

Assessing the relationship of ancient individuals to each other and to modern populations

Joshua G. Schraiber

Started on October 22, 2016. Compiled on February 9, 2017

1 Introduction

Ancient DNA (aDNA) is now ubiquitous in population genetics. Advances in library preparation (CITE MATHIAS) and sequence capture (CITE DAVID) make it possible to obtain genome-wide data from hundreds of samples (CITE DAVID, ESKE). Analysis of these data can provide new insight into recent evolutionary processes which leave faint signatures in modern genomes, including natural selection (CITE) and population replacement (CITE).

One of the most powerful uses of ancient DNA is to assess the continuity of ancient and modern populations. In many cases, it is unclear whether populations that occupied an area in the past are the direct ancestors of the current inhabitants of that area. However, this can be next to impossible to assess using only modern genomes. Questions of population continuity and replacement have particular relevance for the spread of cultures and technology in humans (CITE) and local adaptation (CITE). For instance, recent work showed that modern Native Americans are directly descended from ancestral Native Americans, settling a long-standing debate about the colonization of the Americas (CITE).

Despite its utility in addressing difficult-to-answer questions in evolutionary biology, aDNA also has several limitations. Most strikingly, DNA decays rapidly following the death of an organism, resulting in highly fragmented, degraded starting material when sequencing (CITE SUSANNA/JOHANNES). Thus, ancient data is frequently low coverage and has a significantly higher base-calling error rate than modern samples (CITE). When working with diploid data, as in aDNA extracted from plants and animals, low coverage data prevents genotypes from being called.

Several strategies are commonly used to address the low coverage data. One of the most common approaches is to sample a random read from each covered site and use that as a haploid genotype call (CITE LOTS OF THINGS). Despite the simplicity and appeal of this approach, it has several drawbacks. Primarily, it throws away reads from sites that are covered more than once, resulting in not only a loss of information, but the disregarding of expensive, difficult to acquire data.

On the other hand, several approaches exist to either work with genotype likelihoods or the raw read data. Genotype likelihoods, which are the probability of the read data at a site given each of the three possible diploid genotypes at that site, can be used in calculation of population genetic statistics or likelihood functions to average over uncertainty in the genotype (CITE ANGSD). However, many approaches making use of genotype likelihood require them to be specified *a priori*; with low coverage data, an increase in accuracy is expected if genotype likelihoods are co-estimated with other parameters of interest, due to the covariation between processes that influence read quality and genetic diversity, such as contamination.

A recent method that coestimates genotype likelihoods along (including base-calling error and contamination) along with demographic parameters showed that there can be significant power to assess the relationship of a single ancient sample to a modern population. Nonetheless, they found that for very low coverage data, inferences were not reliable. Thus, they were unable to apply their method to the large number of extremely low coverage ($< 1\times$) genomes that are available. Moreover, they were unable to explore the tradeoffs that come with a limited budget: can we learn more by sequencing fewer individuals to high coverage, or more individuals at lower coverage?

Here, we develop a novel maximum likelihood approach for analyzing low coverage ancient DNA in relation to a modern population. We work directly with raw read data and explicitly model base-calling errors. Crucially, our approach incorporates data from multiple individuals that belong to the same ancient population, which we show substantially increases power and reduces error in parameter estimates. We then apply our new methodology to ancient human data, and show that we can obtain demographic inferences even from very low coverage samples by analyzing them jointly.

2 Results

2.1 Sampling alleles in ancient populations

We assume a scenario in which allele frequencies are known with high accuracy in a modern population. Suppose that an allele is known to be at frequency x in the modern population, and we wish to compute the probability of obtaining k copies of that allele in a sample of n chromosomes from an ancient population. To do so, we assume a simple demographic model in which the ancient individual belongs to a population that split off from the modern population τ_1 generations ago, and subsequently existed as an isolated population for τ_2 . Further, we assume that the modern population has effective size $N_e^{(1)}$ and that the ancient population has effective size $N_e^{(2)}$, and measure time in diffusion units, $t_i = \tau_i/(2N_e^{(i)})$. If we know the conditional probability that an allele is at frequency y in the ancient sample,

given that it is at frequency x , then the sampling probability is simply an integral,

$$\begin{aligned}
P_{n,k}(x) &= \int_0^1 \binom{n}{k} y^k (1-y)^{n-k} f(y; x, t_1, t_2) dy \\
&= \binom{n}{k} \mathbb{E}_x \left(Y^k (1-Y)^{n-k}; t_1, t_2 \right) \\
&\equiv \binom{n}{k} p_{n,k}(t_1, t_2)
\end{aligned} \tag{1}$$

