-specified number (we choose 1,000,000), we can estimate the probability that the private alleles lie on the edges of the size distribution by simply dividing the number of successes by the number of accepted simulations.

Note that the proportion of simulations that have configuration C_6 provides an estimate of $\mathbb{P}[C_6|\theta]$. Through a separate application of 10^6 iterations of steps 1 to 8, we estimate the probabilities of all seven configurations as functions of θ. These estimates appear in Fig. 3. At small values of θ, we see that most simulations produce configuration C_1 ($\{AA/AA\}$), a sensible result because mutations are unlikely to happen for small θ. As θ grows larger, more mutations occur, and we see that configurations with two or more distinct alleles begin to rise in frequency. For large values of θ, mutations happen so often that most trees have configuration C_7 ($\{AB/CD\}$).

Fig. 4 shows, as a function of θ, the probability of interest, $\mathbb{P}[E_1 \cup E_2|C_6, \theta]$, calculated both by simulation and numerically. Because we must truncate the internal sums for the numerical computation, we plot several numerical calculations at varying truncation values. Most of the numerical computations are quite accurate at small θ: we expect few mutations in this case, and the approximation made by truncating the sums will reasonably cover most of the probability mass. We see that as θ becomes large, the numerical results differ from the simulation-based result; at large θ many mutations occur and the numerical approximation is poorer.

We note that the probability of interest appears to level off well above the naïve calculation of 1/3 as the mutation rate grows large. Furthermore, as θ tends toward zero, we see that the probability remains above 1/3 and appears to tend toward 1/2. We can prove this small-θ limiting result by considering a parsimony-style approximation for our probability near $\theta = 0$.

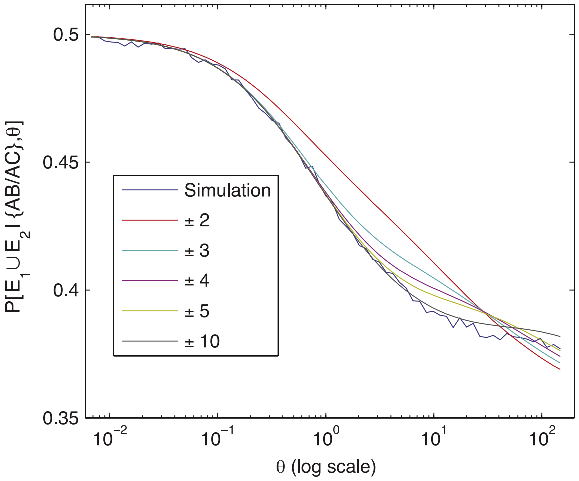


Fig. 4. The probability that the private alleles lie on the edges of the size distribution conditional on production of an $\{AB/AC\}$ configuration, as a function of θ (log scale). This probability is plotted from simulations and for a range of truncations for the infinite sums in numerically approximating Eq. (1). Simulation results are based on 100,000 $\{AB/AC\}$ trees simulated per θ step.

2.9. Small-θ approximation

We can make some simplifications to approximate our calculation of $\mathbb{P}[E_1 \cup E_2|C_6, \theta]$ (Eq. (1)) in the limit as θ becomes small. For small θ, we expect fewer superfluous mutations to occur along a branch with a change of d steps—that is, we expect fewer mutations in one direction to be canceled by mutations in the other direction. Therefore, for very small θ, we can approximate the probability of changing d steps along a branch length t by setting $k = 0$ in Eq. (2) so that no extra mutations occur. Denoting the small-θ approximation to $f(|d|, t, \theta)$ by $f_s(|d|, t, \theta)$, we then obtain

$$f_s(|d|; t, \theta) = e^{-t\theta/2} \frac{(t\theta/4)^{|d|}}{|d|!}.$$

Furthermore, for small θ, we also expect fewer mutations in total to occur on the whole genealogy. The minimum number of mutations needed to provide our pattern of interest, C_6, is two (one mutation on each of two branches). Therefore, for sufficiently small θ, we expect to find no more than two mutations on the entire tree. The probability $f(|d|; t, \theta)$ in Eq. (2) will take one of three forms:

$$f_s(|0|; t, \theta) = e^{-t\theta/2} \tag{10}$$

or

$$f_s(|-1|; t, \theta) = f_s(|1|; t, \theta) = e^{-t\theta/2} \frac{t\theta}{4}. \tag{11}$$

This situation is analogous to a problem in phylogenetics. When rates of change are low, likelihood calculations on trees that consider all possible changes among allelic states converge to calculations of a parsimony score, as only changes of a single unit along a branch have non-trivial likelihood (Felsenstein, 2004). Similarly, our calculation of the probability that the private alleles lie on the edges of the size distribution, considering all possible states for allele sizes, is reduced in the small-θ case to a parsimony-style approximation by replacing $f(|d|; t, \theta)$ with $f_s(|0|; t, \theta)$ and $f_s(|1|; t, \theta)$. This parsimony approximation further eliminates the sums over n_3, n_A, n_B, and n_C, making $\mathbb{P}[E_1 \cup E_2|C_6, \theta]$ (Eq. (1)) tractable to analytically compute.

Examining all the ways of placing two mutations on one of the 18 topologies such that the $\{AB/AC\}$ configuration is produced, each placement will contribute some probability to either the denominator in Eq. (1) or to both the denominator and numerator

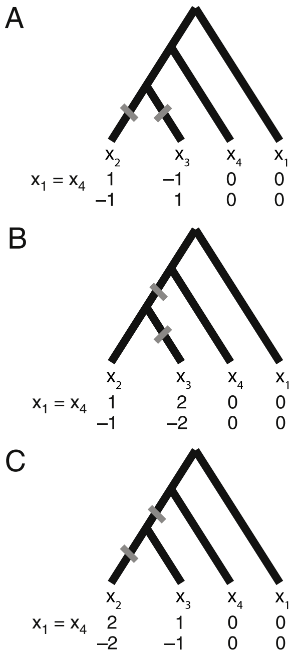


Fig. 5. A small-θ "parsimony" approximation for calculating the probability that private alleles occur on the edges of the size distribution, for the case of $x_1 = x_4$ with history 8. Alleles x_1 and x_2 are in population 1, and alleles x_3 and x_4 are in population 2. In (A), the private alleles (± 1) lie on the edges of the size distribution ($-1 < 0 < 1$); however, in (B) and (C) the private alleles ($\pm 1, \pm 2$) are not on the two edges of the size distribution ($-2 < -1 < 0$ or $0 < 1 < 2$).

in Eq. (1). As an example, consider history 8 from Fig. 1. We first examine the four ways of getting configuration C_6 by assignment of the roles of A, B, and C to the alleles x_1, x_2, x_3, and x_4. We then find all placements of two mutations on the tree that are consistent with this configuration. Each placement will either place the private alleles on both ends of the size distribution, or the shared allele will be on one end. If the private alleles are on both ends, then the term contributes to both the numerator and the denominator. If the shared allele is on an end, then the term contributes to the denominator only. Fig. 5 illustrates this approach for the case of $x_1 = x_4$.

We can substitute f_s for f in our definitions of V^{cat} and V^{sym} to get the following small-θ versions of the probability of an arbitrary set of allele sizes.

$$\begin{aligned}
&V_s^{\mathrm{cat}}(x_1, x_2, x_3, x_4, n_3, \Psi)\\
&\quad = f_s(|x_4|; 2t_2 + t_3 + t_4, \theta) \times f_s(|x_3|; t_3 + t_4, \theta) \times f_s(|n_3|; t_3, \theta)\\
&\qquad \times f_s(|n_3 - x_2|; t_4, \theta) \times f_s(|n_3 - x_1|; t_4, \theta)\\
&V_s^{sym}(x_1, x_2, x_3, x_4, n_3, \Psi)\\
&\quad = f_s(|n_3|; 2t_2 + t_3, \theta) \times f_s(|n_3 - x_4|; t_4, \theta) \times f_s(|n_3 - x_3|; t_4, \theta)\\
&\qquad \times f_s(|x_2|; t_3 + t_4, \theta) \times f_s(|x_1|; t_3 + t_4, \theta).
\end{aligned}$$

Each possible placement of two mutations on one of the 18 labeled histories has a probability that falls into one of 12 equivalence classes as a result of symmetries in V_s^{cat} and V_s^{sym}. We denote these classes by σ_i ($i \in \{1, \ldots, 12\}$), as defined in Table 4.

