PLOS | GENETICS

RESEARCH ARTICLE

Joint Estimation of Contamination, Error and Demography for Nuclear DNA from Ancient Humans

Fernando Racimo[1☯*], **Gabriel Renaud**[2☯], **Montgomery Slatkin**[1]

1 Department of Integrative Biology, University of California, Berkeley, Berkeley, California, United States of America, **2** Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany

☯ These authors contributed equally to this work.
* fernandoracimo@gmail.com

OPEN ACCESS

Citation: Racimo F, Renaud G, Slatkin M (2016) Joint Estimation of Contamination, Error and Demography for Nuclear DNA from Ancient Humans. PLoS Genet 12(4): e1005972. doi:10.1371/journal.pgen.1005972

Editor: Joshua M. Akey, University of Washington, UNITED STATES

Received: August 19, 2015

Accepted: March 11, 2016

Published: April 6, 2016

Data Availability Statement: All the source code for the method is available in GitHub: https://github.com/grenaud/dice.

Funding: This work was supported by the US National Institutes of Health grant to MS (R01-GM40282). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

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 of the contaminating individual(s). 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 contaminant population (e.g. European, East Asian or African). The method is implemented in a C++ program called 'Demographic Inference with Contamination and Error' (DICE). We applied it to simulations and genome data from ancient Neanderthals and modern humans. With reasonable levels of genome sequence coverage (>3X), we find we can recover accurate estimates of all these parameters, even when the contamination rate is as high as 50%.

Author Summary

When extracting and sequencing ancient DNA from human remains, a recurrent problem is the presence of DNA from the paleontologists, archaeologists or geneticists that may have handled the fossil. If a DNA library is highly contaminated, this will introduce biases in downstream analyses, so it is important to determine the amount of extraneous DNA. Different methods exist for this purpose, but few are applicable to the nuclear genome, and none of them can extract reliable genomic information from highly contaminated samples. Thus, samples with high rates of contamination are usually discarded. Here, we present a method to jointly estimate contamination and error rates, along with demographic parameters, like drift times and admixture rates. Our method can serve to uncover

important details about the evolutionary history of archaic and early modern humans from ancient DNA samples, even if those samples are highly contaminated.

Introduction

When sequencing a human genome using ancient DNA (aDNA) recovered from fossils, a common practice is to assess the amount of present-day human contamination in a sequencing library [1, 2, 3, 4, 5, 6]. 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 present-day humans. Then, one counts how many aDNA fragments support the present-day human base at those positions. This is the most popular technique and has been routinely deployed in the sequencing of Neanderthal genomes [7, 1]. However, contamination levels of the mitochondrial genome may sometimes differ drastically from those of the nuclear genome [8, 9].

A second technique involves assessing whether the sample was male or female using the number of fragments that map to the X and the Y chromosomes. 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) [8, 1, 10, 4]. Another method uses a maximum-likelihood approach to estimate contamination, but is only applicable to single-copy chromosomes, like the X chromosome in individuals known *a priori* to be male [11, 12]. Finally, one last technique involves using a maximum-likelihood approach to co-estimate the amount of contamination, sequencing error and heterozygosity in the entire autosomal nuclear genome [1, 3], using an optimization algorithm such as L-BFGS-B [13].

Afterwards, if the aDNA library shows low levels of present-day human contamination ($< \sim 2\%$), 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: contaminating fragments may affect downstream analyses, while discarding the library as a whole may waste precious genomic data that could provide important demographic insights.

One way to address this problem was proposed by skoglund et al. [14], who developed a statistical framework to separate contaminant from endogenous DNA fragments by using the patterns of chemical deamination characteristic of ancient DNA. The method produces a score which reflects the odds that a particular fragment is endogenous or not, based on these chemical patterns. This approach is effective at isolating truly endogenous fragments from contaminant fragments, but at the cost of potentially discarding some fragments that may not have chemical damage and still be endogenous. This becomes more problematic the younger the ancient DNA sample is, because younger samples will tend to have a higher proportion of non-deaminated ancient DNA, and so the method will lead users to discard a larger fraction of endogenous material.

Instead of (or in addition to) attempting to separate the two type of fragments before performing a demographic analysis, one could incorporate the uncertainty stemming from the contaminant fragments into a probabilistic inference framework. Such an approach has already been implemented in the analysis of a haploid mtDNA archaic genome [15]. 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 [16]. The method uses a Bayesian framework to obtain posterior estimates of all parameters of interest, including population-size-scaled divergence times and admixture rates.

Methods

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 fragments observed at site j
- $\mathbf{d}$: vector of d_j counts for all sites $j = \{1, \ldots, N\}$ in a genome
- a_j: number of ancestral fragments observed at site j
- $\mathbf{a}$: vector of a_j counts for all sites $j = \{1, \ldots, N\}$ 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 $j = \{1, \ldots, N\}$ 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 ($\mathbf{w}$) and a set of demographic parameters ($\boldsymbol{\Omega}$). 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 $\boldsymbol{\Omega}$ may simply be the population-scaled times separating the contaminant population and the sample from their common ancestral population. However, $\boldsymbol{\Omega}$ 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 $\boldsymbol{\Omega}$ 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:

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

where

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

Here, i is the true (unknown) genotype of the ancient sample, and $P[i \mid \Omega, \mathbf{O}]$ is the probability of genotype i given the demographic parameters and the data.

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 fragments at a particular site follows a binomial distribution that depends on the genotype, the error rate and the rate of contamination [1, 3]:

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

where

$$q_2 = r_C(w(1-\epsilon) + (1-w)\epsilon) + (1 - r_C)(1 - \epsilon) \quad (4)$$

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

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

In the sections below, we will turn to the more complicated part of the model, which is obtaining the probability $P[i|\Omega, \mathbf{O}]$ 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.

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 contamination from present-day humans, 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 (Fig 1A). 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, $\Omega = \{\tau_\mathbf{C}, \tau_\mathbf{A}\}$, 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

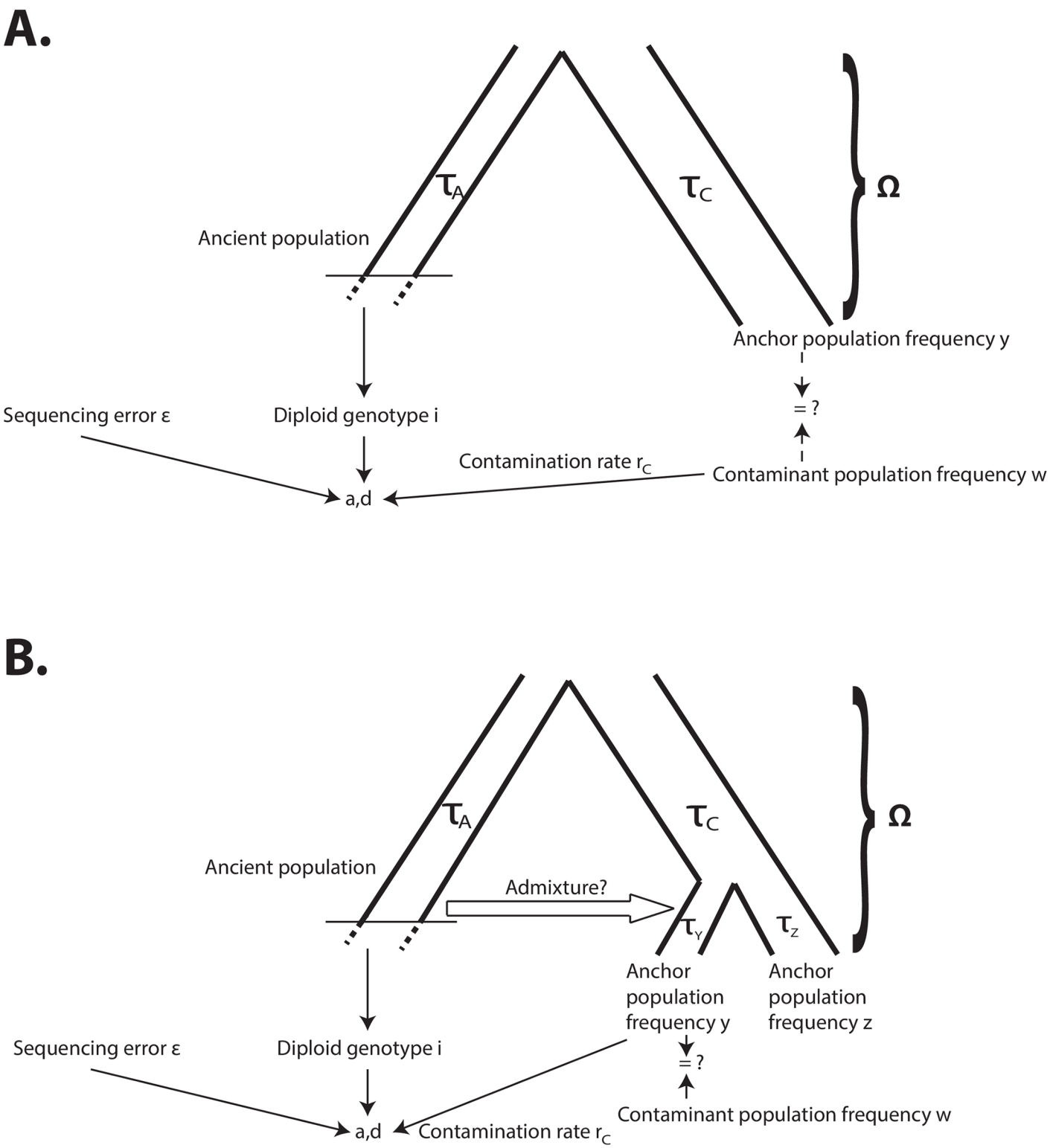

Fig 1. A) Schematic of two-population modeling framework: at each site, derived and ancestral fragments (a, d) are binomially sampled from the true genotype of the archaic individual, with some amount of contamination and error. In turn, the true genotype depends on a demographic model, which can include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

doi:10.1371/journal.pgen.1005972.g001

We need to calculate the conditional probabilities $P[i|\Omega, \mathbf{O}] = \mathbf{P}[\mathbf{i}|\mathbf{y}, \tau_\mathbf{C}, \tau_\mathbf{A}]$ 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 [17]), especially focusing on the two-population site-frequency spectrum (SFS) [18]. The full derivations can be found in the S1 Text, 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 [19] for different choices of τ_C and τ_A (with $\theta = 20$ in each simulation) to verify our analytic expressions were correct (Fig 2). The probability does not depend on θ, so the choice of this value is arbitrary.

