Joint estimation of contamination, error and demography for nuclear DNA from ancient humans

Fernando Racimo[a,1], Gabriel Renaud[b,1], Montgomery Slatkin[a]

[a] *Department of Integrative Biology, University of California, Berkeley, CA, USA*
[b] *Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany*

Abstract

When sequencing an ancient DNA sample from a hominin fossil, DNA from present-day humans involved in excavation and extraction will be sequenced along with the endogenous material. This type of contamination is problematic for downstream analyses as it will introduce a bias towards the population to which the contaminating individuals belong. Quantifying the extent of contamination is a crucial step as it allows researchers to account for possible biases that may arise in downstream genetic analyses. Here, we present an MCMC algorithm to co-estimate the contamination rate, sequencing error rate and demographic parameters - including drift times and admixture rates - for an ancient nuclear genome obtained from human remains, when the putative contaminating DNA comes from present-day humans. We assume we have a large panel representing the putative contaminating population (e.g. European, East Asian or African). The method is implemented in a C++ program called 'Demographic Inference with Contamination and Error' (DICE). The program can also be used to determine the most likely population to which the contaminant DNA belongs. We applied it to simulations and Neanderthal genome data, and we recover accurate estimates of all parameters, even when the average sequencing coverage is low (0.5X) and the per-read contamination rate is high (25%).

Keywords: Ancient DNA, Contamination, MCMC, Human evolution, Demography

Email address: `fernandoracimo@gmail.com` (Fernando Racimo)

[1] These authors contributed equally to this work.

1. Introduction

When attempting to sequence an ancient human genome [1, 2, 3, 4, 5, 6], the common practice is to assess the amount of present-day human contamination in a sequencing library. Several methods exist to obtain a contamination estimate. First, one can look at ’diagnostic positions’ in the mitochondrial genome at which a particular archaic population may be known to differ from all members of the putative contaminant (modern) population. Then, one counts how many ’modern’ reads are observed at those positions in the archaic genome. This is the most popular technique and has been routinely deployed in the sequencing of Neanderthal genomes [7, 1]. However, contamination levels in the mithochondrial genome may differ from those in the rest of the genome. A second technique involves assessing whether the sample was male or female using the ratio of reads that map to the X and the Y chromosomes [1]. After determining the biological sex, the proportion of reads that are non-concordant with the sex of the archaic individual are used to estimate contamination from individuals of the opposite sex (e.g. Y-chr reads in an archaic female genome are indicative of male contamination). A final technique involves using a maximum likelihood approach to co-estimate the amount of contamination, sequencing error and heterozygosity in the autosomal nuclear genome [1, 3], using a likelihood optimization algorithm, like L-BFGS-B [8].

Afterwards, if the sequenced data is assessed to not be highly contaminated ($< \sim 2\%$), demographic analyses are performed on the sequences while ignoring the contamination. If the library is highly contaminated, it is usually treated as unusable and discarded. Neither of these outcomes is optimal: ignoring the contaminating reads may affect downstream analyses, while discarding the library may waste rich genomic data that could provide important demographic insights.

One way to address this problem was proposed by Skoglund et al. [9], who developed a statistical framework to separate contaminant from endogenous DNA reads by using the patterns of chemical deamination characteristic of ancient DNA. The method produces a score which reflects the likelihood that a particular read is endogenous or not. This approach, however, may not be able to make a clean distinction between the two sources of DNA, especially for young ancient DNA samples, as chemical degradation may not have affected all reads belonging to the archaic individual.

Instead of (or in addition to) attempting to separate the two type of reads before performing a demographic analysis, one could incorporate the uncertainty stemming from the contaminant reads into a probabilistic inference framework. Such an approach has already been implemented in the analysis of a haploid mtDNA archaic genome (Renaud et al. in review). However, mtDNA represents a single gene genealogy, and, so far, no equivalent method has been developed for the analysis of the nuclear genome, which contains the richest amount of population genetic information. Here, we present a method to co-estimate the contamination rate, per-base error rate and a simple demography for an autosomal nuclear genome of an ancient hominin. We assume we have a large panel representing the putative contaminant population, for example, European, Asian or African 1000 Genomes data [10]. The method uses a Bayesian framework to obtain posterior probabilities of all parameters of interest, including population-size-scaled divergence times and admixture rates. It can also be used to determine the most probable contaminant population, by running it using different contaminant panels and finding the panel with the highest posterior probability.

2. Methods

2.1. Basic framework for estimation of error and contamination

We will first describe the probabilistic structure of our inference framework. We begin by defining the following parameters:

r_c: contamination rate in the ancient DNA sample coming from the contaminant population

ϵ: error rate, i.e. probability of observing a derived allele when the true allele is ancestral, or vice versa.

i: number of chromosomes that contain the derived allele at a particular site in the ancient individual ($i = 0,\ 1\ or\ 2$)

d_j: number of derived reads observed at site j

$\mathbf{d}$: vector of d_j counts for all sites $j = \{1,\ ...,\ N\}$ in a genome

a_j: number of ancestral reads observed at site j

$\mathbf{a}$: vector of a_j counts for all sites $j = \{1,\ ...,\ N\}$ in a genome

w_j: known frequency of a derived allele in a candidate contaminant panel at site j ($0 \leq w_j \leq 1$)

$\mathbf{w}$: vector of w_j frequencies for all sites $j = \{1,\ ...,\ N\}$ in a genome

K: number of informative SNPs used as input

θ: population-scaled mutation rate. $\theta = 4N_e\mu$, where N_e is the effective population size and μ is the per-generation mutation rate.

We are interested in computing the probability of the data given the contamination rate, the error rate, the derived allele frequencies from the putative contaminant population ($\mathbf{y}$) and a set of demographic parameters ($\mathbf{\Omega}$). We will use only sites that are segregating in the contaminant panel and we will assume that we observe only ancestral or derived alleles at every site (i.e. we ignore triallelic sites). In some of the analyses below, we will also assume that we have additional data ($\mathbf{O}$) from present-day populations that may be related to the population to which the sample belongs. The nature of the data in $\mathbf{O}$ will be explained below, and will vary in each of the different cases we describe. The parameters contained in $\mathbf{\Omega}$ may simply be the drift times separating the contaminant population and the sample from their common ancestral population. However, $\mathbf{\Omega}$ may include additional parameters, such as the admixture rate - if any - between the contaminant and the sample population. The number of parameters we can include in $\mathbf{\Omega}$ will depend on the nature of the data in $\mathbf{O}$.

For all models we will describe, the probability of the data can be defined as:

$$P[\ \mathbf{a},\ \mathbf{d}\ |\ r_C, \epsilon, \mathbf{w}, \Omega, \mathbf{O}] = \prod_{j=1}^{K} P[a_j, d_j | r_C, \epsilon, w_j, \Omega, \mathbf{O}] \tag{1}$$

where

$$P[a_j, d_j | r_C, \epsilon, w_j, \Omega, \mathbf{O}] = \sum_{i=0}^{2} P[a_j, d_j\ |\ i, r_C, \epsilon, w_j] P[i\ | \Omega, \mathbf{O}] \tag{2}$$

We focus now on computation on the likelihood for one site j in the genome. In the following, we abuse notation and drop the subscript j. Given the true genotype of the ancient individual, the number of derived and ancestral reads at a particular site follows a binomial distribution that depends on the genotype, the error rate and the rate of contamination [1, 3]:

$$P[a, d | i, r_C, \epsilon, w] = \binom{a+d}{d} q_i^d (1 - q_i)^a \tag{3}$$

where

$$q_2 = r_C w(1-\epsilon) + r_C(1-w)\epsilon + (1-r_C)(1-\epsilon) \tag{4}$$

$$q_1 = r_C w(1-\epsilon) + r_C(1-w)\epsilon + (1-r_C)(1-\epsilon)/2 + (1-r_C)\epsilon/2 \qquad (5)$$

$$q_0 = r_C w(1-\epsilon) + r_C(1-w)\epsilon + (1-r_C)\epsilon \qquad (6)$$

In the sections below, we will turn to the more complicated part of the model, which is obtaining the probability $P[i|\mathbf{\Omega}, \mathbf{O}]$ for a genotype in the ancient sample, given particular demographic parameters and additional data available. We will do this in different ways, depending on the kind of data we have at hand.

2.2. Diffusion-based likelihood for neutral drift separating two populations

First, we will work with the case in which $\mathbf{O} = \mathbf{y}$, where $\mathbf{y}$ is a vector of frequencies y_j from an "anchor" population that may be closely related to the population of the ancient DNA sample. An example of this scenario would be the sequencing of a Neanderthal sample that is suspected to have present-day human contamination, from which many genomes are available.

For all analyses below, we restrict to sites where $0 < y_j < 1$. Note that it is entirely possible (but not required) that $\mathbf{y} = \mathbf{w}$, meaning that, aside from the ancient DNA sample, the only additional data we have are the frequencies of the derived allele in the putative contaminant population, which we can use as the anchor population too. However, it is also possible to use a contaminant panel that is different from the anchor population (Figure 1.A). We will assume we have sequenced a large number of individuals from a panel of the contaminant population (for example, The 1000 Genomes Project panel) and that the panel is large enough such that the sampling variance is approximately 0. In other words, the frequency we observe in the contaminant panel will be assumed to be equal to the population frequency in the entire contaminant population. In this case, $\mathbf{\Omega} = \{\tau_\mathbf{C}, \tau_\mathbf{A}\}$, where τ_A and τ_C are defined as follows:

τ_A: drift time (i.e. time in generations scaled by twice the haploid effective population size) separating the population to which the ancient individual belongs from the ancestor of both populations

τ_C: drift time separating the anchor population from the ancestor of both populations

We need to calculate the conditional probabilities $P[i|\mathbf{\Omega}, \mathbf{O}] = \mathbf{P}[\mathbf{i}|\mathbf{y}, \tau_\mathbf{C}, \tau_\mathbf{A}]$ for all three possibilities for the genotype in the ancient individual: $i =$

0, 1 or 2. To obtain these expressions, we rely on Wright-Fisher diffusion theory (reviewed in Ewens [11]), especially focusing on the two-population site-frequency spectrum (SFS) [12]. The full derivations can be found in the Appendix, and lead to the following formulas:

$$P[\ i = 0 \mid y, \tau_C, \tau_A\ ] = 1 - y * e^{-\tau_C} - \frac{1}{2} * y * e^{-\tau_A - \tau_C} + y \left(y - \frac{1}{2} \right) e^{-\tau_A - 3\tau_C} \quad (7)$$

$$P[\ i = 1 \mid y, \tau_C, \tau_A\ ] = y * e^{-\tau_A - \tau_C} + y\,(1 - 2y)\, e^{-\tau_A - 3\tau_C} \quad (8)$$

$$P[\ i = 2 \mid y, \tau_C, \tau_A\ ] = y * e^{-\tau_C} - \frac{1}{2} * y * e^{-\tau_A - \tau_C} + y \left(y - \frac{1}{2} \right) e^{-\tau_A - 3\tau_C} \quad (9)$$

We generated 10,000 neutral simulations using msms [13] for different choices of τ_C and τ_A (with $\theta = 20$ in each simulation) to verify our analytic expressions were correct (Figure 2). The probability does not depend on θ, so the choice of this value is arbitrary.

The above probabilities allows us to finally obtain $P[i \mid y_j, \mathbf{\Omega}, \mathbf{O}]$.

2.3. Estimating drift and admixture in a three-population model

Although the above method gives accurate results for a simple demographic scenario, it does not incorporate the possibility of admixture between the contaminant population and the sample population. This is important, as the signal of contamination may mimic the pattern of recent admixture. We will assume that, in addition to the ancient DNA sample, we also have the following data, which constitute $\mathbf{O}$:

1) A large panel from a population suspected to be the contaminant in the ancient DNA sample. The sample frequencies from this panel will be labeled $\mathbf{w}$, as before.

2) Two panels of high-coverage genomes from two “anchor” populations that may be related to the ancient DNA sample. One of these populations - called population Y - may (but need not) be the same population as the contaminant and may (but need not) have received admixture from the ancient population (Figure 1.B). The sample frequencies for this population will be labeled as $\mathbf{y}$. The other population - called Z - will have sample frequencies labeled $\mathbf{z}$. We will assume the drift times separating these two populations

are known (parameters τ_Y and τ_Z in Figure 1.B). This is a reasonable assumption as these parameters can be accurately estimated without the need of using an ancient outgroup sample, as long as admixture is not extremely high.

We can then estimate the remaining drift parameters, the error and contamination rates and the admixture time (β) and rate (α) between the archaic population and modern population Y. The diffusion solution for this three-population scenario with admixture is very difficult to obtain analytically. Instead, we use a numerical approximation, implemented in the program ∂a∂i [14].