Thus, we must compute the binomial moments of the allele frequency distribution in the ancient population. In the Methods, we show that this can be computed using matrix exponentiation,

$$p_{n,k}(t_1, t_2) = \left(e^{Qt_2} e^{Q^\downarrow t_1} \mathbf{h}_n \right)_i, \tag{2}$$

where $(\mathbf{v})_i$ indicates the i th element of the vector \mathbf{v} , $\mathbf{h}_n = ((1-x)^n, x(1-x)^{n-1}, \dots, x^n)^T$ and Q and Q^\downarrow are the sparse matrices

$$Q_{ij} = \begin{cases} \frac{1}{2}i(i-1) & \text{if } j = i-1 \\ -i(n-i) & \text{if } j = i \\ \frac{1}{2}(n-i)(n-i-1) & \text{if } j = i+1 \\ 0 & \text{else} \end{cases}$$

and

$$Q^\downarrow_{ij} = \begin{cases} \frac{1}{2}i(i-1) & \text{if } j = i-1 \\ -i(n-i+1) & \text{if } j = i \\ \frac{1}{2}(n-i+1)(n-i) & \text{if } j = i+1 \\ 0 & \text{else.} \end{cases}$$

This result has an interesting interpretation: the matrix Q^\downarrow can be thought of as evolving the allele frequencies back in time from the modern population to the common ancestor of the ancient and modern populations, while Q evolves the allele frequencies forward in time from the common ancestor to the ancient population (Fig 1).

[Figure 1 about here.]

Because of the fragmentation and degradation of DNA that is inherent in obtaining sequence data from ancient individuals, it is difficult to obtain the high coverage data necessary to make high quality genotype calls from ancient individuals. To address this, we instead work directly with raw read data, and average over all the possible genotypes weighted

by their probability of producing the data. Specifically, we follow (CITE: ANGSD) in modeling the probability of the read data in the ancient population, given the allele frequency at site l as

$$\mathbb{P}(R_l|k) = \sum_{g_{1,l}=0}^2 \dots \sum_{g_{n,l}=0}^2 \mathbb{I}\left(\sum_{i=1}^m g_{i,l} = k\right) \prod_{i=1}^n \binom{2}{g_{i,l}} \mathbb{P}(R_{i,l}|g_{i,l}),$$

where $R_{i,l} = (a_{i,l}, d_{i,l})$ are the counts of ancestral and derived reads in individual i at site l , $g_{i,l} \in \{0, 1, 2\}$ indicates the possible genotype of individual i at site l (i.e. 0 = homozygous ancestral, 1 = heterozygous, 2 = homozygous derived), and $\mathbb{P}(R_{i,l}|g_{i,l})$ is the probability of the read data at site l for individual i , assuming that the individual truly has genotype $g_{i,l}$. We use a binomial sampling with error model, in which the probability that a truly derived site appears ancestral (and vice versa) is given by ϵ . Thus,

$$\mathbb{P}(R|g) = \binom{a+d}{d} p_g^d (1-p_g)^a$$

with

$$\begin{aligned} p_0 &= \epsilon \\ p_1 &= \frac{1}{2} \\ p_2 &= 1 - \epsilon \end{aligned}$$

Combining these two aspects together by summing over possible allele frequencies weighted by their probabilities, we obtain our likelihood of the ancient data,

$$L(D) = \prod_{l=1}^L \sum_{k=0}^n \mathbb{P}(R_l|k) p_{n,k}(x_l). \quad (3)$$

2.2 Impact of coverage and number of samples on inferences

To explore the tradeoff of sequencing more individuals at lower depth compared to fewer individuals at higher coverage, we performed simulations using **msprime** (CITE JEROME). First, we examined the impact of coverage and number of samples on ability to recover the drift times in the modern and the ancient populations. Figure 2 shows results for data simulated with $t_1 = 0.02$ and $t_2 = 0.05$, corresponding to an ancient sample sampled from 300 generations ago from population of effective size 1000. The populations split 400 generations ago, and the modern population has an effective size of 10000. Inferences of t_1 can be relatively accurate even with few low coverage samples (Figure 2A). However,

inferences of t_2 benefit much more from increasing sampling depth, as opposed to coverage (Figure 2B). In particular, two individuals sequenced to 0.5x coverage have a much lower error than a single individual sequenced to 2x coverage. To explore this effect further, we derived the sampling probability of alleles covered by exactly one sequencing read (see Methods). We found that sites covered only once have no information about t_2 , suggesting that heterozygosity is very important for inferences about t_2 .

[Figure 2 about here.]

We next examined the impact of coverage and sampling on the power to reject the hypothesis that the ancient individuals come from a population that is continuous with the modern population. We analyzed both low coverage (0.5x) and higher coverage (4x) datasets consisting of 1 (for both low and high coverage samples) or 5 individuals (only for low coverage). We simulated data with parameters identical to the previous experiment, except we now examined the impact of varying the age of the ancient sample from 0 generations ago through to the split time with the modern population. We then performed a likelihood ratio test comparing the null model of continuity, in which $t_2 = 0$, to a model in which the ancient population is not continuous. Figure 3 shows the power of the likelihood ratio test. For single individuals sequenced at low coverage, we see that the test only has power for very recently sampled ancient individuals (i.e. samples that are highly diverged from the modern population). However, the power increases dramatically as the number of individuals or the coverage per individual is increased. Nonetheless, for samples that are very close to the divergence time, it will be difficult to determine if they are ancestral to the modern population or not.

[Figure 3 about here.]