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

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 from the ancient sample to the contaminant 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 **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 **w**, as before.
2. Two panels of 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 (Fig 1B). The sample frequencies for this population will be labeled as **y**. The other population—called Z—will have sample frequencies labeled **z**. We will assume the drift times separating these two populations are known (parameters τ_Y and τ_Z in Fig 1B). 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 [20].

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

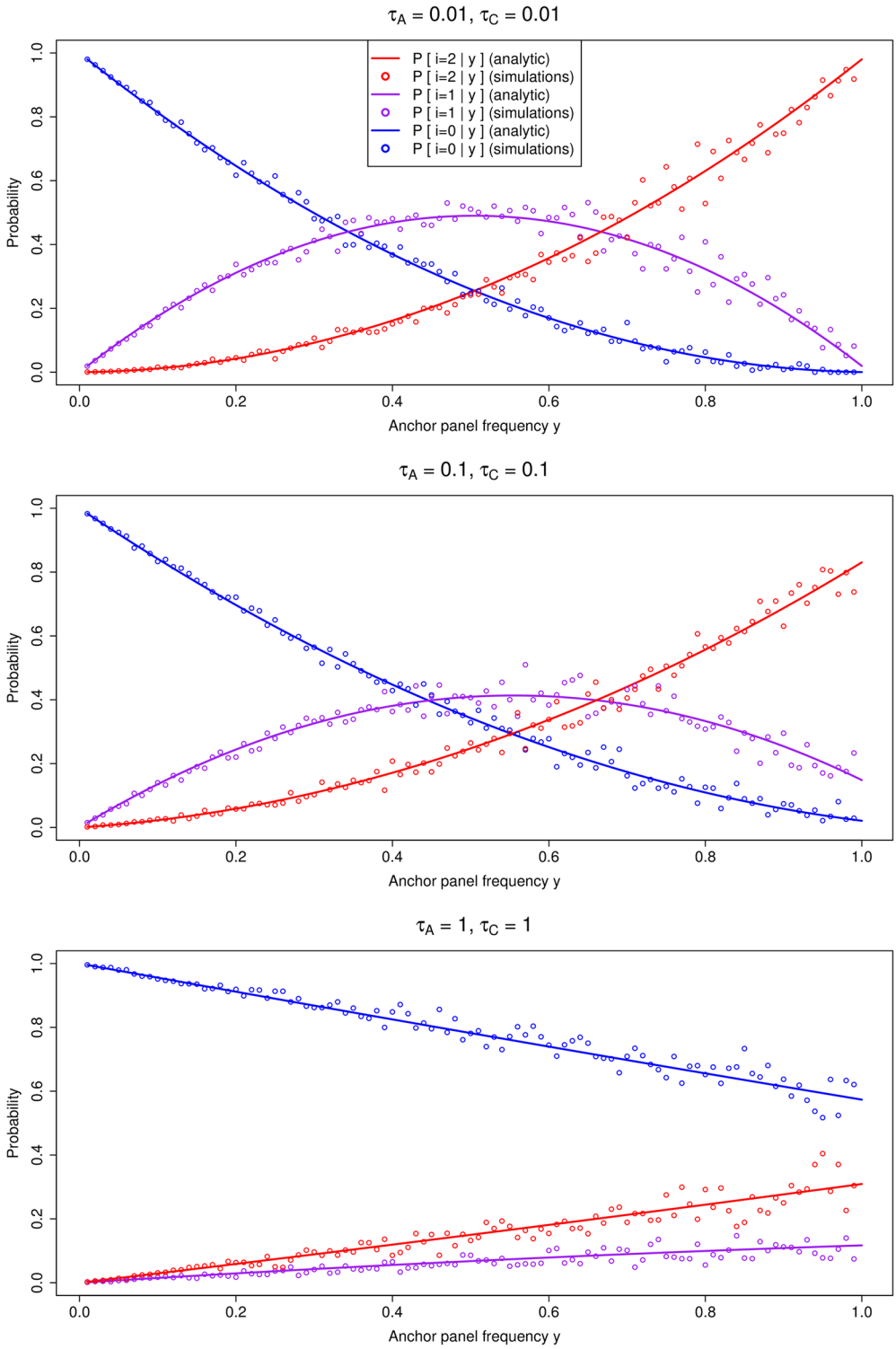

Fig 2. Comparison of analytic solutions to $P[i|y, \tau_C, \tau_A]$ and simulations under neutrality from msms, for different choices of τ_A and τ_C.

doi:10.1371/journal.pgen.1005972.g002

assumed uniform prior distributions for all parameters, and the boundaries of these distributions 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 [21]. By default, this constant is equal to 1,000 for all parameters. The new state is accepted with probability:

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

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

Multiple error rates and ancestral state misidentification

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 genome-wide 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 transition polymorphisms, which are more likely to have occurred due to cytosine deamination in ancient DNA. These sites are therefore likely to be subject to different error rates than those common in present-day sequencing data [22, 23]. Our program can then estimate two error rates separately: one for transitions and one for transversions. Finally, we incorporated an option to include an ancestral state misidentification (ASM) parameter, which should serve to correct for mispolarization of alleles [24].

BAM file functionality

The standard input for DICE is a file containing counts of particular ancestral/derived base combinations and SNP frequencies (see README file online). As an additional feature, we also developed a module for the user to directly input a BAM file and a file containing population allele frequencies for the anchor and contaminant panels, rather than the standard input. The user can either choose to convert the BAM file to native DICE format using a program provided with the software package and then run the program, or run it directly on the BAM file. In the latter case, instead of calculating genome-wide error parameters, the program will calculate error parameters specific to each sequenced fragment, based on mapping qualities, base qualities and estimated deamination rates at each site (see S2 Text).

Results

Two-population method

Simulations. We first used DICE 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* [25], making sure each simulation had at least one usable SNP. 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. In all simulations, the contaminant panel was the same as the anchor population panel. We then applied our method to the combined set of ∼80,000 SNPs.

In Figs 3 and 4, we 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%) (S1 Fig).

In Figs 5, S2, S3 and S4, we show how well the method does at estimating parameters over a wide range of contamination and drift scenarios, by displaying the absolute difference between simulated parameters and their corresponding posterior modes. So long as coverage is high (for example, 5X or 30X), the contamination and anchor drift parameters are accurately estimated even at 75% contamination. The method performs well even if the drift times on both sides of the tree are as small as ≈0.001 or as large as ≈5, but starts becoming inaccurate when contamination is extremely high. In general, the contamination rate and anchor drifts are easier to determine than the drift corresponding to the ancient population.

We find that for samples of very low coverage (0.5X, 1X, 1.5X) we require a larger number of sites to obtain accurate estimates (S5, S6 and S7 Figs). For example, for a sample of 0.5X coverage, we tried different numbers of independent replicate simulations and found that at 800,000 replicates, we obtained approximately 1.6 million valid SNPs for inference, which was enough to reach reasonable levels of accuracy (S14 Fig). 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 $\geq$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. Finally, we note that our failure to recover the true parameters under low coverage in a single MCMC run is partly due to the chain failing to converge. Indeed, when we run the MCMC 10 times and recover the estimates from the chain with the highest posterior probability, we are able to obtain increased accuracy relative to the single run, especially when the drift parameters are extremely low and when the contamination rate is extremely high (S8, S9 and S10 Figs).

