Investigating population continuity and ghost admixture among ancient genomes

Authors:
James McKenna, Carolina Bernhardsson, David Waxman, Mattias Jakobsson, and Per Sjödin

Abstract

Ancient DNA (aDNA) can prove a valuable resource when investigating the evolutionary relationships between ancient and modern populations. Performing demographic inference using datasets that include aDNA samples however, requires statistical methods that explicitly account for the differences in drift expected among a temporally distributed sample. Such drift due to temporal structure can be challenging to discriminate from admixture from an unsampled, or "ghost", population, which can give rise to very similar summary statistics and confound methods commonly used in population genetics. Sequence data from ancient individuals also have unique characteristics, including short fragments, increased sequencing-error rates, and often limited genome-coverage that poses further challenges. Here we present a novel and conceptually simple approach for assessing questions of population continuity among a temporally distributed sample. We note that conditional on heterozygote sites in an individual genome at a particular point in time, the mean proportion of derived variants at those sites in other individuals has different expectations forwards in time and backwards in time. The difference in these processes enables us to construct a statistic that can detect population continuity in a temporal sample of genomes. We show that the statistic is sensitive to historical admixture events from unsampled populations. Simulations are used to evaluate the power of this approach. We investigate a set of ancient genomes from Early Neolithic Scandinavia to assess levels of population continuity to an earlier Mesolithic individual.

Introduction

Advances in DNA sequencing have led to rapidly increasing numbers of ancient genomes available for demographic inference. Understanding the relationships among such temporally distributed genomes can help reveal historical demographic patterns that would be impossible to detect using modern genomes alone (Sjödin, Skoglund, and Jakobsson, 2014; Malmström et al., 2009; Green et al., 2010; Lazaridis et al., 2016; Raghavan et al., 2014; Haak et al., 2015; Rasmussen et al., 2015; Lazaridis et al., 2014; Slatkin, 2016; Schraiber, 2017). In the field of human evolution in particular, ancient genomes have been key to discriminating between models of population continuity, admixture and replacement that have accompanied the emergence and spread of technological innovations, cultures and languages around the world (Lazaridis et al., 2016; Skoglund et al., 2012; Raghavan et al., 2014; Haak et al., 2015; Rasmussen et al., 2015; Olalde et al., 2018).

The primary challenge when investigating population continuity among ancient genomes is that changes in population allele frequencies over time (due to genetic drift) can lead to patterns of genetic differentiation among a temporally distributed sample that obscure continuity. Similar patterns can arise under models of historical admixture, particularly when the source of that admixture

is an unsampled, "ghost" population (Beerli, 2004; Slatkin, 2005; Lawson, Dorp, and Falush, 2018). The confounding effects of such structure have been demonstrated using both model-based methods of inference (Städler et al., 2009; Rosen et al., 2018; Mazet et al., 2016) and more qualitative approaches (Lawson, Dorp, and Falush, 2018; McVean, 2009; François et al., 2019). Repeated findings of such unsampled admixture events in the histories of human populations (Green et al., 2010; Reich et al., 2010; Wall et al., 2019; Durvasula and Sankararaman, 2020) has spurred the development of methods that aim to detect and quantify ghost admixture in the ancestries of modern populations (Skov et al., 2018; Wall et al., 2019; Durvasula and Sankararaman, 2020).

A further challenge is that due to the high levels of DNA fragmentation and contamination present (Allentoft et al., 2012; Skoglund et al., 2014b), ancient genomes are frequently sequenced to low coverage, without the depth necessary to confidently assign diploid genotypes (Dabney and Meyer, 2019). In the absence of the information that resides in patterns of linkage among loci, we are often limited to those inference methods based on genetic distances (Wang, Zollner, and Rosenberg, 2012; Verdu et al., 2014), diversity indices (Skoglund et al., 2014a; Diego-Ortega-Del and Montgomery, 2018), or allele frequency-based summary statistics (Patterson et al., 2012). Although the increasing availability of aDNA sequences has led to some temporally aware population genetic methods that explicitly account for the differences in drift expected among temporally distributed sequences (Sjödin, Skoglund, and Jakobsson, 2014; Yang and Montgomery, 2015; Rasmussen et al., 2015; Racimo, Renaud, and Slatkin, 2016; Schraiber, 2017; Silva et al., 2018; François and Jay, 2020), very often such techniques rely on good diploid calls, and therefore fail to take advantage of the large number of low coverage ancient genomes available. Many commonly used methods do not explicitly account for the genetic drift expected among a temporally distributed sample, leading to contradictory or misleading results (Schraiber, 2017). The popular model-based clustering methods STRUCTURE and ADMIXTURE for instance, do not presently account for different temporal sampling schemes, which can result in distorted patterns of shared ancestry when a sample includes ancient genomes (François et al., 2019). The placement of individuals on Principle Components Analysis (PCA) projections has also been shown to reflect both the temporal and geographic distribution of samples (Skoglund et al., 2014a; Novembre and Stephens, 2008). Figure 1 demonstrates the confounding effect of such temporal structure. Datasets of temporally distributed samples were simulated under alternative demographic models of population continuity and historical admixture from an unsampled population, resulting in similar projections. High levels of branch-specific drift separating an ancient from modern populations can make it difficult to identify genuine cases of population continuity (Slatkin, 2016).