2.3 Application to ancient humans

We applied our approach to ancient human data from (MATHIESON ET AL). Based on sampling location and associated archeological materials, the individuals were grouped into *a priori* populations, which we used to specify population membership when analyzing individuals together. We analyzed all samples for their relationship to the CEU individuals from the 1000 genomes project (CITE). Based on our results that suggested that extremely low coverage samples would yield unreliable estimates, we excluded populations represented by a single individual sequenced to less than 2x coverage.

We computed maximum likelihood estimates of t_1 and t_2 for individuals as grouped into populations (Figure 4A; Table 1). We observe that t_2 is significantly greater than 0 for all populations. Thus, none of these populations are consistent with directly making up a large proportion of the ancestry of modern CEU individuals. Strikingly, we see that $t_2 \gg t_1$, despite the fact that the ancient samples must have existed for fewer generations

since the population split than the modern samples. This suggests that all of the ancient populations are characterized by extremely small effective population sizes.

[Table 1 about here.]

[Figure 4 about here.]

We further explored the relationship between the dates of the ancient samples and the parameters of the model by plotting t_1 and t_2 against the mean sample date of all samples in that population (Figure 4B, C). We expected to find that t_1 correlated with sample age, under the assumption that samples were members of relatively short-lived populations that diverged from the “main-stem” of CEU ancestry. Instead, we see no correlation between t_1 and sample time, suggesting that the relationship of these populations to the CEU is complicated and not summarized well by the age of the samples. On the other hand, we see a strong positive correlation between t_2 and sampling time ($p < 1 \times 10^{-4}$). Because t_2 is a compound parameter, it is difficult to directly interpret this relationship. However, it is consistent with the most ancient samples belonging to populations with the smallest effective sizes, consistent with previous observations (CITE).

We obtained estimates of sequencing error in each of our analyzed ancient individuals. We found that almost all analyzed individuals had inferred error rates less than 1%, with many individuals clustering near 0.5% error (Figure 5A). We hypothesized that higher coverage individuals may have lower error rates, because high coverage may indicate better preserved DNA. We found a marginally significant negative correlation the log of the error rate and the log of coverage when two individual with inferred error rates of 0% were excluded ($p = 0.06$, Figure 5B).

[Figure 5 about here.]

Finally, we examined the impact of grouping individuals into populations in real data. We see that estimates of t_1 for low coverage samples are typically lower than when they are pooled with other individuals of that population (Figure 6A), suggesting a slightly downward bias in estimating t_1 for low coverage samples. On the other hand, there is substantial bias toward overestimating t_2 , particularly for very low coverage samples (Figure 6B). This again shows that for estimates that rely on heterozygosity in ancient populations, pooling many low coverage individuals can significantly improve estimates.

[Figure 6 about here.]

3 Discussion

4 Methods

4.1 Computing allele frequency moments in the ancient population

We wish to compute moments of the form

$$\mathbb{E}_x(g(Y); t_1, t_2) = \int_0^1 g(y) f(y; x, t_1, t_2) dy. \quad (4)$$

To do so, we make use of several results from diffusion theory. To ensure that this paper is self contained, we briefly review those results here. The interested reader may find much of this material covered in (EWENS, KARLIN AND TAYLOR). Several similar calculations can be found in (GRIFFITHS).

Let the probability of an allele going from frequency x to frequency y in τ generations in a population of size N_e be $f(x, y; t)$, where $t = \tau/(2N_e)$. Under a wide variety of models, the change in allele frequencies through time is well approximated by the Wright-Fisher diffusion, which is characterized by its generator,

$$\mathcal{L} = \frac{1}{2}x(1-x)\frac{d^2}{dx^2}.$$

The generator of a diffusion process is useful, because it can be used to define a differential equation for the moments of that process,

$$\frac{d}{dt}\mathbb{E}_x(g(X_t)) = \mathbb{E}_x(\mathcal{L}g(X_t)). \quad (5)$$

We will require the *speed measure* of the Wright-Fisher diffusion, $m(x) = x^{-1}(1-x)^{-1}$, which essentially describes how slow a diffusion at position x is “moving” compared to a Brownian motion at position x . Note that all diffusions are reversible with respect to their speed measures, i.e.

$$m(x)f(x, y; t) = m(y)f(y, x; t).$$

We additionally require the probability of loss, i.e. the probability that the allele currently at frequency x is ultimately lost from the population. This is

$$u_0(x) = 1 - x.$$

Note that it is possible to condition the Wright-Fisher diffusion to eventually be lost. The transition density can be computed as

$$f^\downarrow(x, y; t) = f(x, y; t) \frac{u_0(y)}{u_0(x)}$$

by using Bayes theorem. The diffusion conditioned on loss is characterized by its generator,

$$\mathcal{L}^\downarrow = -x \frac{d}{dx} + \frac{1}{2} x(1-x) \frac{d^2}{dx^2}.$$

In an infinite sites model, in which mutations occur at the times of a Poisson process with rate $\theta/2$ and then each drift according to the Wright-Fisher diffusion, a quasi-equilibrium distribution will be reached, known as the frequency spectrum. The frequency spectrum, $\phi(x)$, predicts the number of sites at frequency x , and can be written in terms of the speed measure and the probability of loss,

$$\phi(x) = \theta m(x) u_0(x).$$

To proceed with calculating (4), note that the conditional probability of an allele being at frequency y in the ancient population given that it's at frequency x in the modern population can be calculated

$$f(y; x, t_1, t_2) = \frac{f(x, y; t_1, t_2)}{\phi(x)}$$

where $f(x, y; t_1, t_2)$ is the joint probability of the allele frequencies in the modern and ancient populations and $\phi(x)$ is the frequency spectrum in the modern population.