Finally, we tested the method on simulations in a more realistic scenario, in which we generated ancient and contaminant fragments based on empirical fragment sizes and then mapped them to a simulated reference genome using BWA [26] with default parameters. We produced DNA sequences from the output of msms [19] via seq-gen v.1.3.3 [27] with the HKY substitution model [28]. This allows for multiple substitutions to occur at the same site since the split from chimpanzee (which could cause ASM). We then simulated ancient DNA fragments that had a fragment size distribution emulating empirical distributions. Contaminant fragments were also sampled from the contaminant population. We used the deamination rates from the single-stranded library from the Loschbour ancient individual [29] (∼8% at the 5' end and ∼34% at the 3' end with a residual deamination rate of ∼1% along the whole fragment) to artificially deaminate the ancient fragments. We simulated sequencing errors on both the ancient and contaminant fragments using empirical sequencing error rates from a PhiX library

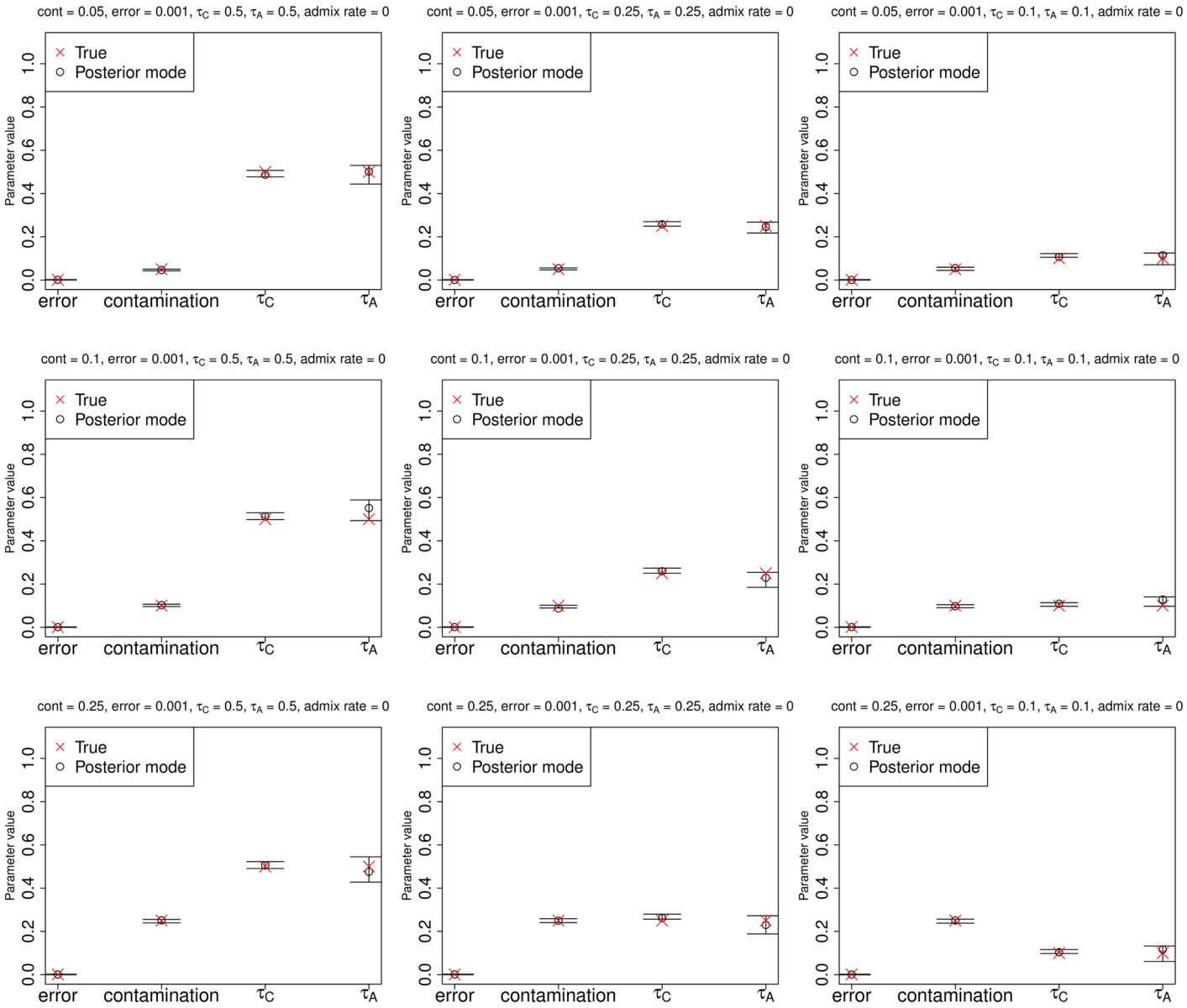

Fig 3. Estimation of parameters for a low-coverage ancient DNA genome (3X) with low sequencing error (0.1%), no admixture and a large anchor population panel (100 haploid genomes). Error bars represent 95% posterior intervals.

doi:10.1371/journal.pgen.1005972.g003

(Illumina Corp.) sequenced at the Max Planck Institute for Evolutionary Anthropology on an Illumina HiSeq, basecalled using freeIbis [30]. With the same empirical PhiX dataset distribution, we generated quality scores for each nucleotide. Fragments were mapped back to a random individual from the contaminant panel. Fig 6 shows DICE's performance on this scenario with different error models. In all cases, we find that the parameters are estimated with high accuracy. As expected, the ts/tv model infers a higher error rate at transitions, due to the additional errors introduced by deamination on the ends of the ancient fragments.

Performance under violations of model assumptions. We evaluated the consequences of different violations of model assumptions. We started by observing the effects of using a small modern human panel. S12 Fig shows results for cases in which the contaminant/anchor panel

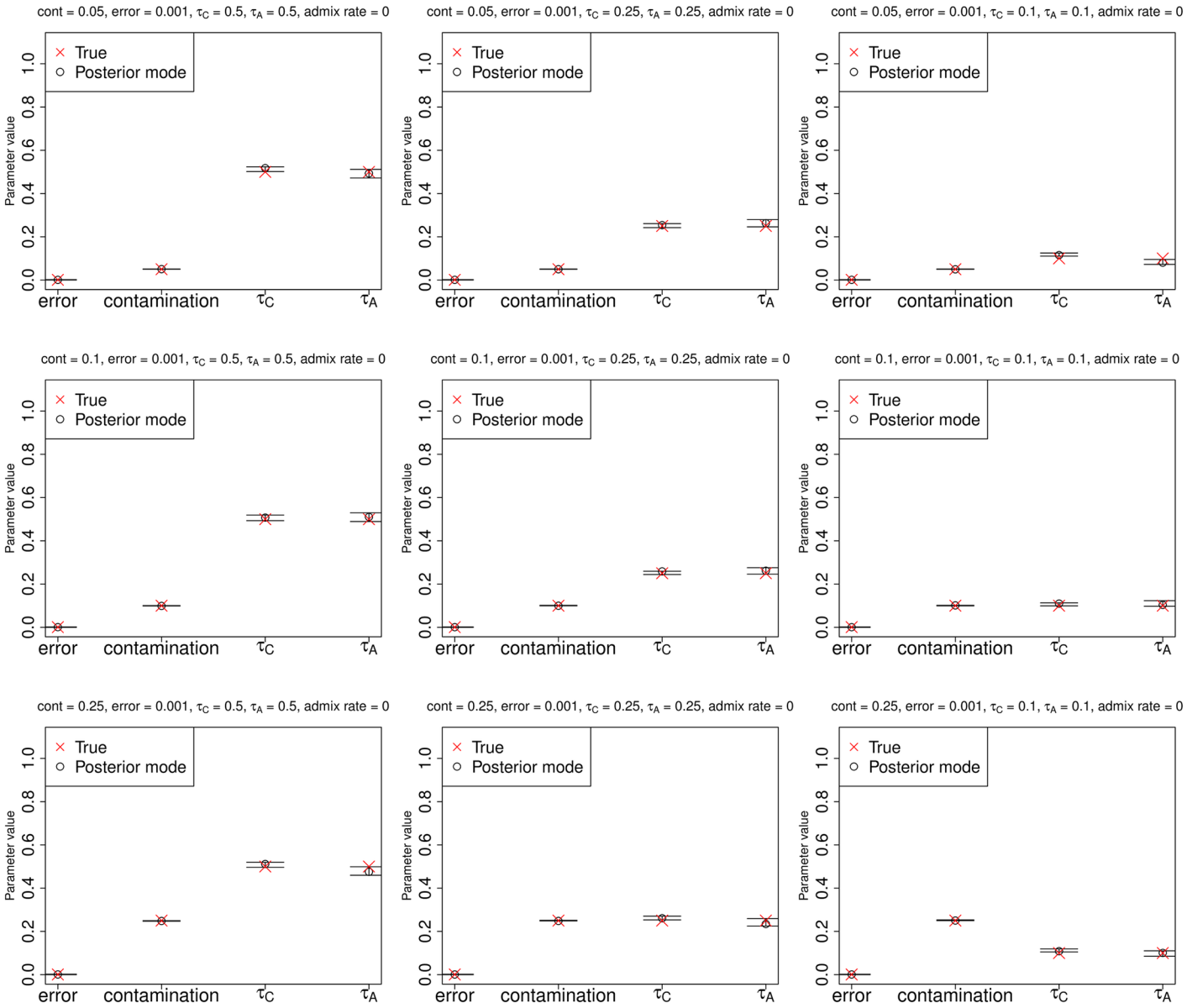

Fig 4. Estimation of parameters for a high-coverage ancient DNA genome (30X) with low sequencing error (0.1%), no admixture and a large anchor population panel (100 haploid genomes). Error bars represent 95% posterior intervals.

doi:10.1371/journal.pgen.1005972.g004

is made up of only 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 fragments from a panel of individuals, we randomly picked a set of two chromosomes at each unlinked site and only drew contaminant fragments from those two chromosomes. S13 Fig shows that inference is robust to this scenario, unless the contamination rate is very high (25%). In that case, the drift of the archaic

