Testing for Ancient Admixture between Closely Related Populations

Eric Y. Durand,*[1] Nick Patterson,[2] David Reich,[2,3] and Montgomery Slatkin[1]

[1]Department of Integrative Biology, University of California, Berkeley

[2]Broad Institute of MIT and Harvard

[3]Department of Genetics, Harvard Medical School

***Corresponding author:** E-mail: eric.durand@berkeley.edu.

Associate editor: Jonathan Pritchard

Research article

Abstract

One enduring question in evolutionary biology is the extent of archaic admixture in the genomes of present-day populations. In this paper, we present a test for ancient admixture that exploits the asymmetry in the frequencies of the two nonconcordant gene trees in a three-population tree. This test was first applied to detect interbreeding between Neandertals and modern humans. We derive the analytic expectation of a test statistic, called the *D* statistic, which is sensitive to asymmetry under alternative demographic scenarios. We show that the *D* statistic is insensitive to some demographic assumptions such as ancestral population sizes and requires only the assumption that the ancestral populations were randomly mating. An important aspect of *D* statistics is that they can be used to detect archaic admixture even when no archaic sample is available. We explore the effect of sequencing error on the false-positive rate of the test for admixture, and we show how to estimate the proportion of archaic ancestry in the genomes of present-day populations. We also investigate a model of subdivision in ancestral populations that can result in *D* statistics that indicate recent admixture.

Key words: admixture, gene genealogies, lineage sorting.

Introduction

Detecting ancient admixture and estimating the extent of archaic ancestry in the genomes of present-day populations are of major importance for many aspects of evolutionary biology (Barton 2001). Admixture occurs during secondary contact between previously isolated populations, and it can play an important role in the speciation process (Barton and Hewitt 1985; Barton 2001). Among anthropologists, there is a long-standing debate on the relationships between modern humans and different groups of archaic humans such as Neandertals. A recent analysis of a draft sequence of the Neandertal genome suggested that 1–4% of genomes from people of Eurasian ancestry are derived from Neandertal (Green et al. 2010). However, there is still no consensus on the overall extent of interbreeding between archaic and modern humans (for a review, see Wall and Hammer 2006).

Methods for detecting and estimating archaic admixture often rely on DNA sequence data from extent populations only. One class of such methods relies on the signature that ancient admixture leaves on patterns of linkage disequilibrium (Plagnol and Wall 2006; Wall et al. 2009). Although they have the advantage of not relying on scarce or damaged ancient DNA samples, they make simplifying assumptions about the demographic history of extant populations that are not testable in general (Wall and Hammer 2006). Another class of methods models genetic introgression using spatially explicit simulations. They aim to make predictions about the proportion of introgression in present-day genomes if ancient admixture had occurred. These methods use either forward-in-time computer simulations (Currat and Excoffier 2004) or deterministic modeling (Forhan et al. 2008).

Recently, direct comparisons of DNA sequences between extant and archaic populations have been made possible by the sequencing of ancient DNA samples. In humans, this line of research began with a comparison of mitochondrial DNA between modern humans and Neandertals (Krings et al. 1997; Serre et al. 2004; Green et al. 2008). It was then extended to fragments of Neandertal nuclear DNA (Green et al. 2006; Noonan et al. 2006; Krause et al. 2007; Lalueza-Fox et al. 2007); the full sequence of Neandertals was finally analyzed in Green et al. (2010). Although data from extinct populations are still limited, direct comparison of ancient and modern sequences has the potential to directly determine the proportion of present-day genomes inherited from archaic populations (Wall 2000). Green et al. (2010) developed a formal test for admixture based on the direct comparison of DNA sequences from Neandertals and modern human populations.

Green et al. (2010) defined a new statistic, called the *D* statistic, to test for admixture between three closely related populations. The expectation of *D* was derived under a simple model of admixture with constant population size and was used to estimate the proportion of Neandertal ancestry in modern humans (Green et al. 2010, Supplementary Online Material 19). However, Green et al. (2010) noted that their test relies on the assumption that the population ancestral to the sampled populations was randomly mating. In this paper, we extend the results of Green et al.

doi:10.1093/molbev/msr048 Advance Access publication February 15, 2011

(2010) by deriving the expectation of D under a model where the ancestral populations are not randomly mating. We also study the statistical properties of D. In addition, we explore the effects of sequencing error and ascertainment bias on D, and we show that D can indicate past admixture even when the admixing population is not sampled.

In a first section, we recall the definition of the D statistic and why it is expected to detect admixture in the case of randomly mating ancestral populations. Then we derive the expectation of D under a model of admixture with arbitrarily varying ancestral population sizes, which generalizes the model of admixture proposed in Green et al. (2010). The derivation reveals a strategy to estimate the proportion of archaic introgression in present-day genomes. In a next section, we derive the expectation of D in a model where the ancestral populations are not randomly mating and show that ancient population structure can confound the test for admixture. Then we determine the effect of sequencing error on D, and we study the power of the test for admixture on synthetic data. When D is computed from genotype data, we determine the impact of ascertainment bias on the test for admixture. Although Green et al. (2010) defined D to compare the Neandertal genome with the sequences of modern humans, we show that D statistics can be used to test for admixture when no archaic sample is available. Finally, we apply the test to the Neandertal data published in Green et al. (2010).

Materials and Methods

A Four-Taxon Statistic to Test for Admixture

Assume that we have sequenced one chromosome from two present-day populations and denote by P_1 and P_2 those populations. Furthermore, assume that we have sampled one chromosome from an archaic population, which we denote as P_3, and one chromosome from an outgroup population, denoted O. Suppose that the four sequences were aligned without error. The null hypothesis that we wish to test is a demographic scenario in which P_1 and P_2 descend from a common ancestral population that diverged from the ancestors of P_3 at an earlier time, without any gene flow between P_3 and P_1 or P_2 after they split. The alternative hypothesis is that P_3 exchanged genes with P_1 or P_2 after these two populations diverged.

We first restrict to positions in the genome where we have coverage for P_1, P_2, P_3, and O. We denote the outgroup allele as "A" and restrict our analysis to biallelic sites at which P_1 and P_2 differ and the alternative allele "B" is seen in P_3. At these sites, we have observed two copies of both alleles, making it less likely that the patterns we analyze have arisen because of sequencing error.

For the ordered set {P_1, P_2, P_3, O}, we call the two allelic configurations of interest "ABBA" or "BABA." The pattern ABBA refers to biallelic sites where P_1 has the outgroup allele and P_2 and P_3 share the derived copy. The pattern BABA corresponds to sites where P_1 and P_3 share the derived allele and P_2 has the outgroup allele. Green et al. (2010) defined a statistic corresponding to the difference in the counts of ABBA and BABA sites across the n base pairs for which we have data of all four samples, normalized by the total number of observations. In this statistic, $C_{ABBA}(i)$ and $C_{BABA}(i)$ are indicator variables; they can be 0 or 1 depending on whether an ABBA or a BABA pattern is seen at base i. Green et al. (2010) denoted this statistic by D, and we have

$$D(P_1, P_2, P_3, O) = \frac{\sum_{i=1}^{n} C_{ABBA}(i) - C_{BABA}(i)}{\sum_{i=1}^{n} C_{ABBA}(i) + C_{BABA}(i)}. \qquad (1)$$

We further denote by $S(P_1, P_2, P_3, O)$ the numerator of $D(P_1, P_2, P_3, O)$.

Under the null hypothesis that P_1 and P_2 descend from a common ancestral population that diverged at an earlier time from the ancestral population of P_3, and if the ancestral population of P_1, P_2, and P_3 was panmictic, then derived alleles in P_3 should match derived alleles in P_1 and P_2 equally often. This is because the patterns ABBA and BABA can only arise from gene trees that are nonconcordant with the population tree of P_1, P_2, and P_3. Under the null hypothesis, the two nonconcordant gene trees should occur with equal frequencies (Tajima 1983; Hudson 1983), and D should equal zero. There are three classes of events that can produce a significant deviation from the null hypothesis. First, P_3 exchanged genes with P_1 or P_2. Then the population ancestral to P_1, P_2, and P_3 may have been structured in such a way that one of the two non-concordant gene trees occurs more often than the other. Alternatively, P_1 or P_2 could have received genes from an unsampled ghost population that we denote as P_G. Note that P_G needs to be at least as diverged as P_3 from (P_1, P_2) for D to differ significantly from zero (supplementary fig. S1, Supplementary Material online). Also note that gene flow between P_1 and P_2, or between P_3 and the ancestor of P_1 and P_2, is not expected to produce a deviation from the null hypothesis.

Although D statistics are primarily designed to be applied on sequence data, they can be readily computed on single nucleotide polymorphism (SNP) data. Assume that n SNPs were genotyped in populations P_1, P_2, P_3, and O. Denote by $\hat{p}_{ij}$ the observed frequency of SNP i in population P_j ($j = 4$ denotes population O). We have

$$D(P_1, P_2, P_3, O) = \frac{\sum_{i=1}^{n}[(1-\hat{p}_{i1})\hat{p}_{i2}\hat{p}_{i3}(1-\hat{p}_{i4}) - \hat{p}_{i1}(1-\hat{p}_{i2})\hat{p}_{i3}(1-\hat{p}_{i4})]}{\sum_{i=1}^{n}[(1-\hat{p}_{i1})\hat{p}_{i2}\hat{p}_{i3}(1-\hat{p}_{i4}) + \hat{p}_{i1}(1-\hat{p}_{i2})\hat{p}_{i3}(1-\hat{p}_{i4})]}. \qquad (2)$$

Thus, computing D on genotype data is straightforward. However, the way SNPs were ascertained can bias estimates of allele frequencies in the different populations (Eberle and Kruglyak 2000; Kuhner et al. 2000; Clark et al. 2005).