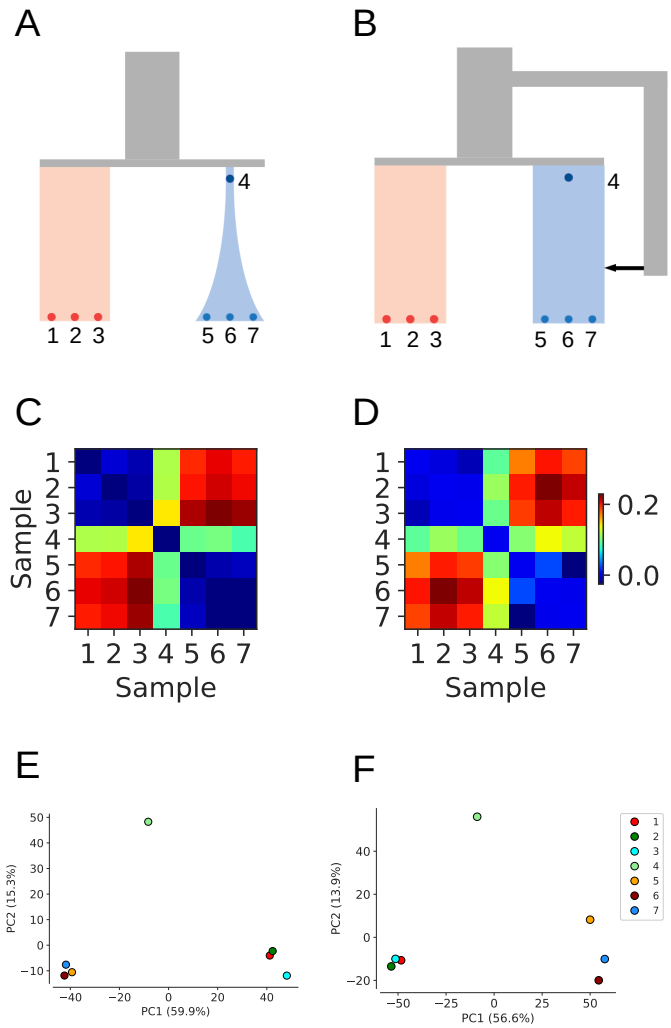


Figure 1: The confounding effects of temporal structure on demographic inference. Datasets simulated under alternative demographic models of A) population branch specific drift, and B) admixture from an unsampled population can result in highly similar pairwise f_2 matrices (C & D) and PCA projections (E & F). In both cases, high levels of drift between ancient and modern samples can obscure relationships of genuine population continuity. Simulations performed using msprime (Kelleher, Etheridge, and McVean, 2016), calculation of f_2-statistics and PCAs performed using scikit-allel (Miles, A. and Murillo, R. and Ralph, P. and Harding, N and Pisupati, R. and Rae, S. and Millar, T., 2010).

Here we propose a novel and conceptually simple approach for investigating population continuity among a temporally distributed sample. The approach is sensitive to historical admixture from unsampled populations. It can further be used with modest coverage ancient genomes and it is robust to missing data. The principle underlying the approach is simple; conditioning on a large number of heterozygote sites in an individual sampled from a population at a particular point in time, the mean proportion of derived alleles at those sites in other individuals is affected differently by the action of genetic drift forwards and backwards in time. The different expectations for these processes allow us to contrast models of population continuity and admixture, and in certain cases, to estimate the proportion of ghost admixture that has occurred. We evaluate the utility and power of this approach using simulations and apply the method to a set of ancient genomes sequenced from two Scandinavian Mesolithic foragers (SHG), five hunter-gatherers from the Neolithic Pitted Ware Culture (PWC) and five contemporaneous farmers from the Funnelbeaker Culture (FBC).

Method

We will first outline the theory behind the model, followed by the derivation of the central summary-statistic and finally describe a set of simulations that illustrate and evaluate the approach.

Theory

Although allele frequencies change over time, it is a well known result of population genetics that given a population frequency of p of a neutral variant, the expected population frequency of this variant will remain indefinitely at p forward in time (Kimura and Ohta, 1969). Consequently, since all (neutral) variants are eventually either lost or fixed in the population, the fixation probability of an allele at present population frequency p is p. This holds regardless of whether the frequency of the derived or the ancestral variant is considered. On the contrary, the expected frequency of a derived variant looking backwards in time cannot be the same as the present frequency simply because we know that it will eventually disappear (at the time of the mutation that gave rise to it). In fact, we show in the Appendix that the expected frequency of a derived variant t generations ago given a present frequency p is $pe^{-\tau}$ where τ is the scaled time or genetic drift between now and t generations ago ($\tau = \int_0^t (2N(x))^{-1}dx = t/2Ne$, where $N(x)$ is the diploid population size at time x and N_e is the harmonic mean of these $N(X)$ between $x = 0, \cdots, t$). To illustrate, in figure 2, conditional on population frequency p in population A, the expected frequency of the derived variant in the branch below population A is p (independent of τ_1) while the expected frequency of the derived variant in the right population branch is $pe^{-\tau_2}$ (independent of τ_3).