Assuming that the ancestral population of the modern and ancient samples was at equilibrium, the joint distribution of allele frequencies can be computed by sampling alleles from the frequency spectrum of the ancestor and evolving them forward in time via the Wright-Fisher diffusion. This can be written mathematically as

$$f(x, y; t_1, t_2) = \int_0^1 f(z, x; t_1) f(z, y; t_2) \phi(z) dz.$$

We now expand the frequency spectrum in terms of the speed measure and the probability of loss and use reversibility with respect to the speed measure to rewrite the equation,

$$\begin{aligned} \int_0^1 f(z, x; t_1) f(z, y; t_2) \phi(z) dz &= \theta \int_0^1 f(z, x; t_1) f(z, y; t_2) m(z) u_0(z) dz \\ &= \theta \int_0^1 \frac{m(x)}{m(z)} f(x, z; t_1) f(z, y; t_2) m(z) u_0(z) dz \\ &= \theta m(x) u_0(x) \int_0^1 f(x, z; t_1) \frac{u_0(z)}{u_0(x)} f(z, y; t_2) dz \\ &= \phi(x) \int_0^1 f^\downarrow(x, z; t_1) f(z, y; t_2) dz. \end{aligned}$$

The third line follows by multiplying by $u_0(x)/u_0(x) = 1$. This equation has the interpretation of sampling an allele from the frequency spectrum in the modern population, then

evolving it *backward* in time to the common ancestor, before evolving it *forward* in time to the ancient population. The interpretation of the diffusion conditioned on loss as evolving backward in time arises by considering the fact that alleles arose from unique mutations at some point in the past; hence, looking backward, alleles must eventually be lost at some point in the past.

To compute the expectation, we substitute this form for the joint probability into (4),

$$\begin{aligned}\int_0^1 g(y) f(y; x, t_1, t_2) dy &= \int_0^1 g(y) \left(\int_0^1 f^\downarrow(x, z; t_1) f(z, y; t_2) dz \right) dy \\ &= \int_0^1 \left(\int_0^1 g(y) f(z, y; t_2) dy \right) f^\downarrow(x, z; t_1) dz,\end{aligned}$$

where the second line follows by rearranging terms and exchanging the order of integration. Note that this formula takes the form of nested expectations. Specifically,

$$\begin{aligned}\int_0^1 g(y) f(z, y; t_2) dy &= \mathbb{E}_z(g(Y_{t_2})) \\ &\equiv h(z)\end{aligned}$$

and

$$\begin{aligned}\int_0^1 h(z) f^\downarrow(x, z; t_1) dz &= \mathbb{E}_x^\downarrow(h(Z_{t_1})) \\ &= \mathbb{E}_x(g(Y); t_1, t_2).\end{aligned}$$

We now use (5) to note that

$$\frac{d}{dt} p_{n,k} = \frac{k(k-1)}{2} p_{n,k-1} - k(n-k) p_{n,k} + \frac{(n-k)(n-k-1)}{2} p_{n,k+1}$$

and

$$\frac{d}{dt} p_{n,k}^\downarrow = \frac{k(k-1)}{2} p_{n,k-1}^\downarrow - k(n-k+1) p_{n,k}^\downarrow + \frac{(n-k+1)(n-k)}{2} p_{n,k+1}^\downarrow$$

with obvious boundary conditions $p_{n,k}(0; z) = z^k(1-z)^{n-k}$ and $p_{n,k}^\downarrow(0; x) = x^k(1-x)^{n-k}$.

These systems of differential equations can be rewritten as matrix differential equations with coefficient matrices Q and Q^\downarrow respectively. Because they are linear, first order equations, they can be solved by matrix exponentiation. Because the expectation of a polynomial in the Wright-Fisher diffusion remains a polynomial, the nested expectations can be computed via matrix multiplication of the solutions to these differential equations, yielding the formula (2).