A Genealogical Argument for the Expected Frequencies of ABBA and BABA Sites

Assume that a single nucleotide substitution occurred on the gene genealogy representing the ancestry of the four samples (P_1, P_2, P_3, and O). There are three possible topologies of the gene genealogy, assuming that O is the

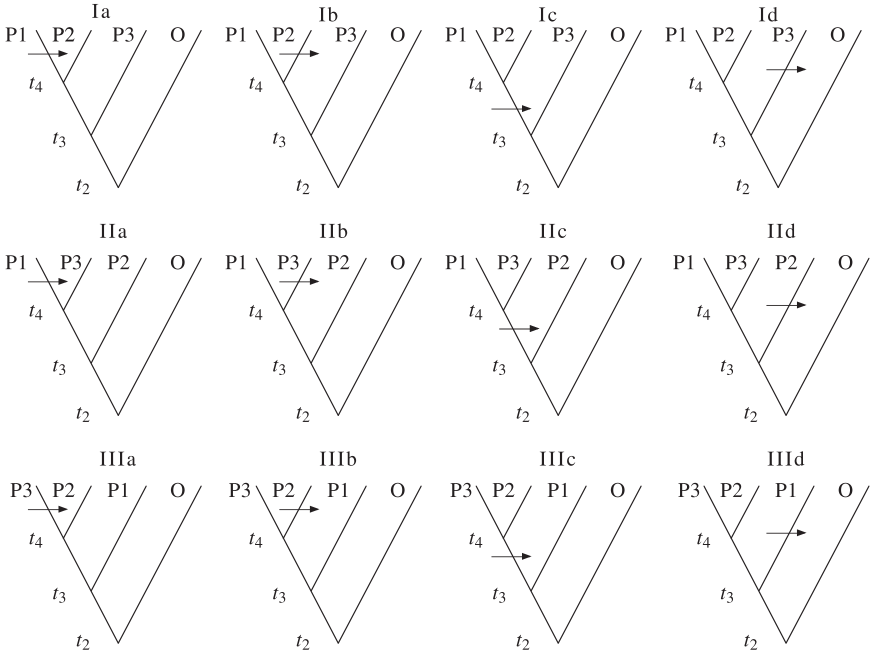

FIG. 1. There are three possible topologies of the gene genealogies, I, II, and III, that can relate samples P_1, P_2, and P_3, and the outgroup O. Assuming that any polymorphisms reflect a single historical mutation, ABBA sites can occur only on a type III gene genealogy due to a mutation that occurs on the branch ancestral to samples P_2 and P_3 (IIIc), and BABA sites can occur only on a type II gene genealogy due to a mutation on the branch ancestral to samples P_1 and P_3 (IIc).

outgroup. Each of these topologies has four branches on which a mutation can create a derived allele in P_1, P_2, and P_3. Figure 1 presents the 12 possible configurations for the four samples P_1, P_2, P_3, and O, assuming that O carries the ancestral allele. The pattern BABA (P_1 and P_3 have the derived nucleotide) is consistent only with tree IIc. The length of the internal branch represents the time during which a mutation can produce the BABA pattern. Similarly, the pattern ABBA (P_2 and P_3 have the derived nucleotide) is consistent only with tree IIIc, and the length of that internal branch is the time during which mutation can create ABBA. The probability that a randomly chosen site will have either pattern is the probability of the appropriate topology multiplied by the expected branch length multiplied by the mutation rate, μ. Thus, to derive the expected frequencies of patterns ABBA and BABA, we need to compute the probabilities of the corresponding tree topologies and average lengths of the relevant branches under the demographic scenario relating P_1, P_2, and P_3.

In what follows, we derive the expected frequencies of patterns ABBA and BABA under a simple demographic model of ancient admixture. We use this model to show that the test is not confounded by demographic events other than admixture, assuming that the population ancestral to P_1, P_2, and P_3 was panmictic.

Instantaneous Unidirectional Admixture

In this model, we assume that there was a single episode of admixture at time t_{GF} in the past ($t = 0$ being the present) from population P_3 to P_2 after the separation of populations P_1 and P_2 (fig. 2). With probability f, the P_2 lineage was derived from a lineage from P_3. The parameter f represents the fraction of the genomes in P_2 that originated from P_3. We define the divergence time of P_1 and P_2 populations as $t_{P2} > t_{GF}$. We denote by $t_{P3} > t_{P2}$ the divergence time of P_3 and the population ancestral to P_1 and P_2, denoted $P_{(12)}$. At generation t in the past, the effective population size of the P_3 population is $N_3(t)$; the effective size of $P_{(12)}$ is $N_{(12)}(t)$; and the effective size of the population ancestral to P_1, P_2, and P_3, denoted $P_{(123)}$, is $N_{(123)}(t)$. All the populations are assumed to be unstructured (i.e., they are randomly mating). We denote by IUA the instantaneous unidirectional admixture model.

Expected Counts of ABBA and BABA under the Model of Instantaneous Admixture

Under the IUA model, there are three classes of events that can produce the patterns ABBA and BABA. The first class corresponds to the case where the P_2 lineage did not originate from P_3. It occurs with probability $1 - f$. In the second class of events, the P_2 lineage traces its ancestry in P_3 (probability f), but the P_2 and P_3 lineages did not coalesce before t_{P3}. The third class corresponds to the case where the P_2 lineage originated from P_3 (probability f) and the P_2 and P_3 lineages coalesced before t_{P3}. The first two classes produce patterns ABBA and BABA with equal frequencies. This is because, assuming that $P_{(123)}$ was panmictic, the P_3 lineage coalesces first with probability 1/3. The third class of event produces only ABBA because the P_2 and P_3 lineages coalesced first. The probabilities of ABBA and BABA are derived in Appendix 1 for a model with arbitrary varying population size in P_3 and $P_{(12)}$ and constant population size in $P_{(123)}$.

Green et al. (2010) studied the simplified case where population size is constant. When we set $N_3(t) = N_{(12)}(t) = N_{(123)} = N$ in Appendix 1, we find, in accordance with Green et al. (2010), that

$$\Pr(\mathrm{ABBA}) = \mu\left[f(t_{P3} - t_{GF}) + (1-f)\left(1 - \frac{1}{2N}\right)^{t_{P3} - t_{P2}}\frac{2N}{3} + f\left(1 - \frac{1}{2N}\right)^{t_{P3} - t_{GF}}\frac{2N}{3}\right] \tag{3}$$

and

$$\Pr(\mathrm{BABA}) = \mu\left[(1-f)\left(1 - \frac{1}{2N}\right)^{t_{P3} - t_{P2}}\frac{2N}{3} + f\left(1 - \frac{1}{2N}\right)^{t_{P3} - t_{GF}}\frac{2N}{3}\right], \tag{4}$$

where μ denotes the per base mutation rate. In this case, the expectation of D reduces to

$$E[D(P_1, P_2, P_3, O)] = \frac{\Pr(\mathrm{ABBA}) - \Pr(\mathrm{BABA})}{\Pr(\mathrm{ABBA}) + \Pr(\mathrm{BABA})} = \frac{3f[t_{P3} - t_{GF}]}{3f[t_{P3} - t_{GF}] + 4N(1-f)(1 - \frac{1}{2N})^{t_{P3} - t_{P2}} + 4Nf(1 - \frac{1}{2N})^{t_{P3} - t_{GF}}}. \tag{5}$$

Note that the test statistic equals 0 when there is no admixture ($f = 0$). If there is admixture ($f > 0$), D tends to one when $t_{P3} - t_{GF}$ becomes large. This is because the P_2 and P_3 lineages have more time to coalesce as $t_{P3} - t_{GF}$ increases. Finally, in the case of constant ancestral population size, D tends to 0 when the effective population size becomes large. This is explained by the fact that when N is large, then the probability that the P_2 and P_3 lineages coalesce in P_3 is small.

A Test for Admixture Insensitive to Many Demographic Assumptions

Under the IUA model, equation (A1.5) shows that $E[D(P_1, P_2, P_3, O)] = 0$ if and only if $f = 0$ or $t_{GF} = t_{P3}$, regardless of the population size fluctuations of P_1, P_2, P_3, $P_{(12)}$, and $P_{(123)}$ (as long as population sizes stay finite) and regardless of the times of population divergences and admixture (as long as $t_{GF} < t_{P3}$). Note that a model with $t_{GF} = t_{P3}$ is equivalent to a model without admixture (see fig. 2). Thus, under the IUA model assumptions, a significant deviation from 0 of $D(P_1, P_2, P_3, O)$ indicates that P_3 exchanged genes with P_1 ($D(P_1, P_2, P_3, O) < 0$) or P_2 ($D(P_1, P_2, P_3, O) > 0$). Although we modeled admixture with an instantaneous episode of gene flow, this conclusion holds for ongoing migration between P_3 and P_2 or P_1.

Sites involved in the computation of D are likely to be in linkage disequilibrium. Therefore, a simple binomial test to assess whether D significantly differs from zero is not appropriate. Instead, the test significance can be assessed using a standard block jackknife procedure (Efron 1981). In this procedure, one removes blocks of adjacent sites one at a time. The size of blocks should be chosen to be larger than the extent of linkage disequilibrium. By computing the variance of the D statistic over the sequences M times leaving each block of the sequence in turn, and then multiplying by M and taking the square root, we can obtain an approximately normally distributed standard error using the theory of the jackknife (Reich et al. 2009).

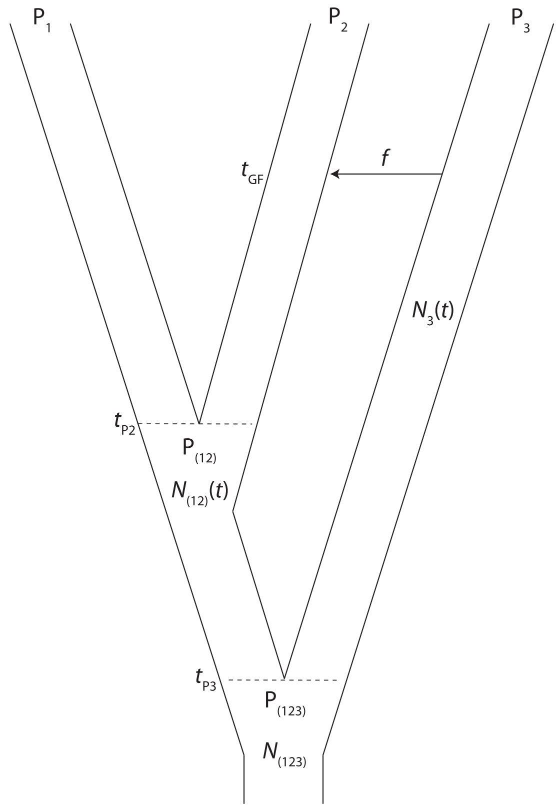

FIG. 2. Model of IUA. P_1 and P_2 split at time t_{P2}. P_3 splits from the ancestral population of P_1 and P_2, denoted $P_{(12)}$, at time t_{P3}. A single episode of admixture takes place from P_3 to P_2 at time t_{GF}. We denote by f the admixture proportion, which is the proportion of P_3 ancestry in P_2 individuals at time t_{GF}. We denote by $N_3(t)$, $N_{(12)}(t)$, and $N_{(123)}(t)$ the size of populations P_3, $P_{(12)}$, and $P_{(123)}$ at generation t in the past. The sizes of populations P_1 and P_2 do not influence D statistics.

