

# 1 Assessing the relationship of ancient and modern populations

2 Joshua G. Schraiber

3 Started on October 22, 2016. Compiled on March 3, 2017

## 4 **Abstract**

5 Genetic material sequenced from ancient samples is revolutionizing our understand-  
6 ing of the recent evolutionary past. However, ancient DNA is often degraded, resulting  
7 in low coverage, error-prone sequencing. Several solutions exist to this problem, rang-  
8 ing from simple approach such as selecting a read at random for each site to more  
9 complicated approaches involving genotype likelihoods. In this work, we present a  
10 novel method for assessing the relationship of an ancient sample with a modern pop-  
11 ulation while accounting for sequencing error by analyzing raw read from multiple  
12 ancient individuals simultaneously. We show that when analyzing SNP data, it is bet-  
13 ter to sequencing more ancient samples to low coverage: two samples sequenced to  
14 0.5x coverage provide better resolution than a single sample sequenced to 2x coverage.  
15 We also examined the power to detect whether an ancient sample is directly ancestral  
16 to a modern population, finding that with even a few high coverage individuals, even  
17 ancient samples that are very slightly diverged from the modern population can be de-  
18 tected with ease. When we applied our approach to European samples, we found that  
19 no ancient samples represent direct ancestors of modern Europeans. We also found  
20 that, as shown previously, the most ancient Europeans appear to have had the small-  
21 est effective population sizes, indicating a role for agriculture in modern population  
22 growth.

23

## 1 Introduction

24

25

26

27

28

29

30

31

Ancient DNA (aDNA) is now ubiquitous in population genetics. Advances in DNA isolation [Dabney et al., 2013], library preparation [Meyer et al., 2012], bone sampling [Pinhasi et al., 2015], and sequence capture Haak et al. [2015] make it possible to obtain genome-wide data from hundreds of samples [Haak et al., 2015, Mathieson et al., 2015, Allentoft et al., 2015, Fu et al., 2016]. Analysis of these data can provide new insight into recent evolutionary processes which leave faint signatures in modern genomes, including natural selection [Schraiber et al., 2016, Jewett et al., 2016] and population replacement [Sjödín et al., 2014, Lazaridis et al., 2014].

32

33

34

35

36

37

38

39

40

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 [Lazaridis et al., 2016]. For instance, recent work showed that modern Native Americans are directly descended from the Clovis culture that inhabited North America over 10,000 years ago, settling a long-standing debate about the colonization of the Americas [Rasmussen et al., 2014].

41

42

43

44

45

46

47

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 [Sawyer et al., 2012]. Thus, ancient data is frequently sequenced to low coverage and has a significantly higher base-calling error rate than modern samples. When working with diploid data, as in aDNA extracted from plants and animals, the low coverage prevents genotypes from being called with confidence.

48

49

50

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 [Haak et al., 2015, Mathieson et al., 2015, Allentoft

51 et al., 2015, Fu et al., 2016, Lazaridis et al., 2016]. Many common approaches to the  
52 analyses of ancient DNA, such as the usage of F-statistics [Patterson et al., 2012], are  
53 designed with this kind of dataset in mind. As shown by Peter [2016], F-statistics can  
54 be interpreted as linear combinations of simpler summary statistics and can often be  
55 understood in terms of testing a tree-like structure relating populations. Nonetheless,  
56 despite the simplicity and appeal of this approach, it has several drawbacks. Primarily,  
57 it throws away reads from sites that are covered more than once, resulting in a potential  
58 loss of information from expensive, difficult-to-acquire data. These approach are also  
59 strongly impacted by sequencing error and contamination.

60 On the other hand, several approaches exist to either work with genotype likelihoods  
61 or the raw read data. Genotype likelihoods are the probabilities of the read data at a  
62 site given each of the three possible diploid genotypes at that site. They can be used  
63 in calculation of population genetic statistics or likelihood functions to average over  
64 uncertainty in the genotype [Korneliussen et al., 2014]. However, many such approaches  
65 assume that genotype likelihoods are fixed by the SNP calling algorithm. However, with  
66 low coverage data, an increase in accuracy is expected if genotype likelihoods are co-  
67 estimated with other parameters of interest, due to the covariation between processes  
68 that influence read quality and genetic diversity, such as contamination.

69 A recent method that coestimates demographic parameters along with error and  
70 contamination rates by using genotype likelihoods showed that there can be significant  
71 power to assess the relationship of a single ancient sample to a modern population  
72 [Racimo et al., 2016]. Nonetheless, they found that for very low coverage data, infer-  
73 ences were not reliable. Thus, they were unable to apply their method to the large  
74 number of extremely low coverage ( $< 1x$ ) genomes that are available. Moreover, they  
75 were unable to explore the tradeoffs that come with a limited budget: can we learn  
76 more by sequencing fewer individuals to high coverage, or more individuals at lower  
77 coverage?

78 Here, we develop a novel maximum likelihood approach for analyzing low coverage  
79 ancient DNA in relation to a modern population. We work directly with raw read data

80 and explicitly model base-calling errors. Crucially, our approach incorporates data  
81 from multiple individuals that belong to the same ancient population, which we show  
82 substantially increases power and reduces error in parameter estimates. We then apply  
83 our new methodology to ancient human data, and show that we can perform accurate  
84 demographic inference even from very low coverage samples by analyzing them jointly.

## 85 2 Results

### 86 2.1 Sampling alleles in ancient populations

87 We assume a scenario in which allele frequencies are known with high accuracy in a  
88 modern population. Suppose that an allele is known to be at frequency  $x \in (0, 1)$  in  
89 the modern population, and we wish to compute the probability of obtaining  $k$  copies  
90 of that allele in a sample of  $n$  ( $0 \leq k \leq n$ ) chromosomes from an ancient population.  
91 Conditioning on the frequency of the allele in the modern population minimizes the  
92 impact of ascertainment, and allows this approach to be used for SNP capture data.