4.2 Sites covered exactly once have no information about drift in the ancient population

Consider a simplified model in which each site has exactly one read. When we have sequence from only a single individual, we have a set l_a of sites where the single read is an ancestral allele and a set l_d of sites where the single read is a derived allele. Thus, we can rewrite (3) as

$$L(D) = \prod_{l \in l_a} \left((1 - \epsilon)P_{2,0}(x_l) + \frac{1}{2}P_{2,1}(x_l) + \epsilon P_{2,2}(x_l) \right) \prod_{l \in l_d} \left(\epsilon P_{2,0}(x_l) + \frac{1}{2}P_{2,1}(x_l) + (1 - \epsilon)P_{2,2}(x_l) \right).$$

We can use formulas from (FERNANDO) to compute $P_{2,k}(x_l)$ for $k \in \{0, 1, 2\}$,

$$\begin{aligned} P_{2,0}(x_l) &= 1 - x_l e^{-t_1} - \frac{1}{2}x_l e^{-(t_1+t_2)} + x_l \left(x_l - \frac{1}{2} \right) e^{-(3t_1+t_2)} \\ P_{2,1}(x_l) &= x_l e^{-(t_1+t_2)} + x_l (1 - 2x_l) e^{-(3t_1+t_2)} \\ P_{2,2}(x_l) &= x_l e^{-t_1} - \frac{1}{2}x_l e^{-(t_1+t_2)} + x_l \left(x_l - \frac{1}{2} \right) e^{-(3t_1-t_2)}. \end{aligned}$$

Note then that

$$(1 - \epsilon)P_{2,0}(x_l) + \frac{1}{2}P_{2,1}(x_l) + \epsilon P_{2,2}(x_l) = 1 - \epsilon - x(1 - 2\epsilon)e^{-t_1}$$

and

$$\epsilon P_{2,0}(x_l) + \frac{1}{2}P_{2,1}(x_l) + (1 - \epsilon)P_{2,2}(x_l) = \epsilon + x(1 - 2\epsilon)e^{-t_1}.$$

Neither of these formulas depend on t_2 ; hence, there is no information about the drift time in the ancient population from data that is exactly 1x coverage.

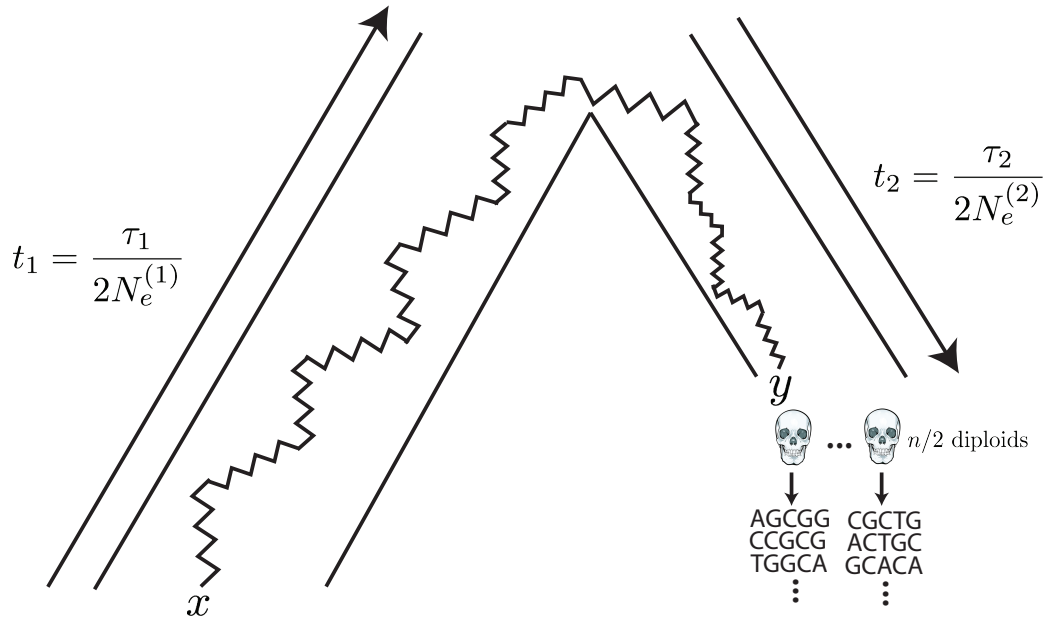


Figure 1: The generative model. Alleles are found at frequency x in the modern population and are at frequency y in the ancient population. The modern population has effective size $N_e^{(1)}$ and has evolved for τ_1 generations since the common ancestor of the modern and ancient populations, while the ancient population is of size $N_e^{(2)}$ and has evolved for τ_2 generations. Ancient diploid samples are taken and sequenced to possibly low coverage, with errors. Arrows indicate that the sampling probability can be calculated by evolving alleles *backward* in time from the modern population and then forward in time to the ancient population.