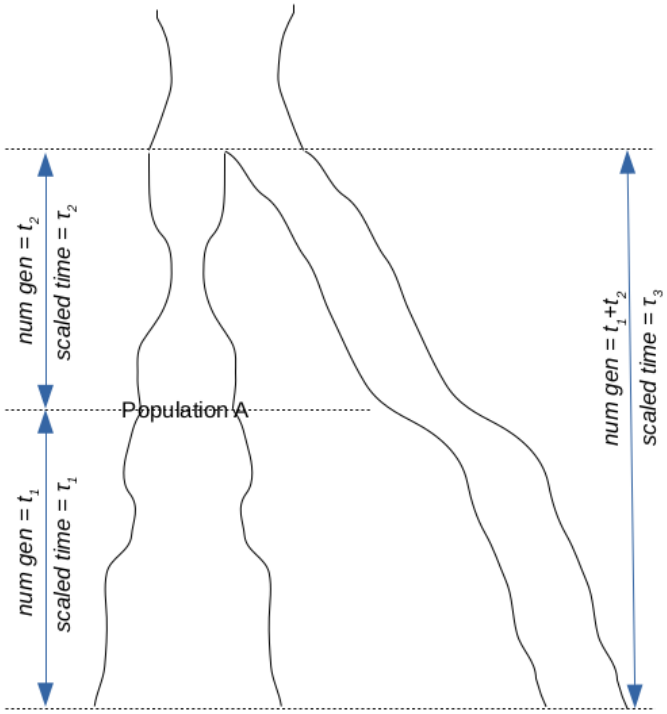


Figure 2: An illustration of the model being analyzed when the two most recent populations are at the same generation. Note however that since scaled time is different from the number of generations, τ_3 may be different from $\tau_1 + \tau_2$.

Now consider that an individual (the *anchor individual* for brevity) has been sequenced and genotypes have been called. If we restrict the analysis to heterozygote sites (and we assume that the ancestral and derived variant is known), we know that the derived variant is at least as old as the anchor individual. We also know that it is neither fixed nor lost at this time point which allows us to use the diffusion theory results of the Appendix. The population frequency of the derived variant is not the same at all of these heterozygote sites, but the conditioned distribution has an expected value which we will denote by $\overline{p_A}$ where

$$\begin{aligned}\overline{p_A} &= E[X_A|1 \text{ derived and } 1 \text{ ancestral variant observed in population}] \\ &= \int_0^1 x f(\text{pop frequency } X_A = x \big| 1 \text{ derived and } 1 \text{ ancestral variant observed in population}) dx\end{aligned}$$

Thus, if we refer to this set of heterozygote sites in the anchor individual as H_A, the probability to draw the derived variant at a site in the set H_A, in the same population that the anchor individual was sampled from, is $\overline{p_A}$. Moreover, for any individual that lived more recent in time and that traces all of its ancestry to

population A, the probability to draw the derived variant among H_A is also $\overline{p_A}$. Such an individual can be said to be completely continuous with the population the anchor individual was sampled from.

The anchor statistic ($R_d(A, x)$)

Given an anchor individual A and a set H_A of heterozygous sites in A, we define the *anchor statistic* for a test individual x as the probability to draw the derived variant in x at sites in H_A and refer to this statistic as $R_d(A, x)$. For empirical data, for low coverage data we do this by adding $k/(k + l)$, where k is the number of reads supporting the derived and l is the number of reads supporting the ancestral, for all sites with $k + l > 0$ and then dividing by the total number of sites (among the H_A sites) with $k + l > 0$. For high coverage data we would just use the number of heterozygous sites (n_{het}), the number of homozygous derived sites (n_{der}) and the number of homozygous ancestral sites (n_{anc}) among the H_A anchor sites and do $(0.5n_{het} + n_{der})/(n_{het} + n_{der} + n_{anc})$.

Estimating admixture

Consider an individual that traces an amount $1 - \gamma$ of its genetic ancestry to the anchor population but a proportion γ of its genetics ancestry comes from a population that diverged from the anchor population branch τ units of genetic drift prior to the anchor population. The average frequency at the time of the split of the latter population (and hence in all of that branch forwards in time) is then $\overline{p_A}e^{-\tau}$ (see above). The probability to draw the derived variant among heterozygote sites in the anchor is therefore no longer $\overline{p_A}$ but

$$(1 - \gamma)\overline{p_A} + \gamma\overline{p_A}e^{-\tau} = \overline{p_A}(1 - \gamma(1 - e^{-\tau}))$$