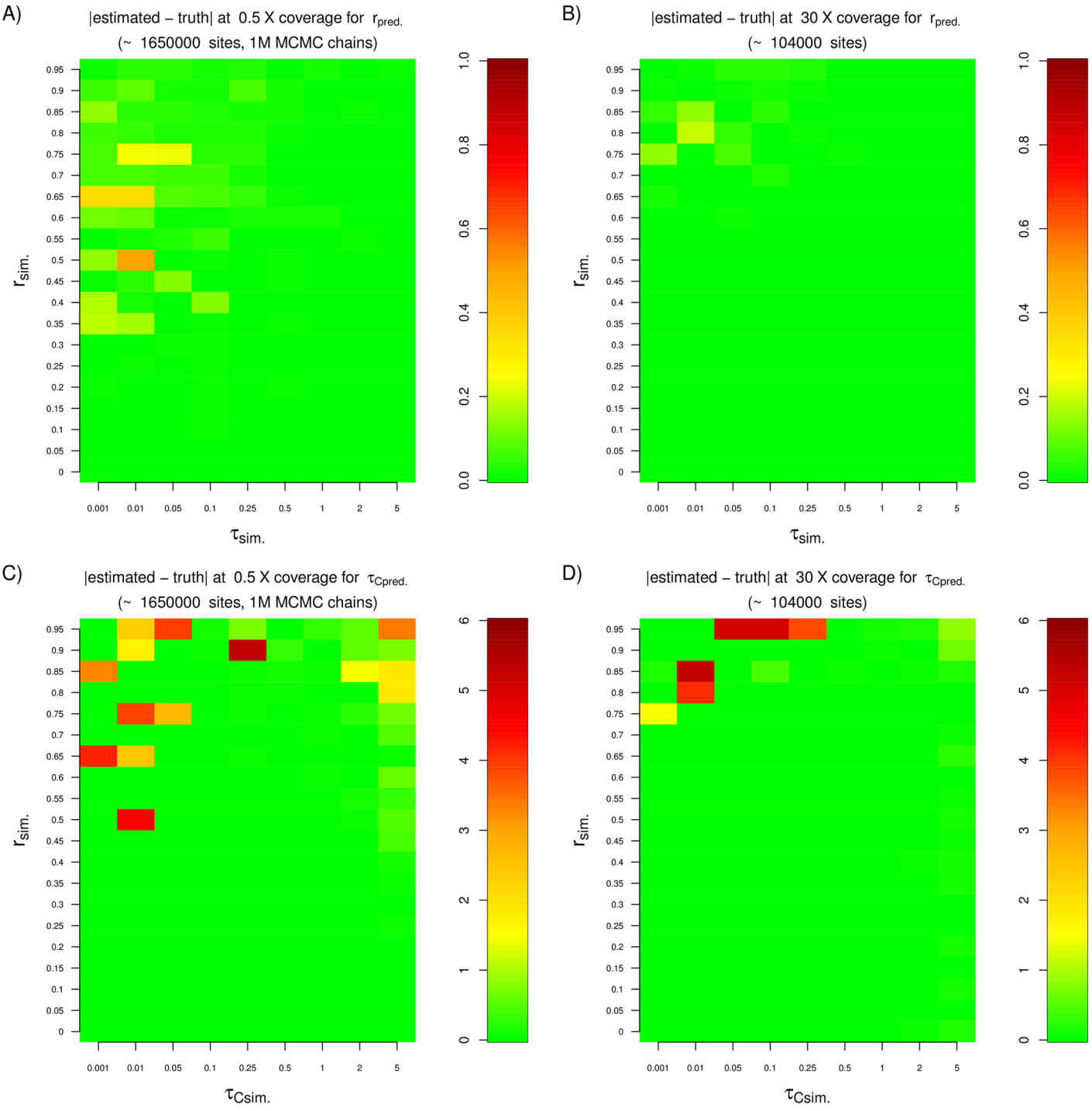

Fig 5. We tested the performance of the two-population method under a variety of drift and contamination scenarios for a sample of very low (0.5X) or very high (30X) coverage. We found that we needed more sites with coverage >0 (≈1.6 million) to obtain accurate estimates from the low coverage sample. The MCMC chain was also run for a longer time (1 million steps). A) Absolute difference between the estimated and the simulated contamination rate for a 0.5X genome. B) Absolute difference between the estimated and estimated and the simulated contamination rate for a 30X genome. C) Absolute difference between the estimated and the simulated anchor drift for a 0.5X genome. D) Absolute difference between the estimated and the simulated anchor drift for a 30X genome. In all simulations, the anchor drift was set to be equal to the ancient sample drift.

doi:10.1371/journal.pgen.1005972.g005

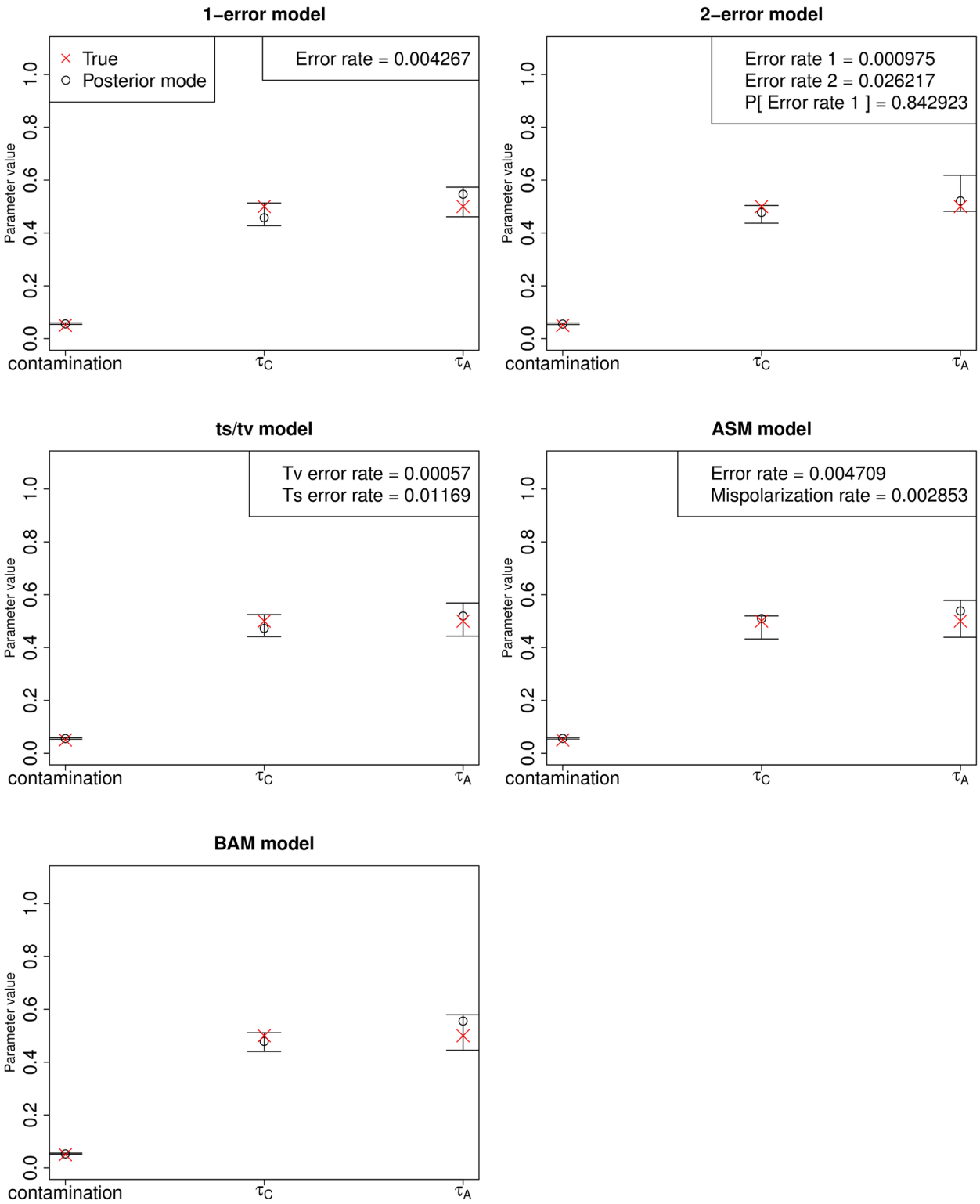

Fig 6. Estimation of parameters for a high-coverage ancient DNA genome (30X) simulated under a realistic scenario in which fragments from the ancient and contaminant genome were generated and then mapped to a reference genome. We allowed for multiple substitutions at the same site after the split from chimp, as well as sequencing errors and post-mortem deamination errors at the ends of the fragments. The five panels show results from inferring parameters under five different error rate models. Top-left: single-error model. Top-right: two-error model [5]. Middle-left: model with separate errors for transitions (ts) and tranversions (tv). Middle-right: single-error model with an ancestral state misidentification parameter. Bottom-left: Model in which errors were inferred individually at each site, using base and mapping qualities obtained from the simulated BAM file. Error bars represent 95% posterior intervals.

doi:10.1371/journal.pgen.1005972.g006

genome is substantially under-estimated, but the error, contamination and anchor drift parameters only show slight inaccuracies in the estimate.

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 (S11 Fig). 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.

Identifying the contaminant population. We sought to see whether we would use our method to identify the contaminant population, from among a set of candidate contaminants (for example, different present-day human panels). Because our MCMC samples are samples from the posterior distribution of the parameters and not the marginal likelihood of the data over the entire parameter space, we cannot perform proper Bayesian model selection. Instead, we used the posterior mode as a heuristic statistic that may suggest which panel is most likely to have contaminated the sample. We validated this choice of statistic using simulations under a variety of demographic scenarios (S15 Fig). We simulated 5-population trees of varying drift times. The outgroup was chosen to be the ancient population and the rest were chosen to be the present-day human populations (A, B, C and D). One of the populations (A) was the true contaminant. To add another layer of complexity, we also allowed for admixture (at 0%, 5% and 50% rate) from the ancient population to the ancestral population of A and B. We then ran our MCMC method four times on each of these demographic scenarios, using D as the anchor and different panels as the putative contaminant in each run.