By tabulating in Table 5 the contributions from each class to the numerator and denominator of the probability for each of the 18 labeled histories, we can now compute the numerator, $2\mathbb{P}[E_1, C_6|\theta]$, in Eq. (1) as

$$\begin{aligned}
&\int_0^\infty \int_0^\infty \int_0^\infty \rho(t_2, t_3, t_4) \frac{1}{18} (32\sigma_1(t_2, t_3, t_4, \theta)\\
&\quad + 32\sigma_2(t_2, t_3, t_4, \theta) + 16\sigma_3(t_2, t_3, t_4, \theta) + 16\sigma_6(t_2, t_3, t_4, \theta)\\
&\quad + 32\sigma_8(t_2, t_3, t_4, \theta) + 8\sigma_9(t_2, t_3, t_4, \theta)\\
&\quad + 8\sigma_{11}(t_2, t_3, t_4, \theta))\, dt_2\, dt_3\, dt_4, \qquad (12)
\end{aligned}$$

Table 4
Definitions for the 12 classes of probability in the small-θ "parsimony" approximation.

Class	Defined probability
1	$\sigma_1(\Psi) = V_s^{\mathrm{cat}}(1, 0, 0, 0, 1, \Psi)$
2	$\sigma_2(\Psi) = V_s^{\mathrm{cat}}(1, 0, 0, 1, 0, \Psi)$
3	$\sigma_3(\Psi) = V_s^{\mathrm{cat}}(0, 0, 0, 1, 1, \Psi)$
4	$\sigma_4(\Psi) = V_s^{\mathrm{cat}}(0, 0, 1, 1, 0, \Psi)$
5	$\sigma_5(\Psi) = V_s^{\mathrm{cat}}(0, 0, 1, 0, 1, \Psi)$
6	$\sigma_6(\Psi) = V_s^{\mathrm{cat}}(1, 1, 0, 0, 0, \Psi)$
7	$\sigma_7(\Psi) = V_s^{\mathrm{cat}}(1, 0, 1, 0, 0, \Psi)$
8	$\sigma_8(\Psi) = V_s^{sym}(1, 0, 0, 1, 0, \Psi)$
9	$\sigma_9(\Psi) = V_s^{sym}(0, 0, 0, 1, 1, \Psi)$
10	$\sigma_{10}(\Psi) = V_s^{sym}(0, 0, 1, 1, 0, \Psi)$
11	$\sigma_{11}(\Psi) = V_s^{sym}(1, 1, 0, 0, 0, \Psi)$
12	$\sigma_{12}(\Psi) = V_s^{sym}(1, 0, 1, 0, 0, \Psi)$

which evaluates to

$$\frac{\theta^2(648 + 990\theta + 489\theta^2 + 79\theta^3)}{18(1+\theta)^2(2+\theta)^3(3+\theta)^3}. \qquad (13)$$

The denominator, $\mathbb{P}[C_6|\theta]$, of Eq. (1) is

$$\begin{aligned}
&\int_0^\infty \int_0^\infty \int_0^\infty \rho(t_2, t_3, t_4) \frac{1}{18} (32\sigma_1(t_2, t_3, t_4, \theta)\\
&\quad + 32\sigma_2(t_2, t_3, t_4, \theta) + 16\sigma_3(t_2, t_3, t_4, \theta)\\
&\quad + 16\sigma_4(t_2, t_3, t_4, \theta) + 16\sigma_5(t_2, t_3, t_4, \theta) + 16\sigma_6(t_2, t_3, t_4, \theta)\\
&\quad + 32\sigma_7(t_2, t_3, t_4, \theta) + 32\sigma_8(t_2, t_3, t_4, \theta) + 8\sigma_9(t_2, t_3, t_4, \theta)\\
&\quad + 16\sigma_{10}(t_2, t_3, t_4, \theta) + 8\sigma_{11}(t_2, t_3, t_4, \theta)\\
&\quad + 16\sigma_{12}(t_2, t_3, t_4, \theta))\, dt_2\, dt_3\, dt_4, \qquad (14)
\end{aligned}$$

which evaluates to

$$\frac{\theta^2(432 + 630\theta + 295\theta^2 + 45\theta^3)}{6(1+\theta)^2(2+\theta)^3(3+\theta)^3}. \qquad (15)$$

Taking the ratio of expressions (13) and (15) and evaluating the limit as θ tends to 0 gives us

$$\begin{aligned}
\lim_{\theta \to 0} \mathbb{P}[E_1 \cup E_2 | C_6, \theta] &= \lim_{\theta \to 0} \frac{(648 + 990\theta + 489\theta^2 + 79\theta^3)}{3(432 + 630\theta + 295\theta^2 + 45\theta^3)}\\
&= \frac{1}{2}. \qquad (16)
\end{aligned}$$

This result shows that, for low mutation rates, we expect the private alleles in an $\{AB/AC\}$ sample of size four to be on the ends of the size distribution approximately 1/2 of the time. This is substantially more often than the value of 1/3 predicted when the relatedness of the alleles was not taken into account.

3. Arbitrary divergence time

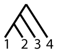
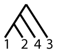
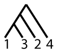
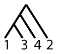
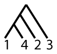
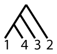
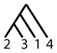
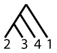
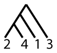
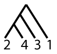
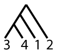
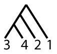
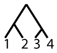
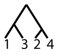
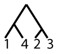
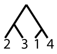
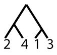
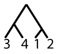
Extending our two-population model, we now consider two populations separated by arbitrary divergence time t_d (Fig. 6). Note that as shown in Fig. 6, the definitions of t_2, t_3, and t_4 differ slightly from those used in the calculations for the $t_d = 0$ case in Fig. 2. We can formulate Eq. (1) for arbitrary divergence time t_d and compute

$$\mathbb{P}[E_1 \cup E_2 | C_6, \theta, t_d] = \frac{2\mathbb{P}[E_1, C_6|\theta, t_d]}{\mathbb{P}[C_6|\theta, t_d]}. \qquad (17)$$

Detailed derivations appear in Appendix A. We calculate Eq. (17) numerically by Gaussian quadrature and by simulation using methods similar to those used for the $t_d = 0$ case (Appendix B).

Fig. 7 shows, as a function of θ and t_d, the probability that the private alleles lie on the edges of the size distribution, as obtained using the simulation in Appendix B. We see that

Table 5
The small-θ approximation contributions to the numerator (Eq. (12)) and denominator (Eq. (14)) of the probability that the private alleles lie on the edges, for each labeled history.

History number	History	Contribution to numerator	Contribution to denominator
1		$4\sigma_1 + 4\sigma_2$	$4\sigma_1 + 4\sigma_2$
2		$4\sigma_1 + 4\sigma_2$	$4\sigma_1 + 4\sigma_2$
3		$2\sigma_1 + 2\sigma_2 + 2\sigma_3 + 2\sigma_6$	$2\sigma_1 + 2\sigma_2 + 2\sigma_3 + 2\sigma_4 + 2\sigma_5 + 2\sigma_6 + 4\sigma_7$
4		$2\sigma_1 + 2\sigma_2 + 2\sigma_3 + 2\sigma_6$	$2\sigma_1 + 2\sigma_2 + 2\sigma_3 + 2\sigma_4 + 2\sigma_5 + 2\sigma_6 + 4\sigma_7$
5		$2\sigma_1 + 2\sigma_2 + 2\sigma_3 + 2\sigma_6$	$2\sigma_1 + 2\sigma_2 + 2\sigma_3 + 2\sigma_4 + 2\sigma_5 + 2\sigma_6 + 4\sigma_7$
6		$2\sigma_1 + 2\sigma_2 + 2\sigma_3 + 2\sigma_6$	$2\sigma_1 + 2\sigma_2 + 2\sigma_3 + 2\sigma_4 + 2\sigma_5 + 2\sigma_6 + 4\sigma_7$
7		$2\sigma_1 + 2\sigma_2 + 2\sigma_3 + 2\sigma_6$	$2\sigma_1 + 2\sigma_2 + 2\sigma_3 + 2\sigma_4 + 2\sigma_5 + 2\sigma_6 + 4\sigma_7$
8		$2\sigma_1 + 2\sigma_2 + 2\sigma_3 + 2\sigma_6$	$2\sigma_1 + 2\sigma_2 + 2\sigma_3 + 2\sigma_4 + 2\sigma_5 + 2\sigma_6 + 4\sigma_7$
9		$2\sigma_1 + 2\sigma_2 + 2\sigma_3 + 2\sigma_6$	$2\sigma_1 + 2\sigma_2 + 2\sigma_3 + 2\sigma_4 + 2\sigma_5 + 2\sigma_6 + 4\sigma_7$
10		$2\sigma_1 + 2\sigma_2 + 2\sigma_3 + 2\sigma_6$	$2\sigma_1 + 2\sigma_2 + 2\sigma_3 + 2\sigma_4 + 2\sigma_5 + 2\sigma_6 + 4\sigma_7$
11		$4\sigma_1 + 4\sigma_2$	$4\sigma_1 + 4\sigma_2$
12		$4\sigma_1 + 4\sigma_2$	$4\sigma_1 + 4\sigma_2$
13		$8\sigma_8$	$8\sigma_8$
14		$4\sigma_8 + 2\sigma_9 + 2\sigma_{11}$	$4\sigma_8 + 2\sigma_9 + 4\sigma_{10} + 2\sigma_{11} + 4\sigma_{12}$
15		$4\sigma_8 + 2\sigma_9 + 2\sigma_{11}$	$4\sigma_8 + 2\sigma_9 + 4\sigma_{10} + 2\sigma_{11} + 4\sigma_{12}$
16		$4\sigma_8 + 2\sigma_9 + 2\sigma_{11}$	$4\sigma_8 + 2\sigma_9 + 4\sigma_{10} + 2\sigma_{11} + 4\sigma_{12}$
17		$4\sigma_8 + 2\sigma_9 + 2\sigma_{11}$	$4\sigma_8 + 2\sigma_9 + 4\sigma_{10} + 2\sigma_{11} + 4\sigma_{12}$
18		$8\sigma_8$	$8\sigma_8$