Next, consider the model set-up shown in figure 3(a). If we sample four genomes, (A_1, A_2, B and C), of which A_2 and B are assumed to be more recent but continuous with the population A_1 is sampled from. C is assumed to be sampled more recently than the others and the admixture event. The admixture event comprises a proportion γ from a population diverging τ time units back from A_1, and $\tau + \tau_A$ drift units back from sample A_2.
If by $E[R_d(A, x)]$ we denote the expected value of the anchor statistic for a test individual x among sites that are heterozygote in an anchor individual A, then

$$\begin{aligned}
E[R_d(A_1, B)] &= E[R_d(A_1, A_2)] = \overline{p_{A_1}} \\
E[R_d(A_1, C)] &= \overline{p_{A_1}}\left(1 - \gamma(1 - e^{-\tau})\right) \\
E[R_d(A_2, B)] &= \overline{p_{A_2}} \\
E[R_d(A_2, A_1)] &= \overline{p_{A_2}}e^{-\tau_A} \\
E[R_d(A_2, C)] &= \overline{p_{A_2}}\left(1 - \gamma(1 - e^{-\tau-\tau_A})\right)
\end{aligned}$$

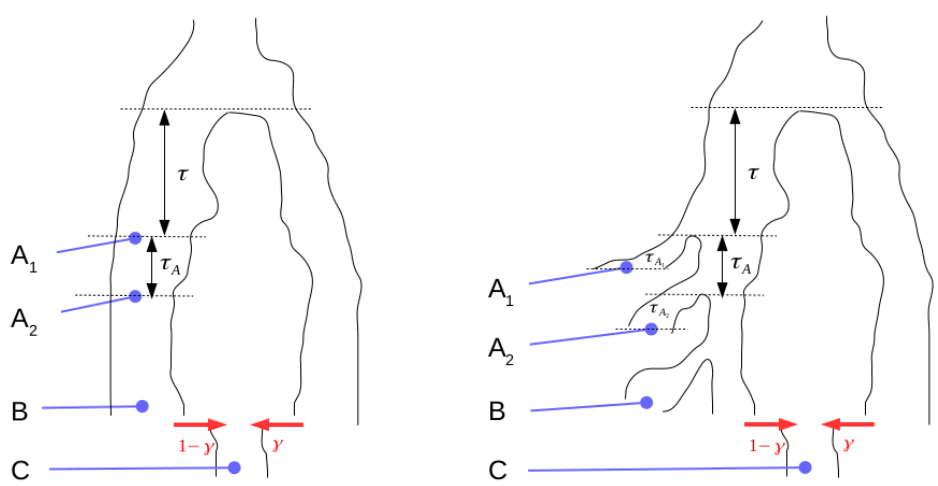
Furthermore, if we allow private drift to each of the sampled individuals so that, in effect, there is no contribution from any of the possible anchor populations to any of the possible test individuals, (see figure 3(b)) as well as sequencing errors specific to the anchors (denoted by ϵ_{A_1} and ϵ_{A_2}) we have

$$\begin{aligned}
E[R_d(A_1,B)] &= E[R_d(A_1,A_2)] = (1-\epsilon_{A_1})e^{-\tau_{A1}}\overline{p_{A_1}} \\
E[R_d(A_1,C)] &= (1-\epsilon_{A_1})e^{-\tau_{A_1}}\overline{p_{A_1}}\left(1-\gamma(1-e^{-\tau})\right) \\
E[R_d(A_2,B)] &= (1-\epsilon_{A_2})e^{-\tau_{A_2}}\overline{p_{A_2}} \\
E[R_d(A_2,A_1)] &= (1-\epsilon_{A_2})e^{-\tau_{A_2}}\overline{p_{A_2}}e^{-\tau_A} \\
E[R_d(A_2,C)] &= (1-\epsilon_{A_2})e^{-\tau_{A_2}}\overline{p_{A_2}}\left(1-\gamma(1-e^{-\tau-\tau_A})\right)
\end{aligned}$$

In effect, this is the same as a rescaling of $\overline{p_{A_1}}$ and $\overline{p_{A_2}}$ and for both figure 3(a) and figure 3(b) we have

$$\widehat{\gamma} = 1 - \frac{R_d(A_1,B)R_d(A_2,C) - R_d(A_1,C)R_d(A_2,A_1)}{R_d(A_1,B)R_d(A_2,B) - R_d(A_1,B)R_d(A_2,A_1)} \tag{1}$$

where $\widehat{\gamma}$ is an estimate of the proportion of ancestry in C from the "ghost" population to the right in figure 3.



(a) (b)

Figure 3: Set up for estimating proportion of admixture (γ) from a population diverging τ time units before anchor sample A_1.

Simulations

Detecting population continuity and quantifying admixture

Here we use simulations to investigate the power of this approach to detect population discontinuity and to quantify admixture from an unsampled population. Datasets were simulated under alternative demographic models of population continuity (figure 4A), and population discontinuity including two pulses of admixture from a population diverging 4000 generations ago (figure 4B). All simulations were performed using msprime (Kelleher, Etheridge, and McVean, 2016), aggregating data across 1000 runs of a 2M base sequence.