S16 Fig shows that the highest posterior mode always corresponds to the run that uses the true contaminant (A), and that the mode decreases the farther the tested contaminant is from the true contaminant in the tree. Additionally, S17, S18 and S19 Figs show the effect of misspecifying the contaminant panel for different admixture scenarios. The error rate and the anchor drift time are correctly estimated, even when the candidate contaminant is highly diverged from the true contaminant, while the other two parameters are more sensitive to misspecification. In general, the correct candidate contaminant produces the highest posterior probability and yields the best parameter estimates.

Empirical data. We first applied our method to published ancient DNA data from a high-coverage genome (52X) from Denisova cave in Siberia (the Altai Neanderthal) [4], and visually ensured that the chain had converged. The demographic, error and contamination estimates are shown in Table 1. We used the African (AFR) 1000 Genomes Phase 3 panel [16] 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 ran our method with different putative contaminant panels: Africans (AFR), East Asians (EAS), Native Americans (AMR), Europeans (EUR), South Asians (SAS). For the Altai sample, we observe a contamination rate of $\sim$1% and an error rate of $\sim$0.1%, regardless of which panel we use. Furthermore, the drift on the Neanderthal side of the tree seems to be 6 times as large as the drift on the modern human side of the tree, reflecting the smaller effective population size of Neanderthals after their divergence. The EUR panel is the one with the highest posterior mode (Table 1).

We then tested a variety of ancient DNA nuclear genome sequences at different levels of coverage, obtained via different methods (shotgun sequencing and SNP capture) and from different hominin groups (modern humans and Neanderthals). We used AFR as the anchor panel and either AFR (S1 Table) or EUR (S2 Table) as the contaminant panel. For samples of high and medium average coverage, the MCMC converges to reasonable values for all parameters.

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.

Contaminant panel	Anchor panel	Error rate	Contamination rate	Modern human drift	Neanderthal drift	Log-posterior mode
EUR	AFR	0.12% (0.119%–0.12%)	0.952% (0.949%–0.956%)	0.414 (0.411–0.414)	2.497 (2.49–2.504)	-6476175.868
AMR	AFR	0.118% (0.118%–0.118%)	0.964% (0.963%–0.967%)	0.414 (0.411–0.414)	2.499 (2.494–2.506)	-6484270.973
SAS	AFR	0.12% (0.12%–0.121%)	0.95% (0.946%–0.951%)	0.411 (0.411–0.414)	2.496 (2.493–2.5)	-6489357.978
EAS	AFR	0.13% (0.129%–0.13%)	0.888% (0.888%–0.891%)	0.414 (0.412–0.414)	2.493 (2.488–2.493)	-6521082.384
AFR	AFR	0.112% (0.111%–0.112%)	0.969% (0.966%–0.973%)	0.412 (0.41–0.413)	2.495 (2.495–2.504)	-6574080.092

doi:10.1371/journal.pgen.1005972.t001

For example, we estimate the ancient population drift parameter (τ_A) to be larger in Neanderthals than in various modern humans sampled across Eurasia, as the effective population size of the former was smaller and their split time to Africans was larger. Of note is a sample from Haak et al. [31] (I0104) which appears to have 14.75% contamination when using Europeans as the contaminant source, but no contamination (0%) when using Africans as the source. This could suggest this sample has elevated levels of European-specific contamination, or that its demographic history may include complex admixture events that are not properly captured by our simple demographic model.

For samples of very low coverage, we observe a failure of some of the parameters to properly converge, as the MCMC seems to get stuck in the boundaries of parameter space. We tested different boundaries and the problem remains. This appears to be less of a problem when using AFR as the putative contaminant panel than when using EUR as the putative contaminant panel, presumably because of the larger amount of SNPs that may be informative for inference. In the former case, we only observe this problem when samples are at lower than $\sim$0.5X coverage. In the latter case, we observe the problem for samples at lower than $\sim$3X coverage.

For example, the low-coverage Neanderthal genome (0.5X) from Mezmaiskaya Cave in Western Russia [4] seems to converge to parameters within the prior boundaries when using AFR as the contaminant panel but the ancient population drift gets stuck in the upper limit of parameter space when any of the other panels are used as contaminants (S3 Table). Regardless of which contaminant panel is used, there is good agreement with the modern human drift parameter obtained when using the Altai Neanderthal genome. However, we note that when using non-African populations as the contaminants, we obtain a higher ($\sim$5%) contamination rate in the Mezmaiskaya Neanderthal than in the Altai Neanderthal. It is currently unclear to us whether this is due to the MCMC failing to properly converge or to a real feature of the data.

We sought to determine the robustness of our results to different levels of GC content. We did this because we initially hypothesized that endogenous DNA might be preserved at lower rates when GC content is low, leading to the presence of proportionally more contaminant DNA. 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 [32]. 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 (S20 Fig). 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 ($P < 2.2e-16$). 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 (S20 Fig), as CpG sites are known to have higher mutation rates than the rest of the genome. For this reason, we recommend filtering them out when testing for contamination on ancient DNA datasets, which is what was done in Tables 1 and S3.

Finally, we 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 (S21 Fig). 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 their closest African sister populations [33].

Three-population method

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 (S22, S23 and S24 Figs, 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 (Fig 7).

As before, we also verified that the posterior mode was a good proxy to identify the true contaminant (A), when running the MCMC using different contaminant panels (A, B, C and D). In all cases, we used D as the unadmixed anchor panel and B as the admixed anchor panel. Results are shown in S25 Fig for all the demographic scenarios from S15 Fig. Again, we observe that the true contaminant (A) is always the one that corresponds to the highest posterior probability, though we again caution that because we do not have the marginal probabilities, we cannot formally perform model selection to favor a particular panel. Furthermore,the admixture rate from the ancient population into the ancestors of A and B is robustly estimated unless the true contaminant (A) is highly diverged from the candidate contaminant (S26, S27 and S28 Figs, for admixture rates of 0%, 5% and 50%, respectively).

Empirical 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

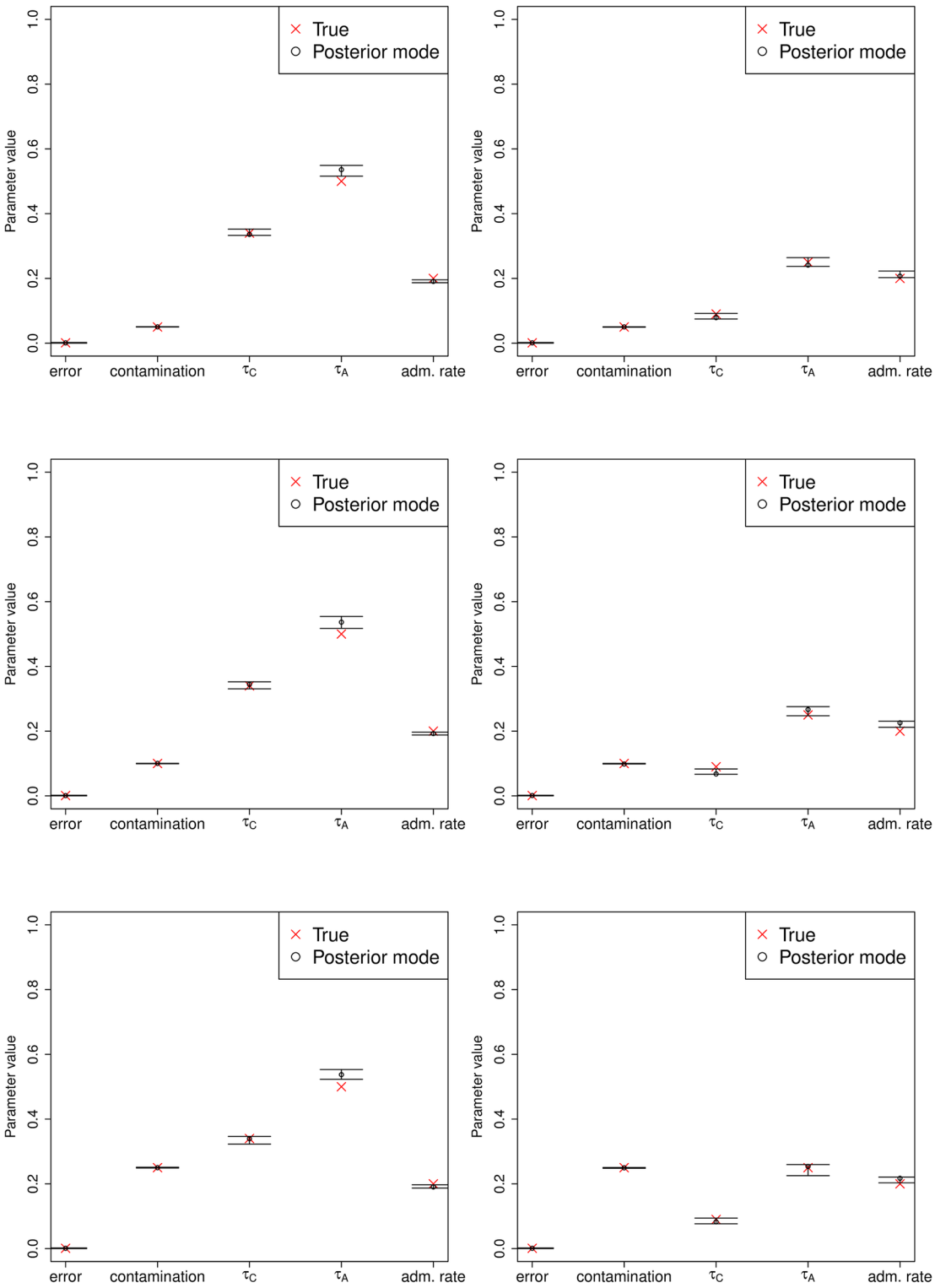

Fig 7. Estimation of error, contamination and demographic parameters in various three-population demographic scenarios, where the admixture rate is 20% and the admixture time was recent (0.005 drift units ago). The prior used for the admixture time was uniform over [0, 0.01]. Error bars represent 95% posterior intervals.

doi:10.1371/journal.pgen.1005972.g007

Table 2. 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 ($\tau_{Afr} = 0.009$, $\tau_{Eur} = 0.255$).