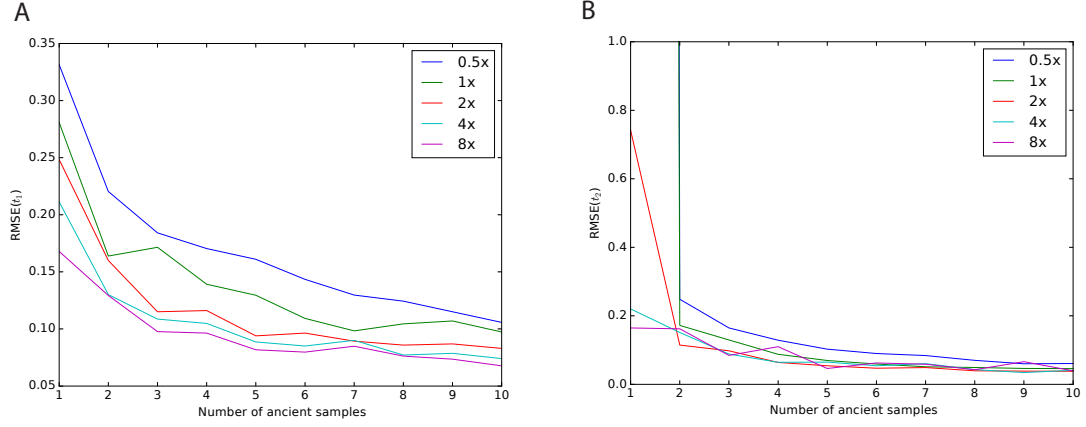


Figure 2: Impact of sampling scheme on parameter estimation error. In each panel, the x axis represents the number of simulated ancient samples, while the y axis shows the relative root mean square error for each parameter. Each different line corresponds to individuals sequenced to different depth of coverage. Panel A shows results for t_1 while panel B shows results for t_2 . Simulated parameters are $t_1 = 0.02$ and $t_2 = 0.05$.

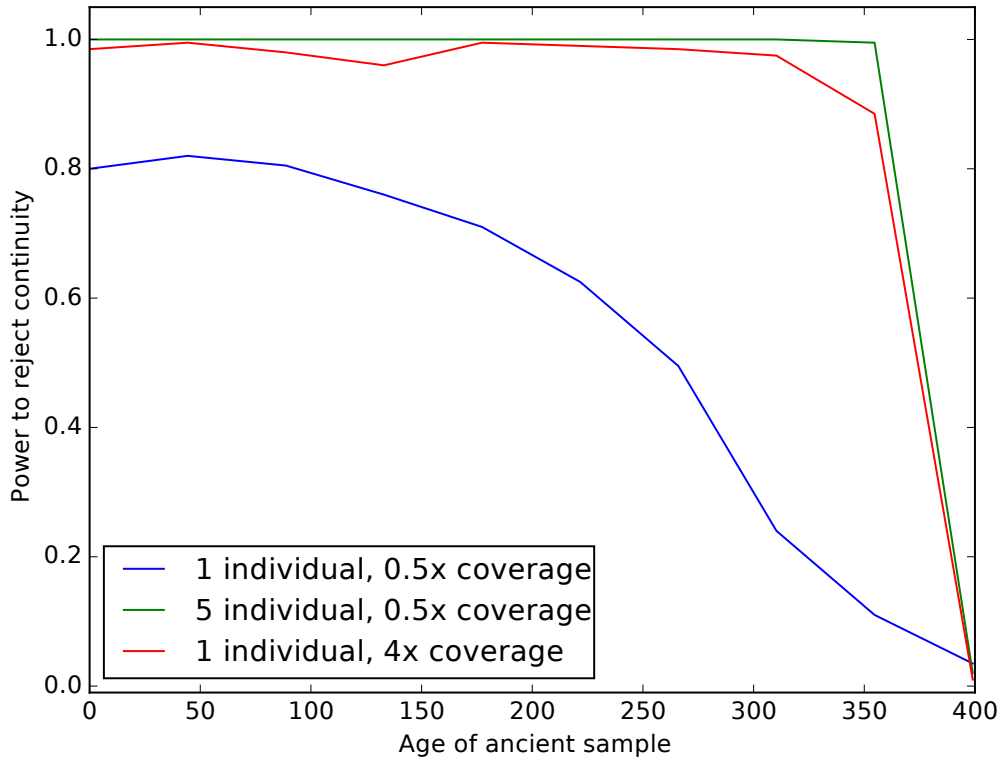


Figure 3: Impact of sampling scheme on rejecting population continuity. The x axis represents the age of the ancient sample in generations, with 0 indicating a modern sample and 400 indicating a sample from exactly at the split time 400 generations ago. The y axis shows the proportion of simulations in which we rejected the null hypothesis of population continuity. Each line shows different sampling schemes, as explained in the legend.