To calculate the sampling probability, we assume a simple demographic model in which the ancient individual belongs to a population that split off from the modern population  $\tau_1$  generations ago, and subsequently existed as an isolated population for  $\tau_2$  generations. 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  $x$ , denoted  $f(y; x, t_1, t_2)$ , then the sampling probability is simply an integral,

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

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

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

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

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

98 and

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

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

103 [Figure 1 about here.]

104 Because of the fragmentation and degradation of DNA that is inherent in obtaining  
 105 sequence data from ancient individuals, it is difficult to obtain the high coverage data  
 106 necessary to make high quality genotype calls from ancient individuals. To address this,  
 107 we instead work directly with raw read data, and average over all the possible genotypes  
 108 weighted by their probability of producing the data. Specifically, we follow Nielsen et al.

109 [2012] in modeling the probability of the read data in the ancient population, given the  
 110 allele frequency at site  $l$  as

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

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

117 Thus,

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

with

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

118 .

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

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

## 121 2.2 Impact of coverage and number of samples on inferences

122 To explore the tradeoff of sequencing more individuals at lower depth compared to fewer  
123 individuals at higher coverage, we performed simulations using `msprime` [Kelleher et al.,  
124 2016] combined with custom scripts to simulate base calling error and low coverage  
125 data. First, we examined the impact of coverage and number of samples on the ability  
126 to recover the drift times in the modern and the ancient populations. Figure 2 shows  
127 results for data simulated with  $t_1 = 0.02$  and  $t_2 = 0.05$ , corresponding to an ancient  
128 individual who died 300 generations ago from population of effective size 1000. The  
129 populations split 400 generations ago, and the modern population has an effective  
130 size of 10000. We simulated approximately 180000 SNPs by simulating 100000 500  
131 base pair fragments. Inferences of  $t_1$  can be relatively accurate even with only one  
132 low coverage ancient sample (Figure 2A). However, inferences of  $t_2$  benefit much more  
133 from increasing the number of ancient samples, as opposed to coverage (Figure 2B). In  
134 particular, two individuals sequenced to 0.5x coverage have a much lower error than a  
135 single individual sequenced to 2x coverage. To explore this effect further, we derived the  
136 sampling probability of alleles covered by exactly one sequencing read (see Methods).  
137 We found that sites covered only once have no information about  $t_2$ , suggesting that  
138 evidence of heterozygosity is very important for inferences about  $t_2$ .

139 [Figure 2 about here.]

140 We next examined the impact of coverage and sampling on the power to reject  
141 the hypothesis that the ancient individuals came from a population that is directly  
142 ancestral to the modern population. We analyzed both low coverage (0.5x) and higher  
143 coverage (4x) datasets consisting of 1 (for both low and high coverage samples) or 5  
144 individuals (only for low coverage). We simulated data with parameters identical to  
145 the previous experiment, except we now examined the impact of varying the age of  
146 the ancient sample from 0 generations ago through to the split time with the modern  
147 population. We then performed a likelihood ratio test comparing the null model of con-  
148 tinuity, in which  $t_2 = 0$ , to a model in which the ancient population is not continuous.

149 Figure 3 shows the power of the likelihood ratio test. For a single individual sequenced  
150 to low coverage, we see that the test only has power for very recently sampled ancient  
151 individuals (i.e. samples that are highly diverged from the modern population). How-  
152 ever, the power increases dramatically as the number of individuals or the coverage per  
153 individual is increased; sequencing 5 individuals to 0.5x coverage results in essentially  
154 perfect power to reject continuity. Nonetheless, for samples that are very close to the  
155 divergence time, it will be difficult to determine if they are ancestral to the modern  
156 population or not, because differentiation is incomplete.

157 [Figure 3 about here.]

### 158 **2.3 Application to ancient humans**

159 We applied our approach to ancient human data from Mathieson et al. [2015], which  
160 is primarily derived from a SNP capture approach that targeted 1.2 million SNPs.  
161 Based on sampling location and associated archeological materials, the individuals  
162 were grouped into *a priori* panels, which we used to specify population membership  
163 when analyzing individuals together. We analyzed all samples for their relationship to  
164 the CEU individuals from the 1000 Genomes Project [Consortium, 2015]. Based on  
165 our results that suggested that extremely low coverage samples would yield unreliable  
166 estimates, we excluded panels that are composed of only a single individual sequenced  
167 to less than 2x coverage.

168 We computed maximum likelihood estimates of  $t_1$  and  $t_2$  for individuals as grouped  
169 into populations (Figure 4A; Table 1). We observe that  $t_2$  is significantly greater than 0  
170 for all populations. Thus, none of these populations are consistent with directly making  
171 up a large proportion of the ancestry of modern CEU individuals. Strikingly, we see  
172 that  $t_2 \gg t_1$ , despite the fact that the ancient samples must have existed for fewer  
173 generations since the population split than the modern samples. This suggests that  
174 all of the ancient populations are characterized by extremely small effective population  
175 sizes.

176 [Table 1 about here.]

177 [Figure 4 about here.]

178 We further explored the relationship between the dates of the ancient samples and  
179 the parameters of the model by plotting  $t_1$  and  $t_2$  against the mean sample date of  
180 all samples in that population (Figure 4B, C). We expected to find that  $t_1$  correlated  
181 with sample age, under the assumption that samples were members of relatively short-  
182 lived populations that diverged from the “main-stem” of CEU ancestry. Instead, we  
183 see no correlation between  $t_1$  and sample time, suggesting that the relationship of  
184 these populations to the CEU is complicated and not summarized well by the age  
185 of the samples. On the other hand, we see a strong positive correlation between  
186  $t_2$  and sampling time ( $p < 1 \times 10^{-4}$ ). Because  $t_2$  is a compound parameter, it is  
187 difficult to directly interpret this relationship. However, it is consistent with the most  
188 ancient samples belonging to populations with the smallest effective sizes, consistent  
189 with previous observations [Skoglund et al., 2014].