Constraining the Admixture Proportion

D statistics depend on the demographic parameters in a complex way. As such, they are not good candidates to estimate demographic parameters. However, Appendix 1 shows that the numerator of D, denoted S, is a much simpler function of demographic parameters; for a model with arbitrary varying population sizes, we have

$$S(P_1, P_2, P_3, O) = f\left[1 - e^{\int_{t_{GF}}^{t_{P3}} \frac{1}{2N_3(x)}dx}\right](t_{P3} + \bar{t}_{P_{(123)}} - (t_{GF} + t^*)), \tag{6}$$

where $\bar{t}_{P(123)}$ is the expected time of coalescence of two lineages in $P_{(123)}$ and t^* is the expected time of coalescence of two lineages in P$_3$, given that they coalesced between t_{GF} and t_{P3}. The time $\bar{t}_{\mathrm{P(123)}}$ is equal to $2N_{(123)}$ in Appendix 1 where we assumed constant population size in P$_{(123)}$. Assume now that we have sampled a second lineage in P$_3$. Denote by P$_{3,1}$ and P$_{3,2}$ the two lineages sampled in P$_3$. It is straightforward to see that $S(\mathrm{P}_1, \mathrm{P}_{3,1}, \mathrm{P}_{3,2}, \mathrm{O})$ is equal to $S(\mathrm{P}_1, \mathrm{P}_2, \mathrm{P}_3, \mathrm{O})$ with $f = 1$ and $t_{\mathrm{GF}} = 0$ because populations P$_2$ and P$_3$ are identical with such parameter values. Therefore,

$$S(\mathrm{P}_1, \mathrm{P}_{3,1}, \mathrm{P}_{3,2}, \mathrm{O}) = \left[1 - \mathrm{e}^{\int_0^{t_{\mathrm{P3}}} \frac{1}{2N_3(x)}\mathrm{d}x}\right](t_{\mathrm{P3}} + \bar{t}_{\mathrm{P}_{(123)}} - (t_{\mathrm{GF}} + t^{**})), \tag{7}$$

where t^{**} is the expected time of coalescence of two lineages in P$_3$, given that they coalesced between $t = 0$ and t_{P3}. Thus, $S(\mathrm{P}_1, \mathrm{P}_2, \mathrm{P}_3, \mathrm{O})/S(\mathrm{P}_1, \mathrm{P}_{3,1}, \mathrm{P}_{3,2}, \mathrm{O})$ is an upper bound for f. In the case of constant ancestral population size, we have

$$\frac{S(\mathrm{P}_1, \mathrm{P}_2, \mathrm{P}_3, \mathrm{O})}{S(\mathrm{P}_1, \mathrm{P}_{3,1}, \mathrm{P}_{3,2}, \mathrm{O})} = f\frac{t_{\mathrm{P3}} - t_{\mathrm{GF}}}{t_{\mathrm{P3}}}. \tag{8}$$

If we assume that t_{GF} is small compared with t_{P3}, then one can estimate f by $\hat{f} = \frac{S(\mathrm{P}_1,\mathrm{P}_2,\mathrm{P}_3,\mathrm{O})}{S(\mathrm{P}_1,\mathrm{P}_{3,1},\mathrm{P}_{3,2},\mathrm{O})}$, which is a conservative minimum.

We note here that the strategy to estimate the admixture proportion can be extended to more populations. Assume that a sample from a sister population of P$_3$ is available. Denote by P$_4$ this population and assume that it diverged from P$_3$ before the admixture event ($t_{\mathrm{P4}} > t_{\mathrm{GF}}$, where t_{P4} is the time of divergence of P$_3$ and P$_4$) (supplementary fig. S2, Supplementary Material online). Then $\hat{f} = \frac{S(\mathrm{P}_1,\mathrm{P}_2,\mathrm{P}_4,\mathrm{O})}{S(\mathrm{P}_1,\mathrm{P}_3,\mathrm{P}_4,\mathrm{O})}$ is an unbiased estimate of the admixture proportion independent of population sizes and divergence times.

Testing for Admixture without Archaic Sample

In practice, it may be very common that no sample from the admixing archaic population is available. On the other hand, it may be fairly easy to sample from other extant populations. Assume that a sample from an outgroup to P$_1$ and P$_2$ is available and denote by P$_0$ this population. Assume that P$_0$ did not exchange genes with P$_1$ or P$_2$. Let t_{P0} be the time of divergence of P$_0$ and the population ancestral to P$_1$ and P$_2$ (see supplementary fig. S3, Supplementary Material online). We can then derive $D(\mathrm{P}_2, \mathrm{P}_1, \mathrm{P}_0, \mathrm{O})$ following the same methodology as in Appendix 1. In the case where population size is constant and equal to N in all ancestral populations, we have

$$\begin{aligned} &E[D(\mathrm{P}_2, \mathrm{P}_1, \mathrm{P}_0, \mathrm{O})] \\ &= \frac{\Pr(\mathrm{ABBA}) - \Pr(\mathrm{BABA})}{\Pr(\mathrm{ABBA}) + \Pr(\mathrm{BABA})} \\ &= \frac{3f[t_{\mathrm{P3}} - t_{\mathrm{P0}}]}{3f[t_{\mathrm{P3}} - t_{\mathrm{P0}}] + 4N(1-f)(1-\frac{1}{2N})^{t_{\mathrm{P0}}-t_{\mathrm{P1}}} + 4Nf(1-\frac{1}{2N})^{t_{\mathrm{P3}}-t_{\mathrm{P0}}}}. \end{aligned} \tag{9}$$

It is remarkable that equation (9) does not depend on the time of admixture, t_{GF}, which is likely to be unknown. It does depend, however, on the time of divergence of the archaic population, t_{P3}. Using two samples from P$_0$ or P$_1$, one can constrain the admixture proportion by using the same strategy as the one using two samples from P$_3$.

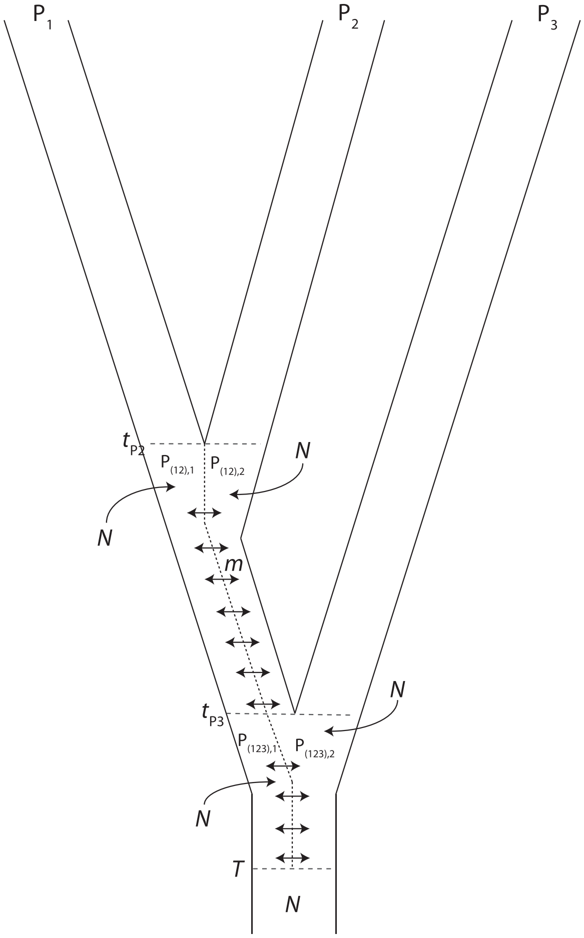

FIG. 3. Model of ancient subdivision (AS). P$_1$ and P$_2$ definitively split at time t_{P2}. The ancestral population of P$_1$ and P$_2$ is subdivided in two subpopulations, denoted P$_{(12),1}$ and P$_{(12),2}$. The two subpopulations exchange migrants at rate m. P$_3$ splits at time t_{P3}. The subdivision persists in the ancestral population of P$_{(12),1}$, P$_{(12),2}$, and P$_3$. We denote by P$_{(123),1}$ the population ancestral to P$_{(12),1}$ and by P$_{(123),2}$ the ancestral population of P$_{(12),2}$. These two subpopulations exchange migrants at rate m. Subpopulations merge at time $T > t_{\mathrm{P3}}$. All population sizes are constant and equal to N.

Ancestral Subdivision

Here we derive the expected frequencies of patterns ABBA and BABA under a simple model in which the ancestral population of P$_1$, P$_2$, and P$_3$ is not panmictic. We assume that the ancestral populations P$_{(12)}$ and P$_{(123)}$ were structured in two random mating subpopulations (fig. 3). We denote by P$_{(12),1}$ and P$_{(12),2}$ the subpopulations in P$_{(12)}$ and by P$_{(123),1}$ and P$_{(123),2}$ in P$_{(123)}$. We assume that subpopulations exchange migrants at symmetrical rate m per generation. At time T in the past, subpopulations merge into one panmictic population. We assume constant population size N in all ancestral populations. A similar model was proposed in Slatkin and Pollack (2008). Slatkin and Pollack (2008) showed that, for large values of T, the nonconcordant gene tree compatible with pattern ABBA occurred with higher frequency than the other nonconcordant gene tree. For small values of m, its frequency is even higher than the frequency of the concordant gene tree. We use AS to denote the ancestral subdivision model.

Five-State Markov Chain in $P_{(12)}$

To analyze the ancestry of the P_1 and P_2 lineages in $P_{(12)}$, we use a Markov chain method similar to the one developed by Slatkin and Pollack (2008). The states of the Markov chain are as follows: state 1, P_1 is in $P_{(12),1}$ and P_2 is in $P_{(12),2}$ (initial state); state 2, P_1 is in $P_{(12),2}$ and P_2 is in $P_{(12),1}$; state 3, P_1 and P_2 are both in $P_{(12),1}$; state 4, P_1 and P_2 are both in $P_{(12),2}$, and state 5, P_1 and P_2 coalesced (absorbing state).

To write the transition probabilities, we assume that m is small enough that only one lineage can migrate to another subpopulation per generation. Moreover, coalescent and migration events cannot occur at the same generation. Assuming constant population size N in $P_{(12),1}$ and $P_{(12),2}$, the nonzero off-diagonal elements of the transition matrix $P^{(12)}$ are

$$p_{13}^{(12)} = p_{14}^{(12)} = p_{23}^{(12)} = p_{24}^{(12)} = p_{31}^{(12)} = p_{34}^{(12)} = p_{41}^{(12)} = p_{42}^{(12)} = m,$$
$$p_{35}^{(12)} = p_{45}^{(12)} = \tfrac{1}{2N}.$$

At time t_{P3}, the distribution of states is $\pi^{(12)}(t_{P3}) = \pi^{(12)}(t_{P2})(P^{(12)})^{t_{P3}-t_{P2}}$, where $\pi^{(12)}(t_{P2}) = \{1, 0, 0, 0, 0\}$.