throughout the parameter space, the probability always exceeds the naïve expectation of 1/3. For all values of θ, we observe that increasing the divergence time between the populations increases the probability of finding the private alleles on the edges of the size distribution. Furthermore, we see that for small θ, the probability that private alleles in a sample of size four are found on the edges of the size distribution quickly tends toward 1 as t_d increases. By applying the small-θ approximation of Eqs. (10) and (11), we can show that this probability does indeed converge to 1 as t_d tends to infinity.

Conditioning on each of the four possible scenarios depicted in Fig. 6, we follow an approach similar to the $t_d = 0$ small-θ derivation to obtain a small-θ approximation for the case of arbitrary divergence time (Appendix C). The resulting limiting expression for this approximation as θ tends to 0 is

$$\lim_{\theta\to 0} \mathbb{P}[E_1 \cup E_2 | C_6, \theta_{small}, t_d] = \frac{3e^{2t_d} - 2t_d - 2}{3e^{2t_d} - 1}. \tag{18}$$

Eq. (18) is sensible in that it agrees with the small-θ result of 1/2 at $t_d = 0$ (Eq. (16)), and it approaches the conditional result $\mathbb{P}[E_1 \cup E_2 | C_6, \theta_{small}, t_d, E_{11}] = 1$ as t_d increases without bound (Eq. (C.6)). In Fig. 8, we plot the function of t_d in Eq. (18) along with simulated results at increasingly small θ. We see that for each θ, the probability that the private alleles lie on the edges of the size distribution increases monotonically as a function of the divergence time, and that the simulated probability approaches the limiting expression as θ approaches 0.

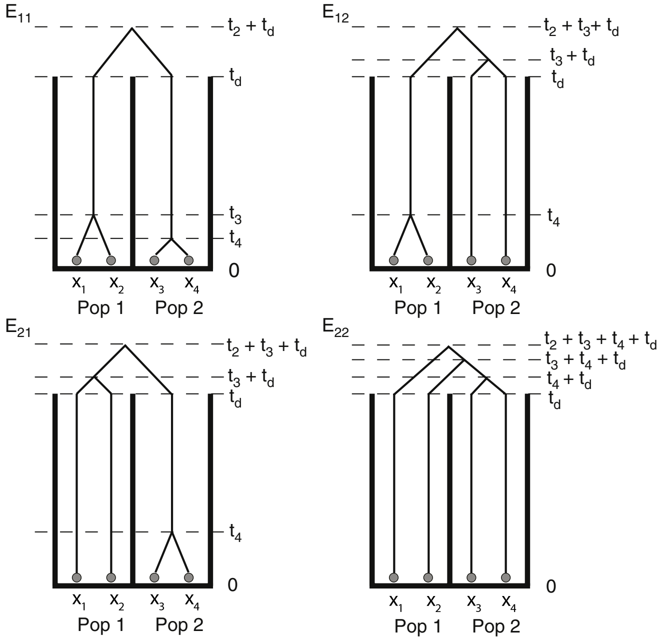


Fig. 6. The four types of coalescent scenarios with $t_d > 0$, with their coalescent time parameterizations. In scenario E_{11}, t_3 is defined as the time to coalescence of the two lineages in population 1, and t_4 is defined as the time to coalescence of the two lineages in population 2.

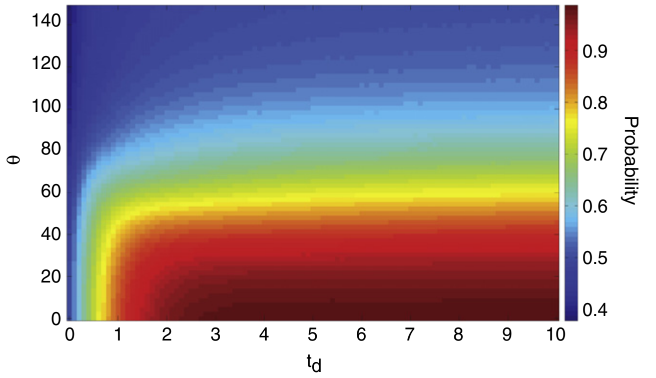


Fig. 7. Simulated probability that the private alleles lie on the edges of the size distribution, conditional on production of an $\{AB/AC\}$ configuration. The plot shows this probability as a function of θ (log scale) and t_d. 10^6 $\{AB/AC\}$ trees are simulated for each choice of θ and t_d.

4. Properties of the probability that private alleles lie on the edges

In order to investigate the probability that private alleles lie on the edges of the size distribution, we started with a naïve argument that suggests that this should happen 1/3 of the time in a sample of four alleles, two from each of two populations. However, this naïve argument ignored the relatedness of the four alleles. We have presented a calculation of the desired probability using a coalescent framework for gene relatedness, together with the symmetric stepwise mutation model. When fixing t_d, we see a monotonic decrease in the probability that the private alleles lie on the edges as θ grows, but for every collection of parameter values evaluated (scaled mutation rate θ and divergence time t_d between the two populations), the probability remains greater than 1/3.

Furthermore, the probability appears to stay well above 1/3 even for very large θ. For large θ, we might expect so many mutations to occur on the tree that the allele sizes would not be correlated, effectively "erasing" the genealogical relatedness. In this case, we would expect the naïve prediction of 1/3 to hold. However, in order to observe a C_6 configuration, two alleles must be identical by state. Thus, when conditioning on configuration C_6, the distribution of branch lengths is biased toward shorter branches compared to the unconditional distribution, and even for large θ, the number of mutations tends to be small enough that genealogical relatedness remains important.

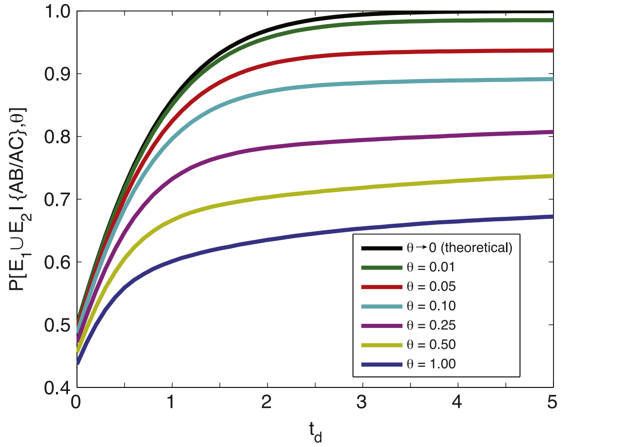


Fig. 8. Simulated small-θ probabilities that the private alleles lie on the edges of the size distribution conditional on production of an $\{AB/AC\}$ configuration, and the corresponding limiting probability computed analytically for $\theta \to 0$ (Eq. (18)), as functions of t_d. The simulation approach follows that of Fig. 7 and is described in Appendix B.