190 Our model provided estimates of sequencing error in each of our analyzed ancient  
191 individuals. We found that almost all analyzed individuals had inferred error rates  
192 less than 1%, with many individuals clustering near 0.5% error (Figure 5A). We hy-  
193 pothesized that higher coverage individuals may have lower error rates, because high  
194 coverage may indicate better preserved DNA. We found a marginally significant neg-  
195 ative correlation between the log of the error rate and the log of coverage when two  
196 outlier individuals with inferred error rates of 0% were excluded ( $p = 0.06$ , Figure 5B).

197 [Figure 5 about here.]

198 Finally, we examined the impact of grouping individuals into populations in real  
199 data. We see that estimates of  $t_1$  for low coverage samples are typically lower when an-  
200 alyzed individually than when pooled with other individuals of the same panel (Figure  
201 6A), suggesting a slightly downward bias in estimating  $t_1$  for low coverage samples. On  
202 the other hand, there is substantial bias toward overestimating  $t_2$  when analyzing sam-

203 ples individually, particularly for very low coverage samples (Figure 6B). This again  
204 shows that for estimates that rely on heterozygosity in ancient populations, pooling  
205 many low coverage individuals can significantly improve estimates.

206 [Figure 6 about here.]

### 207 **3 Discussion**

208 Ancient DNA (aDNA) presents unique opportunities to enhance our understanding  
209 of demography and selection in recent history. However, it also comes equipped with  
210 several challenges, due to postmortem DNA damage [Sawyer et al., 2012]. Several  
211 strategies have been developed to deal with the low quality of aDNA data, from rela-  
212 tively simple options like sampling a read at random at every site [Green et al., 2010]  
213 to more complicated methods making use of genotype likelihoods [Racimo et al., 2016].  
214 Here, we presented a novel maximum likelihood approach for making inferences about  
215 how ancient populations are related to modern populations by analyzing read counts  
216 from multiple ancient individuals and explicitly modeling relationship between the two  
217 populations. Using this approach, we examined some aspects of sampling strategy for  
218 aDNA analysis and we applied our approach to ancient humans.

219 We found that sequencing many individuals from an ancient population to low cov-  
220 erage ( .5-1x) can be a significantly more cost effective strategy than sequencing fewer  
221 individuals to relatively high coverage. For instance, we saw from simulations that far  
222 more accurate estimates of the drift time in an ancient population can be obtained by  
223 pooling 2 individuals at 0.5x coverage than by sequencing a single individual to 2x cov-  
224 erage (Figure 2). We saw this replicated in our analysis of the real data: low coverage  
225 individuals showed a significant amount of variation and bias in estimating the model  
226 parameters that was substantially reduced when individuals were analyzed jointly in a  
227 population (Figure 6). To explore this further, we showed that sites sequenced to 1x  
228 coverage in a single individual retain no information about the drift time in the ancient  
229 population. This can be intuitively understood because the drift time in the ancient

230 population is strongly related the amount of heterozygosity in the ancient population:  
231 an ancient population with a longer drift time will have lower heterozygosity at sites  
232 shared with a modern population. When a site is only sequenced once in a single indi-  
233 vidual, there is no information about the heterozygosity of that site. We also observed  
234 a pronounced upward bias in estimates of the drift time in the ancient population from  
235 low coverage samples. We speculate that this is due to the presence of few sites covered  
236 more than once being likely to be homozygous, thus deflating the estimate of heterozy-  
237 gosity in the ancient population. Thus, for analysis of SNP data, we recommend that  
238 aDNA sampling be conducted to maximize the number of individuals from each ancient  
239 population that can be sequenced to  $\sim 1x$ , rather than attempting to sequence fewer  
240 individuals to high coverage.

241 Of particular interest in many studies of ancient populations is the question of  
242 direct ancestry: are the ancient samples members of a population that contributed  
243 substantially to a modern population? Several methods have been proposed to test this  
244 question, but thus far they have been limited to many individuals sequenced at a single  
245 locus [Sjödín et al., 2014] or to a single individual with genome-wide data [Rasmussen  
246 et al., 2014]. Our approach provides a rigorous, maximum likelihood framework for  
247 testing questions of population continuity using multiple low coverage ancient samples.  
248 We saw from simulations (Figure 3) that data from single, low coverage individuals  
249 result in very little power to reject the null hypothesis of continuity unless the ancient  
250 sample is very recent (i.e. it has been diverged from the modern population for a long  
251 time). Nonetheless, when low coverage individuals are pooled together, or a single  
252 high coverage individual is used, there is substantial power to reject continuity for all  
253 but the most ancient samples (i.e. samples dating from very near the population split  
254 time).

255 When we applied our approach to European history, we made several noteworthy  
256 observations. Primarily, we rejected continuity for all populations that we analyzed,  
257 suggesting that most ancient samples thus far collected come from populations that  
258 had diverged from main-stem European ancestry, and did not contribute a substantial

259 portion of ancestry directly to modern Europeans. Secondly, we observed that the  
260 drift time in the ancient population was much larger than the drift time in the modern  
261 population. This supports the observation that ancient Europeans were often members  
262 of small, isolated populations [Skoglund et al., 2014]. We obtained further support for  
263 this by examining the relationship between the age of the ancient populations and  
264 the drift time leading to them (Figure 4C). We saw a strong positive correlation, and  
265 although this drift time is a compound parameter, which complicates interpretations, it  
266 appears that the oldest Europeans were members of extremely small populations, and  
267 that effective population size has grown through time as agriculture spread through  
268 Europe.