Thirteen-State Markov Chain in $P_{(123)}$

To model the ancestry of P_1, P_2, and P_3 in $P_{(123)}$, we use a 13-state Markov chain with states defined by state 1, P_3, $P_2 \mid P_1$; state 2, P_1, $P_3 \mid P_2$; state 3, P_1, $P_2 \mid P_3$; state 4, P_1, P_2, P_3; state 5, $(P_2, P_3) \mid P_1$; state 6, $(P_1, P_3) \mid P_2$; state 7, $(P_1, P_2) \mid P_3$; state 8, (P_2, P_3), P_1; state 9, (P_1, P_3), P_2; state 10, (P_1, P_2), P_3; state 11, $((P_2, P_3), P_1)$; state 12, $((P_1, P_3), P_2)$, and state 13, $((P_1, P_2), P_3)$. We used the symbol "|" to denote that lineages are in different subpopulations and brackets denote coalescence events. For example, state 6 takes into account the case where P_1 and P_3 lineages have coalesced, and the resulting ancestral lineage is in a different subpopulation than the P_2 lineage.

Again we assume that m is small enough that only one lineage can migrate to another subpopulation per generation and that coalescent and migration events cannot occur at the same generation. Assuming constant population size N in $P_{(123),1}$ and $P_{(123),2}$, the nonzero off-diagonal elements of the transition matrix $P^{(123)}$ are

$$\begin{aligned} p_{12}^{(123)} &= p_{13}^{(123)} = p_{14}^{(123)} = p_{21}^{(123)} = p_{23}^{(123)} = p_{24}^{(123)} = p_{31}^{(123)} \\ &= p_{32}^{(123)} = p_{34}^{(123)} = p_{58}^{(123)} = p_{85}^{(123)} = p_{69}^{(123)} = p_{96}^{(123)} \\ &= p_{7,10}^{(123)} = p_{10,7}^{(123)} = m, \end{aligned}$$

$$p_{15}^{(123)} = p_{26}^{(123)} = p_{37}^{(123)} = p_{48}^{(123)} = p_{49}^{(123)} = p_{4,10}^{(123)} = \frac{1}{2N}.$$

At time T, the distribution of states is $\pi^{(123)}(t_{P3}) = \pi^{(123)}(t_{P3})(P^{(123)})^{T-t_{P3}}$, where $\pi^{(123)}(t_{P3}) = \{\pi_1^{(12)}, \pi_2^{(12)}, \pi_3^{(12)}, \pi_4^{(12)}, 0, 0, \pi_5^{(12)}, 0, 0, 0, 0, 0, 0\}$.

Events (coalescence and/or migration) in $P_{(12)}$ only affect the expected value of D by changing the starting probabilities in $P_{(123)}$. Appendix 2 presents the details of all the possible cases leading to ABBA and BABA patterns at time T, where all subpopulations merge into a single random mating population.

The frequencies of patterns ABBA and BABA have no simple analytical expressions under the AS model. However, they can be easily computed numerically for any set of parameter values.

Expected Coalescence Time Conditional on Gene Tree Topology

In this section, we derive the expected coalescence time of P_2 and P_3, given that they coalesced first, under the IUA and the AS models. This is a quantity of interest because it determines the length of the internal branch of tree topology III (fig. 1), and therefore the frequency of pattern ABBA. We use these derivations to illustrate the confounding effect of ancestral subdivision on the test for admixture.

We denote by $\tau_{23}^{(IUA)}$ the expected coalescence time of P_2 and P_3, given that they coalesced first, under the IUA model with constant ancestral population size. It can be decomposed as the sum of two terms, depending on whether the P_2 lineage originated from P_3 or not:

$$\begin{aligned} \tau_{23}^{(IUA)} = & f\left[\left(1 - \left(1 - \frac{1}{2N}\right)^{t_{P3} - t_{GF}}\right)t* \right. \\ & \left. + \frac{1}{3}\left(1 - \frac{1}{2N}\right)^{t_{P3} - t_{GF}}\left(\frac{2N}{3} + t_{P3}\right)\right] \\ & + \frac{(1-f)}{3}\left(1 - \frac{1}{2N}\right)^{t_{P3} - t_{P2}}\left(\frac{2N}{3} + t_{P3}\right), \end{aligned} \tag{10}$$

where t^* is the expected time of coalescence of P_2 and P_3 given that they coalesce between t_{GF} and t_{P3} (Appendix 1, eq. A1.3).

Then we derive the conditional expected time of coalescence of P_2 and P_3 under the AS model, denoted $\tau_{23}^{(AS)}$. We start by conditioning on whether coalescence occurs before time T:

$$\begin{aligned} \tau_{23}^{(AS)} = & (\tau_{23}^{(AS)} | \tau_{23}^{(AS)} \le T)\Pr(\tau_{23}^{(AS)} \le T) \\ & + (\tau_{23}^{(AS)} | \tau_{23}^{(AS)} > T)\Pr(\tau_{23}^{(AS)} > T). \end{aligned} \tag{11}$$

The first term is equal to the expected time for the Markov chain to first hit state 5, 8, or 11, conditional on this time being lower than T; we compute it using standard Markov chain theory. It can be shown that this conditional time tends to $4N + 1/(2m)$ when T grows to infinity, which is equal to the expected coalescence time of two individuals in different demes in a two-island model (Slatkin 1991). Because the ancestral population is assumed to be panmictic after time T, the second term of equation (10) is equal to

$$(\tau_{23}^{(AS)} | \tau_{23}^{(AS)} > T)\Pr(\tau_{23}^{(AS)} > T) = \frac{1}{3}\left(t_{P3} + \frac{2N}{3}\right)\sum_{i=1}^{4}\pi_i^{(123)}(T). \tag{12}$$

Although there is no simple analytical expression for $\tau_{23}^{(AS)}$, it can be computed numerically for any set of parameter values.

Effect of Sequencing Error on *D* Statistics

In this section, we study the effect of sequencing error on *D*. In what follows, we use capital letters to denote the observed count of a pattern and small letters to denote the true counts in the absence of sequencing error. We assume that sequencing error is uniform along the sequence. We denote by e_1, e_2, e_3, and e_4 the probability that a base was incorrectly read in population P$_1$, P$_2$, P$_3$, and O, respectively. Let e_{ij} be the probability that a site was read with an error in populations P$_i$ and P$_j$ at the same time ($j = 4$ denotes population O). We neglect terms of higher order. A sequencing error on O will cause the reads for P$_1$, P$_2$, and P$_3$ to be mislabeled. For example, if O was sequenced with an error at a site where the true pattern is "aaaa", we would read pattern "BBBA." Thus, for the two patterns of interest, we have

$$\begin{aligned} n_{\mathrm{ABBA}} = {} & e_1 n_{\mathrm{bbba}} + e_2 n_{\mathrm{aaba}} + e_3 n_{\mathrm{abaa}} + e_4 n_{\mathrm{baaa}} + n_{\mathrm{abba}} \\ & + e_{12} n_{\mathrm{baba}} + e_{13} n_{\mathrm{bbaa}} + e_{14} n_{\mathrm{aaaa}} \\ & + e_{23} n_{\mathrm{aaaa}} + e_{24} n_{\mathrm{bbaa}} \\ & + e_{34} n_{\mathrm{baba}} \end{aligned} \tag{13}$$

and

$$\begin{aligned} n_{\mathrm{BABA}} = {} & e_1 n_{\mathrm{aaba}} + e_2 n_{\mathrm{bbba}} + e_3 n_{\mathrm{baaa}} + e_4 n_{\mathrm{abaa}} + n_{\mathrm{baba}} \\ & + e_{12} n_{\mathrm{abba}} + e_{13} n_{\mathrm{aaaa}} + e_{14} n_{\mathrm{bbaa}} \\ & + e_{23} n_{\mathrm{bbaa}} + e_{24} n_{\mathrm{aaaa}} \\ & + e_{34} n_{\mathrm{abba}} \end{aligned} \tag{14}$$

assuming that error rates are small.

False-Positive Rate in the Presence of Sequencing Error

Here we look at the effect of sequencing error on the false-positive rate of the test for admixture. We assume a null model with no admixture and randomly mating ancestral populations. In this model, the true expected counts verify $n_{\mathrm{abba}} = n_{\mathrm{baba}}$ and $n_{\mathrm{abaa}} = n_{\mathrm{baaa}}$. The second equality assumes that the mutation rates in P$_1$ and P$_2$ are equal. Under these assumptions, we have

$$\begin{aligned} n_{\mathrm{ABBA}} - n_{\mathrm{BABA}} = {} & (e_1 - e_2)(n_{\mathrm{bbba}} - n_{\mathrm{aaba}}) + (e_{14} + \\ & e_{23} - e_{13} - e_{24})(n_{\mathrm{aaaa}} - n_{\mathrm{bbaa}}). \end{aligned} \tag{15}$$

The effect of sequencing error on *D*(P$_1$, P$_2$, P$_3$, O) is then

$$D(\mathrm{P}_1, \mathrm{P}_2, \mathrm{P}_3, \mathrm{O}) = \frac{(e_1 - e_2)(n_{\mathrm{bbba}} - n_{\mathrm{aaba}}) + (e_{14} + e_{23} - e_{13} - e_{24})(n_{\mathrm{aaaa}} - n_{\mathrm{bbaa}})}{n_{\mathrm{abba}} + n_{\mathrm{baba}} + (e_1 + e_2)(n_{\mathrm{bbba}} + n_{\mathrm{aaba}}) + (e_{14} + e_{23} + e_{13} + e_{24})(n_{\mathrm{aaaa}} + n_{\mathrm{bbaa}}) + 2(e_3 + e_4) n_{\mathrm{abaa}}}. \tag{16}$$

If the sequencing error rates are the same for the four samples, then sequencing error will have no influence on the false-positive rate of the test as long as they are small enough that third-order terms can be ignored.

To illustrate the effect of different sequencing error rates on the false-positive rate of the test, note that in the case where the outgroup divergence time, t_{O}, is large compared with the divergence time of P$_3$, the two dominant patterns will be bbba and aaaa. We considered a case where 98% of sites were aaaa, 0.02% abba and baba, and 1.97% bbba. All other configurations represented 0.01% of sites. We took 10^9 sites in total. This settings roughly correspond to a population tree where P$_1$ and P$_2$ are modern human populations, P$_3$ is a Neandertal population, and Chimpanzee is the outgroup. Indeed, Green et al. (2010) estimated that the time of divergence of Neandertal and modern humans was less than 440,000 years before present, which is small compared with divergence time of modern humans and chimpanzee.

Under the simplifying assumption that sites are independent, true pattern counts abba and baba follow a binomial distribution with parameters $p = 0.5$ and $n = n_{\mathrm{abba}} + n_{\mathrm{baba}}$. In this case, the standard deviation (SD) of *D* is equal to $2\sqrt{0.25/(n_{\mathrm{abba}} + n_{\mathrm{baba}})} = 0.0022$.