Contaminant panel	Unadmixed anchor panel	Admixed anchor panel	Error rate	Contamination rate	Ancestral human drift	Neanderthal drift	Admixture rate	Log-posterior mode
EUR	AFR	EUR	0.119% (0.119%–0.12%)	0.967% (0.954%–0.967%)	0.411 (0.405–0.414)	2.669 (2.656–2.689)	1.72% (1.682%–1.805%)	-7452958.125
AMR	AFR	EUR	0.119% (0.118%–0.12%)	0.967% (0.962%–0.974%)	0.407 (0.402–0.412)	2.677 (2.651–2.708)	1.661% (1.618%–1.696%)	-7461041.325
SAS	AFR	EUR	0.122% (0.122%–0.123%)	0.95% (0.944%–0.955%)	0.399 (0.398–0.406)	2.682 (2.677–2.695)	1.469% (1.422%–1.48%)	-7465214.726
EAS	AFR	EUR	0.13% (0.129%–0.132%)	0.896% (0.884%–0.903%)	0.421 (0.413–0.428)	2.702 (2.658–2.706)	2.388% (2.009%–2.447%)	-7509504.053
AFR	AFR	EUR	0.117% (0.117%–0.119%)	0.957% (0.945%–0.964%)	0.409 (0.409–0.418)	2.681 (2.66–2.702)	1.837% (1.766%–1.961%)	-7554080.773

doi:10.1371/journal.pgen.1005972.t002

L-BFGS-B likelihood optimization algorithm [13], 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 [34]. The error rate and contamination rates we obtain are similar to those obtained under the two-population method, and we estimate an admixture rate from Neanderthals into modern humans of 1.72% for the choice of contaminant panel with the highest posterior mode—which is again EUR (Table 2).

We also applied the method to the low-coverage Mezmaiskaya Neanderthal genome. As before, we are able to reach convergence for all parameters (including the admixture rate) with the exception of the Neanderthal drift, which gets stuck in the upper boundary of parameter space (S4 Table).

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, [35]). Although we cannot perform proper model testing, we found via extensive simulations that the posterior mode of an MCMC run was a robust heuristic statistic to help detect which panel was most likely to have contaminated the sample. We caution, however, that the fact that a particular panel yields a higher posterior mode than another is no guarantee that it is a better fit to the data for demographic scenarios that may be different from the ones we simulated.

We also applied our method to empirical data, specifically to two Neanderthal genomes at high and low coverage, a present-day high-coverage Yoruba genome, and several ancient genome sequences of varying degrees of coverage, some obtained via shotgun-sequencing and some via SNP capture. For the high-coverage 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 Altai Neanderthal 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 approximately 0.411 + 0.009 = 0.42 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 use our drift time in 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 531,486 years. This number roughly agrees with the most recent estimates obtained via other methods [4]. Additionally, the Neanderthal-specific drift time is approximately 6.5 times as large as the modern human drift time, which is expected as Neanderthals had much smaller population sizes than modern humans [36, 4]. The admixture rate from archaic to modern humans that we estimate is 1.72%, which is consistent with the rate estimate obtained via methods that do not jointly model contamination (1.5–2.1%) [4]. In the case of the Altai Neanderthal, we observe that the sample was probably contaminated by one or more individuals with European ancestry.

When testing modern human and Neanderthal ancient genomes of lower coverage than the Altai Neanderthal, we obtain reasonable parameter estimates for samples of medium to high-coverage. However, we run into problems in estimation when the samples are of low coverage. For these reasons, and from our simulation results, we recommend that our method should be used on nuclear genomes with >3X coverage. The method may converge under certain conditions at coverages as low as 0.5X (for example, in the case of the Mezmaiskaya genome under the two-population model when using AFR as the anchor and contaminant panel), but, in such cases, we caution the user to check convergence is achieved before drawing any conclusions from the estimates. For SNP capture data, we obtain reliable estimates for samples with a minimum coverage of 500,000 sites that are polymorphic in the anchor panel.

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 [20] to handle more than 3 populations at a time. In recent years, several studies have made advances in the development of methods to compute the likelihood of an SFS for larger numbers of populations using coalescent theory [37, 38, 39], with multiple population size changes and admixture events. We hope that some of these techniques could be incorporated in future versions of our inference framework.

Supporting Information

S1 Table. We applied the two-population method to ancient Neanderthal and modern human genomes ranging from 52X to 0.054X coverage. We tested both shotgun-sequencing data and SNP capture data. We used AFR as both the anchor panel and the putative contaminant panel. Samples are sorted by decreasing mean coverage. We define Convergence to be true (T) if all the parameters stably converged in a region of parameter space that does not include the upper parameter boundary. Otherwise Convergence is false (F). A line separates the two Convergence classes. SNPs = number of SNPs overlapping with anchor panel. Observations = total number of base observations analyzed. SC = SNP capture. SS = shotgun sequencing. HG = hunter-gatherer. LBK = Linear Pottery culture. MN = Middle Neolithic. LN = Late Neolithic. NEA = Neanderthal. MH = Modern Human. LogPos = Log-posterior mode. Reported Cov. = Mean read coverage reported in corresponding study. For SNP capture, this is the mean coverage of the targeted SNPs.
(PDF)

S2 Table. We applied the two-population method to ancient Neanderthal and modern human genomes ranging from 52X to 0.054X coverage. We tested both shotgun-sequencing data and SNP capture data. We used AFR as the anchor panel and EUR as the putative contaminant panel. Samples are sorted by decreasing mean coverage. We define Convergence to be true (T) if all the parameters stably converged in a region of parameter space that does not include the upper parameter boundary. Otherwise Convergence is false (F). A line separates the two Convergence classes. SNPs = number of SNPs overlapping with anchor panel. Observations = total number of base observations analyzed. SC = SNP capture. SS = shotgun sequencing. HG = hunter-gatherer. LBK = Linear Pottery culture. MN = Middle Neolithic. LN = Late Neolithic. NEA = Neanderthal. MH = Modern Human. LogPos = Log-posterior mode. Reported Cov. = Mean read coverage reported in corresponding study. For SNP capture, this is the mean coverage of the targeted SNPs.
(PDF)

S3 Table. 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. AFR were the anchor population in all cases, so the modern human drift is with respect to Africans. Values in parentheses are 95% posterior quantiles. Except when using AFR as the contaminant, the Neanderthal drift parameter gets stuck at the upper boundary (5 drift units) of parameter space.
(PDF)

S4 Table. Posterior modes of parameter estimates under the three-population inference framework for the Mezmaiskaya 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 ($\tau_{Afr} = 0.009$, $\tau_{Eur} = 0.255$). In all cases, the Neanderthal drift parameter gets stuck at the upper boundary (5 drift units) of parameter space.
(PDF)

S1 Fig. 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). Error bars represent 95% posterior intervals.
(TIFF)

S2 Fig. Absolute difference between estimated and simulated contamination rates for a variety of anchor drift and contamination scenarios, for different levels of coverage. In all simulations, the anchor drift was set to be equal to the ancient sample drift. A) 0.5X coverage (800,000 simulations). B) 1.5X coverage (200,000 simulations). C) 5X coverage (200,000 simulations). D) 30X coverage (200,000 simulations). The number of sites with coverage > 0 is denoted at the top of each panel.
(TIFF)

S3 Fig. Absolute difference between estimated and simulated anchor drifts for a variety of anchor drift and contamination scenarios, for different levels of coverage. In all simulations, the anchor drift was set to be equal to the ancient sample drift. A) 0.5X coverage (800,000 simulations). B) 1.5X coverage (200,000 simulations). C) 5X coverage (200,000 simulations). D) 30X coverage (200,000 simulations). The number of sites with coverage > 0 is denoted at the top of each panel.
(TIFF)

S4 Fig. Absolute difference between estimated and simulated ancient sample drifts for a variety of anchor drift and contamination scenarios, for different levels of coverage. In all simulations, the anchor drift was set to be equal to the ancient sample drift. A) 0.5X coverage (800,000 simulations). B) 1.5X coverage (200,000 simulations). C) 5X coverage (200,000 simulations). D) 30X coverage (200,000 simulations). The number of sites with coverage > 0 is denoted at the top of each panel.
(TIFF)

S5 Fig. Absolute difference between estimated and simulated contamination rates for a variety of anchor drift and contamination scenarios, when coverage is low (0.5X, 1X or 1.5X). Here, we used a large number of sites and run the MCMC chain for 1 million steps. In all simulations, the anchor drift was set to be equal to the ancient sample drift. A) 0.5X coverage (800,000 simulations). B) 1X coverage (400,000 simulations). C) 1.5X coverage (200,000 simulations). The number of sites with coverage > 0 is denoted at the top of each panel.
(TIFF)