269 We anticipate the further development of methods that explicitly account for dif-  
270 ferential drift times in ancient and modern samples will become important as aDNA  
271 research becomes even more integrating into population genomics. This is because  
272 many common summary methods, such as the use of Structure [Pritchard et al., 2000]  
273 and Admixture [Alexander et al., 2009], are sensitive to differential amounts of drift  
274 between populations [Falush et al., 2016]. As we've shown in ancient Europeans, an-  
275 cient samples tend to come from isolated subpopulations with a large amount of drift,  
276 thus confounding such summary approaches. Moreover, standard population genetics  
277 theory shows that allele frequencies are expected to be deterministically lower in an-  
278 cient samples, even if they are direct ancestors of a modern population. Intuitively,  
279 this arises because the alleles must have arisen at some point from new mutations, and  
280 thus were at lower frequencies in the past. A potentially fruitful avenue to combine  
281 these approaches moving forward may be to separate regions of the genome based on  
282 ancestry components, and assess the ancestry of ancient samples relative to specific  
283 ancestry components, rather than to genomes as a whole.

284 Our current approach leaves several avenues for improvement. Although our method  
285 is robust to non-constant demography because we consider only alleles that are seg-  
286 regating in both the modern and the ancient population, we are losing information  
287 by not modeling new mutations that arise in the ancient population. Similarly, we

288 only consider a single ancient population at a time, albeit with multiple samples. Ide-  
289 ally, ancient samples would be embedded in complex demographic models that include  
290 admixture, detailing their relationships to each other and to modern populations [Pat-  
291 terson et al., 2012, Lipson and Reich, 2017]. However, inference of such complex models  
292 is difficult, and though there has been some progress in simplified cases [Lipson et al.,  
293 2014, Pickrell and Pritchard, 2012], it remains an open problem due to the difficulty of  
294 simultaneously inferring a non-tree-like topology along with demographic parameters.  
295 Software such as *mom* [Kamm et al., 2016] that can compute the likelihood of SNP data  
296 in an admixture graph may be able to be used to integrate over genotype uncertainty  
297 in larger settings than considered here.

## 298 4 Methods

### 299 4.1 Computing allele frequency moments in the ancient popu- 300 lation

301 We wish to compute moments of the form

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

302 To do so, we make use of several results from diffusion theory. To ensure that this  
303 paper is self contained, we briefly review those results here. The interested reader may  
304 find much of this material covered in Ewens [2012], Karlin and Taylor [1981]. Several  
305 similar calculations can be found in Griffiths [2003].

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

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

310 The generator of a diffusion process is useful, because it can be used to define a differ-  
311 ential equation for the moments of that process,

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

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

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

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

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

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

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

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

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

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

326 written in terms of the speed measure and the probability of loss,

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

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

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

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

332 Assuming that the ancestral population of the modern and ancient samples was at  
333 equilibrium, the joint distribution of allele frequencies can be computed by sampling  
334 alleles from the frequency spectrum of the ancestor and evolving them forward in time  
335 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,

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

336 The third line follows by multiplying by  $u_0(x)/u_0(x) = 1$ . This equation has the inter-  
 337 pretation of sampling an allele from the frequency spectrum in the modern population,  
 338 then evolving it *backward* in time to the common ancestor, before evolving it *forward*  
 339 in time to the ancient population. The interpretation of the diffusion conditioned on  
 340 loss as evolving backward in time arises by considering the fact that alleles arose from  
 341 unique mutations at some point in the past; hence, looking backward, alleles must  
 342 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 inte-  
 gration. 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

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

343 We now use (5) to note that

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

344 and

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

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

347 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).

## 353 4.2 Sites covered exactly once have no information about drift 354 in the ancient population

355 Consider a simplified model in which each site has exactly one read. When we have  
356 sequence from only a single individual, we have a set  $l_a$  of sites where the single read is  
357 an ancestral allele and a set  $l_d$  of sites where the single read is a derived allele. Thus,  
358 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 Racimo et al. [2016] to compute  $P_{2,k}(x_l)$  for  $k \in \{0, 1, 2\}$ ,

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

$$P_{2,1}(x_l) = x_l e^{-(t_1+t_2)} + x_l(1-2x_l)e^{-(3t_1+t_2)}$$

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

359 Note then that

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

360 and

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

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

## 363 5 Software Availability

364 Python implementations of the described methods are available at [www.github.com/schraiber/continuity/](http://www.github.com/schraiber/continuity/)

## 365 6 Acknowledgments

366 We wish to thank Melinda Yang, Iain Mathieson, Pontus Skoglund, and Fernando  
367 Racimo for several discussions during the conception of this work that greatly improved  
368 its scope and rigor. We are grateful to Fernando Racimo and Benjamin Vernot for  
369 comments on early versions of this work that significantly improved its quality. We  
370 would also like to thank Tamara Broderick for several stimulating discussions about  
371 clustering that ultimately didn't result in any interesting application to ancient DNA.

## 372 References

373 David H Alexander, John Novembre, and Kenneth Lange. Fast model-based estimation  
374 of ancestry in unrelated individuals. *Genome research*, 19(9):1655–1664, 2009.