2.4. Markov Chain Monte Carlo method for inference

We incorporated the likelihood functions defined above into a Markov Chain Monte Carlo (MCMC) inference method, to obtain posterior probability distributions for the contamination rate, the sequencing error rate, the drift times and the admixture rate. Our program - which we called 'DICE' - is coded in C++ and is freely available at: `http://grenaud.github.io/dice/`

We assumed uniform prior distributions for all parameters. By default, we limit the maximum contamination rate to 50% and the maximum sequencing error rate per read to 10%. When incorporating admixture, we also capped the maximum possible admixture rate to 50% and generally chose realistic admixture time boundaries when analyzing real data. Although these are the default boundaries, they can be modified by the user.

For the starting chain at step 0, an initial set of parameters $X_0 = \{ r_{C0}, \epsilon_0, \Omega_0 \}$ is sampled randomly from their prior distributions. At step k, a new set of values for step $k+1$ is proposed by drawing values for each of the parameters from Normal distributions. The mean of each of those distributions is the value for each parameter at state X_k and the standard deviation is the difference between the upper and lower boundary of the prior, divided by a constant that can be increased or decreased to achieve a desired rate of acceptance of new states [15]. By default, this constant is equal to 1,000 for all parameters. The new state is accepted with probability:

$$P[accept] = min\left(1, \frac{P[\mathbf{a},\mathbf{d} \mid X_{k+1}]}{P[\mathbf{a},\mathbf{d} \mid X_k]}\right) \tag{10}$$

where $P[\mathbf{a},\mathbf{d} \mid X_k]$ is the likelihood defined in Equation 1.

Unless otherwise stated below, we ran the MCMC chain for 100,000 steps in all analyses, with a burn-in period of 40,000 and sampling every 100 steps.

The sampled values were then used to construct posterior distributions for each parameter.

2.5. BAM file functionality and multiple error rates

The standard input for DICE is a file containing counts of particular ancestral / derived read combinations and SNP frequency configurations (see README file online). As an additional feature of DICE, we also incorporated a module for the user to directly input a BAM file and a file containing population frequencies for the anchor and contaminant panels, rather than the standard input.

Fu et al. [5] showed that, when estimating contamination, ancient DNA data can be better fit by a two-error model than a single-error model. In that study, the authors co-estimate the two error rates along with the proportion of the data that is affected by each rate. Therefore, we also included this error model as an option that the user can choose to incorporate when running our program. Furthermore, we developed an alternative error estimation method that allows the user to flag sites that are likely to undergo cytosine deamination in ancient DNA, and therefore suffer from different types of errors than those commonly found in present-day sequencing data. Our program can then estimate the two error rates separately, for sites that are prone to be deaminated and those that are not.

3. Results: two-population method

3.1. Simulations

We first used the MCMC implementation described above to obtain posterior distributions from simulated data, under the two-population inference framework. We simulated two populations (i.e. an archaic and a modern human population) with constant population size that split a number of generations ago. For each demographic scenario tested, we generated 20,000 independent replicates (theta=1) in *ms* [16], making sure each simulation had at least one usable SNP (i.e. segregating in the anchor population(s)). In general, this yielded ∼80,000 usable SNPs in total. We then proceeded to sample derived and ancestral allele counts using the same binomial sampling model we use in our inference framework, under different sequencing coverage and contamination conditions. Our simulation framework does not include correlated or base-specific sequencing errors, but allows us to concentrate on the strengths and limitations of our method in inferring contamination and demographic parameters, rather than on sequencing-specific limitations that may vary across platforms and samples. In all simulations, the contaminant panel was the same as the anchor population panel.

Figure 3 and 4 show parameter estimation results from various demographic and contamination scenarios for a low-coverage (3X) and a high-coverage (30X) archaic genome, respectively, with low sequencing error (0.1%), and a contaminant/anchor population panel of 100 haploid genomes. In both cases, the method accurately estimates the error rate, the contamination rate and the drift parameters. All parameters are also accurately estimated for the same scenarios even if the sequencing error rate is high (10%) (Figure S1).

3.2. Performance under violations of model assumptions

We also checked what would happen if the modern human panel used was small. Figure S4 shows results for cases in which the contaminant/anchor panel is made up of 20 haploid genomes. In this case, all parameters are estimated accurately, with only a slight bias towards overestimating the drift parameters, presumably because the low sampling of individuals acts as a population bottleneck, artificially increasing the drift time parameters estimated.

Additionally, we simulated a scenario in which only a single human contaminated the sample. That is, rather than drawing contaminant reads from a panel of individuals, we randomly picked a set of two chromosomes at each unlinked site and only drew contaminant reads from those two chromosomes. Figure S5 shows that inference is robust to this scenario, unless the contamination rate is very high (25%). In that case, the drift of the archaic genome is substantially under-estimated, but the error, contamination and anchor drift parameters only show slight inaccuracies in estimation.

We then investigated the effect of admixture in the anchor/contaminant population from the archaic population, occurring after their divergence, which we did not account for in the simple, two-population model (Figure S2). In this case, the error and the contamination rates are accurately estimated, but both drift times are underestimated. This is to be expected, as admixture will tend to homogenize allele frequencies and thereby reduce the apparent drift separating the two populations.

Finally, we tested performance when the sample is of extremely low average coverage (0.5X). We tried different numbers of independent replicate simulations and found that the number of sites needed to obtain accurate inferences is higher than when using a sample of higher coverage. At 800,000 replicates with theta= 20, we obtained approximately 1.6 million valid SNPs for inference, which was enough to reach reasonable levels of accuracy (Figure S3). We note that this number of SNPs is approximately the same as what is available, for example, in the low-coverage (0.5X) Mezmaiskaya Neanderthal genome [4], which contains about 1.55 million valid sites with coverage ≥ 1, and which we analyze below. We also observed that the MCMC chain in some of these simulations needed a longer time to converge than when testing samples of higher coverage, especially when contamination is very high, and so in this set of simulations, we ran it for 1 million steps instead of 100,000, with a burn-in of 940,000 steps and sampling every 100 steps.

3.3. Real data

We first applied our method to published ancient DNA data from two Neanderthals: a low-coverage genome (0.5X) from Mezmaiskaya Cave in Western Russia and a high-coverage genome (52X) from Denisova cave in Siberia (the Altai Neanderthal) [4]. In both cases, we visually ensured that the chain had converged. The demographic, error and contamination estimates are shown in Tables 1 and 2, respectively. We used the African (AFR) 1000 Genomes phase 3 panel [10] as the anchor population. The drift times estimated for both samples are consistent with the known demographic history of Neanderthals and modern humans, and the contamination rates largely agree with previous estimates (see Discussion below). We observe a higher error rate and a lower contamination rate in the Mezmaiskaya sample than in the Altai sample.

We ran our method with different putative contaminant panels (AFR, EAS, AMR, EUR, SAS). For the Altai sample, the most probable contaminant is of European ancestry, as the EUR panel has a much larger posterior probability than the other panels (Table 1). For the Mezmaiskaya sample, all panels have very similar posterior probabilities (Table 2): the low coverage in this case precludes us from clearly distinguishing which was the contaminant population.

We sought to determine the robustness of our results to different levels of GC content. We partitioned the Altai Neanderthal genome into three different regions of low ($0\% - 30\%$), medium ($31\% - 69\%$) and high ($70\% - 100\%$) GC content, using the 'GC content' track downloaded from the UCSC genome browser [17]. We then used the two-population method to infer contamination, error and drift parameters, using Africans as the anchor population and Europeans as the contaminant population (Figure S6). We observe that contamination rates are higher in low-GC regions than in medium-GC regions (Welch one-sided t-test on the posterior samples, $P < 2.2e\text{-}16$), which in turn have higher contamination rates than high-GC regions ($P < 2.2e\text{-}16$). The opposite trend occurs in the error estimates, while the drift parameters are largely unaffected. However, we find that the differences we observe across GC levels are almost entirely eliminated by removing CpG sites from the input dataset (Figure S6). CpG sites are known to have higher mutation rates than the rest of the genome, and are more likely to lead to ancestral state misidentification (ASM, Hernandez et al. [18]). For this reason, we recommend either filtering them out when testing for contamination on ancient DNA datasets (which is what we did in Tables 1 and 2) or developing new models that can account for ASM, which we do not pursue here.

As a negative control, we also tested a present-day Yoruba genome (HGDP00936) sequenced to high coverage [4], which should not contain any contamination. Indeed, when applying our method, we find this to be the case (Figure S7). We infer 0% contamination, regardless of whether we use EUR or AFR as the candidate contaminant. Furthermore, the anchor drift time is very close to 0 when using AFR as the anchor population (as the sample belongs to that same population), while it is non-zero ($= 0.22$) when using EUR, which is consistent with the drift time separating Europeans from the ancestor of Europeans and Africans [19]. This also indicates that the method is useful for testing samples that have shorter drift times than Neanderthal, like ancient modern humans.

4. Results: three-population method

4.1. Simulations

We applied our three-population method to estimate both drift times and admixture rates. We simulated a high-coverage (30X) archaic human genome under various demographic and contamination scenarios. Each of the two anchor population panels contained 20 haploid genomes. The admixture time was 0.08 drift units ago, which under a constant population size of 2N=20,000 would be equivalent to 1,600 generations ago. When running our inference program, we set the admixture time prior boundaries to be between 0.06 and 0.1 drift units ago.

We find that the admixture time is inaccurately estimated under this implementation - likely due to lack of information in the site-frequency spectrum - so we do not show estimates for that parameter below. For admixture rates of 0%, 5% or 20%, the error and contamination parameters are estimated accurately in all cases (Figures 5, S8 and S9, respectively). The method is less accurate when estimating the demographic parameters, especially the admixture rate which is sometimes under-estimated. Importantly though, the accuracy of the contamination rate estimates are not affected by incorrect estimation of the demographic parameters.

We also tested what would happen if the admixture time was simulated to be recent: 0.005 drift units ago, or 100 generations ago under a constant population size of 2N=20,000. When estimating parameters, we set the prior for the admixture time to be between 0 and 0.01 drift units ago. In this last case, we observe that the drift times and the admixture rate (20%) are more accurately estimated than when the admixture event is ancient (Figure 6).

4.2. Real data

We also applied the three-population inference framework to the high-coverage Altai Neanderthal genome. We first estimated the two drift times specific to Europeans and Africans after the split from each other (τ_Y and τ_Z, respectively), using ∂a∂i and the L-BFGS-B likelihood optimization algorithm [8], but without using the archaic genome ($\tau_{Afr} = 0.009$, $\tau_{Eur} = 0.255$). Then, we used our MCMC method to estimate the rest of the drift times, the archaic admixture rate and the contamination and error parameters in the Neanderthal genome. We set the admixture time prior boundaries to be between 0.06 and 0.1 drift units ago, which is a realistic time frame given knowledge about modern human - Neanderthal cohabitation in Eurasia [20]. As before, we tested different populations for the putative contaminant and find Europeans to be the most probable contaminant population.

Although we attempted to apply the three-population method to the low-coverage Mezmaiskaya Neanderthal genome, different contaminant panels resulted in highly inconsistent drift parameters, even when using the same anchor population. This is due to the larger number of parameters that have to be explored in the three-population method, which requires more data than available in the Mezmaiskaya sample. Therefore, we conclude the two-population method is better suited than the three-population method for samples of very low coverage

5. Discussion

We have developed a new method to jointly infer demographic parameters, along with contamination and error rates, when analyzing an ancient DNA sample. The method can be deployed using a C++ program (DICE) that is easy to use and freely downloadable. We therefore expect it to be highly applicable in the field of paleogenomics, allowing researchers to derive useful information from previously unusable (highly contaminated) samples, including archaic humans like Neanderthals, as well as ancient modern humans.

Applications to simulations show that the error and contamination parameters are estimated with high accuracy, and that demographic parameters can also be estimated accurately so long as enough information (e.g. a large panel of modern humans) is available. The drift time estimates reflect how much genetic drift has acted to differentiate the archaic and modern populations since the split from their common ancestral population, and can be converted to divergence times in generations if an accurate history of population size changes is also available (for example, via methods like PSMC, [21]).

We also applied our method to real data, specifically to two Neanderthal genomes at high and low coverage, and a present-day Yoruba genome. For the Yoruba genome, we infer no contamination, as would be expected from a modern-day sample, and drift times indicating the Yoruba sample indeed belongs to an African population.

The contamination and sequencing error estimates we obtained for the Neanderthals are roughly in accordance with previous estimates [4]. The drift times we obtain under the three population model for the African population ($\tau_C + \tau_{Afr}$) are all approximately $0.483 + 0.009 = 0.492$ drift units. The geometric mean of the history of population sizes from the PSMC results in Prüfer et al. [4] give roughly that $N_e \approx 21,818$ since the African population size history started differing from that of Neanderthals, assuming a mutation rate of $1.25 * 10^{-8}$ per bp per generation. If we assume a generation time of 29 years, and plug in our drift time into the equation relating divergence time in generations to drift time ($t/(2N_e) \approx \tau$), this gives an approximate human-Neanderthal population divergence time of 622,598 years. This number agrees with the most recent estimates obtained via other methods [4]. Additionally, the Neanderthal-specific drift time is approximately 5.5 times as large as the modern human drift time, which is expected as Neanderthals had much smaller population sizes than modern humans [22, 4]. The admixture rate from archaic to modern humans that we estimate is 1.29%, which is roughly consistent with the rate estimate obtained via methods that do not jointly model contamination ($1.5-2.1\%$) [4]. Our method also allows us to obtain the most probable ancestry of the individual(s) who contaminated the sample, so long as the sample has high coverage. In the case of the Altai Neanderthal, we observe that this corresponds to one or more individuals with European ancestry.