S6 Fig. Absolute difference between estimated and simulated anchor drifts for a variety of anchor drift and contamination scenarios, when coverage is low (0.5X, 1X or 1.5X). Here, we used a large number of sites and run the MCMC chain for 1 million steps. In all simulations, the anchor drift was set to be equal to the ancient sample drift. A) 0.5X coverage (800,000 simulations). B) 1X coverage (400,000 simulations). C) 1.5X coverage (200,000 simulations). The number of sites with coverage > 0 is denoted at the top of each panel.
(TIFF)

S7 Fig. Absolute difference between estimated and simulated ancient sample drifts for a variety of anchor drift and contamination scenarios, when coverage is low (0.5X, 1X or 1.5X). Here, we used a large number of sites and run the MCMC chain for 1 million steps. In all simulations, the anchor drift was set to be equal to the ancient sample drift. A) 0.5X coverage (800,000 simulations). B) 1X coverage (400,000 simulations). C) 1.5X coverage (200,000 simulations). The number of sites with coverage > 0 is denoted at the top of each panel.
(TIFF)

S8 Fig. Absolute difference between estimated and simulated contamination rates for a variety of anchor drift and contamination scenarios, when coverage is low (0.5X, 1X or 1.5X). We used a large number of sites and run 10 MCMC chains for 1 million steps each. To ensure convergence, we then selected the chain with the highest posterior probability, and here show estimates from that chain. In all simulations, the anchor drift was set to be equal to the ancient sample drift. A) 0.5X coverage (800,000 simulations). B) 1X coverage (400,000 simulations). C) 1.5X coverage (200,000 simulations). The number of sites with coverage > 0 is denoted at the top of each panel.
(TIFF)

S9 Fig. Absolute difference between estimated and simulated anchor drifts for a variety of anchor drift and contamination scenarios, when coverage is low (0.5X, 1X or 1.5X). We used a large number of sites and run 10 MCMC chains for 1 million steps each. To ensure convergence, we then selected the chain with the highest posterior probability, and here show estimates from that chain. In all simulations, the anchor drift was set to be equal to the ancient sample drift. A) 0.5X coverage (800,000 simulations). B) 1X coverage (400,000 simulations). C) 1.5X coverage (200,000 simulations). The number of sites with coverage > 0 is denoted at the top of each panel.
(TIFF)

S10 Fig. Absolute difference between estimated and simulated ancient sample drifts for a variety of anchor drift and contamination scenarios, when coverage is low (0.5X, 1X or 1.5X). We used a large number of sites and run 10 MCMC chains for 1 million steps each. To ensure convergence, we then selected the chain with the highest posterior probability, and here show estimates from that chain. In all simulations, the anchor drift was set to be equal to the ancient sample drift. A) 0.5X coverage (800,000 simulations). B) 1X coverage (400,000 simulations). C) 1.5X coverage (200,000 simulations). The number of sites with coverage > 0 is denoted at the top of each panel.
(TIFF)

S11 Fig. 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. Error bars represent 95% posterior intervals.
(TIFF)

S12 Fig. 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). Error bars represent 95% posterior intervals.
(TIFF)

S13 Fig. Estimation of parameters for a high-coverage ancient DNA genome (30X), when the contaminant fragments are exclusively drawn from a single diploid individual from the contaminant panel. Error bars represent 95% posterior intervals.
(TIFF)

S14 Fig. 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. Error bars represent 95% posterior intervals.
(TIFF)

S15 Fig. Three demographic models used to test the method when the contaminant is misspecified. When testing the two-population method, we set panel A as the true contaminant and panel D as the anchor. When testing the three-population method, we set panel A as the true contaminant, panel D as the unadmixed anchor and panel B as the admixed anchor. The numbers on the branches represent the drift parameters. The parameter α represents the admixture rate from the ancient population into the ancestor of A and B.
(TIFF)

S16 Fig. When testing different putative contaminants, the highest mode of the posterior likelihoods from the MCMC under the two-population model corresponds to the true contaminant (panel A). The y-axis shows the difference between the log-posterior for contaminant panel A and the log-posterior for different candidate contaminant panels (A, B, C, D), so low values correspond to high posterior probabilities for each of the candidates. We added a 1 to the difference to be able to plot the difference on a logarithmic scale. The three panels contain results for three admixture scenarios (from left to right: admixture rate of 0%, 5% and 50%) and each panel shows the difference under different contamination rates and demographic models (the population relationships of panels A, B, C and D can be found in S15 Fig).
(TIFF)

S17 Fig. Parameters estimates under the two-population model using different putative contaminants, when the true contaminant is panel A. Each row of panels represents a different set of drift parameters, keeping the contamination rate fixed at 25% and the error rate at 0.1%. In this case, the admixture rate from the ancient population to the ancestor of A and B was kept at 0%. The anchor panel used was panel D (the population relationships of panels A, B, C and D can be found in S15 Fig).
(TIFF)

S18 Fig. Parameters estimates under the two-population model using different putative contaminants, when the true contaminant is panel A. Each row of panels represents a different set of drift parameters, keeping the contamination rate fixed at 25% and the error rate at 0.1%. In this case, the admixture rate from the ancient population to the ancestor of A and B was kept at 5%. The anchor panel used was panel D (the population relationships of panels A, B, C and D can be found in S15 Fig).
(TIFF)

S19 Fig. Parameters estimates under the two-population model using different putative contaminants, when the true contaminant is panel A. Each row of panels represents a different set of drift parameters, keeping the contamination rate fixed at 25% and the error rate at 0.1%. In this case, the admixture rate from the ancient population to the ancestor of A and B was kept at 50%. The anchor panel used was panel D (the population relationships of panels A, B, C and D can be found in S15 Fig).
(TIFF)

S20 Fig. 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. Error bars represent 95% posterior intervals.
(TIFF)

S21 Fig. 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. Error bars represent 95% posterior intervals.
(TIFF)

S22 Fig. 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]. Error bars represent 95% posterior intervals.
(TIFF)

S23 Fig. 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]. Error bars represent 95% posterior intervals.
(TIFF)

S24 Fig. 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]. Error bars represent 95% posterior intervals.
(TIFF)

S25 Fig. When testing different putative contaminants, the highest mode of the posterior likelihoods from the MCMC under the three-population model corresponds to the true contaminant (panel A). The y-axis shows the difference between the log-posterior for contaminant panel A and the log-posterior for different candidate contaminant panels (A, B, C, D), so low values correspond to high posterior probabilities for each of the candidates. We added a 1 to the difference to be able to plot the difference on a logarithmic scale. The three panels contain results for three admixture scenarios (from left to right: admixture rate of 0%, 5% and 50%) and each panel shows the difference under different contamination rates and demographic models (see S15 Fig).
(TIFF)

S26 Fig. Parameters estimates under the three-population model using different putative contaminants, when the true contaminant is panel A. Each row of panels represents a different set of drift parameters, keeping the contamination rate fixed at 25% and the error rate at 0.1%. In this case, the admixture rate from the ancient population to the ancestor of A and B was kept at 0%. The unadmixed anchor panel used was panel D and the admixed anchor panel was panel B (see S15 Fig).
(TIFF)

S27 Fig. Parameters estimates under the three-population model using different putative contaminants, when the true contaminant is panel A. Each row of panels represents a different set of drift parameters, keeping the contamination rate fixed at 25% and the error rate at 0.1%. In this case, the admixture rate from the ancient population to the ancestor of A and B was kept at 5%. The unadmixed anchor panel used was panel D and the admixed anchor panel was panel B (see S15 Fig).
(TIFF)

S28 Fig. Parameters estimates under the three-population model using different putative contaminants, when the true contaminant is panel A. Each row of panels represents a different set of drift parameters, keeping the contamination rate fixed at 25% and the error rate at 0.1%. In this case, the admixture rate from the ancient population to the ancestor of A and B was kept at 50%. The unadmixed anchor panel used was panel D and the admixed anchor panel was panel B (see S15 Fig).
(TIFF)

S1 Text. Genotype probabilities conditional on a demography. Derivation of formulas to obtain the probabilities of particular genotype states given a demographic history and an anchor population allele frequency, using diffusion theory.
(PDF)

S2 Text. Probabilistic inference using BAM files. Explanation of methodology for inferring fragment-specific error parameters in the optional BAM mode of DICE.
(PDF)

Acknowledgments

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

Author Contributions

Conceived and designed the experiments: FR GR MS. Performed the experiments: FR GR. Analyzed the data: FR GR. Contributed reagents/materials/analysis tools: FR GR MS. Wrote the paper: FR GR MS.