Synthetic Data Analysis

Power of the Test for Admixture

In order to assess the power of the test for admixture, we simulated data under different demographic scenarios. To simulate counts for ABBA and BABA patterns, we simulated trees using the *ms* software (Hudson 2002). We then placed a mutation on the tree; the branch it fell on was selected with probability proportional to its length. The descendants of the mutated branch then define which pattern was generated. Under this simulation scheme, the number of independent *ms* replicates corresponds to the number of unlinked polymorphic sites in the three populations P$_1$, P$_2$, and P$_3$. For all the demographic scenarios described below, we varied this number between 10,000 and 1,000,000.

We simulated counts for ABBA and BABA patterns under the IUA model. The time of divergence of P$_1$ and P$_2$ was set to $t_{\mathrm{P2}} = 3{,}000$ generations. P$_3$ diverged from the ancestral population P$_{(123)}$ at time $t_{\mathrm{P3}} = 12{,}000$ generations. Admixture occurred at time $t_{\mathrm{GF}} = 2{,}500$ generations. We considered different cases for the ancestral population sizes: constant size $N = 10{,}000$ in all populations; a bottleneck in P$_3$ prior to the admixture event; and a bottleneck in P$_{(12)}$. Both bottlenecks involved a 100-fold reduction in population size and lasted for 1,000 generations. They both started at 3,500 generations in the past. We also considered a case where P$_1$ and P$_2$ exchanged migrants at rate $2Nm = 0.5$ between $t = 0$ and t_{P2}. We varied the admixture proportion *f* from 0 (no admixture) to 0.1 (10% introgression from P$_3$ into P$_2$).

Because replicates are independent, D is approximately normally distributed, and its SD can be computed from the binomial distribution: $\mathrm{SD}(D) = 2\sqrt{0.25/(n_{\mathrm{ABBA}} + n_{\mathrm{BABA}})}$. Each simulation class was repeated 1,000 times. We computed the power of the test for admixture as the proportions of the 1,000 repetitions were $|D|$ was larger than twice its standard error. For the simulations without admixture, we expect that roughly 95% of the simulated D statistics will be comprised between plus or minus twice the SD.

Effect of Ascertainment Bias

When computed on genotype data, the way SNPs were ascertained can bias D statistics. Typically, an ascertainment scheme consists of two phases. First, SNPs are discovered from the genetic material of a small group of individuals, called the discovery panel. Then the discovered SNPs are typed in a larger sample (Eberle and Kruglyak 2000). The extent to which the ascertainment scheme will bias D statistics depends on many parameters, such as the genetic composition of the discovery panel and its relationship to the larger samples. Ascertainment bias can be corrected when one assumes that the discovery panel is a random sample from the studied population (Nielsen et al. 2004). However, in the context of D statistics, the relationship between the discovery sample and the four typed populations P_1, P_2, P_3, and O can be complex.

To illustrate the effect of a simple ascertainment scheme on D statistics, we simulated genotype data under the IUA model using 100,000 independent replicates of *ms* (Hudson 2002). The time of divergence of P_1 and P_2 was set to $t_{P2} = 3{,}000$ generations. P_3 diverged from the ancestral population $P_{(123)}$ at time $t_{P3} = 12{,}000$ generations. Admixture occurred at time $t_{GF} = 2{,}500$ generations. All population sizes were set equal to $N = 10{,}000$. We simulated 100 individuals in P_1 and P_2 and 1 individual in P_3. To simulate ascertainment bias, we subsampled individuals in P_1 or P_2 and kept only sites that were variable in those individuals. In total, we explored five subsampling scenarios: 1) 2 individuals from P_1; 2) 20 individuals from P_1; 3) 2 individuals from P_2; 4) 20 individuals from P_2; and 5) no subsampling (no ascertainment bias). Each of these five subsampling scenarios was repeated for $f = 0$ (no admixture) and $f = 0.05$ (5% introgression from P_3 into P_2).

Application to Neandertal Data

The Neandertal genome was recently sequenced to ~1× coverage, producing a total of ~4 billon base pairs (Green et al. 2010). Green et al. (2010) analyzed data from five present-day human males from the CEPH-Human Genome Diversity Project panel (French, Han, Papuan, San, and Yoruba) that were sequenced to ~5× coverage. For each pair of modern humans P_1 and P_2, Green et al. (2010) computed D(P_1, P_2, Neandertal, Chimpanzee). The chimpanzee individual was represented by the reference chimpanzee genome (PanTro2).

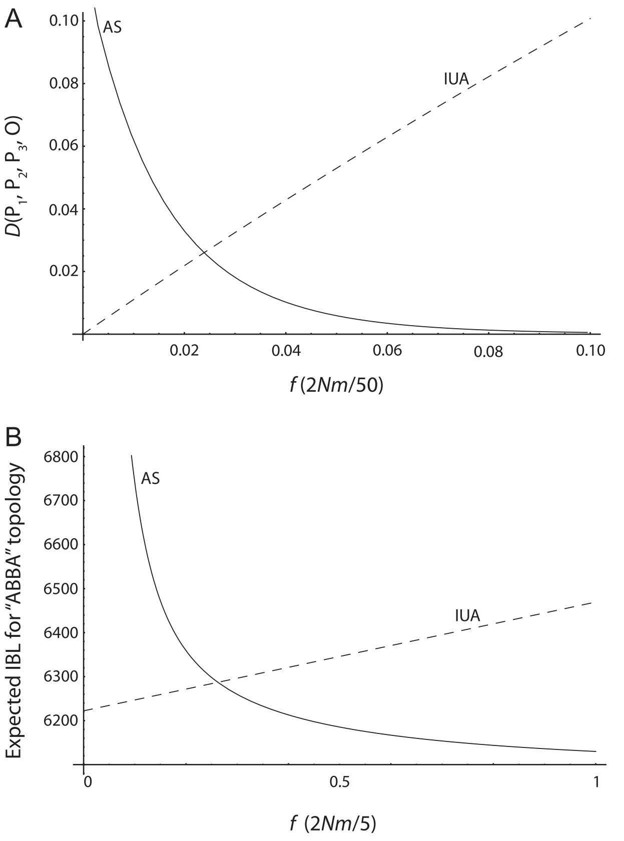

FIG. 4. D statistics and expected coalescence time of P_2 and P_3 under the IUA and AS models. We assumed parameter values of $t_{GF} = 2{,}500$, $t_{P2} = 3{,}000$, $t_{P3} = 12{,}000$, $T = 25{,}000$ (times in generations), and a constant ancestral population size equal to $N = 10{,}000$. (*A*) D(P_1, P_2, P_3, O) as a function of f (IUA) and $2Nm$ (AS). (*B*) Expected internal branch length (IBL) for topologies where P_2 and P_3 coalesced first as a function of f (IUA) and $2Nm$ (AS). The rescaling for $2Nm$ on (*A*) and (*B*) was chosen so that $2Nm$ and f vary in the same range.

Results

Ancient Subdivision Confounds the Test for Admixture

We computed D(P_1, P_2, P_3, O) under the IUA and the AS models (Appendices 1 and 2) for parameter values $t_{GF} = 2{,}500$, $t_{P2} = 3{,}000$, $t_{P3} = 12{,}000$, $T = 25{,}000$ (times in generations), and a constant ancestral population size equal to $N = 10{,}000$. Figure 4*A* displays D statistics as a function of f (IUA) and $2Nm$ (AS). We also computed the expected time of coalescence of P_2 and P_3 given that they coalesced first. Figure 4*B* displays $\tau_{23}^{(\mathrm{IUA})}$ as a function of f and $\tau_{23}^{(\mathrm{AS})}$ as a function of $2Nm$. We used a computer algebra program to compute numerically quantities under the AS model.

The D statistic curves intersect, which shows that both the IUA and the AS models can be made to fit data if P_3 shares more derived allele with P_1 or P_2. Therefore, the test for admixture based on D alone will always be confounded by the presence of ancestral subdivision.

More generally, figure 4*B* shows that the IUA and the AS models predict the same expected coalescence times for some parameter values. Thus, any test for admixture solely based on expected coalescence times between one lineage

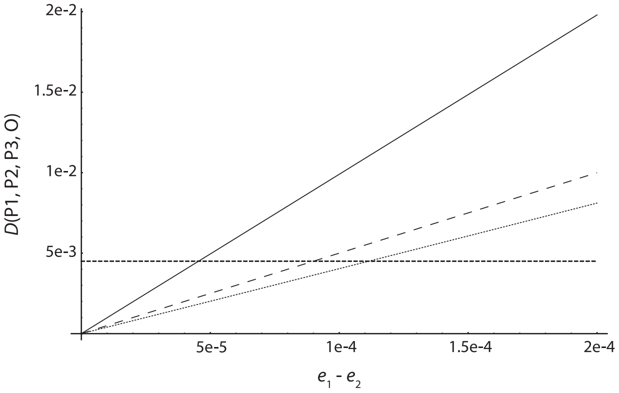

Fig. 5. $D(P_1, P_2, P_3, O)$ as a function of the sequencing error difference $e_1 - e_2$. The dotted curve assumes that the probability to have an error in two populations is equal to the product of sequencing errors in the two populations ($e_{ij} = e_i \times e_j$). The dashed curve corresponds to $e_{ij} = \min(e_i, e_j)/10$, and the solid line assumes $e_{ij} = \min(e_i, e_j)$. The horizontal dashed line corresponds to twice the standard error of D, assuming that sites are independent.

from different populations will be confounded by nonrandom mating in ancestral populations.

Robustness to Sequencing Error

We assume a background sequencing error of 0.001 in each sample. Figure 5 plots D against $e_1 - e_2$, keeping $e_1 + e_2 = 0.002$ and $e_3 = e_4 = 0.001$. If we assume that the sequencing error rate in two populations is the product of error rates in each of the two populations (i.e., errors are independent in the two populations), then a difference in sequencing error rates $e_1 - e_2 = 0.00011$ will cause the test for admixture to produce false-positive results. If we assume that the probability to sequence two populations with an error at the same base is only one order of magnitude lower than the error rate in one population, then differences in error rates of 0.00008 will create false positives. In the worst-case scenario, where sequencing error in two populations is of the same order of magnitude as sequencing error in one populations, then $e_1 - e_2 = 0.000045$ is enough to create false positives.