The demographic models used in our approach are simple, involving no more than three populations and a single admixture event. This is partly due to limitations of known theory about the diffusion-based likelihood of an arbitrarily complex demography for the 2-D site-frequency spectrum - in the case of the two-population method - and to the inability of ∂a∂i [14] to handle more than 3 populations at a time. In recent years, several papers have made advances in the development of methods to compute the likelihood of an SFS for larger numbers of populations using coalescent theory [23, 24, 25], with multiple population size changes and admixture events. We hope to incorporate some of these techniques in future versions of our inference framework.

6. Acknowledgments

We thank Kelley Harris, Philip Johnson, Graham Coop, Nicolas Duforet-Frebourg, Sergi Castellano, Christoph Theunert, Janet Kelso, Rasmus Nielsen and members of the Slatkin and Nielsen labs for helpful advice and discussions.

7. References

[1] R. E. Green, J. Krause, A. W. Briggs, T. Maricic, U. Stenzel, M. Kircher, N. Patterson, H. Li, W. Zhai, M. H.-Y. Fritz, et al., A draft sequence of the neandertal genome, science 328 (2010) 710–722.

[2] D. Reich, R. E. Green, M. Kircher, J. Krause, N. Patterson, E. Y. Durand, B. Viola, A. W. Briggs, U. Stenzel, P. L. Johnson, et al., Genetic history of an archaic hominin group from denisova cave in siberia, Nature 468 (2010) 1053–1060.

[3] M. Meyer, M. Kircher, M.-T. Gansauge, H. Li, F. Racimo, S. Mallick, J. G. Schraiber, F. Jay, K. Prüfer, C. de Filippo, et al., A high-coverage genome sequence from an archaic denisovan individual, Science 338 (2012) 222–226.

[4] K. Prüfer, F. Racimo, N. Patterson, F. Jay, S. Sankararaman, S. Sawyer, A. Heinze, G. Renaud, P. H. Sudmant, C. de Filippo, et al., The complete genome sequence of a neanderthal from the altai mountains, Nature 505 (2014) 43–49.

[5] Q. Fu, H. Li, P. Moorjani, F. Jay, S. M. Slepchenko, A. A. Bondarev, P. L. Johnson, A. Aximu-Petri, K. Prüfer, C. de Filippo, et al., Genome sequence of a 45,000-year-old modern human from western siberia, Nature 514 (2014) 445–449.

[6] A. Seguin-Orlando, T. S. Korneliussen, M. Sikora, A.-S. Malaspinas, A. Manica, I. Moltke, A. Albrechtsen, A. Ko, A. Margaryan, V. Moiseyev, et al., Genomic structure in europeans dating back at least 36,200 years, Science 346 (2014) 1113–1118.

[7] R. E. Green, A.-S. Malaspinas, J. Krause, A. W. Briggs, P. L. Johnson, C. Uhler, M. Meyer, J. M. Good, T. Maricic, U. Stenzel, et al., A complete neandertal mitochondrial genome sequence determined by high-throughput sequencing, Cell 134 (2008) 416–426.

[8] R. H. Byrd, P. Lu, J. Nocedal, C. Zhu, A limited memory algorithm for bound constrained optimization, SIAM Journal on Scientific Computing 16 (1995) 1190–1208.

[9] P. Skoglund, B. H. Northoff, M. V. Shunkov, A. P. Derevianko, S. Pääbo, J. Krause, M. Jakobsson, Separating endogenous ancient dna from modern day contamination in a siberian neandertal, Proceedings of the National Academy of Sciences 111 (2014) 2229–2234.

[10] . G. P. Consortium, et al., An integrated map of genetic variation from 1,092 human genomes, Nature 491 (2012) 56–65.

[11] W. J. Ewens, Mathematical Population Genetics 1: I. Theoretical Introduction, volume 27, Springer Science & Business Media, 2004.

[12] H. Chen, R. E. Green, S. Pääbo, M. Slatkin, The joint allele-frequency spectrum in closely related species, Genetics 177 (2007) 387–398.

[13] G. Ewing, J. Hermisson, Msms: a coalescent simulation program including recombination, demographic structure and selection at a single locus, Bioinformatics 26 (2010) 2064–2065.

[14] R. N. Gutenkunst, R. D. Hernandez, S. H. Williamson, C. D. Bustamante, Inferring the joint demographic history of multiple populations from multidimensional snp frequency data, PLoS genetics 5 (2009) e1000695.

[15] G. O. Roberts, A. Gelman, W. R. Gilks, et al., Weak convergence and optimal scaling of random walk Metropolis algorithms, The Annals of Applied Probability 7 (1997) 110–120.

[16] R. R. Hudson, Generating samples under a wright–fisher neutral model of genetic variation, Bioinformatics 18 (2002) 337–338.

[17] K. R. Rosenbloom, J. Armstrong, G. P. Barber, J. Casper, H. Clawson, M. Diekhans, T. R. Dreszer, P. A. Fujita, L. Guruvadoo, M. Haeussler, et al., The ucsc genome browser database: 2015 update, Nucleic acids research 43 (2015) D670–D681.

[18] R. D. Hernandez, S. H. Williamson, C. D. Bustamante, Context dependence, ancestral misidentification, and spurious signatures of natural selection, Molecular biology and evolution 24 (2007) 1792–1800.

[19] M. Lipson, P.-R. Loh, A. Levin, D. Reich, N. Patterson, B. Berger, Efficient moment-based inference of admixture parameters and sources of gene flow, Molecular biology and evolution 30 (2013) 1788–1802.

[20] T. Higham, K. Douka, R. Wood, C. B. Ramsey, F. Brock, L. Basell, M. Camps, A. Arrizabalaga, J. Baena, C. Barroso-Ruíz, et al., The timing and spatiotemporal patterning of neanderthal disappearance, Nature 512 (2014) 306–309.

[21] H. Li, R. Durbin, Inference of human population history from individual whole-genome sequences, Nature 475 (2011) 493–496.

[22] S. Castellano, G. Parra, F. A. Sánchez-Quinto, F. Racimo, M. Kuhlwilm, M. Kircher, S. Sawyer, Q. Fu, A. Heinze, B. Nickel, et al., Patterns of coding variation in the complete exomes of three neandertals, Proceedings of the National Academy of Sciences 111 (2014) 6666–6671.

[23] H. Chen, The joint allele frequency spectrum of multiple populations: a coalescent theory approach, Theoretical population biology 81 (2012) 179–195.

[24] E. M. Jewett, N. A. Rosenberg, Theory and applications of a deterministic approximation to the coalescent model, Theoretical population biology 93 (2014) 14–29.

[25] J. A. Kamm, J. Terhorst, Y. S. Song, Efficient computation of the joint sample frequency spectra for multiple populations, arXiv preprint arXiv:1503.01133 (2015).

[26] M. Kimura, Solution of a process of random genetic drift with a continuous model, Proceedings of the National Academy of Sciences of the United States of America 41 (1955) 144.

[27] M. Abramowitz, I. A. Stegun, Handbook of mathematical functions, Dover New York, 1965.

[28] J. F. Crow, M. Kimura, An introduction to population genetics theory., An introduction to population genetics theory. (1970).

8. Figures

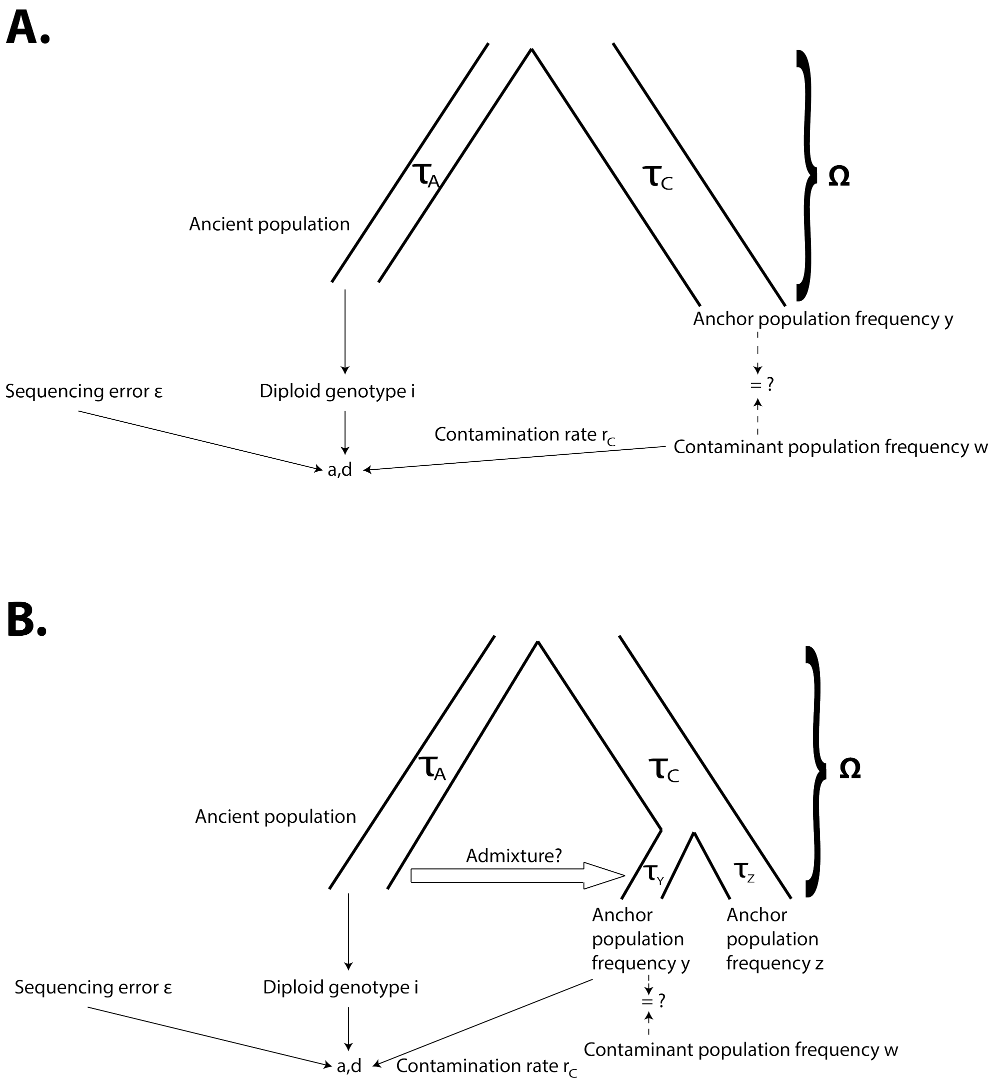

Figure 1. A) Schematic of two-population modeling framework: at each site, derived and ancestral reads (a, d) are binomially sampled from the true genotype of the archaic individual, with some amount of contamination and error. In turn, the true genotype depends on a demographic model, which can include the contaminant population. B) Schematic of three-population modeling framework, incorporating admixture between the archaic population and one of two anchor populations.