In each model, 7 diploid individuals were sampled at times ranging from present day to 720 generations in the past. The generation time was set to 29 years, with mutation and recombination rates fixed at 1.45×10^{-8} and 1.25×10^{-8} respectively. Both models include a population divergence event $4,000$ generations in the past, with a population experiencing an expansion from $N = 1,000$ at divergence time to $N = 10,000$ at present day. All other branches of the models are fixed and constant at $N = 10,000$. Note however that the anchor framework requires no assumption of fixed or constant population sizes through time to detect population continuity or ghost admixture. Model B in Figure 4 includes two pulses of 25% admixture from the diverging population at 180 and 420 generations respectively. We can condition on heterozygote sites in the oldest individual from our sample (individual 1 sampled at 720 generations) in both models, and compare the mean proportion of derived at those sites in more recent individuals.

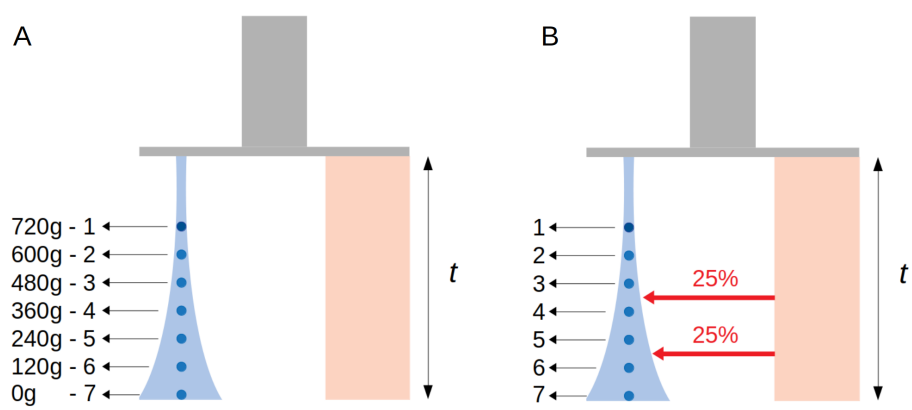


Figure 4: Illustrating the set-up used for simulation of data under models of A) population continuity and B) pulsed admixture from an unsampled population. Samples $1-7$ are taken at different time points from one population undergoing a population expansion from initial size at population divergence of $1,000$ diploid individuals, to final size at present day of $10,000$. Times shown in generations with 1 gen=29 years. Population divergence time $t = 4,000$ generations.

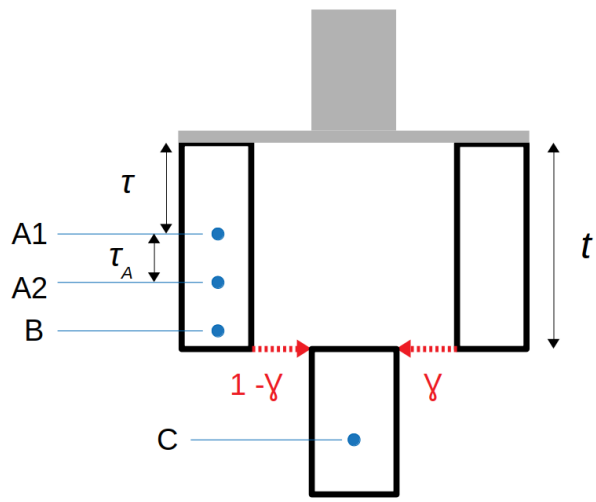


Figure 5: Simulation model set up for estimating proportion of admixture (γ) from a population diverging τ time units before anchor sample A_1.

In order to evaluate the power of this framework to estimate admixture proportions from an unsampled population, a dataset was simulated under the demographic model of a pulse admixture event shown in figure 5. Simulations were performed in an identical manner as before, but this time with four individuals (A_1, A_2, B and C) sampled at $1,000$, 100, 50 and 0 generations respectively. A single pulse of admixture into this temporal sample was included at 25 generations in the past (between sampling times of individuals B and C). The performance of in estimating admixture proportions was assessed when three parameters of this model were allowed to vary; admixture proportions (γ) varying from 0% to 90%, population divergence time t ranging from $1,000$ to $5,500$ generations (genetic drift between 0.05 to 0.2), and the degree of drift in the branch separating anchor samples (number of generations separating A_1 and A_2 varied from 100 to 900 generation which together with a diploid population size of $10,000$ gives that the genetic drift varied from 0.005 to 0.045).

Results

Simulations

Population continuity and ghost admixture

Figure 6 shows the mean proportion derived variants in individuals more recent than the anchor-individual ($R_d(Anchor, test)$), for simulated data from models of population continuity and admixture. Forward in time, the mean proportion derived variants remains constant under a model of population continuity even

in the presence of strong branch specific genetic drift followed by population expansion (figure 4A). Under a model of historical admixture however (figure 4B), there are two clear successive reductions in the mean proportion derived, each associated with separate admixture events. The magnitude of the reduction in proportion derived decreases with the degree of drift along the branch between the anchor-individual (individual 1) and the population divergence event ($t = 4,000$) with the branch leading to the population that mixes into the sampled population.