375 Morten E Allentoft, Martin Sikora, Karl-Göran Sjögren, Simon Rasmussen, Morten  
376 Rasmussen, Jesper Stenderup, Peter B Damgaard, Hannes Schroeder, Torbjörn

- 377 Ahlström, Lasse Vinner, et al. Population genomics of bronze age eurasia. *Nature*, 522(7555):167–172, 2015.  
378
- 379 1000 Genomes Project Consortium. A global reference for human genetic variation.  
380 *Nature*, 526(7571):68–74, 2015.
- 381 Jesse Dabney, Matthias Meyer, and Svante Pääbo. Ancient dna damage. *Cold Spring*  
382 *Harbor perspectives in biology*, 5(7):a012567, 2013.
- 383 Warren J Ewens. *Mathematical population genetics 1: theoretical introduction*, vol-  
384 ume 27. Springer Science & Business Media, 2012.
- 385 Daniel Falush, Lucy van Dorp, and Daniel Lawson. A tutorial on how (not) to over-  
386 interpret structure/admixture bar plots. *bioRxiv*, page 066431, 2016.
- 387 Qiaomei Fu, Cosimo Posth, Mateja Hajdinjak, Martin Petr, Swapan Mallick, Daniel  
388 Fernandes, Anja Furtwängler, Wolfgang Haak, Matthias Meyer, Alissa Mittnik, et al.  
389 The genetic history of ice age europe. *Nature*, 2016.
- 390 Richard E Green, Johannes Krause, Adrian W Briggs, Tomislav Maricic, Udo Stenzel,  
391 Martin Kircher, Nick Patterson, Heng Li, Weiwei Zhai, Markus Hsi-Yang Fritz, et al.  
392 A draft sequence of the neandertal genome. *science*, 328(5979):710–722, 2010.
- 393 RC Griffiths. The frequency spectrum of a mutation, and its age, in a general diffusion  
394 model. *Theoretical population biology*, 64(2):241–251, 2003.
- 395 Wolfgang Haak, Iosif Lazaridis, Nick Patterson, Nadin Rohland, Swapan Mallick,  
396 Bastien Llamas, Guido Brandt, Susanne Nordenfelt, Eadaoin Harney, Kristin Stew-  
397 ardsen, et al. Massive migration from the steppe was a source for indo-european  
398 languages in europe. *Nature*, 522(7555):207–211, 2015.
- 399 Ethan M Jewett, Matthias Steinrücken, and Yun S Song. The effects of population  
400 size histories on estimates of selection coefficients from time-series genetic data.  
401 *Molecular Biology and Evolution*, page msw173, 2016.

- 402 John A Kamm, Jonathan Terhorst, and Yun S Song. Efficient computation of the joint  
403 sample frequency spectra for multiple populations. *Journal of Computational and*  
404 *Graphical Statistics*, (just-accepted):1–37, 2016.
- 405 Samuel Karlin and Howard E Taylor. *A second course in stochastic processes*. Elsevier,  
406 1981.
- 407 Jerome Kelleher, Alison M Etheridge, and Gilean McVean. Efficient coalescent sim-  
408 ulation and genealogical analysis for large sample sizes. *PLoS Comput Biol*, 12(5):  
409 e1004842, 2016.
- 410 Thorfinn Sand Korneliussen, Anders Albrechtsen, and Rasmus Nielsen. Angsd: analysis  
411 of next generation sequencing data. *BMC bioinformatics*, 15(1):356, 2014.
- 412 Iosif Lazaridis, Nick Patterson, Alissa Mittnik, Gabriel Renaud, Swapan Mallick,  
413 Karola Kirsanow, Peter H Sudmant, Joshua G Schraiber, Sergi Castellano, Mark  
414 Lipson, et al. Ancient human genomes suggest three ancestral populations for  
415 present-day europeans. *Nature*, 513(7518):409–413, 2014.
- 416 Iosif Lazaridis, Dani Nadel, Gary Rollefson, Deborah C Merrett, Nadin Rohland, Swa-  
417 pan Mallick, Daniel Fernandes, Mario Novak, Beatriz Gamarra, Kendra Sirak, et al.  
418 Genomic insights into the origin of farming in the ancient near east. *Nature*, 536  
419 (7617):419–424, 2016.
- 420 Mark Lipson and David Reich. Working model of the deep relationships of diverse  
421 modern human genetic lineages outside of africa. *Molecular Biology and Evolution*,  
422 page msw293, 2017.
- 423 Mark Lipson, Po-Ru Loh, Nick Patterson, Priya Moorjani, Ying-Chin Ko, Mark  
424 Stoneking, Bonnie Berger, and David Reich. Reconstructing austronesian popu-  
425 lation history in island southeast asia. *Nature communications*, 5, 2014.
- 426 Iain Mathieson, Iosif Lazaridis, Nadin Rohland, Swapan Mallick, Nick Patterson,  
427 Songül Alpaslan Roodenberg, Eadaoin Harney, Kristin Stewardson, Daniel Fernan-

- 428 des, Mario Novak, et al. Genome-wide patterns of selection in 230 ancient eurasians.  
429 *Nature*, 528(7583):499–503, 2015.
- 430 Matthias Meyer, Martin Kircher, Marie-Theres Gansauge, Heng Li, Fernando Racimo,  
431 Swapan Mallick, Joshua G Schraiber, Flora Jay, Kay Prüfer, Cesare De Filippo, et al.  
432 A high-coverage genome sequence from an archaic denisovan individual. *Science*, 338  
433 (6104):222–226, 2012.
- 434 Rasmus Nielsen, Thorfinn Korneliussen, Anders Albrechtsen, Yingrui Li, and Jun  
435 Wang. Snp calling, genotype calling, and sample allele frequency estimation from  
436 new-generation sequencing data. *PloS one*, 7(7):e37558, 2012.
- 437 Nick Patterson, Priya Moorjani, Yontao Luo, Swapan Mallick, Nadin Rohland, Yiping  
438 Zhan, Teri Genschoreck, Teresa Webster, and David Reich. Ancient admixture in  
439 human history. *Genetics*, 192(3):1065–1093, 2012.
- 440 Benjamin M Peter. Admixture, population structure, and f-statistics. *Genetics*, 202  
441 (4):1485–1501, 2016.
- 442 Joseph K Pickrell and Jonathan K Pritchard. Inference of population splits and mix-  
443 tures from genome-wide allele frequency data. *PLoS Genet*, 8(11):e1002967, 2012.
- 444 Ron Pinhasi, Daniel Fernandes, Kendra Sirak, Mario Novak, Sarah Connell, Songül  
445 Alpaslan-Roodenberg, Fokke Gerritsen, Vyacheslav Moiseyev, Andrey Gromov, Pál  
446 Raczky, et al. Optimal ancient dna yields from the inner ear part of the human  
447 petrous bone. *PloS one*, 10(6):e0129102, 2015.
- 448 Jonathan K Pritchard, Matthew Stephens, and Peter Donnelly. Inference of population  
449 structure using multilocus genotype data. *Genetics*, 155(2):945–959, 2000.
- 450 Fernando Racimo, Gabriel Renaud, and Montgomery Slatkin. Joint estimation of  
451 contamination, error and demography for nuclear dna from ancient humans. *PLoS*  
452 *Genet*, 12(4):e1005972, 2016.