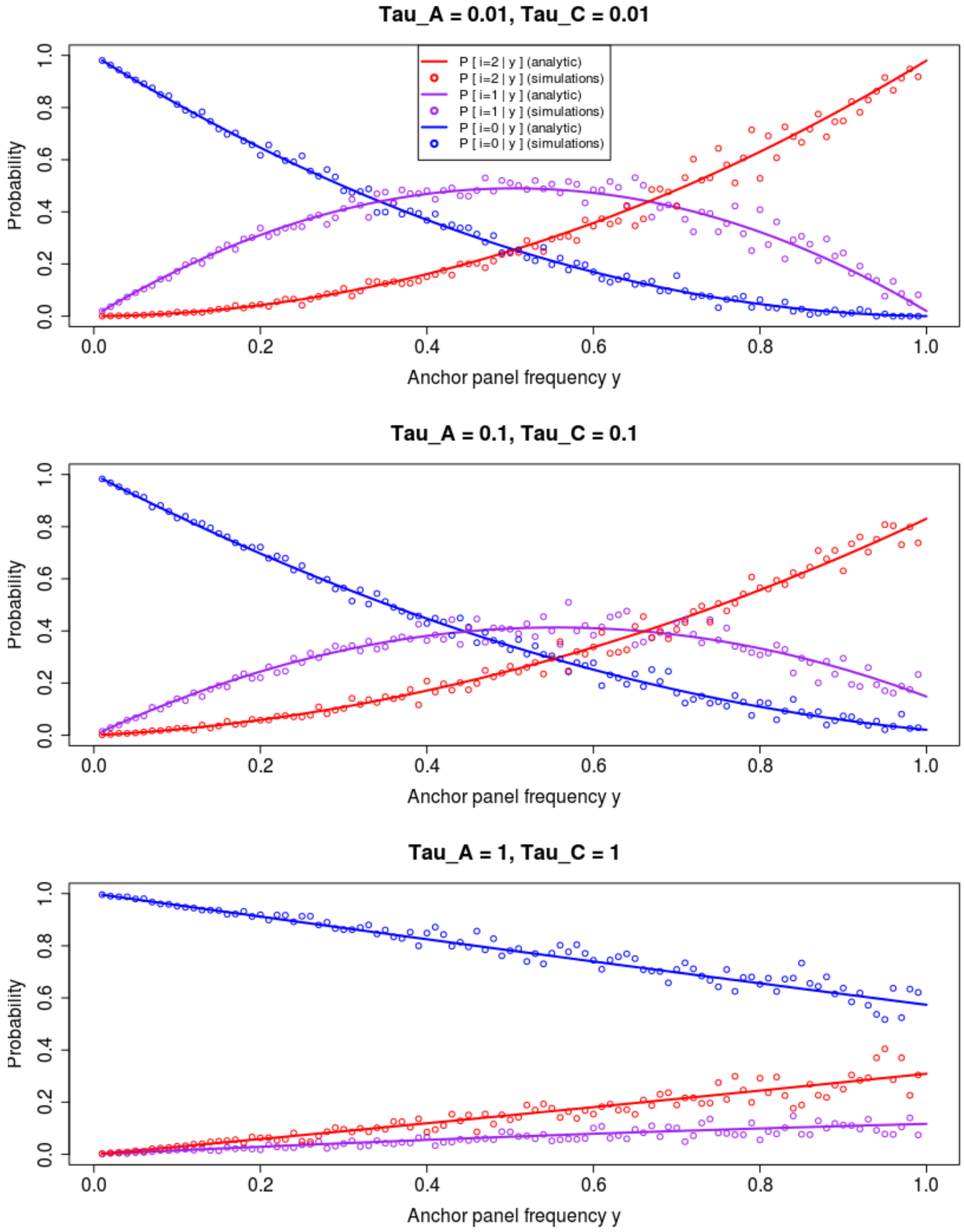

Figure 2. Comparison of analytic solutions to $P[i|y, \tau_C, \tau_A]$ and simulations under neutrality from msms, for different choices of τ_A and τ_C.

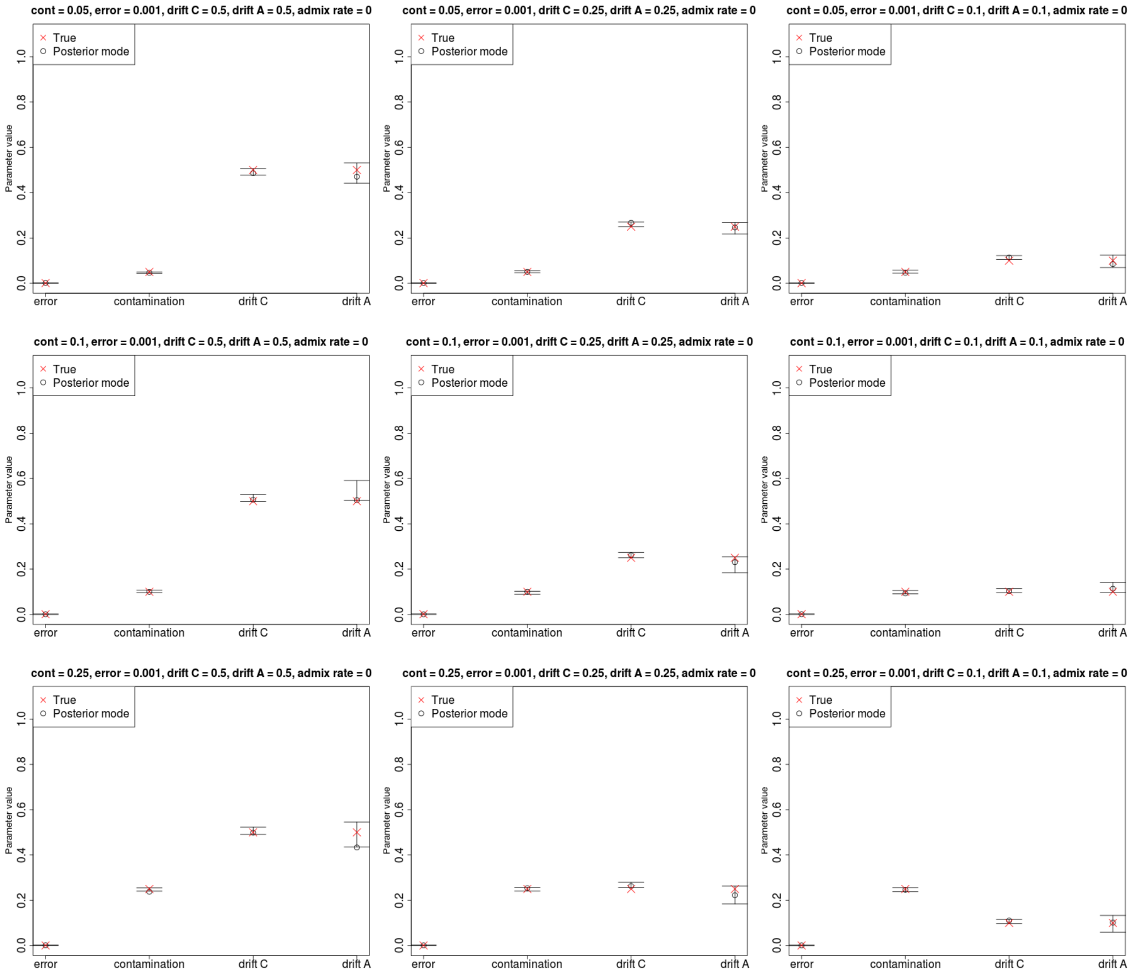

Figure 3. Estimation of parameters for a low-coverage ancient DNA genome (3X) with low sequencing error (0.1%), no admixture and a large anchor population panel (100 haploid genomes).

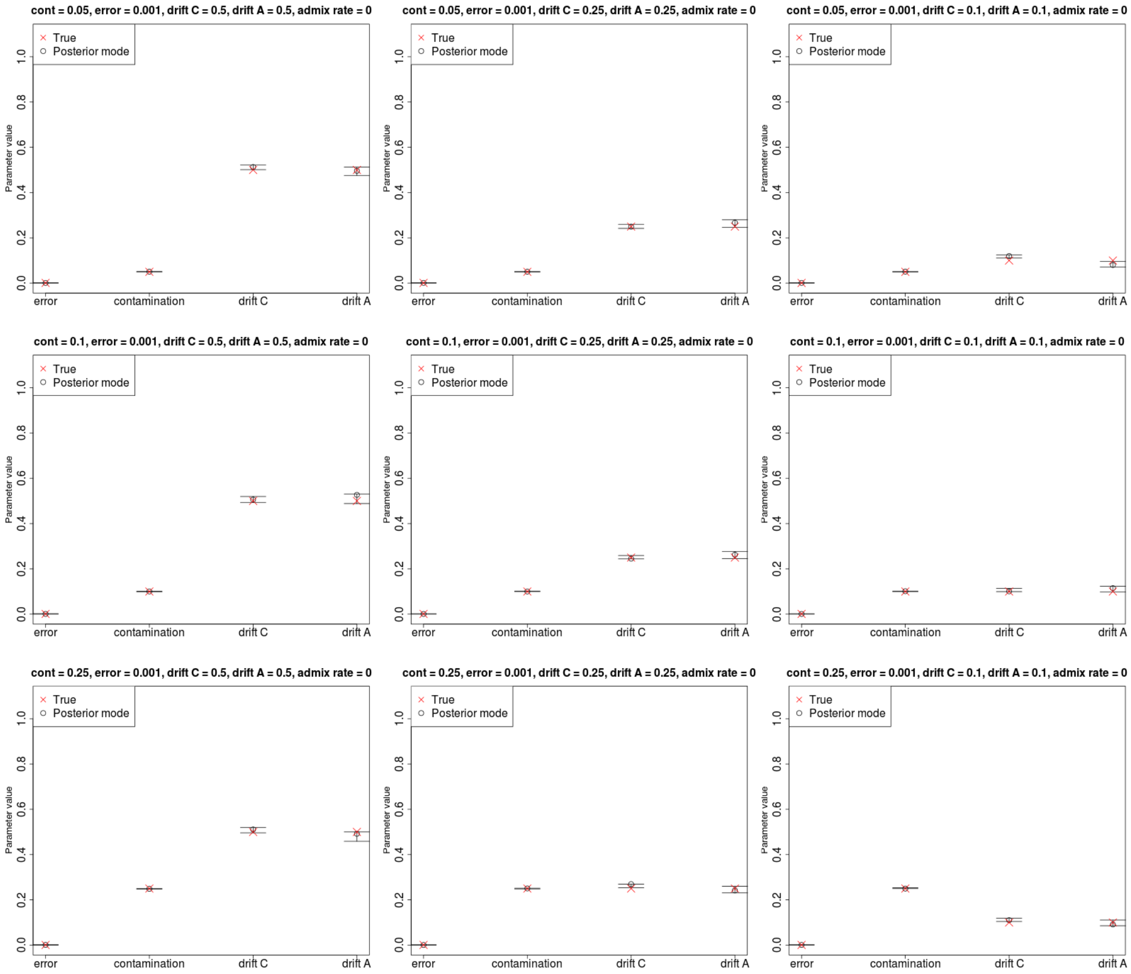

Figure 4. Estimation of parameters for a high-coverage ancient DNA genome (30X) with low sequencing error (0.1%), no admixture and a large anchor population panel (100 haploid genomes).

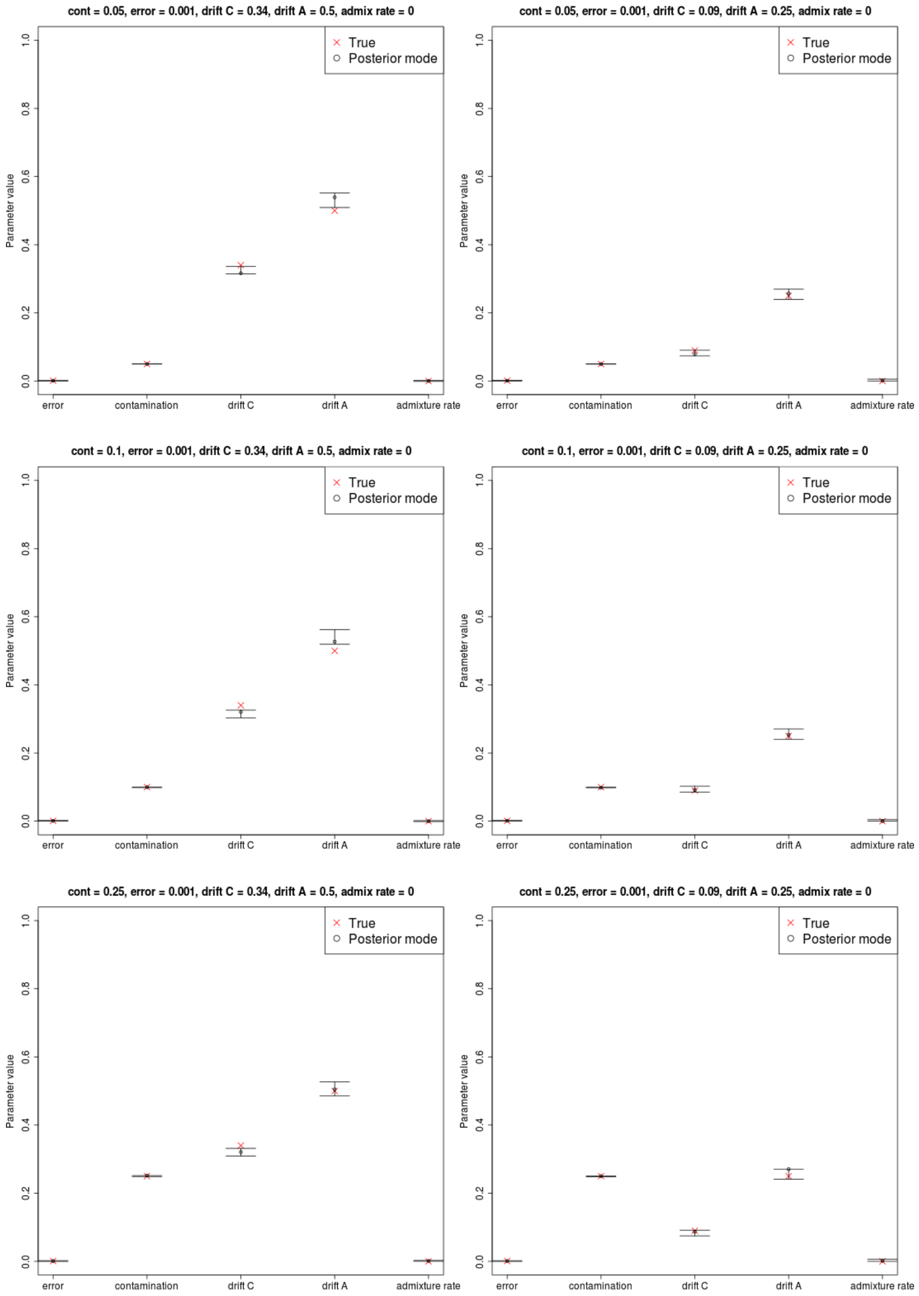

Figure 5. Estimation of error, contamination and demographic parameters in various three-population demographic scenarios, where the admixture rate is 0%. The prior used for the admixture time was uniform over $[0.06, 0.1]$.