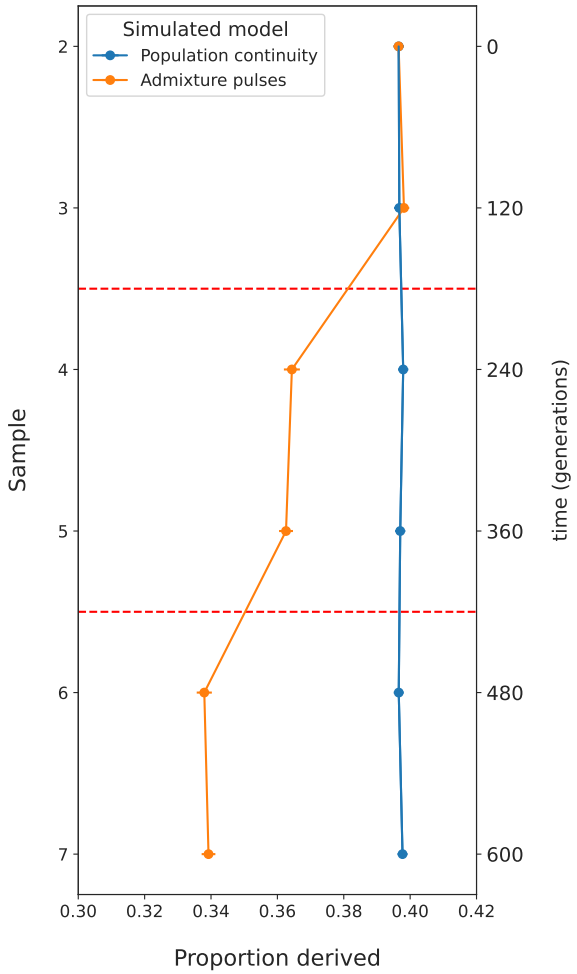


Figure 6: Anchor statistic under simulated models of population continuity and admixture. In each case the oldest sampled individual (1) is used as the anchor-individual, with the mean proportion derived at anchor heterozygote sites counted in all more recent individuals. Red dashed lines at $t = 180$ & $t = 420$ generations indicate admixture events (25%) in the Admixture Pulse model. The mean proportion derived variants remains constant forwards in time for individuals sampled from a model of population continuity. The mean proportion derived variants shows successive reductions for individuals sampled from a model that includes admixture. Error bars show two weighted-block jackknife standard deviations.

Estimating admixture proportions

The anchor statistic can also be used to estimate admixture as outlined in the Methods section. From simulations, we show that our estimate of admixture proportion is highly correlated to the true admixture proportion ($r = 0.98$, Pearson correlation). Furthermore, the accuracy of the admixture-estimate does not appear to be affected by the true proportion of admixture itself (figure 7a). Varying the population divergence time (t) from 1000 to 5500 generations does not change to accuracy of the estimation (figure 7b). Figure 7c shows that admixture estimation uncertainty increases as the degree of drift between anchor-individuals is reduced from 0.045 to 0.005 in drift units. These observations suggest that it is not the genetic drift prior to the anchor-individuals, but rather the degree of drift between the two anchor samples that influence accuracy of admixture estimation.

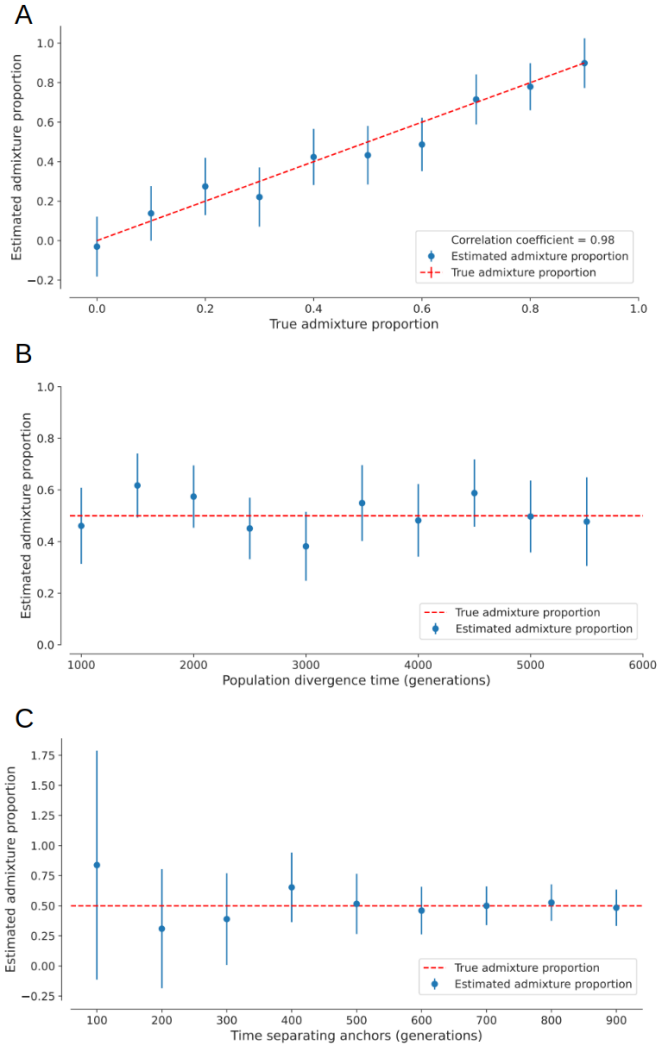


Figure 7: Estimating admixture proportions from an unsampled population with A) a fixed divergence time (t) of $4,000$ generations and admixture proportions varying from 0 to 0.9, B) a fixed admixture proportion of 0.5 and t varying from $1,000$ to $5,500$ generations, and C) a fixed admixture proportion of 0.5 and t of $4,000$ generations, with time separating anchors A_1 and A_2 ranging from 100 to 900 generations. In each case, sampling time of anchor-individual A_1 is $1,000$ generations and population sizes are constant at $10,000$ diploid individuals. Error bars show two weighted-block jackknife standard deviations.