453 Morten Rasmussen, Sarah L Anzick, Michael R Waters, Pontus Skoglund, Michael De-  
454 Giorgio, Thomas W Stafford Jr, Simon Rasmussen, Ida Moltke, Anders Albrechtsen,  
455 Shane M Doyle, et al. The genome of a late pleistocene human from a clovis burial  
456 site in western montana. *Nature*, 506(7487):225–229, 2014.

457 Susanna Sawyer, Johannes Krause, Katerina Guschanski, Vincent Savolainen, and  
458 Svante Pääbo. Temporal patterns of nucleotide misincorporations and dna frag-  
459 mentation in ancient dna. *PloS one*, 7(3):e34131, 2012.

460 Joshua G Schraiber, Steven N Evans, and Montgomery Slatkin. Bayesian inference of  
461 natural selection from allele frequency time series. *Genetics*, 203(1):493–511, 2016.

462 Per Sjödin, Pontus Skoglund, and Mattias Jakobsson. Assessing the maximum con-  
463 tribution from ancient populations. *Molecular biology and evolution*, page msu059,  
464 2014.

465 Pontus Skoglund, Helena Malmström, Ayça Omrak, Maanasa Raghavan, Cristina  
466 Valdiosera, Torsten Günther, Per Hall, Kristiina Tambets, Jüri Parik, Karl-Göran  
467 Sjögren, et al. Genomic diversity and admixture differs for stone-age scandinavian  
468 foragers and farmers. *Science*, 344(6185):747–750, 2014.

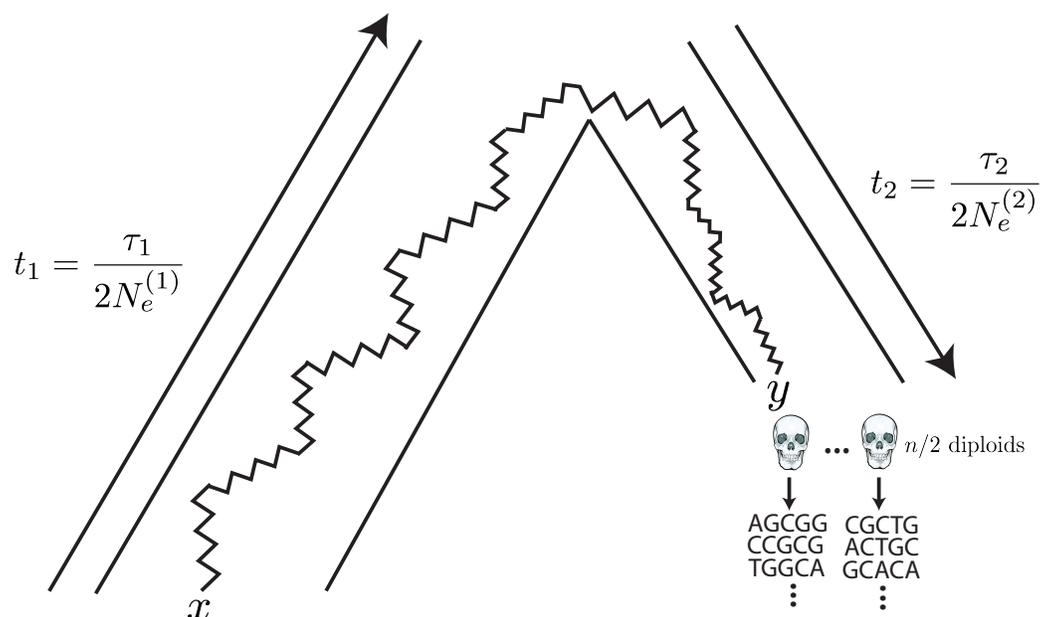


Figure 1: The generative model. Alleles are found at frequency  $x$  in the modern population and are at frequency  $y$  in the ancient population. The modern population has effective size  $N_e^{(1)}$  and has evolved for  $\tau_1$  generations since the common ancestor of the modern and ancient populations, while the ancient population is of size  $N_e^{(2)}$  and has evolved for  $\tau_2$  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.