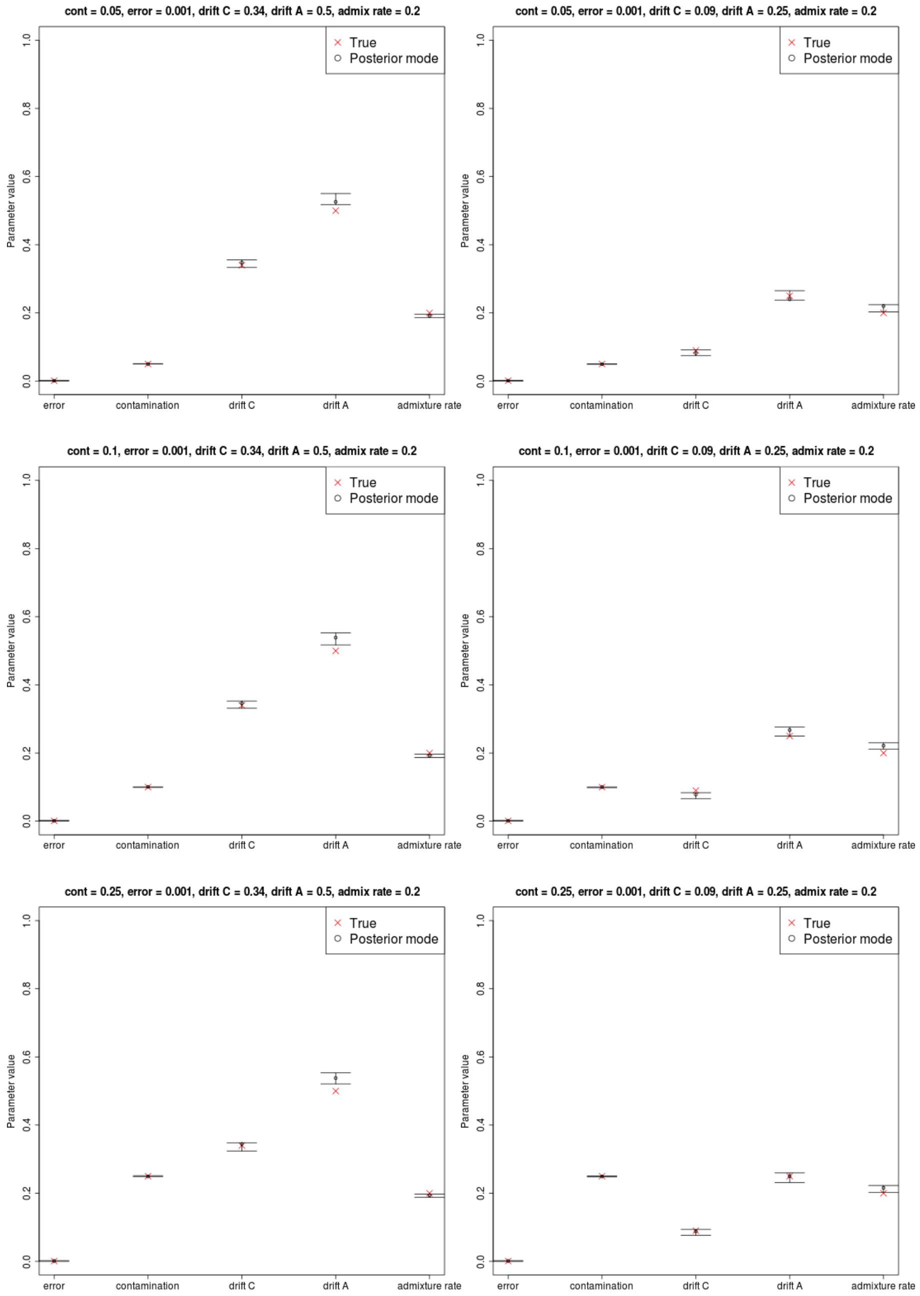

Figure 6. Estimation of error, contamination and demographic parameters in various three-population demographic scenarios, where the admixture rate is 20% and the admixture time was recent (0.005 drift units ago). The prior used for the admixture time was uniform over $[0, 0.01]$.

9. Tables

Table 1. Posterior modes of parameter estimates under the two-population inference framework for the Altai Neanderthal autosomal genome. We used different 1000G populations as candidate contaminants. Africans were the anchor population in all cases, so the modern human drift is with respect to Africans. Values in parentheses are 95% posterior quantiles. The panel with the highest posterior probability for being the contaminant (EUR) is in bold font.

Contaminant panel	Anchor panel	Error rate	Contamination rate	Modern human drift	Neanderthal drift	Log-posterior
AFR	AFR	0.234% (0.232% − 0.234%)	0.75% (0.747% − 0.755%)	0.455 (0.453 − 0.456)	2.481 (2.471 − 2.488)	-4822974.862
AMR	AFR	0.134% (0.134% − 0.137%)	0.917% (0.911% − 0.919%)	0.455 (0.453 − 0.456)	2.48 (2.469 − 2.485)	-3553563.224
EAS	AFR	0.198% (0.196% − 0.199%)	0.724% (0.723% − 0.729%)	0.454 (0.452 − 0.456)	2.481 (2.47 − 2.488)	-3579030.145
EUR	**AFR**	**0.133%** **(0.132% − 0.134%)**	**0.915%** **(0.912% − 0.918%)**	**0.455** **(0.453 − 0.456)**	**2.479** **(2.469 − 2.491)**	**-3546071.79**
SAS	AFR	0.138% (0.137% − 0.14%)	0.898% (0.892% − 0.899%)	0.456 (0.452 − 0.456)	2.478 (2.473 − 2.488)	-3557872.703

Table 2. Posterior modes of parameter estimates under the two-population inference framework for the Mezmaiskaya Neanderthal autosomal genome. We used different 1000G populations as candidate contaminants. Africans were the anchor population in all cases, so the modern human drift is with respect to Africans. Values in parentheses are 95% posterior quantiles.

Contaminant panel	Anchor panel	Error rate	Contamination rate	Modern human drift	Neanderthal drift	Log-posterior
AFR	AFR	2.662% (2.625% − 2.697%)	0.001% (0.001% − 0.015%)	0.464 (0.459 − 0.465)	2.576 (2.525 − 2.659)	-790458.8989
AMR	AFR	2.629% (2.623% − 2.702%)	0.002% (0.002% − 0.048%)	0.459 (0.459 − 0.465)	2.579 (2.522 − 2.709)	-790459.797
EAS	AFR	2.666% (2.625% − 2.698%)	0.002% (0.001% − 0.027%)	0.463 (0.459 − 0.466)	2.604 (2.532 − 2.677)	-790461.3417
EUR	AFR	2.672% (2.614% − 2.692%)	0.016% (0.002% − 0.096%)	0.462 (0.459 − 0.466)	2.593 (2.528 − 2.692)	-790459.0857
SAS	AFR	2.657% (2.622% − 2.70%)	0.002% (0.001% − 0.024%)	0.46 (0.459 − 0.465)	2.594 (2.52 − 2.69)	-790461.2111

Table 3. Posterior modes of parameter estimates under the three-population inference framework for the Altai Neanderthal autosomal genome. We used different 1000G populations as candidate contaminants. In all cases, Africans were the unadmixed anchor population and Europeans were the admixed anchor population. The ancestral human drift refers to the drift in the modern human branch before the split of Europeans and Africans. The post-split European-specific and African-specific drifts were estimated separately without the archaic genome ($\tau_{Afr} = 0.009$, $\tau_{Eur} = 0.255$). The panel with the highest posterior probability for being the contaminant (EUR) is in bold font.

Contaminant panel	Unadmixed anchor panel	Admixed anchor panel	Error rate	Contamination rate	Ancestral human drift	Neanderthal drift	Admixture rate	Log-posterior
AFR	AFR	EUR	0.39% (0.388% – 0.393%)	0.626% (0.619% – 0.637%)	0.481 (0.476 – 0.484)	2.542 (2.53 – 2.562)	1.255% (1.158% – 1.275%)	-4813930.165
AMR	AFR	EUR	0.291% (0.286% – 0.292%)	0.814% (0.809% – 0.824%)	0.481 (0.479 – 0.486)	2.54 (2.532 – 2.562)	1.286% (1.264% – 1.334%)	-4780699.543
EAS	AFR	EUR	0.344% (0.342% – 0.347%)	0.649% (0.643% – 0.657%)	0.483 (0.479 – 0.486)	2.543 (2.532 – 2.56)	1.293% (1.241% – 1.331%)	-4801849.357
EUR	**AFR**	**EUR**	**0.283%** **(0.282% – 0.287%)**	**0.827%** **(0.815% – 0.83%)**	**0.483** **(0.479 – 0.486)**	**2.547** **(2.529 – 2.561)**	**1.29%** **(1.265% – 1.337%)**	**-4774875.983**
SAS	AFR	EUR	0.292% (0.288% – 0.294%)	0.802% (0.794% – 0.809%)	0.481 (0.477 – 0.485)	2.558 (2.533 – 2.568)	1.264% (1.241% – 1.328%)	-4782524.26

10. Supplementary Materials

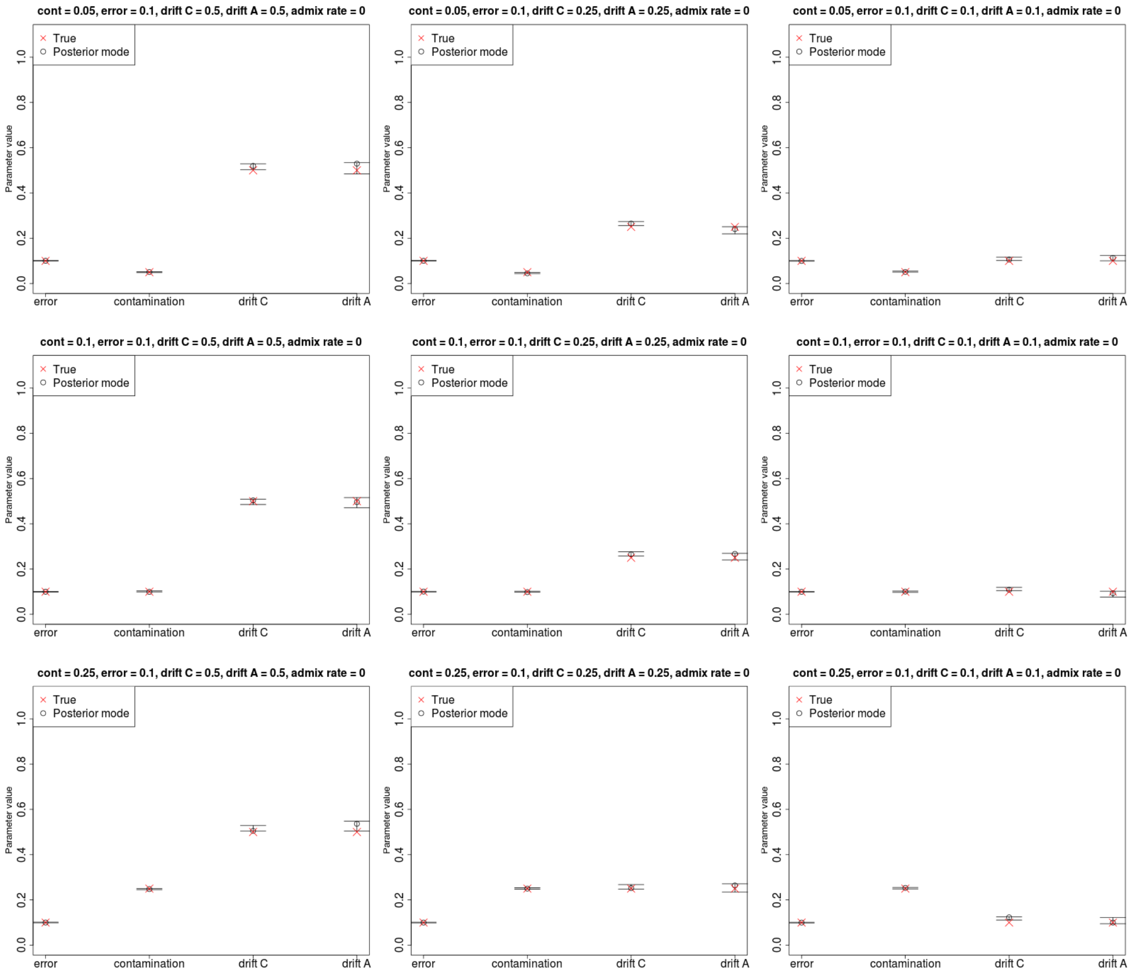

Figure S1. Estimation of parameters for a high-coverage ancient DNA genome (30X) with high sequencing error (10%), no admixture and a large anchor population panel (100 haploid genomes).