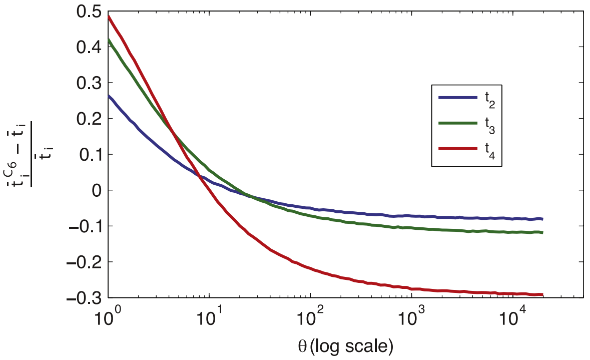


Fig. 9. Relative difference between mean coalescence times conditional on obtaining configuration C_6 ($\bar{t}_i^{C_6}$) and unconditional mean coalescence times ($\bar{t}_i$), as a function of θ (log scale). The mean conditional coalescence times were calculated by taking the mean of 10^6 simulated coalescence times in scenarios that produced configuration C_6.

Holding t_d fixed at 0, Fig. 9 plots $(\bar{t}_i^{C_6} - \bar{t}_i)/\bar{t}_i$ versus θ, where $\bar{t}_i$ is the unconditional expectation of t_i under the coalescent and $\bar{t}_i^{C_6}$ is the conditional expectation given configuration C_6, as obtained in 10^6 simulations that produced this configuration. We see that as θ increases, the relative difference between the conditional mean coalescence times given configuration C_6 and the unconditional mean coalescence times becomes increasingly negative. Most notably, t_4 becomes particularly short, reflecting the observation that for large θ, scenarios with configuration C_6 often have a "cherry" with short external branches of length t_4 on which no mutations occur.

In the small-θ case, we find that for $t_d = 0$, the probability that the private alleles lie on the edges in a sample of size four approaches 1/2 as θ tends to zero. By letting the divergence time between the two populations exceed zero ($t_d > 0$), we see a monotonic increase in this probability. In fact, in the small-θ limit, the probability that the private alleles lie on the edges in a sample of size four tends to 1 as t_d tends to infinity.

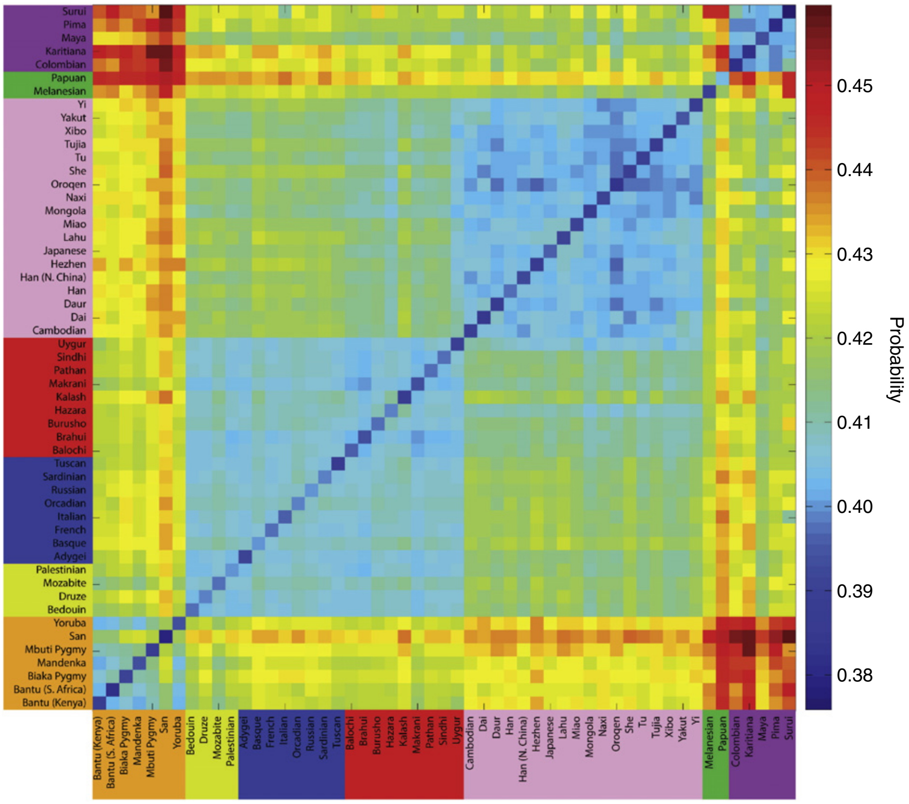


Fig. 10. The empirical probability that private alleles lie on the edges of the size distribution in a sample of size four from a pair of populations. Plotted are pairwise calculations of this frequency for all 53 worldwide populations from the Human Genome Diversity Panel, arranged in major geographic regions. African, Middle Eastern, European, Central/South Asian, East Asian, Oceanian, and American populations are arranged by color in the labels. Blue represents a lower probability, and red represents a higher probability.

These results show that the genealogical history of a set of microsatellite alleles is an important factor in determining the prevalence of private alleles in the ends of the allele size distribution, even under circumstances in which we might expect the genealogy to be relatively unimportant. Our calculations also predict that the probability that private alleles lie on the edges of the allele size distribution grows as the divergence time between populations increases.

5. Application to data

To test the prediction that the probability that private alleles lie on the edges of the allele size distribution grows as the divergence time between populations grows, we analyzed data on microsatellites at 783 loci covering 1048 individuals in 53 worldwide populations from the Human Genome Diversity Panel (Rosenberg et al., 2005). Computations with these microsatellites have established a general increase of genetic differentiation (and hence, divergence time) with increasing geographical distance between a pair of populations (Ramachandran et al., 2005). Thus, although a strict divergence model is only an approximation to the population histories, we can consider the pairwise comparisons of populations that are geographically near each other to represent populations that diverged recently. Similarly, we can consider the pairwise comparisons of populations that are geographically distant from each other to represent populations that diverged relatively farther in the past. Pairwise comparisons of a population with itself can be interpreted as the case in which a population divergence happened at time $t_d = 0$ in the past. Based on the theory we have developed, we expect that pairs of geographically separated populations will produce a higher probability that the private alleles will lie on the edges of the size distribution. Similarly, we expect smaller probabilities for pairs of geographically proximate populations and the smallest probabilities for comparisons of populations with themselves. We further expect that measures of genetic differentiation such as F_{ST} will correlate with this probability as well, since these measures can be taken as a loose proxy for the divergence time between two populations.

To estimate the empirical frequency that the private alleles in a sample of size four lie on the edges of the size distribution, we perform the following analysis. For each population at each locus, we estimate the allele frequency distribution by counting the total number of observations of each distinct allele size and dividing by the total number of observations in the population. For a pair of populations, we then draw two alleles from the empirical allele frequency distribution in each population. If the set of four alleles has an $\{AB/AC\}$ configuration, we accept the draw and determine if the private alleles lie on the edges of the size distribution. If so, then the draw is counted as a success. We repeatedly draw sets of four alleles until 100,000 draws are accepted. Finally, we calculate the empirical frequency that the private alleles lie on the edges of the size distribution for a locus by dividing the number of successes by the number of acceptances, and we calculate the mean of this empirical frequency across loci. By performing this analysis, we get an estimate for the mean frequency that private alleles lie on the edges of the size distribution.

The results of this analysis are plotted in Fig. 10, and we find that real populations do indeed follow the expected theoretical trend. The probabilities that private alleles lie on the edges range from 0.3759 to 0.4595. African populations paired with each other have lower probabilities, and a trend toward higher probabilities occurs as African populations are paired with other populations that are more geographically distant. The pairings of African

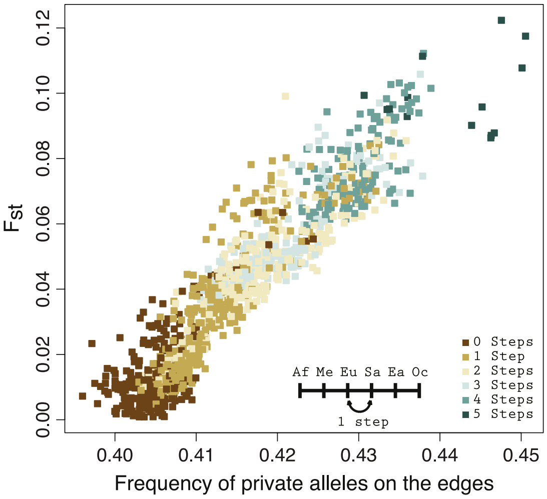


Fig. 11. F_{ST} vs. the empirical frequency with which private alleles in a sample of size four lie on the edges of the size distribution. Each point represents a pair among 53 worldwide populations from the Human Genome Diversity Panel, excluding comparisons involving native American populations and comparisons of populations with themselves. Pearson's $r = 0.9333$.

populations with native American populations (representing the most genetically distant pairs) have the highest probabilities. Furthermore, pairings close to the diagonal in Fig. 10 tend to be more closely related than pairings farther away from the diagonal, and for these pairs, we see mostly low probabilities. Finally, the main diagonal represents the analysis of a population paired with itself; this is interpreted as comparing two populations with a divergence time of $t_d = 0$. We find that probabilities along the diagonal are the lowest among all pairs considered.