Power of the Test for Admixture

Figure 6 reports the power analysis for the simulations of instantaneous admixture with constant population size, a bottleneck in P_3, a bottleneck in $P_{(12)}$, and migration between P_1 and P_2. Adding a bottleneck in P_3 increased the power of the test substantially; this is because, if the P_2 lineage originated from P_3, the bottleneck increases the probability that the P_2 and P_3 lineages coalesce before t_{P3}. Adding a bottleneck in $P_{(12)}$ improved the power of the test even more. This is because, in the case where admixture did not happen (eq. A1.1), the probability that the P_1 and P_2 lineages do not coalesce before t_{P3} is decreased by the bottleneck. Therefore, this case contributes less to the denominator of D. Migration between P_1 and P_2 decreased the power of the test for admixture.

For the simulations without admixture ($f = 0$), figure 6 reports the false-positive rate of the test for admixture at a 5% level. In the case of constant population size, the false-positive rate was 0.050 for 10,000 sites and 0.046 for 100,000 sites. This is an expected result as D is approximately normally distributed for independent sites. Plus or minus twice the SD represents roughly a 95% confidence interval for D. As predicted, adding bottlenecks in the ancestral populations, or migration between P_1 and P_2, did not significantly affect the false-positive rate of the test.

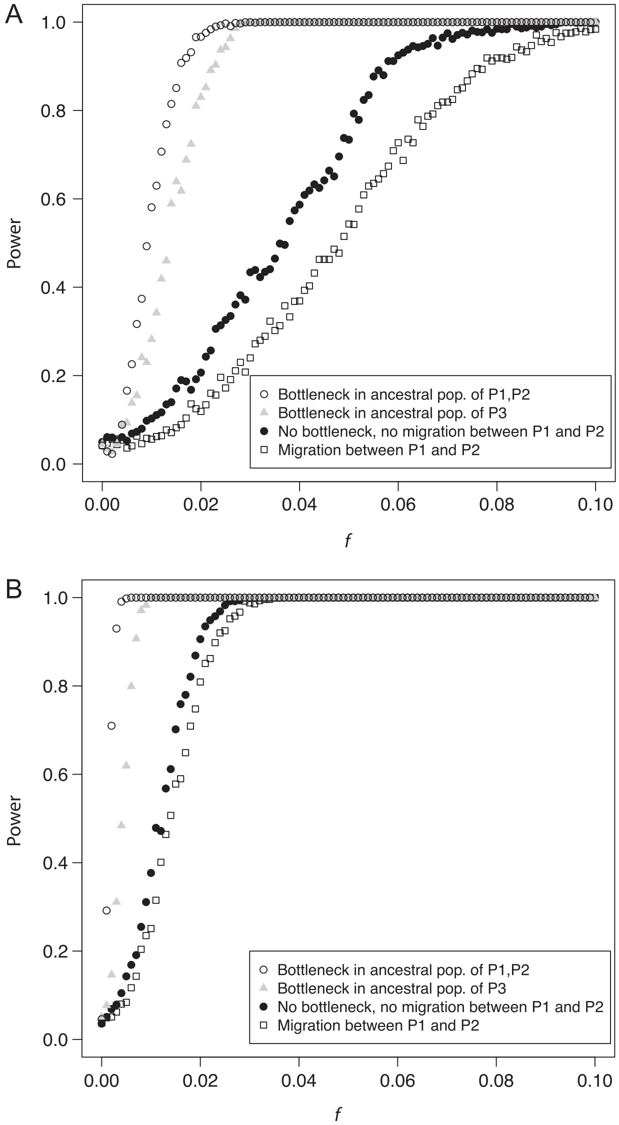

Fig. 6. Power of the test for admixture. We simulated (*A*) 10,000 unlinked polymorphic sites and (*B*) 100,000 unlinked polymorphic sites. Dots: no bottleneck and no migration between P_1 and P_2. Circles: bottleneck in the ancestral population of P_1 and P_2. Grey triangles: bottleneck in P_3. Squares: ongoing migration between P_1 and P_2 at rate $2Nm = 0.5$. All bottlenecks involved a 100-fold reduction in population size and lasted 1,000 generations. They started at 3,500 generations in the past.

Robustness to Ascertainment Bias

For data simulated without admixture ($f = 0$), $D(P_1, P_2, P_3, O)$ did not differ significantly from 0 regardless of the ascertainment scheme used. More precisely, we obtained the following values for $D(P_1, P_2, P_3, O)$ under each scheme: 1)

Table 1. Neandertals Share More Genetic Variants With Non-Africans than With Africans (Green et al. 2010).

	P_1	P_2	$D(P_1, P_2, N, C)$, %
African–African	San	Yoruba	−0.1 ± 0.4
African–non-African	San	French	4.2 ± 0.4
	San	Papuan	3.9 ± 0.5
	San	Han	5.0 ± 0.5
	Yoruba	French	4.5 ± 0.4
	Yoruba	Papuan	4.4 ± 0.6
	Yoruba	Han	5.3 ± 0.5
Non-African–non-African	French	Papuan	0.1 ± 0.5
	French	Han	1.0 ± 0.6
	Papuan	Han	0.7 ± 0.6

-7.2×10^{-4} (SD = 5.7×10^{-3}); 2) 6.2×10^{-4} (SD = 3.4×10^{-3}); 3) -2.5×10^{-5} (SD = 5.5×10^{-3}); 4) 3.8×10^{-4} (SD = 3.4×10^{-3}); and 5) -7.9×10^{-5} (SD = 3.5×10^{-3}).

The IUA model with 5% admixture ($f = 0.05$) predicts $D(P_1, P_2, P_3, O) = 0.053$ for the demographic parameters we used (eq. A1.5). The simulated data yielded the following values: 1) 0.021 (SD = 4.8×10^{-3}); 2) 0.024 (SD = 3.4×10^{-3}); 3) 0.037 (SD = 6.0×10^{-3}); 4) 0.048 (SD = 3.5×10^{-3}); and 5) 0.053 (SD = 3.5×10^{-3}).

Therefore, the simple ascertainment scheme we used here did not affect the false-positive rate of the test for admixture. However, it lowered the value of $D(P_1, P_2, P_3, O)$ for scenarios with 5% admixture. We checked that this result was consistent for other admixture proportions (not reported). The difference between the true value and the observed one was larger when the discovery sample was taken in P_1.

Application to Neandertal

Testing for Admixture between Neandertals and Modern Humans

Table 1 shows D statistics from Green et al. (2010). $D(P_1, P_2, P_3, O)$ were significantly positive if P_1 is an African individual (San or Yoruba) and P_2 is a non-African individual (French, Han, or Papuan). When comparing two African individuals (P_1 = San and P_2 = Yoruba) or two non-African individuals (P_1 and P_2 = French, Han, or Papuan), D did not differ significantly from 0. Furthermore, D statistics involving one African and one non-African did not differ significantly from one another. Significance was assessed using a standard block jackknife method (Efron 1981). A D value was considered to be significantly different from 0 when its absolute value was greater than three times its standard error. The D statistics are compatible with a scenario where Neandertals admixed with modern humans outside of Africa, before Europeans, East Asians, and Melanesians split (Green et al. 2010).

We note that Green et al. (2010) used about 7,475,200 biallelic sites for the D statistics analysis (Green et al. 2010, SOM 15). Therefore, we expect the power of the test to be extremely high. Green et al. (2010) also computed D statistics using genotype data from the CEPH-Human Genome Diversity Project panel. D statistics were lower when computed on SNPs than when computed on sequencing data, consistent with the predicted bias of a simple ascertainment scheme. However, it is remarkable that although raw values of D statistics changed, the general conclusion that non-Africans were significantly closer to Neandertals than Africans held.

Estimating the Proportion of Neandertal Ancestry in Non-Africans

Taking advantage of the fact that bones from different Neandertal individuals were available, Green et al. (2010) estimated the proportion of Neandertal ancestry in non-Africans as

$$\hat{f} = \frac{S(\mathrm{Afr}, \mathrm{OOA}, \mathrm{Nea}_1, \mathrm{Chimp})}{S(\mathrm{Afr}, \mathrm{Nea}_1, \mathrm{Nea}_2, \mathrm{Chimp})}, \qquad (15)$$

where Afr is an African sample (San or Yoruba), OOA is a non-African sample (French, Han, or Papuan), and Nea_1 and Nea_2 are two Neandertal individuals. The S statistics were averaged over the different possible combinations of African and non-African samples. Using the standard errors of S statistics (estimated by block jackknife), this yielded an estimate of $\hat{f} = 1-4\%$. However, equation (7) shows that this is a conservative minimum. Green et al. (2010) estimated the population divergence time of Neandertals and modern humans to be $t_N \approx 270{,}000 - 440{,}000$ years before present. The time of gene flow has to occur after modern humans went out of Africa, an event that took place an estimated $t_{\mathrm{OOA}} \approx 45{,}000 - 100{,}000$ years before present (Forster and Matsumura 2005). Assuming a constant population size, the bias in the estimator of f is most important when $t_N - t_{\mathrm{OOA}}$ is small. In the worst-case scenario, the bias is equal to $(270{,}000 - 100{,}000)/100{,}000 = 0.063$. Therefore, a corrected estimate of f in the case of constant ancestral population size is $f^* = 1$–6%.

The data are also compatible with a scenario of ancestral subdivision. Using $t_{P3} = 12{,}000$ generations, $t_{P2} = 3{,}000$ generations, and a constant population size of $N = 10{,}000$, migration rates $1.0 < 2Nm < 2.0$ and $15{,}000 < T < 25{,}000$ generations were compatible with the data. However, we note that the ancestral subdivision has to last for thousands of generations in order for D to be on the order of 5% (supplementary fig. S4, Supplementary Material online). Such subdivision would require that local populations and the geographic barriers that separate them persist for very long times, much longer than seems reasonable for highly mobile and adaptable hominins.

Discussion

Asymmetry in three population gene trees is known to occur in the presence of gene flow (Meng and Kubatko 2009) or ancestral subdivision (Slatkin and Pollack 2008). Here we show that asymmetry can be exploited to design a test for admixture between three closely related populations. The test requires only the assumption that the ancestral population is randomly mating; it is robust to other demographic assumptions such as variations in population size. Moreover, the test is robust to sequencing error if one assumes that the error rate is the same in the sequences analyzed. However, if sequencing error rates differ, then sequencing error can create false positives and that is more likely to be true if sequencing errors are not independent.

Finally, we illustrated how to constrain the admixture proportion in a simple model of admixture.

Duration and Direction of Gene Flow

The model of admixture as a one-way instantaneous gene flow event is not intended to be realistic. Instead, instantaneous gene flow is the simplest possible admixture model that enables to derive the *D* statistic expectation and estimate the admixture proportion. In the more realistic case of ongoing migration, *D* statistics can be computed numerically using the theory of Markov chains, as illustrated in the ancient structure model.