Estimating admixture proportion among a temporal sample therefore requires two individuals of sufficiently high coverage that heterozygotes can be

confidently called, and for those anchor individuals to be separated by enough genetic drift to gain the necessary power for accurate estimation.

Empirical data

β-drift and a convenient transformation of $R_d(A, x)$

Empirical data will often have a proportion erroneously called heterozygote anchor sites and it is hard to completely rule out some level of private drift associated with the anchor population. For these reasons we argue that studying $-ln(R_d(A, x))$ is preferable to $R_d(A, x)$ (where A is an anchor and x is any test individual. The incentive for this is that the distance between different estimates then has a clear biological interpretation as the difference in backward drift in the path from the anchor population to the population of a test individual. For brevity, we will refer to this drift as β-drift which is always relative to some anchor population. To illustrate, consider two test individuals B and C with β-drift τ_B and τ_C relative to anchor A. If ϵ_A represents the probability of an anchor site being due to an error in the anchor sequence and not representing a a true heterozygote site and τ_A the amount of private drift associated with the anchor, then

$$E[R_d(A, B)] = (1 - \epsilon_A)\overline{p_A}e^{-\tau_A}e^{-\tau_B}$$
$$E[R_d(A, C)] = (1 - \epsilon_A)\overline{p_A}e^{-\tau_A}e^{-\tau_C}$$

so that

$-ln(R_d(A, C)) + ln(R_d(A, B))$ is an estimate of the difference in β-drift between C and B: $\tau_C - \tau_B$

Investigating population continuity from Mesolithic foragers to different Neolithic groups in Scandinavia

Scandinavia as a region holds special interest when considering the roles that population continuity and admixture have had on the emergence and spread of human cultures. It was the last region of Europe to become free of ice after the Last Glacial Maximum, and it harboured some of the last populations of hunter-gatherers in Europe. The earliest expressions of Neolithic material culture also emerged relatively late, with the arrival of the Funnelbeaker culture (FBC) circa 3500BCE to southern Scandinavia (Midgley, 1992; Lillie et al., 2000; Skoglund et al., 2012; Fraser et al., 2018). Paleogenetic data show a close connection between these early farmers of Scandinavia and other Early and Middle Neolithic farmers of Europe (Malmström et al., 2015; Malmström et al., 2019), providing strong support for a model of population discontinuity between the Scandinavian hunter-gatherers of the Mesolithic and the arriving farmers. The Neolithic Pitted Ware Culture (PWC) represents an intriguing case of a hunter-gathering culture in Scandinavia that emerge circa 3200BCE and post-dates the migration of the farmers that spread the FBC culture into

Scandinavia. Despite this, the results of several studies have indicated that the PWC culture predominantly shows genetic affinity to earlier Mesolithic hunter-gatherers rather than contemporary Neolithic farmers (Skoglund et al., 2012; Skoglund et al., 2014a). Furthermore, archaeological data has been used to paint a picture of two cultures overlapping in both time and space, coexisting in parallel for several hundred years and yet still maintaining distinct material cultures and dietary patterns, in addition to maintaining distinct genetic make-ups (Malmström et al., 2009; Fraser et al., 2018; Coutinho et al., 2020; Malmström et al., 2019).

We assembled a panel of fully UDG (uracil-DNA glycosylase) treated high-coverage ancient genomes from an ongoing effort to increase genome coverage from Scandinavian Stone Age humans, including a Mesolithic hunter-gatherer from the Baltic Island Stora Karlsö (sf12) and five PWC hunter-gatherers from the Baltic island of Gotland, together with five FBC farmers from the Megalithic passage tombs of Gökhem and Rössberga. In order to evaluate the impact genome coverage has on the statistics, we also included two relatively low coverage blunt-end screening genomes. One of these comes from the same ancient remains as one of the high-coverage PWC samples (ajv058) while the other one (sf9) is contemporaneous with and from the same site as sf12 (supplementary table 1, Günther et al., 2018; Fraser et al., 2018; Coutinho et al., 2020; Malmström et al., 2019). Although from distinct material cultures, all PWC and FBC individuals are broadly contemporaneous from the same period of the Nordic Middle Neolithic (circa $3200 - 2300$BCE). The raw sequence data was processed according to Alves et al. (2022) and Günther et al. (2018). See Appendix section "Data processing" for how genotypes were called.