Because we also expect the frequency of private alleles on the edges to correlate with measures of genetic differentiation, we calculate pairwise F_{ST} between populations using Eq. (5.3) from Weir (1996). In Fig. 11, F_{ST} values are plotted against the frequency with which private alleles occur on the edges of the size distribution, and we find a very tight correlation ($r = 0.9333$). Thus, our empirical calculations show that our model for explaining the size distribution of private microsatellite alleles is able to predict phenomena observed in real data.

6. Discussion

We have modeled the phenomenon of private microsatellite alleles lying on the edges of the allele size distribution in order to explain an observation by Wang et al. (2007) that they occupy these locations more often than is expected by chance. Using a simple two-population model with sample size four, we have provided a naïve argument, in which we expect the probability that private microsatellite alleles lie on the edges of the size distribution to be 1/3. Using a coalescent model with symmetric stepwise mutation to explicitly calculate this probability as a function of two parameters (mutation rate θ and divergence time t_d), we find that this probability appears to always exceed 1/3. Furthermore, the model predicts that the probability that private alleles lie on the edges of the size distribution grows larger as the divergence time between populations increases. We have found that this prediction holds in an analysis of worldwide microsatellite data in humans.

Intuitively, we can understand why $\mathbb{P}[E_1 \cup E_2 | C_6, \theta, t_d]$ might be expected to exceed the naïve expectation by considering the process by which private alleles are generated. When an ancestral population splits into two groups, all allele sizes present in the population become shared alleles in the descendant populations, and these shared alleles define the center of the allele size distribution. As allele sizes diffuse away from the center in the separate descendant populations, mutations in either population toward the edges of the size distribution are likely to generate alleles that are novel and therefore private. Conversely, mutations that push alleles toward the center of the size distribution are likely to produce sizes that already exist in both populations, as a result of the shared descent of central allele sizes. Furthermore, to produce shared alleles on an edge of the size distribution, unless the edge allele size is inherited by descent from the ancestral population in both descendant groups, alleles from each population must separately mutate to the same size on the edge. Because more mutations in total are required for producing such a shared allele on the edge compared to the number required in one population to produce a private allele on the edge, we expect private alleles to lie on the edges of the size distribution more often than is predicted under the assumption that there is no relationship between privacy and allele size.

This work augments the coalescent theory of microsatellite markers by providing predictions about the properties of private alleles in a simple model with sample size four. Previous work has examined additional quantities in the case of a four-allele sample. For example, Kimmel and Chakraborty (1996) and Pritchard and Feldman (1996) studied the expectation $E[(X_i - X_k)^2(X_j - X_\ell)^2]$ for random allele sizes X_i, X_j, X_k, and X_ℓ in a stepwise mutation model. Zhang and Rosenberg (2007) studied the genealogies of duplicated microsatellites in a model with four sampled alleles, two each for two paralogous microsatellite loci. Together with these other efforts, our work demonstrates that analytical formulas can sometimes be obtained in coalescent-based microsatellite models of non-trivial size.

While our main goal has been to explore the properties of our simple model, the model may potentially enable the inference of θ and t_d. For each of a collection of loci whose mutational characteristics are assumed to be identical, the probability that private alleles lie on the edges of the size distribution could be estimated from data by repeatedly sampling alleles from the observed allele frequency distributions for pairs of populations. Using this empirical estimate, a likelihood surface could then be constructed to jointly estimate θ and t_d. This approach might not produce identifiable estimates; however, if θ has already been estimated by another method or if additional summary statistics are combined with a private allele statistic, a potentially viable method for estimating t_d might be constructed, considering the dramatic effect that this parameter has on the probability that private alleles lie on the edges of the size distribution.

We conclude with a discussion of model limitations. Because of the complexity of the probability calculations, we have restricted our attention to a sample of size four. We have assumed a simple demographic model of two populations, in which population sizes are equal and no migration occurs after the populations diverge. The simple stepwise mutation model assumes symmetry in the direction of mutation and independence of the mutation rate with allele size, and both the demographic model and the mutation model likely reflect conditions that are not strictly met in the human population example that we consider. Indeed, more complex mutation models, allowing for directional bias, multistep mutations, length-dependent mutation rates, or a combination of these factors could potentially be considered (e.g. Calabrese and Durrett, 2003; Whittaker et al., 2003; Watkins, 2007). In general, however, we did not need a more complex model to explain the core observation that private alleles frequently lie on the edges of the size distribution. While the true demographic and mutational phenomena are undoubtedly more complicated than our model captures, we are still able to observe that as predicted, the probability that private microsatellite alleles lie on the edges of the size distribution in a sample of four alleles correlates with the genetic differentiation between pairs of populations.

Acknowledgments

Support for this work was provided by NIH grants R01 GM081441 and T32 HG000040, NSF grant DEB-0716904, and a grant from the Burroughs Wellcome Fund.

Appendix A. Derivation for arbitrary divergence time

The expression that must be calculated in order to obtain the probability that the private alleles lie on the edges of the size distribution for arbitrary t_d appears in Eq. (17). To perform the calculation in Eq. (17), we must utilize the probability that two lineages reduce to one lineage during time t_d as well as the probability that two lineages survive until t_d. Under the coalescent (Wakeley, 2009), these probabilities are $g_{21}(t_d) = 1 - e^{-t_d}$ and $g_{22}(t_d) = e^{-t_d}$, where $g_{ij}(t_d)$ denotes the probability under the coalescent that i lineages reduce to j lineages during time t_d.

We can partition our probability calculation into four pieces corresponding to the four coalescent scenarios possible by time t_d (Fig. 6). First, in each population, the two lineages could coalesce more recently than t_d (event E_{11}). Second, the two lineages in population 1 could coalesce more recently than t_d, and the two lineages in population 2 could survive to t_d (event E_{12}). Third, the two lineages in population 1 could survive to t_d, and the two lineages in population 2 could coalesce more recently than t_d (event E_{21}). Finally, in each population, the two lineages could survive to t_d (event E_{22}). These four events happen with the following probabilities:

$$\mathbb{P}[E_{11}|t_d] = (g_{21}(t_d))^2 = (1 - e^{-t_d})^2, \tag{A.1}$$

$$\mathbb{P}[E_{12}|t_d] = g_{21}(t_d)g_{22}(t_d) = (1 - e^{-t_d})e^{-t_d}, \tag{A.2}$$

$$\mathbb{P}[E_{21}|t_d] = g_{22}(t_d)g_{21}(t_d) = e^{-t_d}(1 - e^{-t_d}), \tag{A.3}$$

$$\mathbb{P}[E_{22}|t_d] = (g_{22}(t_d))^2 = e^{-2t_d}. \tag{A.4}$$

We then calculate $\mathbb{P}[E_1 \cup E_2|C_6, \theta, t_d]$ by separately conditioning on E_{11}, E_{12}, E_{21}, and E_{22} to get

$$\mathbb{P}[E_1 \cup E_2|C_6, \theta, t_d] = \frac{2\sum_{i=1}^{2}\sum_{j=1}^{2}\mathbb{P}[E_1, C_6|\theta, t_d, E_{ij}]\mathbb{P}[E_{ij}|\theta, t_d]}{\sum_{i=1}^{2}\sum_{j=1}^{2}\mathbb{P}[C_6|\theta, t_d, E_{ij}]\mathbb{P}[E_{ij}|\theta, t_d]}, \tag{A.5}$$

in which

$$\mathbb{P}[E_1, C_6|\theta, t_d, E_{ij}] = \int_0^\infty\int_0^\infty\int_0^\infty \mathbb{P}[E_1, C_6|t_2, t_3, t_4, \theta, t_d, E_{ij}] \times \rho_{ij}(t_2, t_3, t_4) dt_2\, dt_3\, dt_4, \tag{A.6}$$

$$\mathbb{P}[C_6|\theta, t_d, E_{ij}] = \int_0^\infty\int_0^\infty\int_0^\infty \mathbb{P}[C_6|t_2, t_3, t_4, \theta, t_d, E_{ij}] \times \rho_{ij}(t_2, t_3, t_4)\, dt_2\, dt_3\, dt_4, \tag{A.7}$$

$$\mathbb{P}[E_1, C_6|t_2, t_3, t_4, \theta, t_d, E_{ij}] = \sum_{n_3=-\infty}^{\infty}\sum_{n_A=-\infty}^{\infty}\sum_{n_B=-\infty}^{n_A-1}\sum_{n_C=n_A+1}^{\infty} V_{E_{ij}}(n_A, n_B, n_C, n_3, \Psi), \tag{A.8}$$

and

$$\mathbb{P}[C_6|t_2, t_3, t_4, \theta, t_d, E_{ij}] = \sum_{n_3=-\infty}^{\infty}\sum_{n_A=-\infty}^{\infty}\sum_{n_B=-\infty}^{\infty}\sum_{n_C=-\infty}^{\infty} V_{E_{ij}}(n_A, n_B, n_C, n_3, \Psi). \tag{A.9}$$