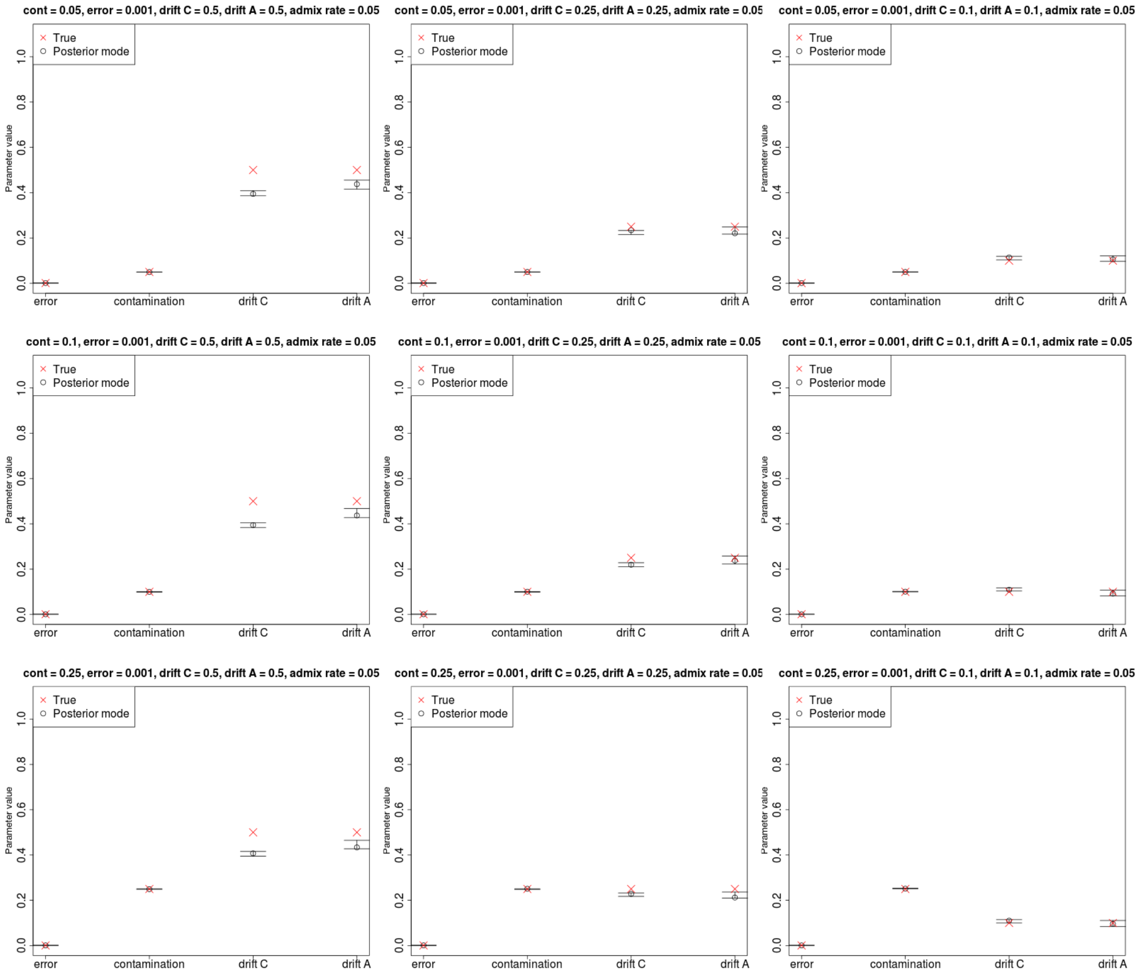

Figure S2. Estimation of parameters for a high-coverage ancient DNA genome (30X) with low sequencing error (0.1%), a large anchor population panel (100 haploid genomes) and admixture in the anchor population from the archaic population (5%), using the two-population inference framework, which does not model admixture.

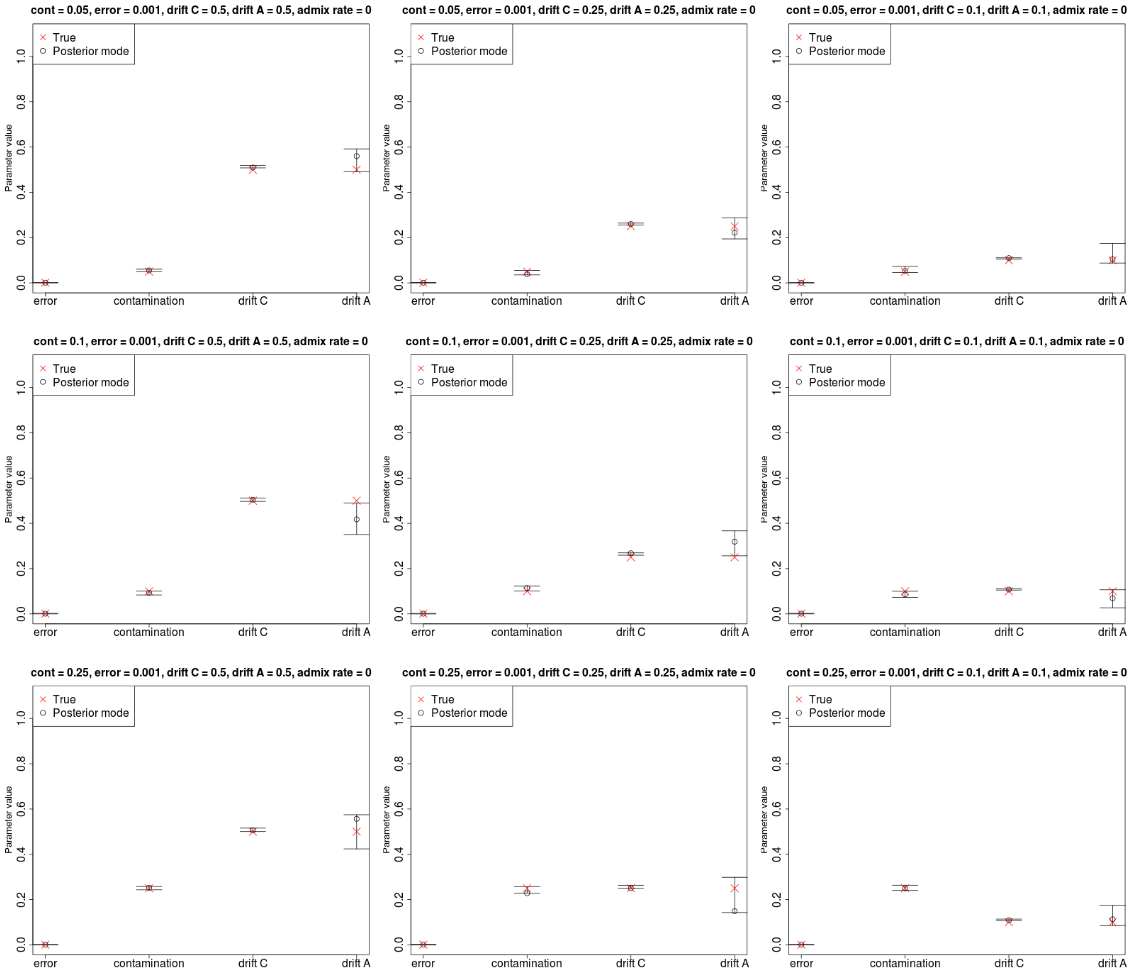

Figure S3. Estimation of parameters for an ancient DNA genome of very low coverage (0.5X) with low sequencing error (0.1%) and a large anchor population panel (100 haploid genomes). Note that unlike the rest of the simulations, the number of SNPs used in this case was approximately 1.6 million instead of 80,000, and the MCMC chain was run for 1 million steps instead of 100,000. Using a lower number of SNPs or running the chain for a shorter time resulted in inaccurate inferences.

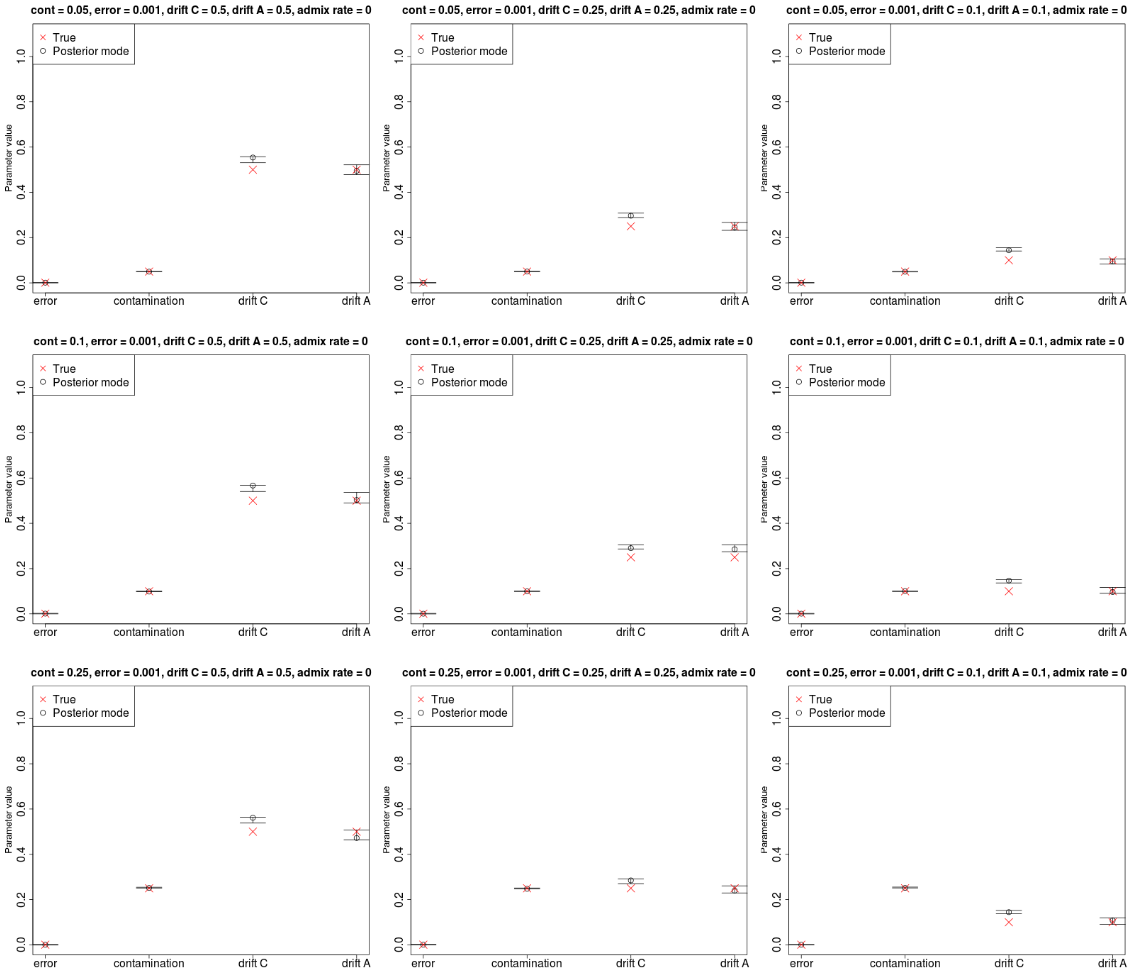

Figure S4. Estimation of parameters for a high-coverage ancient DNA genome (30X) with low sequencing error (0.1%), no admixture and a small anchor population panel (20 haploid genomes).