Introgression is often asymmetrical between hybridizing species (Barton and Hewitt 1985; Orive and Barton 2002). When an expanding population colonizes a new habitat and hybridizes with previous residents, breeding events at the front of expansion can result in the substantial introduction of genes in the expanding population (Currat et al. 2008). As a consequence, detectable ancient admixture is more likely to be found in extant populations when introgression with archaic populations occurred during a range expansion, as is thought to be the case for modern humans and Neandertals (Currat et al. 2008; Green et al. 2010).

However, there are also cases known where detectable introgression occurs from the invasive population to the resident species (Hastings et al. 2005). We note that deriving *D* statistics when gene flow is bidirectional or in the other direction is not more difficult, and the same methodology can be applied. If there was gene flow from the extant population to the archaic one, then any *D* statistic involving individuals only from the extant populations would not be affected. This provides a framework to test for the direction of gene flow. Taking advantage that Eurasians are more closely related to some African populations than to others, Green et al. (2010) used a similar argument to rule out substantial gene flow from modern humans to Neandertals.

Effect of Ascertainment Bias

Although *D* statistics are primarily designed for sequence data, they can be readily applied to genotype data. We showed that the false-positive rate of the test for admixture was not affected by a very simplified ascertainment scheme where we subsampled individuals in either P_1 or P_2. This is because, in this case, the bias in allele frequencies introduced by the ascertainment compensated itself in patterns ABBA and BABA. However, even this very simple ascertainment scheme lowers the power of the test because the ascertainment biases sampled SNPs toward higher frequencies, increasing the counts of pattern ABBA and BABA to a similar extent, and therefore increasing the denominator of *D*.

We caution that we used an oversimplified ascertainment scheme. More complex scenarios can introduce bias in an unforeseen way. For example, if the discovery sample is from neither P_1 nor P_2, then the bias will depend on the genealogical relationships between the population where SNPs were discovered and P_1, P_2, and P_3. However, we note that the main conclusion regarding the test for admixture of Green et al. (2010) held when applied to genotype data. Still, we stress that the test is likely to be more powerful using sequencing data. An investigator wishing to apply the test using genotype data should try to correct allele frequencies for ascertainment bias. Several approaches have been proposed to do so (e.g., Clark et al. 2005).

Effect of Recurrent Mutation

We assumed that each polymorphism was created by a single mutation. If mutations were recurrent, then different copies of the derived allele would arise independently. Hence, many of the gene genealogies in figure 1 that we ignored in computing the expectation of *D* would have to be accounted for. If the outgroup diverged from the lineage leading to the other populations a long time in the past, as it is the case for modern humans, Neandertals, and chimpanzees, recurrent mutations are likely to occur on the long branch leading to the outgroup. As a consequence, genealogies Ib, IId, and IIIb would create the pattern BABA and genealogies Ia, IIa, and IIId would create the pattern ABBA (see fig. 1). If there is no gene flow from P_3 to P_1 or P_2, and if the mutation rate is the same in the three populations, then Ia = Ib, IId = IIId, and IIIb = IIa (where the equal sign denotes that corresponding trees occur with equal frequencies). Therefore, recurrent mutations are not expected to increase the false-positive rate of the test for Neandertal admixture. In the case where P_1, P_2, P_3, and O are more diverged from each other (e.g., modern humans, chimpanzee, gorilla, and orangutan), recurrent mutations are likely to occur on other branches as well. Recurrent mutations on the terminal branches leading to P_1 and P_2 could potentially increase the false-positive rate of the test for admixture.

In the presence of gene flow, it is difficult to predict whether recurrent mutation increases or decreases *D*. The exact result depends on details of the demographic model, but the effect is proportional to the probability that a second mutation occurs and the constant of proportionality is less than one because many of the gene genealogies still have no effect on *D*. Still, recurrent mutations can affect the power of the test for admixture.

We also note that differing mutation rates in the P_1 and P_2 populations since their divergence are not expected to bias *D*. The reason for this is that by restricting to sites where P_3 is derived relative to the outgroup O, we are restricting the analysis to mutations that arose prior to divergence of the ancestral populations of P_1, P_2, and P_3.

Effect of Mapping Errors

We assumed that the sequences from P_1, P_2, P_3, and O were aligned to a common sequence without error. Depending on the nature of the sampled populations and the sequencing technology, different alignment strategies may be used. Green et al. (2010) aligned the Neandertal sequence both to the reference human and to the chimpanzee genomes (hg18 and panTro2). Different aligners were used depending on the technology used to sequence the Neandertal and modern humans genomes used in the study. Because of the variety of the alignment strategies, it is difficult to model the effect of mapping error on D statistics in a general framework.

The extent to which D statistics will be affected depends on which sample is more prone to mapping error. Let us assume that all sequences were aligned to the outgroup O. Mapping errors in the P_3 sequence are likely to have a negligible effect on the false-positive rate of the test for admixture. This is because mismapping in P_3 will equally affect patterns ABBA and BABA. However, alignment error in P_3 can affect the power of the test if it artificially increases the counts of patterns ABBA and BABA. Mapping errors in P_1 or P_2 can have a strong effect on the false-positive rate of the test. Mismapping in P_1 will cause P_2 and P_3 to match each other artificially too often (therefore increasing the count of the ABBA pattern). Thus, it is crucial to test the effect of mapping error in P_1 or P_2. One possible approach is to stratify the data into categories based on the read coverage in sequences P_1 and P_2 and to compute $D(P_1, P_2, P_3, O)$ in each category. If mapping is not an issue, then D should remain stable in categories where the coverage is high enough.

Effect of Natural Selection

We implicitly assumed selective neutrality in all our analyses. For natural selection to bias the test for admixture, it must increase the number of derived alleles shared by P_3 and P_1 or P_2. The ABBA and BABA patterns in Green et al. (2010) were distributed genome wide. Therefore, it seems unlikely that the observed D statistics in Green et al. (2010) were biased by natural selection.

However, when the data consist of a few loci, natural selection could potentially bias D statistics by increasing the number of derived alleles in, say, P_2. It would result in P_2 matching P_3 more often than P_1, even if there was no admixture. The exact effect of natural selection on D statistics remains a subject of further investigation.

Distinguishing between Gene Flow and AS

We showed that D statistics do not allow us to distinguish between the models of admixture and ancestral subdivision because ancestral subdivision and admixture can produce the same expected coalescence time between two individuals from different populations. However, it is known that ancestral population subdivision results in a higher than expected variance in coalescence times (Wall et al. 2009). Therefore, ancestral subdivision is likely to result in more variation in gene tree depth when using several samples from the extant population. This in turn will affect the site frequency spectrum of the extant population (Harpending et al. 1998). Designing a statistic to distinguish between the two scenarios will require using more than one sample per population.

Concluding Remarks

Inference of ancient admixture is usually limited by the lack of archaic DNA sequence data. Most methods so far have focused on detecting the expected signature of ancient admixture in extant genomes. Here we studied the behavior of a test statistic for ancient admixture based on the direct comparison of DNA sequences of three closely related populations. Our derivation of this test statistic under alternative demographic models showed how to robustly estimate the archaic contribution to extant populations. We demonstrated that our approach can also be applied when no archaic population has been sampled. This paper illustrates the advantage of using one sample per population, which greatly simplifies inference by removing the influence of many demographic assumptions that are not testable in most situations.

Supplementary Material

Supplementary figures S1–S4 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We thank Anna-Sapfo Malaspinas for helpful discussions that helped improve the manuscript. E.Y.D and M.S were supported by a US National Institute of Health grant (R01-GM40282).

References

Barton N. 2001. The role of hybridization in evolution. *Mol Ecol.* 10:551–568.

Barton N, Hewitt G. 1985. Analysis of hybrid zones. *Annu Rev Ecol Syst.* 16:113–148.

Clark AG, Hubisz MJ, Bustamante CD, Williamson SH, Nielsen R. 2005. Ascertainment bias in studies of human genome-wide polymorphism. *Genome Res.* 15(11):1496–1502.

Currat M, Excoffier L. 2004. Modern humans did not admix with Neanderthals during their range expansion into Europe. *PLoS Biol.* 2(12):e421.

Currat M, Ruedi M, Petit RJ, Excoffier L. 2008. The hidden side of invasions: massive introgression by local genes. *Evolution.* 62(8):1908–1920.

Eberle M, Kruglyak L. 2000. An analysis of strategies for discovery of single-nucleotide polymorphisms. *Genet Epidemiol.* 19(Suppl 1): S29–S35.

Efron B. 1981. Nonparametric estimates of standard error: the jackknife, the bootstrap and other methods. *Biometrika.* 68(3): 589–599.

Forhan G, Martiel J, Blum M. 2008. A deterministic model of admixture and genetic introgression: the case of Neanderthal and Cro-Magnon. *Math Biosci.* 216:71–76.

Forster P, Matsumura S. 2005. Evolution—did early humans go north or south? *Science* 308(5724):965–966.

Green RE, Krause J, Briggs AW, et al. (56 co-authors). 2010. A draft sequence of the Neandertal genome. *Science.* 328(5979):710–722.

Green RE, Krause J, Ptak SE, et al. (11 co-authors). 2006. Analysis of one million base pairs of Neanderthal DNA. *Nature* 444(7117):330–336.

Green RE, Malaspinas AS, Krause J, et al. (25 co-authors). 2008. A complete Neandertal mitochondrial genome sequence determined by high-throughput sequencing. *Cell* 134(3):416–426.

Harpending HC, Batzer MA, Gurven M, Jorde LB, Rogers AR, Sherry ST. 1998. Genetic traces of ancient demography. *Proc Natl Acad Sci U S A.* 95(4):1961–1967.

Hastings A, Cuddington K, Davies K. 2005. The spatial spread of invasions: new developments in theory and evidence. *Ecol Lett.* 8(1):91–101.

Hudson RR. 1983. Testing the constant-rate neutral allele model with protein-sequence data. *Evolution* 37(1):203–217.

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

Krause J, Lalueza-Fox C, Orlando L, et al. (13 co-authors). 2007. The derived FOXP2 variant of modern humans was shared with Neandertals. *Curr Biol.* 17(21):1908–1912.

Krings M, Stone A, Schmitz RW, Krainitzki H, Stoneking M, Pääbo S. 1997. Neandertal DNA sequences and the origin of modern humans. *Cell* 90(1):19–30.

Kuhner MK, Beerli P, Yamato J, Felsenstein J. 2000. Usefulness of single nucleotide polymorphism data for estimating population parameters. *Genetics* 156(1):439–447.

Lalueza-Fox C, Rompler H, Caramelli D, et al. (17 co-authors). 2007. A melanocortin 1 receptor allele suggests varying pigmentation among Neanderthals. *Science* 318(5855):1453–1455.

Meng C, Kubatko LS. 2009. Detecting hybrid speciation in the presence of incomplete lineage sorting using gene tree incongruence: a model. *Theor Popul Biol.* 75(1):35–45.