We can determine the values of the conditional probability $V_{E_{ij}}$ of the node allele sizes and the conditional coalescence time density ρ_{ij} by examining which labeled histories are possible for each E_{ij}. For example, for event E_{11} both pairs of lineages coalesce more recently than time t_d, and only symmetric histories are possible. Furthermore, x_1 will always coalesce with x_2 and x_3 will always coalesce with x_4 in this scenario, leaving only two possible equiprobable histories (histories 13 and 18 in Fig. 1). Therefore, we only sum over the V^{sym} terms that are associated with these histories.

Table A.1
The reparameterizations of Ψ for the events E_{ij}.

Event	$\Psi = (\tau, \theta)$
E_{11}	$\tau = (t_2+t_d-\max(t_3, t_4), \max(t_3, t_4)-\min(t_3, t_4), \min(t_3, t_4))$
E_{12}	$\tau = (t_2, t_3 + t_d - t_4, t_4)$
E_{21}	$\tau = (t_2, t_3 + t_d - t_4, t_4)$
E_{22}	$\tau = (t_2, t_3, t_4 + t_d)$

In addition, for each event, compared to the case of $t_d = 0$, we must reparameterize the branch lengths of the histories to account for changes due to forced survival of lineages to time t_d. For event E_{11}, we reparameterize by setting $\Psi = (\tau, \theta)$ with $\tau = (t_2 + (t_d - \max(t_3, t_4)), \max(t_3, t_4)-\min(t_3, t_4), \min(t_3, t_4))$, as illustrated in Fig. 6 and tabulated in Table A.1. By conditioning on one of the four events E_{11}, E_{12}, E_{21}, or E_{22}, the density of coalescence times differs from the corresponding density $\rho(t_2, t_3, t_4)$ defined in the $t_d = 0$ case.

For event E_{11}, the distribution of coalescence times is $\rho_{11}(t_2, t_3, t_4) = \rho_{11t_2}(t_2)\rho_{11t_3}(t_3)\rho_{11t_4}(t_4)$, where $\rho_{11t_2}(t) = e^{-t}$ and $\rho_{11t_3}(t) = \rho_{11t_4}(t) = \mathbf{1}_{t<t_d}e^{-t}/(1 - e^{-t_d})$. We can then write

$$\begin{aligned} V_{E_{11}} = \frac{1}{2}(&V^{sym}(n_A, n_B, n_A, n_C, n_3, \Psi) \\ &+ V^{sym}(n_A, n_B, n_C, n_A, n_3, \Psi) + V^{sym}(n_B, n_A, n_A, n_C, n_3, \Psi) \\ &+ V^{sym}(n_B, n_A, n_C, n_A, n_3, \Psi) + V^{sym}(n_A, n_A, n_B, n_C, n_3, \Psi) \\ &+ V^{sym}(n_A, n_C, n_B, n_A, n_3, \Psi) + V^{sym}(n_B, n_A, n_A, n_C, n_3, \Psi) \\ &+ V^{sym}(n_B, n_C, n_A, n_A, n_3, \Psi)). \end{aligned} \tag{A.10}$$

We proceed with similar arguments for events E_{12}, E_{21}, and E_{22}. The corresponding values for Ψ are tabulated in Table A.1, and the values for $V_{E_{ij}}$ and ρ_{ij} are tabulated in Table A.2.

Appendix B. Implementing the computation for arbitrary divergence time

To implement the calculation of $\mathbb{P}[E_1 \cup E_2|C_6, \theta, t_d]$ (Eq. (17)) derived in Appendix A, we use Gaussian quadrature and a simulation-based approach. These approaches are analogous to the approaches that we used in the case of $t_d = 0$.

B.1. Numerical computation

As in the $t_d = 0$ case, we use Gaussian quadrature to numerically evaluate $\mathbb{P}[E_1, C_6|\theta, t_d, E_{ij}]$ (Eq. (A.6)) and $\mathbb{P}[C_6|\theta, t_d, E_{ij}]$ (Eq. (A.7)), once again estimating the expression $e^{-t\theta/2}I_{|d|}(t\theta/2)$ using the GNU Scientific Library (GSL) function `gsl_sf_bessel_In_scaled`$(|d|, t\theta/2)$. We use the same value as in the $t_d = 0$ case (±10) to truncate the infinite sums in Eqs. (A.8) and (A.9). Additionally, we again integrate all time dimensions in Eqs. (A.6) and (A.7) from 0 to 10 rather than from 0 to ∞. As in the case of $t_d = 0$, these calculations are very accurate for small values of θ and less accurate for large values of θ (not shown).

Table A.2
The probabilities of node allele sizes and the coalescence time densities conditional on events E_{ij}.

Event	$V_{E_{ij}}$	$\rho_{ij}(t_2, t_3, t_4, t_d) = \rho_{ijt_2}(t_2, t_d)\rho_{ijt_3}(t_3, t_d)\rho_{ijt_4}(t_4, t_d)$		
		$\rho_{ijt_2}(t, t_d)$	$\rho_{ijt_3}(t, t_d)$	$\rho_{ijt_4}(t, t_d)$
E_{11}	$\frac{1}{2}(4V^{sym}(n_A, n_B, n_A, n_C, n_3, \Psi) + 4V^{sym}(n_A, n_C, n_A, n_B, n_3, \Psi))$	e^{-t}	$\mathbf{1}_{t<t_d}e^{-t}/(1-e^{-t_d})$	$\mathbf{1}_{t<t_d}e^{-t}/(1-e^{-t_d})$
E_{12}	$\frac{1}{3}(4V^{\text{cat}}(n_A, n_B, n_A, n_C, n_3, \Psi) + 4V^{\text{cat}}(n_A, n_B, n_C, n_A, n_3, \Psi)$ $+ 4V^{sym}(n_A, n_C, n_A, n_B, n_3, \Psi))$	e^{-t}	$3e^{-3t}$	$\mathbf{1}_{t<t_d}e^{-t}/(1-e^{-t_d})$
E_{21}	$\frac{1}{3}(4V^{\text{cat}}(n_A, n_B, n_A, n_C, n_3, \Psi) + 4V^{\text{cat}}(n_A, n_B, n_C, n_A, n_3, \Psi)$ $+ 4V^{sym}(n_A, n_C, n_A, n_B, n_3, \Psi))$	e^{-t}	$3e^{-3t}$	$\mathbf{1}_{t<t_d}e^{-t}/(1-e^{-t_d})$
E_{22}	$\frac{1}{18}(4V^{\text{cat}}(n_A, n_A, n_B, n_C, n_3, \Psi) + 4V^{\text{cat}}(n_A, n_A, n_C, n_B, n_3, \Psi)$ $+ 8V^{\text{cat}}(n_A, n_B, n_A, n_C, n_3, \Psi) + 8V^{\text{cat}}(n_A, n_B, n_C, n_A, n_3, \Psi)$ $+ 8V^{\text{cat}}(n_A, n_C, n_A, n_B, n_3, \Psi) + 8V^{\text{cat}}(n_A, n_C, n_B, n_A, n_3, \Psi)$ $+ 8V^{\text{cat}}(n_B, n_C, n_A, n_A, n_3, \Psi) + 8V^{sym}(n_A, n_B, n_A, n_C, n_3, \Psi)$ $+ 8V^{sym}(n_A, n_C, n_A, n_B, n_3, \Psi) + 4V^{sym}(n_A, n_A, n_B, n_C, n_3, \Psi)$ $+ 4V^{sym}(n_B, n_C, n_A, n_A, n_3, \Psi))$	e^{-t}	$3e^{-3t}$	$6e^{-6t}$

B.2. Simulation-based computation

As in the $t_d = 0$ case, we are able to accurately estimate the quantity $\mathbb{P}[E_1 \cup E_2|C_6, \theta, t_d]$ (Eq. (17)), directly obtaining the ratio $2\mathbb{P}[E_1, C_6|\theta, t_d]/\mathbb{P}[C_6|\theta, t_d]$ by simulating the coalescent and mutation processes and counting the outcomes of interest. The simulation proceeds as follows.