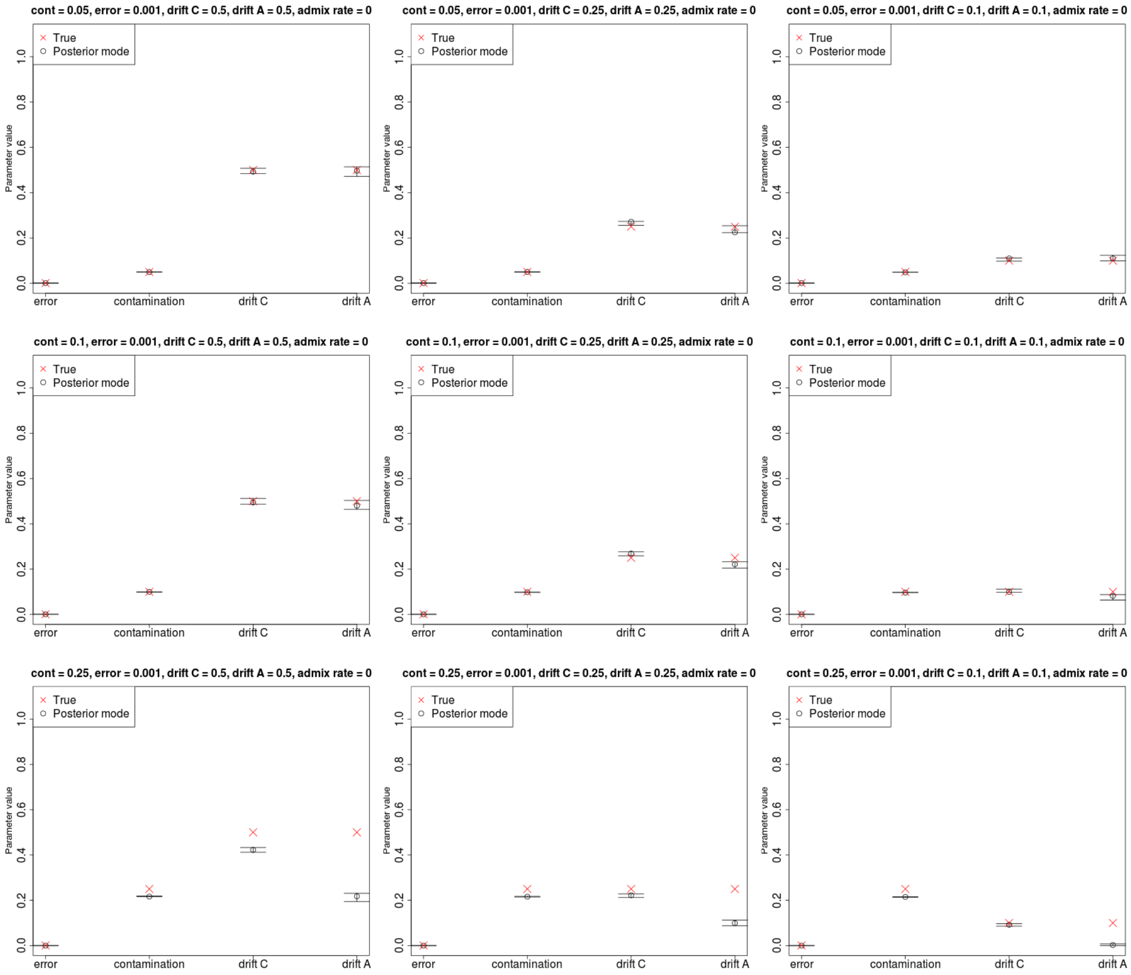

Figure S5. Estimation of parameters for a high-coverage ancient DNA genome (30X), when the contaminant reads are exclusively drawn from a single diploid individual from the contaminant panel.

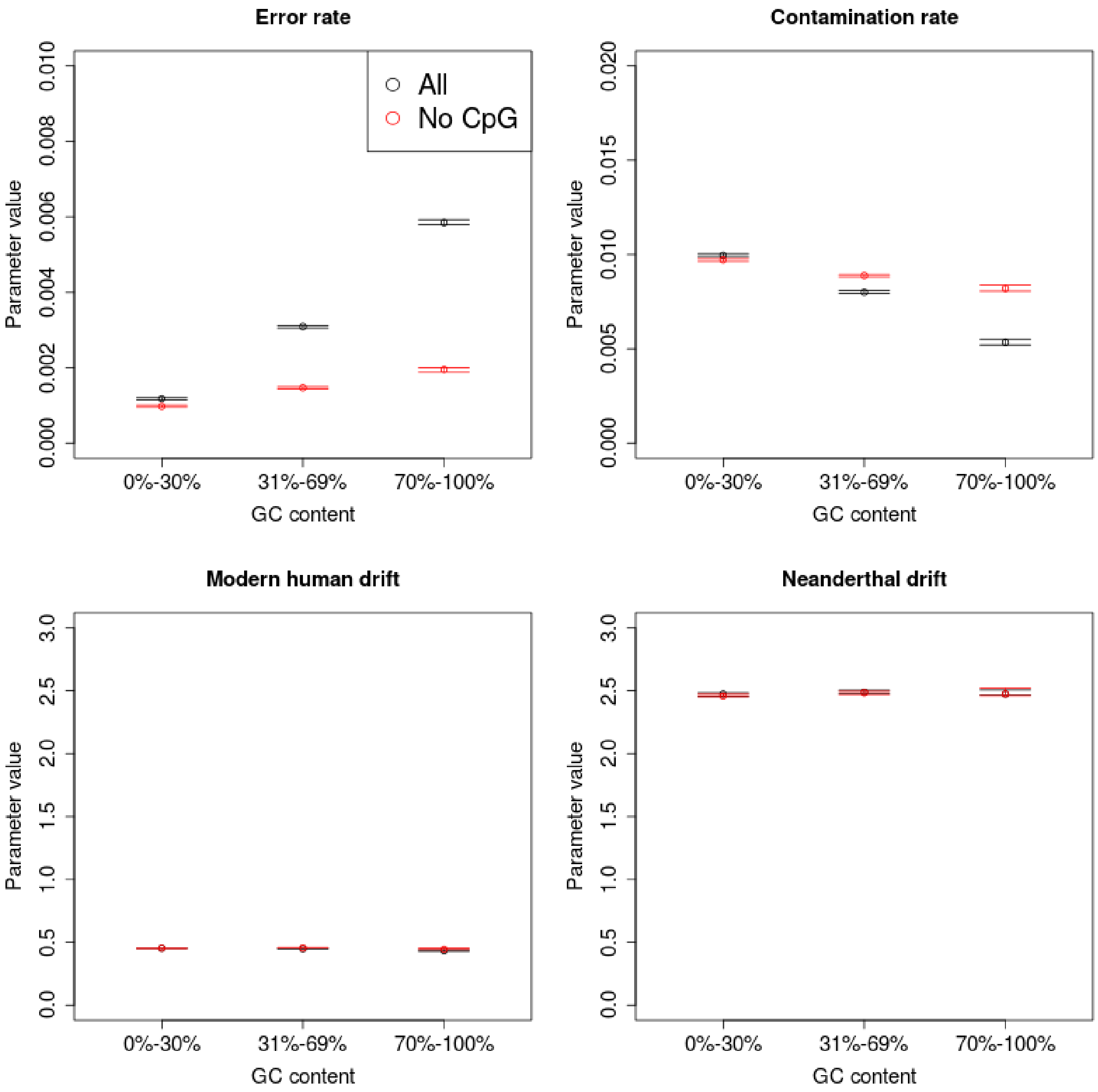

Figure S6. Estimation of parameters for the Altai Neanderthal genome across different GC levels using the two-population model, while keeping (black) or removing (red) CpG sites from the input dataset.

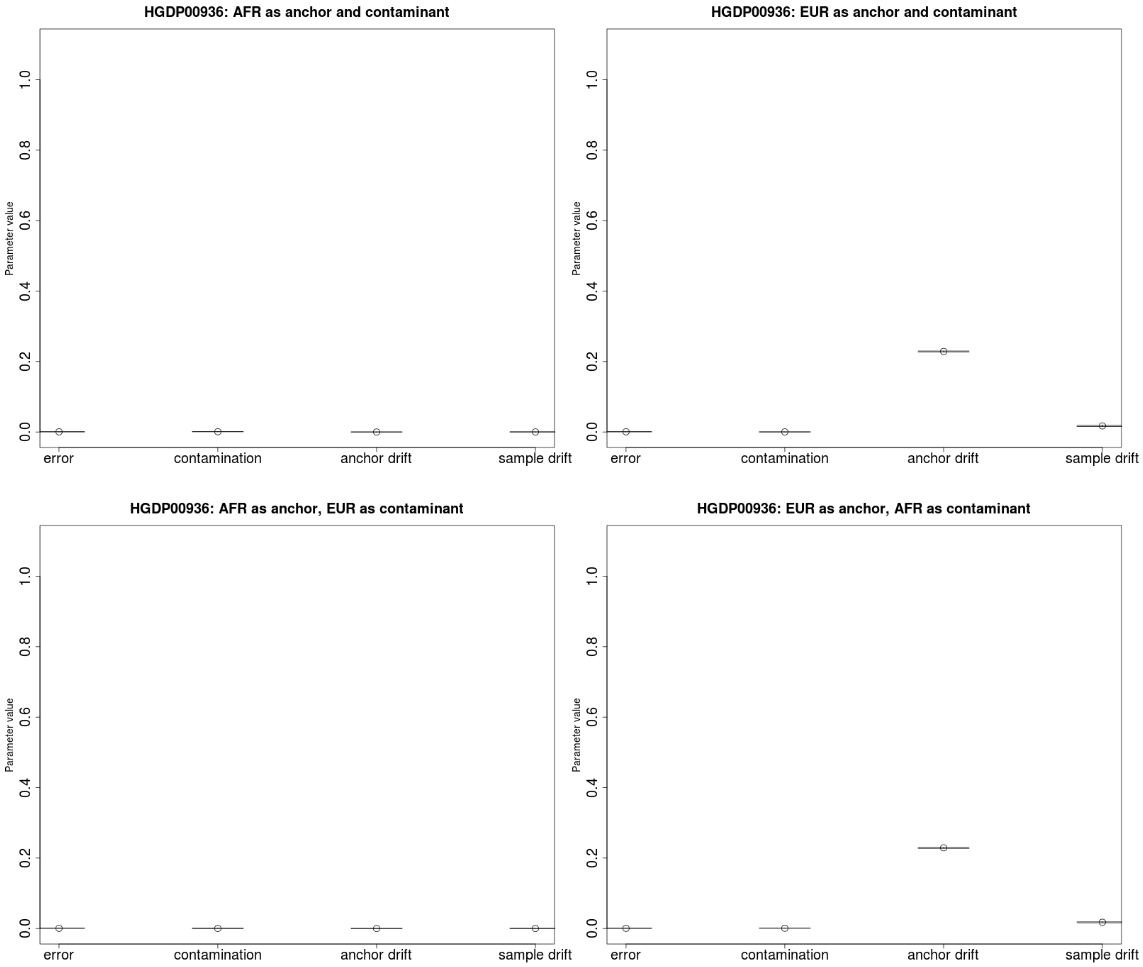

Figure S7. We tested one of the Yoruba genomes from Prüfer et al. [4] and obtain an estimate of 0% contamination, regardless of whether we use Europeans or Africans as the candidate contaminant. The anchor drift time is close to 0 when using Africans as the anchor population, as the sample belongs to that same population, while it is non-zero (= 0.22) when using Europeans.

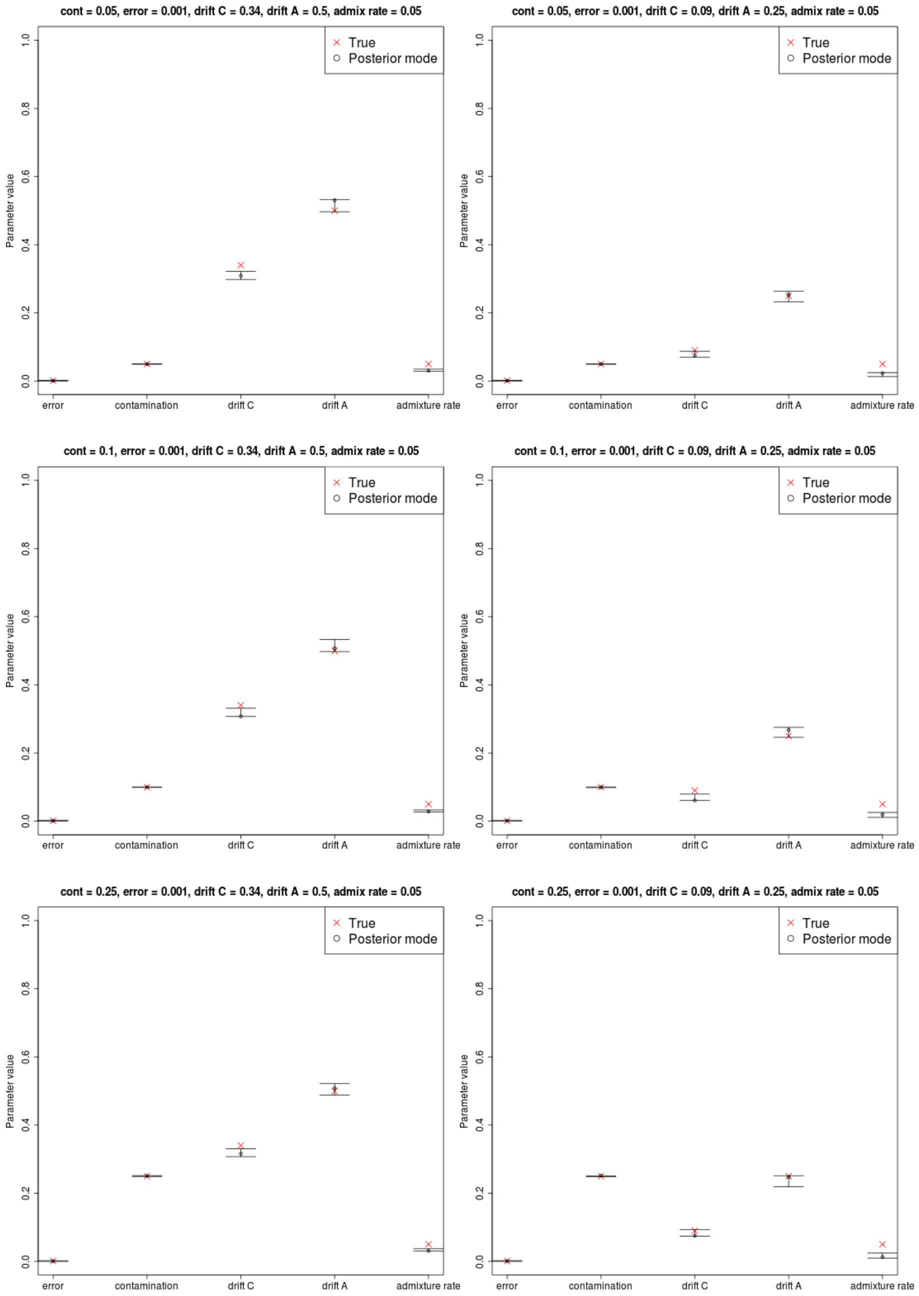

Figure S8. Estimation of error, contamination and demographic parameters in various three-population demographic scenarios, where the admixture rate is 5% and the admixture time is ancient (0.08 drift units ago). The prior used for the admixture time was uniform over $[0.06, 0.1]$.