Nielsen R, Hubisz M, Clark A. 2004. Reconstituting the frequency spectrum of ascertained single-nucleotide polymorphism data. *Genetics* 168(4):2373–2382.

Noonan JP, Coop G, Kudaravalli S, et al. (11 co-authors). 2006. Sequencing and analysis of Neanderthal genomic DNA. *Science* 314(5802):1113–1118.

Orive ME, Barton NH. 2002. Associations between cytoplasmic and nuclear loci in hybridizing populations. *Genetics.* 162(3): 1469–1485.

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

Reich D, Thangaraj K, Patterson N, Price AL, Singh L. 2009. Reconstructing Indian population history. *Nature* 461(7263):489–494.

Serre D, et al. 2004. No evidence of Neandertal mtDNA contribution to early modern humans. *PLoS Biol.* 2(3):E57.

Slatkin M. 1991. Inbreeding coefficients and coalescence times. *Genet Res.* 58(2):167–175.

Slatkin M, Pollack JL. 2008. Subdivision in an ancestral species creates asymmetry in gene trees. *Mol Biol Evol.* 25(10):2241–2246.

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

Wall JD. 2000. Detecting ancient admixture in humans using sequence polymorphism data. *Genetics* 154(3):1271–1279.

Wall JD, Hammer MF. 2006. Archaic admixture in the human genome. *Curr Opin Genet Dev.* 16(6):606–610.

Wall JD, Lohmueller KE, Plagnol V. 2009. Detecting ancient admixture and estimating demographic parameters in multiple human populations. *Mol Biol Evol.* 26(8):1823–1827.

Appendix A. Derivation of *D* under the Model of Gene Flow

Here we present the detailed derivation of $D(\mathrm{P}_1, \mathrm{P}_2, \mathrm{P}_3, \mathrm{O})$ under the IUA model. For simplicity of notations, we assume that the population size in $\mathrm{P}_{(123)}$ is constant and equal to $N_{(123)}$. However, the derivations can be easily extended for arbitrary varying population size in $\mathrm{P}_{(123)}$. There are three classes of event that can produce patterns ABBA and BABA.

1. *The* P_2 *lineage traces its ancestry through the* $\mathrm{P}_{(12)}$ *side of the phylogeny (probability 1−f), and between* t_{P2} *and* t_{P3}, *the* P_1 *and* P_2 *lineages do not coalesce* $\left(\textit{probability } e^{-\int_{t_{\mathrm{P2}}}^{t_{\mathrm{P3}}} \frac{1}{2N_{(12)}(x)} dx}\right)$.

$$\Pr_1(\mathrm{ABBA}) = \Pr_1(\mathrm{BABA})$$
$$= \frac{2N_{(123)}\mu(1-f)}{3} e^{-\int_{t_{\mathrm{P2}}}^{t_{\mathrm{P3}}} \frac{1}{2N_{(12)}(x)} dx}. \tag{A1.1}$$

2. *The* P_2 *lineage traces its ancestry through the* P_3 *side of the phylogeny (probability f), and between* t_{GF} *and* t_{P3}, *the two lineages do not coalesce* $\left(\textit{probability } e^{-\int_{t_{\mathrm{GF}}}^{t_{\mathrm{P3}}} \frac{1}{2N_3(x)} dx}\right)$.

$$\Pr_2(\mathrm{ABBA}) = \Pr_2(\mathrm{BABA}) = \frac{2N_{(123)}\mu f}{3} e^{-\int_{t_{\mathrm{GF}}}^{t_{\mathrm{P3}}} \frac{1}{2N_3(x)} dx}. \tag{A1.2}$$

3. *The* P_2 *lineage traces its ancestry through the* P_3 *side of the phylogeny (probability f), and between* t_{GF} *and* t_{P3}, *the two lineages coalesce.* This history creates gene genealogies only of type III and results only in ABBA sites (never BABA sites in the absence of recurrent mutation in the same genealogy). The probability that there is a coalescence before t_{P3} is $\left[1 - e^{-\int_{t_{\mathrm{GF}}}^{t_{\mathrm{P3}}} \frac{1}{2N_3(x)} dx}\right]$. Once the coalescence occurs, the ancestral lineage cannot coalesce with P_1 before t_{P3}. After t_{P3}, the average coalesce time is $2N_{(123)}$. Therefore, the expected length of the internal branch is $t_{\mathrm{P3}} + 2N_{(123)} - (t_{\mathrm{GF}} + \bar{t})$, where $\bar{t}$ is the expected coalescence time between P_2 and P_3, given that the coalescence occurs before t_{P3}. A short analysis shows that

$$\bar{t} = \frac{1}{1 - e^{-\int_{t_{\mathrm{GF}}}^{t_{\mathrm{P3}}} \frac{1}{2N_3(x)} dx}} \int_{t_{\mathrm{GF}}}^{t_{\mathrm{P3}}} \frac{t}{2N_3(x)} e^{-\int_{t_{\mathrm{GF}}}^{t} \frac{1}{2N_3(x)} dx} dt. \tag{A1.3}$$

Thus,

$$\Pr_3(\mathrm{ABBA}) = \mu f \left[1 - e^{-\int_{t_{\mathrm{GF}}}^{t_{\mathrm{P3}}} \frac{1}{2N_3(x)} dx}\right] \left[t_{\mathrm{P3}} + 2N_{(123)} - t_{\mathrm{GF}} - \frac{\int_{t_{\mathrm{GF}}}^{t_{\mathrm{P3}}} \frac{t}{2N_3(t)} e^{-\int_{t_{\mathrm{GF}}}^{t_{\mathrm{P3}}} \frac{1}{2N_3(x)} dx} dt}{1 - e^{-\int_{t_{\mathrm{GF}}}^{t_{\mathrm{P3}}} \frac{1}{2N_3(x)} dx}}\right]. \tag{A1.4}$$

The overall probability of ABBA and BABA is obtained by adding, and the mutation rate parameter μ cancels. Therefore,Appendix B

Here we present the detailed derivation of $D(\mathrm{P}_1, \mathrm{P}_2, \mathrm{P}_3, \mathrm{O})$ under the AS model. At time T, there are six possibilities depending on the states of the Markov chain running in $\mathrm{P}_{(123)}$.

$$E[D] = \frac{\Pr(\mathrm{ABBA}) - \Pr(\mathrm{BABA})}{\Pr(\mathrm{ABBA}) + \Pr(\mathrm{BABA})}$$

$$= \frac{3f\left[1 - e^{-\int_{t_{\mathrm{GF}}}^{t_{\mathrm{P3}}} \frac{1}{2N_3(x)}dx}\right](2N_{(123)} + t_{\mathrm{P3}} - t_{\mathrm{GF}} - \bar{t})}{3f\left[1 - e^{-\int_{t_{\mathrm{GF}}}^{t_{\mathrm{P3}}} \frac{1}{2N_3(x)}dx}\right](2N_{(123)} + t_{\mathrm{P3}} - t_{\mathrm{GF}} - \bar{t}) + 2fN_{(123)}e^{-\int_{t_{\mathrm{GF}}}^{t_{\mathrm{P3}}} \frac{1}{2N_3(x)}dx} + 2(1-f)N_{(123)}e^{-\int_{t_{\mathrm{GF}}}^{t_{\mathrm{P3}}} \frac{1}{2N_3(x)}dx}}. \quad \text{(A1.5)}$$

Appendix B

1. *No coalescence event occurred*

This case corresponds to states 1, 2, 3, and 4 of the Markov chain in $P_{(123)}$. It contributes equally to ABBA and BABA:

$$\Pr_1(\mathrm{ABBA}) = \Pr_1(\mathrm{BABA}) = \frac{2N\mu}{3}\sum_{i=1}^{4}\pi_i^{(123)}(T). \quad \text{(A2.1)}$$

2. P_1 *and* P_2 *lineages coalesced*

This case does not contribute to ABBA and BABA.

3. P_2 *and* P_3 *lineages coalesced, but not the* P_1 *lineage*

This case corresponds to states 5 and 8 of the Markov chain in $P_{(123)}$. It does not contribute to BABA. Its contribution to ABBA is

$$\Pr_3(\mathrm{ABBA}) = (\pi_5^{(123)} + \pi_8^{(123)})(T)\mu(2N + T - t_{\mathrm{P3}} - t_{23}^*), \quad \text{(A2.2)}$$

where t_{23}^* is the expected time of coalescence of the P_2 and P_3 lineages given that coalescence occurs between t_{P3} and T.

4. P_3 *and* P_1 *lineages coalesced, but not the* P_2 *lineage*

This case corresponds to states 6 and 9 of the Markov chain in $P_{(123)}$. It does not contribute to ABBA. Its contribution to BABA is, similarly to case (3),

$$\Pr_4(\mathrm{ABBA}) = (\pi_6^{(123)} + \pi_9^{(123)})(T)\mu(2N + T - t_{\mathrm{P3}} - t_{13}^*). \quad \text{(A2.3)}$$

5. P_3 *and* P_2 *lineages coalesced first and then with the* P_1 *lineage.*

This case corresponds to state 11 of the Markov chain in $P_{(123)}$. It does not contribute to BABA. Its contribution to ABBA is

$$\Pr_5(\mathrm{ABBA}) = \pi_{11}^{(123)}(T)\mu(t_1^{**} - t_{23}^*), \quad \text{(A2.4)}$$

where t_1^{**} is the expected time of coalescence of the P_1 lineage, given that coalescence occurs between t_{23}^* and T.

6. P_3 *and* P_1 *lineages coalesced first and then with the* P_2 *lineage.*

This case corresponds to state 12 of the Markov chain in $P_{(123)}$. It does not contribute to ABBA. Similarly to case (5), its contribution to BABA is

$$\Pr_6(\mathrm{ABBA}) = \pi_{12}^{(123)}(T)\mu(t_2^{**} - t_{13}^*), \quad \text{(A2.5)}$$

where t_2^{**} is the expected time of coalescence of the P_2 lineage, given that coalescence occurs between t_{13}^* and T.

The conditional expected coalescence times, t_{23}^*, t_{13}^*, t_1^{**}, and t_2^{**} are computed using standard Markov chain theory as the expected times of first hit of corresponding states given that the hit occurred before time T. We have

$$t_{23}^* = \sum_{i=1}^{4}\pi_i^{(12)}(t_{\mathrm{P3}})\sum_{t=t_{\mathrm{P3}}}^{T} t\left[\frac{(P^{(123)}(t) - P^{(123)}(t-1))[i,5]}{\pi_5^{(123)}(T)} + \frac{(P^{(123)}(t) - P^{(123)}(t-1))[i,8]}{\pi_8^{(123)}(T)}\right]. \quad \text{(A2.6)}$$

The other conditional expected coalescence times are computed similarly. There is no simple analytic expression for D.