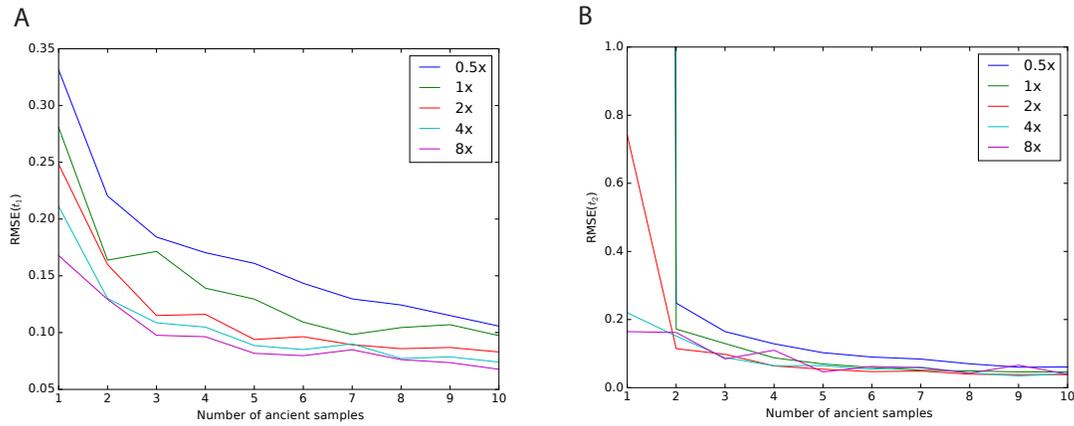


Figure 2: Impact of sampling scheme on parameter estimation error. In each panel, the  $x$  axis represents the number of simulated ancient samples, while the  $y$  axis shows the relative root mean square error for each parameter. Each different line corresponds to individuals sequenced to different 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  $t_2 = 0.05$ .

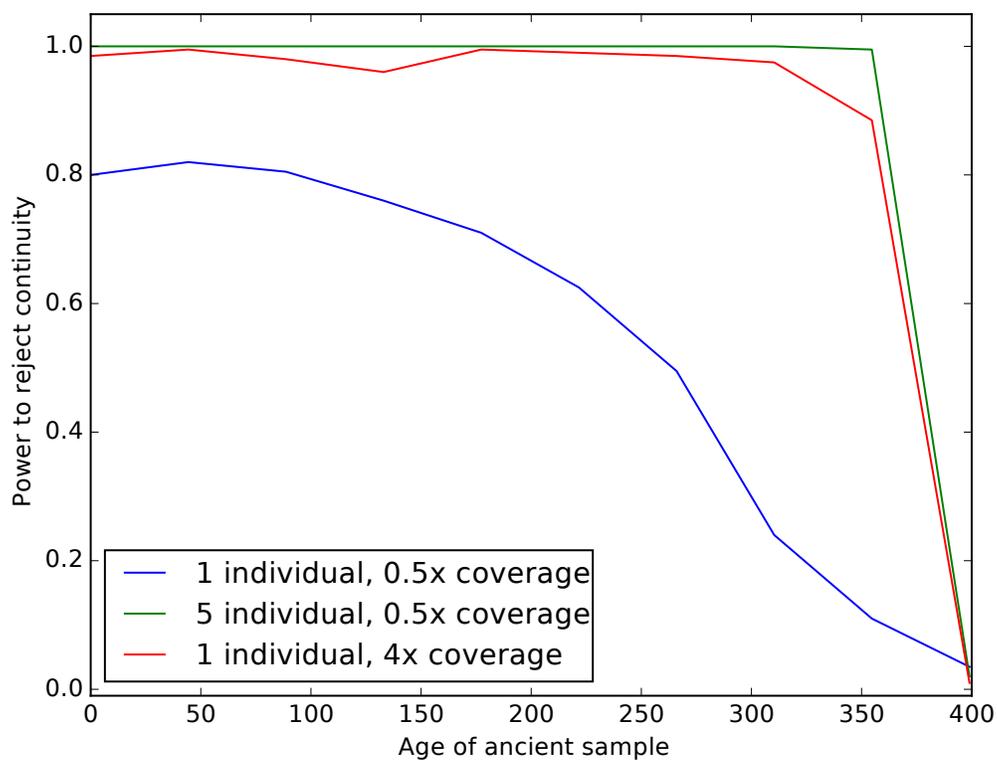


Figure 3: Impact of sampling scheme on rejecting population continuity. The  $x$  axis represents the age of the ancient sample in generations, with 0 indicating a modern sample and 400 indicating a sample from exactly at the split time 400 generations ago. The  $y$  axis shows the proportion of simulations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