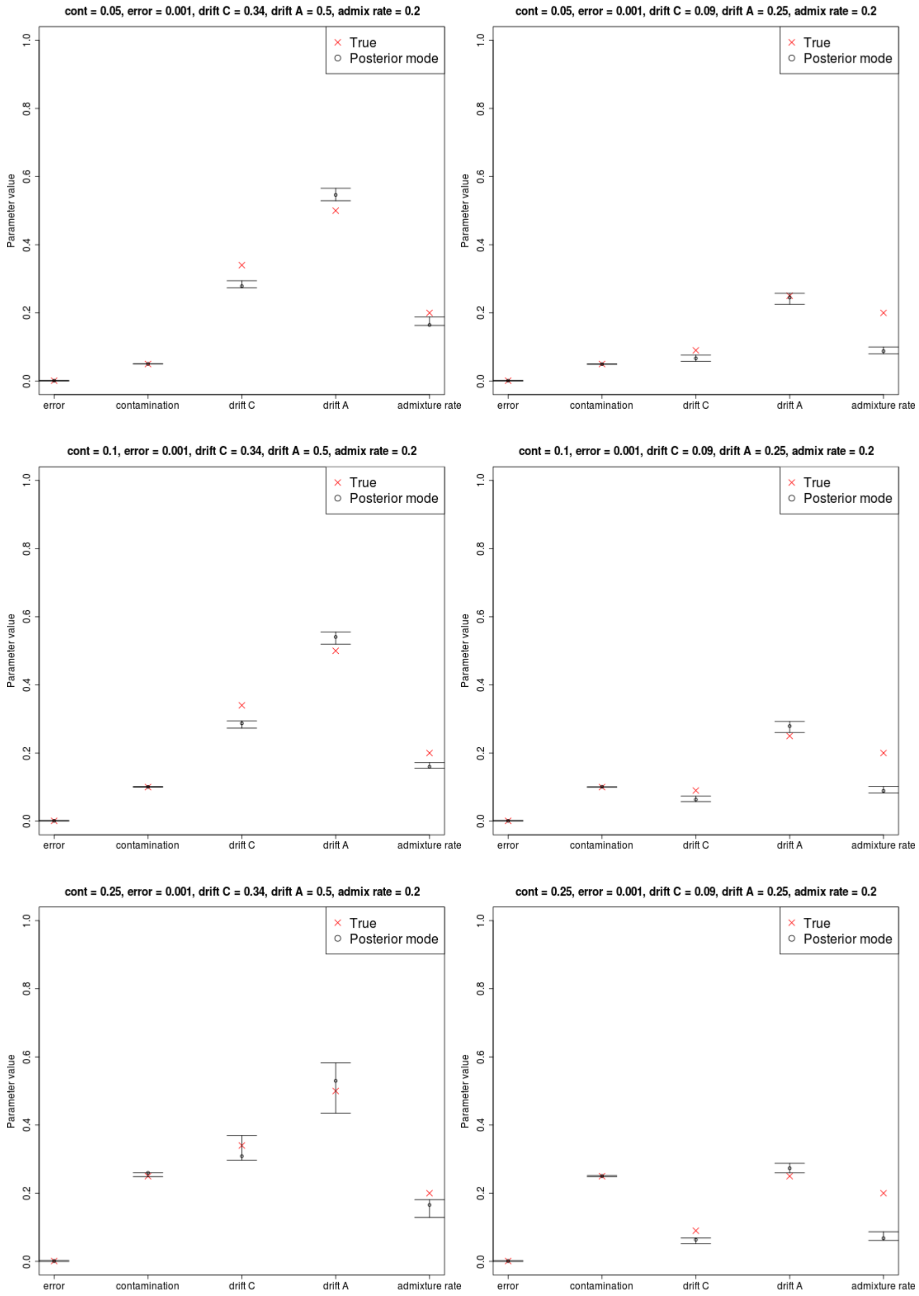

Figure S9. Estimation of error, contamination and demographic parameters in various three-population demographic scenarios, where the admixture rate is 20% and the admixture time is ancient (0.08 drift units ago). The prior used for the admixture time was uniform over $[0.06, 0.1]$.

Appendix A. Genotype probabilities conditional on a demography

Below we derive formulas 7, 8 and 9. Recall that we are interested in calculating the conditional probabilities $P[i|\mathbf{\Omega}, \mathbf{O}] = \mathbf{P}[\mathbf{i}|\mathbf{y}, \tau_{\mathbf{C}}, \tau_{\mathbf{A}}]$ for all three possibilities for the genotype in the ancient individual: $i = 0,\ 1$ or 2. These can be obtained from the definition of conditional probability. Let f_y^{DD} be the joint probability that a site has frequency y $(0 < y < 1)$ in the contaminant panel and is homozygous for the derived allele in the ancient individual. Let f_y^{DA} be the joint probability that a site has frequency y in the contaminant panel and is heterozygous in the ancient individual. Finally, let f_y^{AA} be the joint probability that a site has frequency y in the anchor panel and is homozygous for the ancient allele in the ancient individual. Then:

$$P[\ i = 0 \mid y, \tau_C, \tau_A\ ] = \frac{f_y^{AA}}{f_y} = \frac{f_y^{AA}}{f_y^{AA} + f_y^{DA} + f_y^{DD}} \tag{A.1}$$

$$P[\ i = 1 \mid y, \tau_C, \tau_A\ ] = \frac{f_y^{DA}}{f_y} = \frac{f_y^{DA}}{f_y^{AA} + f_y^{DA} + f_y^{DD}} \tag{A.2}$$

$$P[\ i = 2 \mid y, \tau_C, \tau_A\ ] = \frac{f_y^{DD}}{f_y} = \frac{f_y^{DD}}{f_y^{AA} + f_y^{DA} + f_y^{DD}} \tag{A.3}$$

In the above expressions, the functions f depend on τ_C and τ_A, but we omit this conditioning for ease of notation. As can be seen, all we need to find is the joint probabilities f_y^{AA}, f_y^{DA} and f_y^{DD}. Here is where diffusion theory comes into play. Let $\phi(\bullet, \tau|x, 0)$ be the Kimura solution to the neutral forward diffusion equation in the absence of mutation [26], given a frequency x at time 0 and an elapsed drift time τ:

$$\phi(y, \tau|x, 0) = 4x(1-x)\sum_{h=1}^{\infty}\frac{2j+1}{j(j+1)}C_{h-1}^{3/2}(1-2x)C_{h-1}^{3/2}(1-2y)e^{-j(j+1)\tau/2} \tag{A.4}$$

Here, x is the unknown population frequency of the derived allele in the ancestral population and $C_{h-1}^{(3/2)}(\bullet)$ is the Gegenbauer polynomial of order h-1 [27].

Assuming the ancestral population follows an equilibrium frequency distribution $g(x) = \theta/x$, we can write f_y^{DD} as follows:

$$f_y^{DD} = \int_0^1 \phi(y, \tau_C|x, 0)g(x)\left(\int_0^1 z^2\phi(z, \tau_A|x, 0)dz\right)dx \tag{A.5}$$

where z is the unknown population frequency of a derived allele in the population to which the ancient individual belongs.

The expression in parentheses is the second moment of the transition density and its solution is known [28]:

$$\int_0^1 z^2\phi(z, \tau_A|x, 0)dz = x - x(1-x)e^{-\tau_A} \tag{A.6}$$

This results in:

$$f_y^{DD} = \theta\int_0^1 \phi(y, \tau_C|x, 0)[1-(1-x)e^{-\tau_A}]dx \tag{A.7}$$

$$f_y^{DD} = \theta\left[\int_0^1 \phi(y, \tau_C|x, 0)dx - e^{-\tau_A}\int_0^1 \phi(y, \tau_C|x, 0)dx + e^{-\tau_A}\int_0^1 x\ \phi(y, \tau_C|x, 0)dx\right] \tag{A.8}$$

The integral of the first two terms of the sum was solved in Chen et al. [12]:

$$\int_0^1 \phi(y, \tau_C|x, 0)dx = e^{-\tau_C} \tag{A.9}$$

The third term of the sum can be solved by noting that, though the integrand is an infinite sum (i.e. formula A.4 multiplied by x), only the integrals of the first two terms of that infinite sum are not equal to 0. This can be seen by integrating the parts of the terms of that infinite sum that depend on x:

$$\int_0^1 x^2(1-x)C_{h-1}^{(3/2)}(1-2x)dx = \begin{cases} 1/12 & h=1 \\ -1/20 & h=2 \\ 0 & h \geq 3 \end{cases}$$

Therefore, after integrating the first two terms of the infinite sum, we obtain:

$$\int_0^1 x\phi(y, \tau_C|x, 0)dx = \frac{1}{2}e^{-\tau_C} + \left(y - \frac{1}{2}\right)e^{-3\tau_C} \tag{A.10}$$

So we finally arrive at:

$$f_y^{DD} = \theta\left[e^{-\tau_C} - \frac{1}{2}e^{-\tau_A-\tau_C} + \left(y-\frac{1}{2}\right)e^{-\tau_A-3\tau_C}\right] \tag{A.11}$$

We can obtain f_y^{DA} in a similar fashion:

$$f_y^{DA} = \int_0^1 \phi(y,\tau_C|x,0)g(x)\left(\int_0^1 2z(1-z)\phi(z,\tau_A|x,0)dz\right)dx \tag{A.12}$$

Solving the term in the parentheses:

$$\int_0^1 2z(1-z)\phi(z,\tau_A|x,0)dz = 2\left(\int_0^1 z\phi(z,\tau_A|x,0)dz - \int_0^1 z^2\phi(z,\tau_A|x,0)dz\right) \tag{A.13}$$

The first term of the difference is the first moment of the transition density, which is equal to x [28], while the second term is the second moment (formula A.6). Therefore:

$$f_y^{DA} = 2\theta e^{-\tau_A}\left[\int_0^1 \phi(y,\tau_C|x,0)(1-x)dx\right] \tag{A.14}$$

$$f_y^{DA} = 2\theta e^{-\tau_A}\left[\int_0^1 \phi(y,\tau_C|x,0)dx - \int_0^1 x\ \phi(y,\tau_C|x,0)\ dx\right] \tag{A.15}$$

And after using formulas A.9 and A.10, we obtain:

$$f_y^{DA} = \theta\left[e^{-\tau_A-\tau_C} + (1-2y)\,e^{-\tau_A-3\tau_C}\right] \tag{A.16}$$

To obtain f_y^{AA}, we know that, assuming the anchor population to be at equilibrium:

$$f_y = g(y) \tag{A.17}$$

And therefore:

$$f_y^{AA} + f_y^{DA} + f_y^{DD} = \frac{\theta}{y} \tag{A.18}$$

So we finally obtain:

$$f_y^{AA} = \theta \left[\frac{1}{y} - e^{-\tau_C} - \frac{1}{2} e^{-\tau_A - \tau_C} + \left(y - \frac{1}{2} \right) e^{-\tau_A - 3\tau_C} \right] \tag{A.19}$$

We now have all the elements necessary to obtain the conditional probabilities from formulas A.1, A.2 and A.3, which immediately lead us to formulas 7, 8 and 9.