1. Beginning with $k = 4$ alleles, arbitrarily define two alleles to be in one population and the other two alleles to be in the other population.
2. Randomly choose an event E_{11}, E_{12}, E_{21}, or E_{22} based on their relative probabilities conditional on t_d (Eqs. (A.1)–(A.4)).
3. If event E_{11} is chosen:
   (a) Generate a random time to coalescence from an exponential $\left(\binom{2}{2}\right)$ distribution conditional on being less than t_d.
   (b) Coalesce the pair of lineages in population 1; set $k = k - 1$.
   (c) Generate a random time to coalescence from an exponential $\left(\binom{2}{2}\right)$ distribution conditional on being less than t_d.
   (d) Coalesce the pair of lineages in population 2; set $k = k - 1$.
4. If event E_{12} or E_{21} is chosen:
   (a) Generate a random time to coalescence from an exponential $\left(\binom{2}{2}\right)$ distribution conditional on being less than t_d.
   (b) Coalesce a pair of lineages in population 1 (if event E_{12}) or population 2 (if event E_{21}); set $k = k - 1$.
5. Extend all remaining lineages up to t_d.
6. Generate a random time to coalescence from an exponential $\left(\binom{k}{2}\right)$ distribution.
7. Randomly choose two lineages to coalesce; set $k = k - 1$.
8. If $k \neq 1$, go to 6.
9. For each branch of the genealogy, generate a random number of mutation events, x, from a Poisson distribution with rate $\theta t/2$, where t is the branch length.
10. Assign each mutation a value of $+1$ or -1 by sampling the number of $+1$ mutations from a binomial(x,1/2) distribution. Those mutations not chosen to be $+1$ are assigned a value of -1.
11. Determine the allele size of each of the four sampled alleles by summing the net value of mutations from the root (allele size 0) down to the leaves.
12. Classify the collection of four alleles into one of the seven allele configurations (Table 1).
13. If the alleles are in the C_6 configuration, accept the simulation and determine if the sizes of the private alleles (B and C) are on the ends of the distribution ($n_C < n_A < n_B$ or $n_B < n_A < n_C$). If yes, count a success.

As in the $t_d = 0$ case, after the number of accepted simulations reaches some pre-specified number (we choose 1,000,000), we estimate the probability of the private alleles occurring on the edges of the size distribution by dividing the number of successes by the number of accepted simulations.

Appendix C. Small-θ approximation for arbitrary divergence time

With $t_d > 0$, we can consider a small-θ approximation to the probability that the private alleles lie on the edges in a similar way to the corresponding calculation with $t_d = 0$. By considering a fixed t_d, we proceed as before, counting the contributions of each labeled history to the numerator and denominator in Eq. (17). The probability distribution of labeled histories depends on t_d, and the 18 histories are no longer equiprobable when $t_d > 0$. Conditional on one of the events $\{E_{11}, E_{12}, E_{21}, E_{22}\}$, however, we can determine the possible histories and weight the probability contributions of these histories to the numerator and denominator as before.

Thus, following Eq. (A.5) for the small-θ case, we wish to calculate

$$\mathbb{P}[E_1 \cup E_2|C_6, \theta_{small}, t_d] = \frac{\sum_{i=1}^{2}\sum_{j=1}^{2} \mathbb{P}[E_1 \cup E_2, C_6|\theta_{small}, t_d, E_{ij}]\mathbb{P}[E_{ij}|\theta_{small}, t_d]}{\sum_{i=1}^{2}\sum_{j=1}^{2} \mathbb{P}[C_6|\theta_{small}, t_d, E_{ij}]\mathbb{P}[E_{ij}|\theta_{small}, t_d]}. \quad \text{(C.1)}$$

Note that although E_1 and E_2 have the same probability, in this calculation it is convenient to calculate $E_1 \cup E_2$ directly. We do this by tabulating contributions to the numerator and denominator conditional on each event E_{ij} (Table 5), reparameterizing Ψ to augment certain branch lengths by amounts dependent on t_d (Table A.1).

First, consider event E_{11}. If both pairs of lineages coalesce more recently than the population divergence time, then the only possible histories are 13 and 18, and the conditional contribution to the denominator of Eq. (C.1) is

$$\mathbb{P}[C_6|\theta_{small}, t_d, E_{11}] = \int_0^\infty \int_0^\infty \int_0^\infty \frac{1}{2} 16\sigma_8(t_2 + t_d - t_{\max}, t_{\max} - t_{\min}, t_{\min}, \theta)\rho_{E_{11}}(t_2, t_3, t_4, t_d)dt_2\, dt_3\, dt_4, \quad \text{(C.2)}$$

where $t_{\max} = \max(t_3, t_4)$ and $t_{\min} = \min(t_3, t_4)$. Here, we obtain the coefficients for each σ_i by referencing histories 13 and 18 in Table 5, and we use the conditional density of coalescence times $\rho_{E_{11}}(t_2, t_3, t_4, t_d)$ from Table A.2. Eq. (C.2) also provides the $\mathbb{P}[E_1 \cup E_2, C_6|\theta_{small}, t_d, E_{11}]$ term in the numerator, because for histories 13 and 18, at small θ, the private alleles always lie on the edges of the size distribution.

Next, consider event E_{ij} ($i \neq j$). If the two lineages in one population coalesce more recently than the divergence time, and the two lineages in the other population survive to the divergence time, then the only possible histories are 1, 2, and 18 for E_{12}

or 11, 12, and 13 for E_{21}. Because E_{12} and E_{21} differ only in which population contains the coalescence more recent than the population divergence, they have the same probability. The conditional contribution to the denominator for either event is then

$$\begin{aligned}\mathbb{P}[C_6|\theta_{small}, t_d, E_{ij}]\\ &= \int_0^\infty \int_0^\infty \int_0^\infty \frac{1}{3}(8\sigma_1(t_2, t_3 + t_d - t_4, t_4, \theta)\\ &\quad + 8\sigma_2(t_2, t_3 + t_d - t_4, t_4, \theta) + 8\sigma_8(t_2, t_3 + t_d - t_4, t_4, \theta))\\ &\quad \times \rho_{E_{ij}}(t_2, t_3, t_4, t_d) dt_2\, dt_3\, dt_4, \end{aligned} \tag{C.3}$$

where the σ_i coefficients are taken from Table 5 using either set of histories (1, 2, and 18 for E_{12} or 11, 12, and 13 for E_{21}) and $\rho_{E_{ij}}(t_2, t_3, t_4, t_d)$ is taken from Table A.2. Eq. (C.3) is also equal to the $\mathbb{P}[E_1 \cup E_2, C_6|\theta_{small}, t_d, E_{ij}]$ term in the numerator, because for either set of histories, at small θ, the private alleles always lie on the edges of the size distribution.

For event E_{22}, if in both populations the two lineages survive to the divergence time, then all 18 histories are possible. The conditional contribution to the denominator is

$$\begin{aligned}\mathbb{P}[C_6|\theta_{small}, t_d, E_{22}] &= \int_0^\infty \int_0^\infty \int_0^\infty \frac{1}{18}(32\sigma_1(t_2, t_3, t_4 + t_d, \theta)\\ &\quad + 32\sigma_2(t_2, t_3, t_4 + t_d, \theta) + 16\sigma_3(t_2, t_3, t_4 + t_d, \theta)\\ &\quad + 16\sigma_4(t_2, t_3, t_4 + t_d, \theta) + 16\sigma_5(t_2, t_3, t_4 + t_d, \theta)\\ &\quad + 16\sigma_6(t_2, t_3, t_4 + t_d, \theta) + 32\sigma_7(t_2, t_3, t_4 + t_d, \theta)\\ &\quad + 32\sigma_8(t_2, t_3, t_4 + t_d, \theta) + 8\sigma_9(t_2, t_3, t_4 + t_d, \theta)\\ &\quad + 16\sigma_{10}(t_2, t_3, t_4 + t_d, \theta) + 8\sigma_{11}(t_2, t_3, t_4 + t_d, \theta)\\ &\quad + 16\sigma_{12}(t_2, t_3, t_4 + t_d, \theta))\rho_{E_{22}}(t_2, t_3, t_4) dt_2\, dt_3\, dt_4 \end{aligned} \tag{C.4}$$

and the conditional contribution to the numerator is

$$\begin{aligned}\mathbb{P}[E_1 \cup E_2, C_6|\theta_{small}, t_d, E_{22}]\\ &= \int_0^\infty \int_0^\infty \int_0^\infty \frac{1}{18}(32\sigma_1(t_2, t_3, t_4 + t_d, \theta)\\ &\quad + 32\sigma_2(t_2, t_3, t_4 + t_d, \theta) + 16\sigma_3(t_2, t_3, t_4 + t_d, \theta)\\ &\quad + 16\sigma_6(t_2, t_3, t_4 + t_d, \theta) + 32\sigma_8(t_2, t_3, t_4 + t_d, \theta)\\ &\quad + 8\sigma_9(t_2, t_3, t_4 + t_d, \theta) + 8\sigma_{11}(t_2, t_3, t_4 + t_d, \theta))\\ &\quad \times \rho_{E_{22}}(t_2, t_3, t_4) dt_2\, dt_3\, dt_4, \end{aligned} \tag{C.5}$$

where the σ_i coefficients are from Table 5 and $\rho_{E_{22}}(t_2, t_3, t_4)$ is from Table A.2.