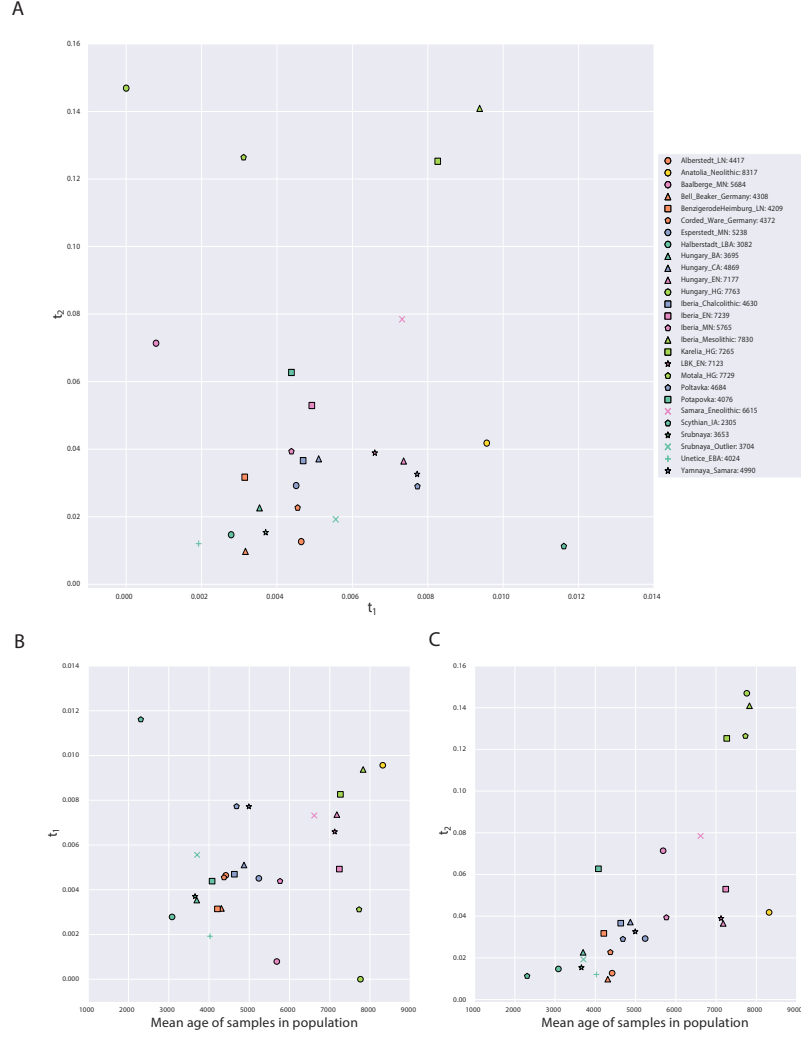


Figure 4: Parameters of the model inferred from ancient West Eurasian samples. Panel A shows t_1 on the x-axis and t_2 on the y-axis, with each point corresponding to a population as indicated in the legend. Numbers in the legend correspond to the mean date of all samples in the population. Panels B and C show scatterplots of the mean age of the samples in the population (x-axis) against t_1 and t_2 , respectively. Points are described by the same legend as Panel A.

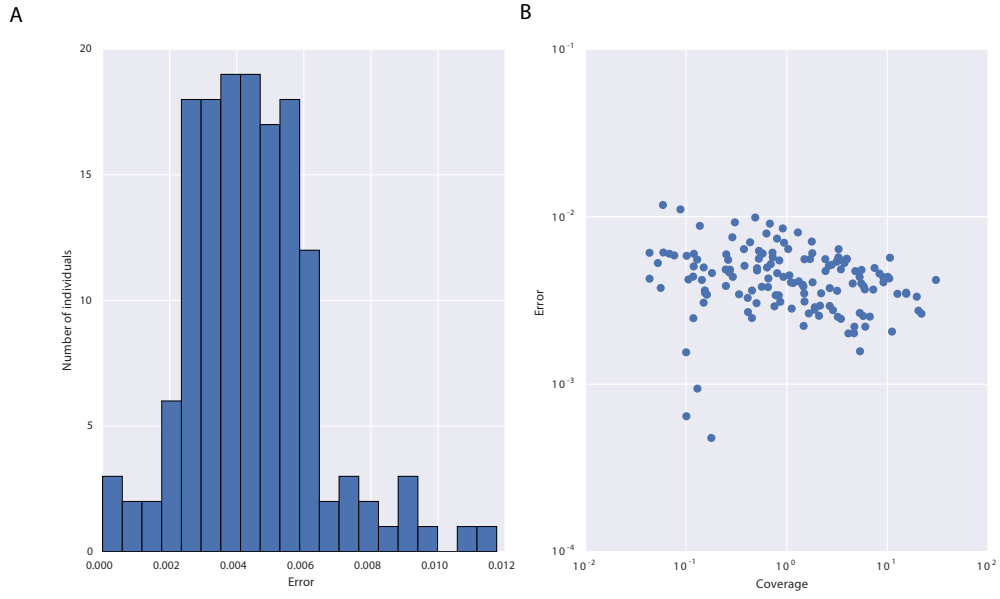


Figure 5: Properties of sequencing error inferred from ancient samples. Panel A shows a histogram of error rates estimated across all samples. Panel B is a scatterplot of coverage (x-axis) against inferred error rate (y-axis) for all individuals, except two individuals for whom an error rate of 0 was estimated.