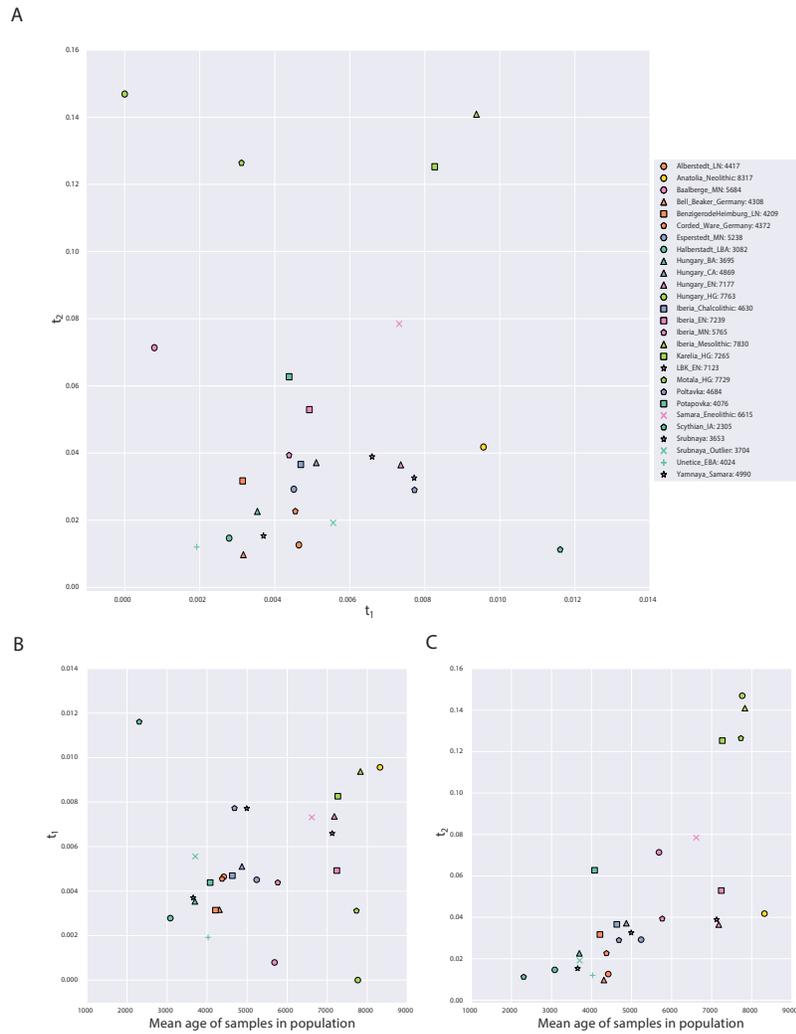


Figure 4: Parameters of the model inferred from ancient West Eurasian samples. Panel A shows  $t_1$  on the x-axis and  $t_2$  on the y-axis, with each point corresponding to a population as indicated in the legend. Numbers in the legend correspond to the mean date of all samples in the population. Panels 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.

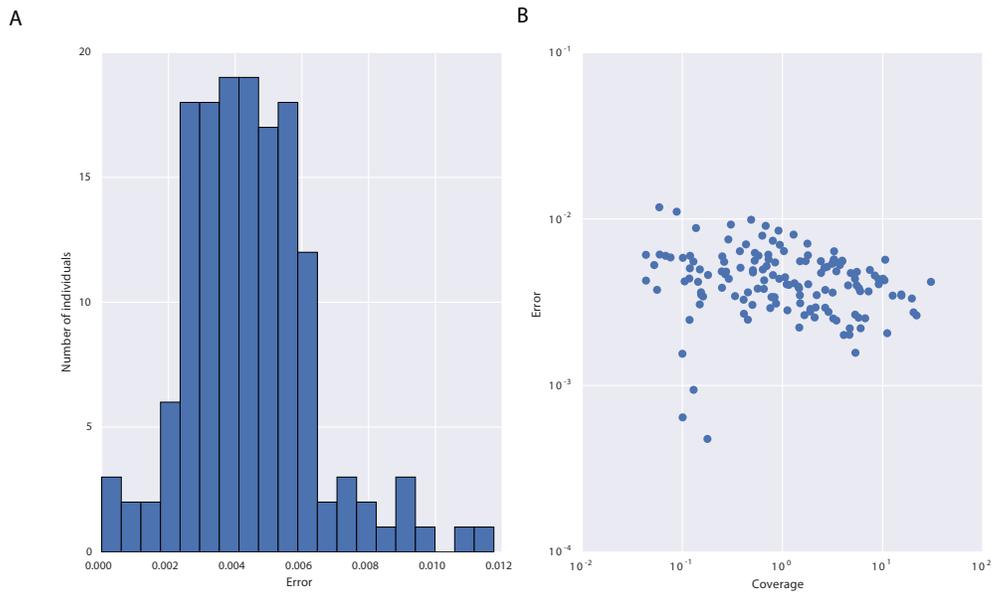


Figure 5: Properties of sequencing error inferred from ancient samples. Panel A shows a histogram of error rates estimated across all samples. Panel B is a scatterplot of coverage (x-axis) against inferred error rate (y-axis) for all individuals, except two individuals for whom an error rate of 0 was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

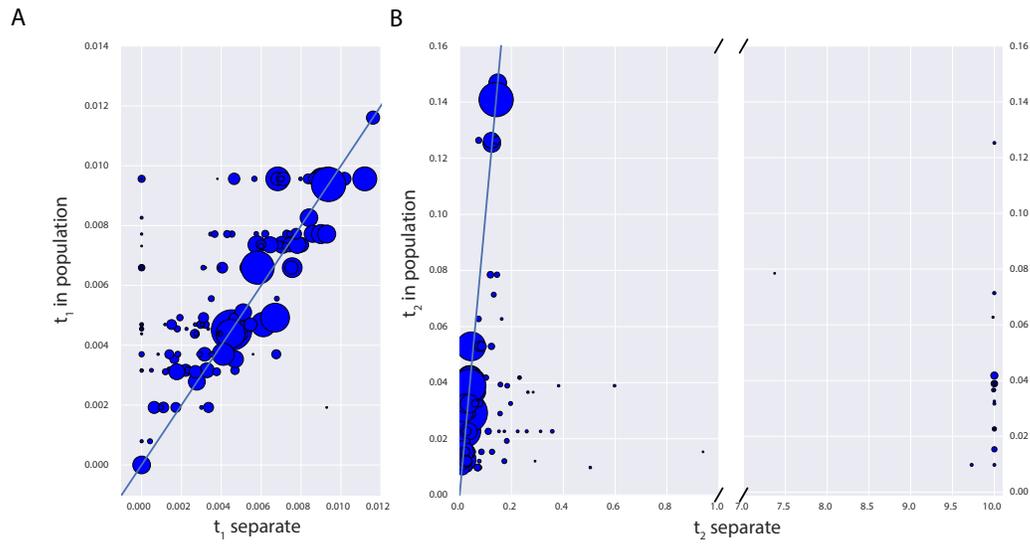


Figure 6: Impact of pooling individuals into populations when estimating model parameters from real data. In both panels, the x-axis indicates the parameter estimate when individuals are analyzed separately, while the y-axis indicates the parameter estimate when individuals are grouped into populations. Size of points is proportional to the coverage of each individual. 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
BenzigerodeHeimburg_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  $t_2 = 0$ .