We can understand how Eq. (C.1) will behave for large values of t_d by considering the behavior of $\mathbb{P}[E_{ij}|t_d]$ (Eqs. (A.1)–(A.4)) as t_d tends toward ∞. Independently of the value of θ, when the divergence time between populations grows very large, we expect each pair of lineages to always coalesce before the population divergence (event E_{11}). Taking the limits of Eqs. (A.1)–(A.4), $\lim_{t_d \to \infty} \mathbb{P}[E_{11}|t_d] = 1$ and $\lim_{t_d \to \infty} \mathbb{P}[E_{12}|t_d] = \lim_{t_d \to \infty} \mathbb{P}[E_{21}|t_d] = \lim_{t_d \to \infty} \mathbb{P}[E_{22}|t_d] = 0$. Thus as t_d tends to ∞, Eq. (C.1) reduces to

$$\mathbb{P}[E_1 \cup E_2|C_6, \theta_{small}, t_d] = \frac{\mathbb{P}[E_1 \cup E_2, C_6|\theta_{small}, t_d, E_{11}]}{\mathbb{P}[C_6|\theta_{small}, t_d, E_{11}]} = 1. \tag{C.6}$$

Therefore, for large t_d, we intuitively expect the small-θ probability that the private alleles lie on the edges of the size distribution to tend to unity.

Note that Eqs. (C.4) and (C.5) differ from Eqs. (14) and (12) only in the definitions of the time parameters and densities of coalescence times. Using the conditional contributions in Eqs. (C.2)–(C.5) together with $\mathbb{P}[E_{ij}|t_d]$ in Eqs. (A.1)–(A.4), we can calculate Eq. (C.1). The resulting expression is unwieldy (not shown), but taking its limit as θ tends to 0, we obtain Eq. (18).

References

Barton, N.H., Slatkin, M., 1986. A quasi-equilibrium theory of the distribution of rare alleles in a subdivided population. Heredity 56, 409–415.

Calabrese, P., Durrett, R., 2003. Dinucleotide repeats in the Drosophila and human genomes have complex, length-dependent mutation processes. Molecular Biology and Evolution 20, 715–725.

Calafell, F., Shuster, A., Speed, W.C., Kidd, J.R., Kidd, K.K., 1998. Short tandem repeat polyorphism evolution in humans. European Journal of Human Genetics 6, 38–49.

Felsenstein, J., 2004. Inferring Phylogenies. Sinauer Associates, Inc., Sunderland, MA, USA.

Fiumera, A.C., Parker, P.G., Fuerst, P.A., 2000. Effective population size and maintenance of genetic diversity in captive-bred populations of a Lake Victoria cichlid. Conservation Biology 14, 886–892.

Gradshteyn, I.S., Ryzhik, I.M., 2000. Table of Integrals, Series, and Products, 6th ed. Academic Press, London.

Kalinowski, S.T., 2004. Counting alleles with rarefaction: private alleles and hierarchical sampling designs. Conservation Genetics 5, 539–543.

Kimmel, M., Chakraborty, R., 1996. Measures of variation at DNA repeat loci under a general stepwise mutation model. Theoretical Population Biology 50, 345–367.

Neel, J.V., 1973. Private genetic variants and the frequency of mutation among South American Indians. Proceedings of the National Academy of Sciences of the United States of America 70, 3311–3315.

Neel, J.V., 1978. Rare variants, private polymorphisms, and locus heterozygosity in Amerindian populations. American Journal of Human Genetics 30, 465–490.

Neel, M.C., Cummings, M.P., 2003. Effectiveness of conservation targets in capturing genetic diversity. Conservation Biology 17, 219–229.

Neel, J.V., Thompson, E.A., 1978. Founder effect and number of private polymorphisms observed in Amerindian tribes. Proceedings of the National Academy of Sciences of the United States of America 75, 1904–1908.

Parker, K.M., Sheffer, R.J., Hedrick, P.W., 1999. Molecular variation and evolutionarily significant units in the endangered Gila topminnow. Conservation Biology 13, 108–116.

Petit, R., Mousadik, A.E., Pons, O., 1998. Identifying populations for conservation on the basis of genetic markers. Conservation Biology 12, 844–855.

Pritchard, J.K., Feldman, M.W., 1996. Statistics for microsatellite variation based on coalescence. Theoretical Population Biology 50, 325–344.

Ramachandran, S., Deshpande, O., Roseman, C.C., Rosenberg, N.A., Feldman, M.W., Cavalli-Sforza, L.L., 2005. Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. Proceedings of the National Academy of Sciences of the United States of America 102, 15942–15947.

Rosenberg, N.A., Mahajan, S., Ramachandran, S., Zhao, C., Pritchard, J.K., Feldman, M.W., 2005. Clines, clusters, and the effect of study design on the inference of human population structure. PLoS Genetics 1, 660–671.

Schroeder, K.B., Schurr, T.G., Long, J.C., Rosenberg, N.A., Crawford, M.H., Tarskaia, L.A., Osipova, L.P., Zhadanov, S.I., Smith, D.G., 2007. A private allele ubiquitous in the Americas. Biology Letters 3, 218–223.

Slatkin, M., 1985. Rare alleles as indicators of gene flow. Evolution 39, 53–65.

Szpiech, Z.A., Jakobsson, M., Rosenberg, N.A., 2008. ADZE: a rarefaction approach for counting alleles private to combinations of populations. Bioinformatics 24, 2498–2504.

Torres, E., Iriondo, J.M., Pérez, C., 2003. Genetic structure of an endangered plant, *Antirrhinum microphyllum* (Scrophulariaceae): allozyme and RAPD analysis. American Journal of Botany 90, 85–92.

Wakeley, J., 2009. Coalescent Theory: An Introduction. Roberts and Company Publishers, Greenwood Village, CO, USA.

Wang, S., Lewis Jr., C.M., Jakobsson, M., Ramachandran, S., Ray, N., Bedoya, G., Rojas, W., Parra, M.V., Molina, J.A., Gallo, C., Mazzotti, G., Poletti, G., Hill, K., Hurtado, A.M., Labuda, D., Klitz, W., Barrantes, R., Bortolini, M.C., Salzano, F.M., Petzl-Erler, M.L., Tsuneto, L.T., Llop, E., Rothhammer, F., Excoffier, L., Feldman, M.W., Rosenberg, N.A., Ruiz-Linares, A., 2007. Genetic variation and population structure in native Americans. PLoS Genetics 3, e185.

Watkins, J.C., 2007. Microsatellite evolution: Markov transition functions for a suite of models. Theoretical Population Biology 71, 147–159.

Wehrhahn, C.F., 1975. The evolution of selectively similar electrophoretically detectable alleles in finite natural populations. Genetics 80, 375–394.

Weir, B.S., 1996. Genetic Data Analysis II. Sinauer Associates, Inc., Sunderland, MA, USA.

Whittaker, J.C., Harbord, R.M., Boxall, N., Mackay, I., Dawson, G., Sibly, R.M., 2003. Likelihood-based estimation of microsatellite mutation rates. Genetics 164, 781–787.

Wilson, I.J., Balding, D.J., 1998. Genealogical inference from microsatellite data. Genetics 150, 499–510.

Zhang, K., Rosenberg, N.A., 2007. On the genealogy of a duplicated microsatellite. Genetics 177, 2109–2122.