References

1. Green R. E., Krause J., Briggs A. W., Maricic T., Stenzel U., Kircher M., Patterson N., Li H., Zhai W., Fritz M. H.-Y., et al., A draft sequence of the Neandertal genome, Science 328 (2010) 710–722. doi: 10.1126/science.1188021 PMID: 20448178
2. Reich D., Green R. E., Kircher M., Krause J., Patterson N., Durand E. Y., Viola B., Briggs A. W., Stenzel U., Johnson P. L., et al., Genetic history of an archaic hominin group from Denisova Cave in Siberia, Nature 468 (2010) 1053–1060. doi: 10.1038/nature09710 PMID: 21179161
3. Meyer M., Kircher M., Gansauge M.-T., Li H., Racimo F., Mallick S., Schraiber J. G., Jay F., Prüfer K., de Filippo C., et al., A high-coverage genome sequence from an archaic Denisovan individual, Science 338 (2012) 222–226. doi: 10.1126/science.1224344 PMID: 22936568
4. Prüfer K., Racimo F., Patterson N., Jay F., Sankararaman S., Sawyer S., Heinze A., Renaud G., Sudmant P. H., de Filippo C., et al., The complete genome sequence of a neanderthal from the altai mountains, Nature 505 (2014) 43–49. doi: 10.1038/nature12886 PMID: 24352235
5. Fu Q., Li H., Moorjani P., Jay F., Slepchenko S. M., Bondarev A. A., Johnson P. L., Aximu-Petri A., Prüfer K., de Filippo C., et al., Genome sequence of a 45,000-year-old modern human from western Siberia, Nature 514 (2014) 445–449. doi: 10.1038/nature13810 PMID: 25341783
6. Seguin-Orlando A., Korneliussen T. S., Sikora M., Malaspinas A.-S., Manica A., Moltke I., Albrechtsen A., Ko A., Margaryan A., Moiseyev V., et al., Genomic structure in Europeans dating back at least 36,200 years, Science 346 (2014) 1113–1118. doi: 10.1126/science.aaa0114 PMID: 25378462
7. Green R. E., Malaspinas A.-S., Krause J., Briggs A. W., Johnson P. L., Uhler C., Meyer M., Good J. M., Maricic T., Stenzel U., et al., A complete Neandertal mitochondrial genome sequence determined by high-throughput sequencing, Cell 134 (2008) 416–426. doi: 10.1016/j.cell.2008.06.021 PMID: 18692465
8. Green R. E., Briggs A. W., Krause J., Prüfer K., Burbano H. A., Siebauer M., Lachmann M., Pääbo S., The Neandertal genome and ancient DNA authenticity, The EMBO journal 28 (2009) 2494–2502. doi: 10.1038/emboj.2009.222 PMID: 19661919

9. Sawyer S., Renaud G., Viola B., Hublin J.-J., Gansauge M.-T., Shunkov M. V., Derevianko A. P., Prüfer K., Kelso J., Pääbo S., Nuclear and mitochondrial dna sequences from two denisovan individuals, Proceedings of the National Academy of Sciences 112 (2015) 15696–15700.
10. Skoglund P., Storå J., Götherström A., Jakobsson M., Accurate sex identification of ancient human remains using DNA shotgun sequencing, Journal of Archaeological Science 40 (2013) 4477–4482. doi: 10.1016/j.jas.2013.07.004
11. Rasmussen M., Guo X., Wang Y., Lohmueller K. E., Rasmussen S., Albrechtsen A., Skotte L., Lindgreen S., Metspalu M., Jombart T., et al., An aboriginal australian genome reveals separate human dispersals into asia, Science 334 (2011) 94–98. doi: 10.1126/science.1211177 PMID: 21940856
12. Korneliussen T. S., Albrechtsen A., Nielsen R., ANGSD: analysis of next generation sequencing data, BMC Bioinformatics 15 (2014) 356. doi: 10.1186/s12859-014-0356-4 PMID: 25420514
13. Byrd R. H., Lu P., Nocedal J., Zhu C., A limited memory algorithm for bound constrained optimization, SIAM Journal on Scientific Computing 16 (1995) 1190–1208. doi: 10.1137/0916069
14. Skoglund P., Northoff B. H., Shunkov M. V., Derevianko A. P., Pääbo S., Krause J., Jakobsson M., Separating endogenous ancient DNA from modern day contamination in a Siberian Neandertal, Proceedings of the National Academy of Sciences 111 (2014) 2229–2234. doi: 10.1073/pnas.1318934111
15. Renaud G., Slon V., Duggan A. T., Kelso J., Schmutzi: estimation of contamination and endogenous mitochondrial consensus calling for ancient dna, Genome biology 16 (2015) 1–18. doi: 10.1186/s13059-015-0776-0
16. G. P. Consortium, et al., A global reference for human genetic variation, Nature 526 (2015) 68–74. doi: 10.1038/nature15393 PMID: 26432245
17. Ewens W. J., Mathematical Population Genetics 1: I. Theoretical Introduction, volume 27, Springer Science & Business Media, 2004.
18. Chen H., Green R. E., Pääbo S., Slatkin M., The joint allele-frequency spectrum in closely related species, Genetics 177 (2007) 387–398. doi: 10.1534/genetics.107.070730 PMID: 17603120
19. Ewing G., Hermisson J., MSMS: a coalescent simulation program including recombination, demographic structure and selection at a single locus, Bioinformatics 26 (2010) 2064–2065. doi: 10.1093/bioinformatics/btq322 PMID: 20591904
20. Gutenkunst R. N., Hernandez R. D., Williamson S. H., Bustamante C. D., Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data, PLoS Genetics 5 (2009) e1000695. doi: 10.1371/journal.pgen.1000695 PMID: 19851460
21. Roberts G. O., Gelman A., Gilks W. R., et al., Weak convergence and optimal scaling of random walk Metropolis algorithms, The Annals of Applied Probability 7 (1997) 110–120. doi: 10.1214/aoap/1034625254
22. Hofreiter M., Jaenicke V., Serre D., von Haeseler A., Pääbo S., Dna sequences from multiple amplifications reveal artifacts induced by cytosine deamination in ancient dna, Nucleic acids research 29 (2001) 4793–4799. doi: 10.1093/nar/29.23.4793 PMID: 11726688
23. Briggs A. W., Stenzel U., Meyer M., Krause J., Kircher M., Pääbo S., Removal of deaminated cytosines and detection of in vivo methylation in ancient dna, Nucleic acids research 38 (2010) e87–e87. doi: 10.1093/nar/gkp1163 PMID: 20028723
24. Hernandez R. D., Williamson S. H., Bustamante C. D., Context dependence, ancestral misidentification, and spurious signatures of natural selection, Molecular Biology and Evolution 24 (2007) 1792–1800. doi: 10.1093/molbev/msm108 PMID: 17545186
25. Hudson R. R., Generating samples under a wright–fisher neutral model of genetic variation, Bioinformatics 18 (2002) 337–338. doi: 10.1093/bioinformatics/18.2.337 PMID: 11847089
26. Li H., Durbin R., Fast and accurate short read alignment with burrows–wheeler transform, Bioinformatics 25 (2009) 1754–1760. doi: 10.1093/bioinformatics/btp324 PMID: 19451168
27. Rambaut A., Grass N. C., Seq-gen: an application for the monte carlo simulation of dna sequence evolution along phylogenetic trees, Computer applications in the biosciences: CABIOS 13 (1997) 235–238. PMID: 9183526
28. Hasegawa M., Kishino H., Yano T.-a., Dating of the human-ape splitting by a molecular clock of mitochondrial dna, Journal of molecular evolution 22 (1985) 160–174. doi: 10.1007/BF02101694 PMID: 3934395
29. Lazaridis I., Patterson N., Mittnik A., Renaud G., Mallick S., Kirsanow K., Sudmant P. H., Schraiber J. G., Castellano S., et al., Ancient human genomes suggest three ancestral populations for present-day Europeans, Nature 513 (2014) 409–413. doi: 10.1038/nature13673 PMID: 25230663
30. Renaud G., Kircher M., Stenzel U., Kelso J., freeibis: an efficient basecaller with calibrated quality scores for illumina sequencers, Bioinformatics 29 (2013) 1208–1209. doi: 10.1093/bioinformatics/btt117 PMID: 23471300

31. Haak W., Lazaridis I., Patterson N., Rohland N., Mallick S., Llamas B., Brandt G., Nordenfelt S., Harney E., Stewardson K., et al., Massive migration from the steppe was a source for indo-european languages in europe, Nature (2015). doi: 10.1038/nature14317

32. Rosenbloom K. R., Armstrong J., Barber G. P., Casper J., Clawson H., Diekhans M., Dreszer T. R., Fujita P. A., Guruvadoo L., Haeussler M., et al., The ucsc genome browser database: 2015 update, Nucleic acids research 43 (2015) D670–D681. doi: 10.1093/nar/gku1177 PMID: 25428374

33. Lipson M., Loh P.-R., Levin A., Reich D., Patterson N., Berger B., Efficient moment-based inference of admixture parameters and sources of gene flow, Molecular biology and evolution 30 (2013) 1788–1802. doi: 10.1093/molbev/mst099 PMID: 23709261

34. Higham T., Douka K., Wood R., Ramsey C. B., Brock F., Basell L., Camps M., Arrizabalaga A., Baena J., Barroso-Ruíz C., et al., The timing and spatiotemporal patterning of Neanderthal disappearance, Nature 512 (2014) 306–309. doi: 10.1038/nature13621 PMID: 25143113

35. Li H., Durbin R., Inference of human population history from individual whole-genome sequences, Nature 475 (2011) 493–496. doi: 10.1038/nature10231 PMID: 21753753

36. Castellano S., Parra G., Sánchez-Quinto F. A., Racimo F., Kuhlwilm M., Kircher M., Sawyer S., Fu Q., Heinze A., Nickel B., et al., Patterns of coding variation in the complete exomes of three Neandertals, Proceedings of the National Academy of Sciences 111 (2014) 6666–6671. doi: 10.1073/pnas.1405138111

37. Chen H., The joint allele frequency spectrum of multiple populations: a coalescent theory approach, Theoretical Population Biology 81 (2012) 179–195. doi: 10.1016/j.tpb.2011.11.004 PMID: 22155588

38. Jewett E. M., Rosenberg N. A., Theory and applications of a deterministic approximation to the coalescent model, Theoretical Population Biology 93 (2014) 14–29. doi: 10.1016/j.tpb.2013.12.007 PMID: 24412419

39. 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).