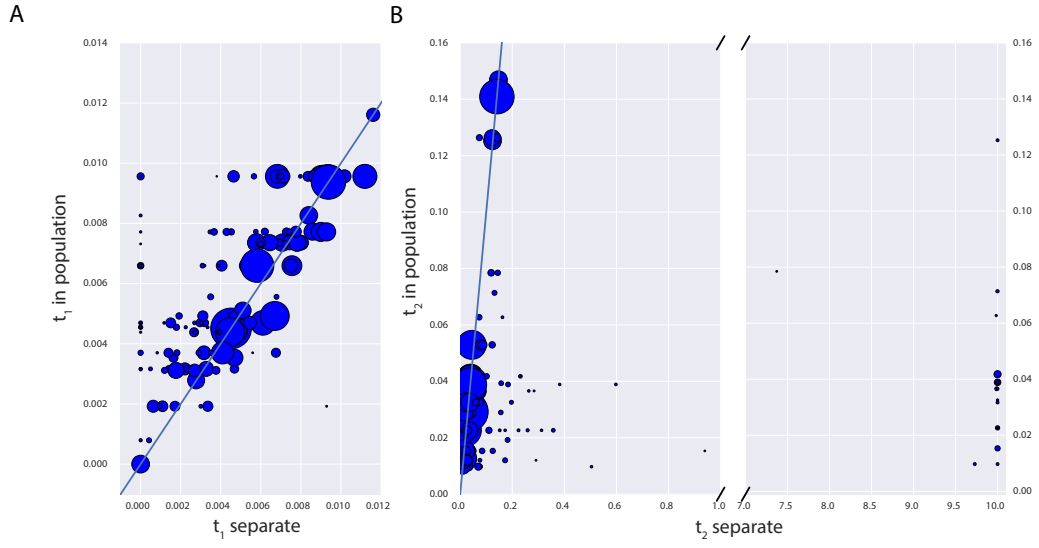


Figure 6: Impact of pooling individuals into populations when estimating model parameters. In both panels, the x-axis indicates the parameter estimate when individuals are analyzed separately, while the y-axis indicates the parameter estimate when individuals are grouped into populations. Panel A reports the impact on estimation of t_1 , while Panel B reports the impact on t_2 . Note that Panel B has a broken x-axis. Solid lines in each figure indicate $y = x$.

pop	cov	date	t_1	t_2	lnL	t_1 (cont)	lnL (cont)
Alberstedt_LN	12.606	4417.000	0.005	0.013	-779411.494	0.006	-779440.143
Anatolia_Neolithic	3.551	8317.500	0.010	0.042	-9096440.714	0.044	-9106156.877
Baalberge_MN	0.244	5684.333	0.001	0.071	-201575.306	0.007	-201750.419
Bell_Beaker_Germany	1.161	4308.444	0.003	0.010	-1834486.744	0.008	-1834652.858
BenzigerodeHeimbürg_LN	0.798	4209.750	0.003	0.032	-346061.545	0.007	-346134.356
Corded_Ware_Germany	2.250	4372.833	0.005	0.023	-2139002.723	0.017	-2139858.192
Esperstedt_MN	30.410	5238.000	0.005	0.029	-975890.329	0.009	-976047.889
Halberstadt_LBA	5.322	3082.000	0.003	0.015	-558966.522	0.004	-558993.078
Hungary_BA	3.401	3695.750	0.004	0.023	-789754.969	0.010	-789939.889
Hungary_CA	5.169	4869.500	0.005	0.037	-504413.094	0.010	-504549.603
Hungary_EN	4.033	7177.000	0.007	0.036	-3478429.262	0.033	-3481855.461
Hungary_HG	5.807	7763.000	0.000	0.147	-469887.471	0.015	-471652.083
Iberia_Chalcolithic	1.686	4630.625	0.005	0.037	-2351769.869	0.028	-2354249.543
Iberia_EN	4.875	7239.500	0.005	0.053	-1483274.628	0.030	-1485675.934
Iberia_MN	5.458	5765.000	0.004	0.039	-1491407.962	0.023	-1492793.179
Iberia_Mesolithic	21.838	7830.000	0.009	0.141	-720759.133	0.030	-723091.935
Karelia_HG	2.953	7265.000	0.008	0.125	-652952.676	0.033	-655352.439
LBK_EN	2.894	7123.429	0.007	0.039	-3656617.954	0.033	-3660838.639
Motala_HG	2.207	7729.500	0.003	0.126	-1477338.076	0.068	-1489573.895
Poltavka	2.211	4684.500	0.008	0.029	-1334662.071	0.020	-1335358.630
Potapovka	0.267	4076.500	0.004	0.063	-220112.816	0.011	-220251.379
Samara_Eneolithic	0.463	6615.000	0.007	0.078	-362161.674	0.020	-362689.209
Scythian_IA	3.217	2305.000	0.012	0.011	-492961.306	0.013	-492973.694
Srubnaya	1.662	3653.273	0.004	0.015	-2578065.957	0.013	-2578645.731
Srubnaya_Outlier	0.542	3704.500	0.006	0.019	-285828.766	0.008	-285851.523
Unetice_EBA	1.320	4024.786	0.002	0.012	-1676798.610	0.008	-1677026.310
Yamnaya_Samara	1.937	4990.500	0.008	0.033	-2440183.354	0.028	-2442192.801

Table 1: Details of populations included in analysis. “pop” is population name, “cov” is mean coverage of individuals in the population, “date” is mean date of individuals in the population, “ t_1 ” is the maximum likelihood estimate of t_1 in the full model, “ t_2 ” is the maximum likelihood estimate of t_2 in the full model, “LnL” is the maximum likelihood value in the full model, “ t_1 (cont)” is the maximum likelihood estimate of t_1 in the model where $t_2 = 0$, “LnL” is the maximum likelihood value in the model where $t_2 = 0$.