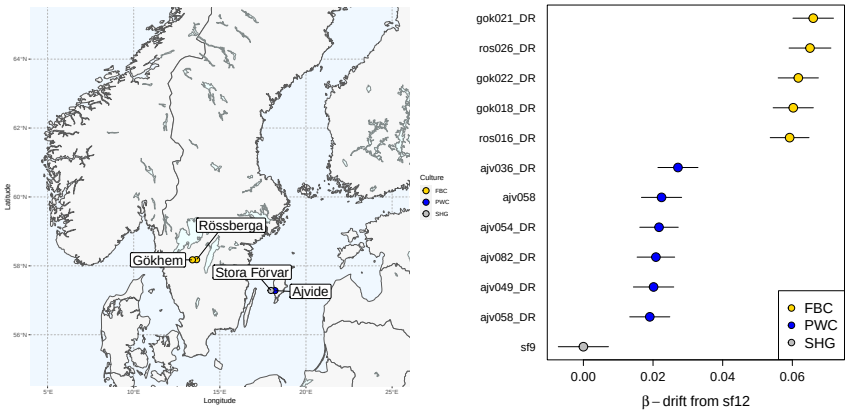


Figure 8: A) Map showing distributions of the sampled individuals that were either representatives of the Funnel Beaker culture (FBC) or the sub-Neolithic Pitted Ware culture (PWC) during the Nordic Middle Neolithic period. All PWC individuals from Ajvide burial, Gotland. FBC burial sites sampled include Gökhem passage grave, and Rössberga passage grave. Mesolithic individuals sf9 and sf12 from Stora Förvar on Stora Karlsö. B) β-drift relative to sf9 for different individuals with Scandinavian Mesolithic hunter-gatherer sf12 used as anchor. Error bars show two weighted-block jackknife standard deviations. The suffix "_DR" in the sample names indicate damage-repair through full UDG treatment of the sequenced libraries.

The Mesolithic hunter-gatherer sf12 was used as the anchor-individual so as to measure differences in β-drift between this individual and a set of other, more recent, individuals (figure 8B). All PWC individuals show a similar mean β-drift that is lower compared to the FBC individuals but higher than for sf9. These results are consistent with previous findings of modest gene-flow between the contemporaneous PWC and FBC groups (Malmström et al., 2009; Fraser et al., 2018; Coutinho et al., 2020) but also that PWC is not a direct descendent population from SHG.

Discussion

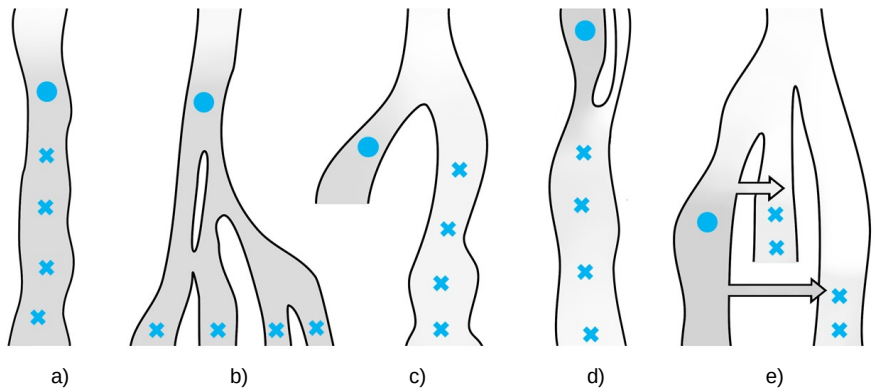


Figure 9: In each of the scenarios a)-d) all test individuals (blue crosses) have the same expected value of the anchor statistic when calculated relative to the anchor (blue filled circle). e): although it will depend on the details of the two admixture events (the arrows), this scenario can result in all test individuals (blue crosses) having the same expected value of the anchor statistic. In all scenarios, the darker the grey, the lower the backward drift in the average path from the anchor population.

Ancient genomes have the capacity to revolutionize our understanding of the demographic processes contributing to patterns of genetic variation among current-day humans as well as among other species. aDNA can reveal population continuity through time and aid in the detection of historical admixture events and population replacements. Particularly in studies of human demographic history, aDNA has proven an important resource in understanding the pre-historic movements of people that have spread cultures, languages, and technologies to new areas (*e.g.* Arcos and Raghavan, 2023). In order to take full advantage of this resource however, it is essential to develop population genetic tools capable of utilizing samples of temporally distributed genomic data. Here we have outlined a novel and conceptually simple approach that is capable of elucidating questions of population continuity through time. It is sensitive to admixture from unsampled ("ghost") populations and can take advantage of the increasing numbers of low coverage genomes available.

Although the concept of continuity is generally clear from the context (*e.g.* Mattila et al., 2023), for the purpose of discussing the relationship between the anchor statistic and continuity, we will with a statement such as "an individual x has continuity level c with population A (that existed at time point t)"

mean that a proportion c of all ancestors to x living at time point t belonged to population A. This in turn implies that a proportion c of x's autosomal genetic material is expected to trace back to individuals living in population A. With respect to the different scenarios in figure 9, this definition would imply that all test individuals in scenarios a) and b) have continuity level 1 with the anchor population while scenario c) has continuity level 0. In d), all test individuals would have some continuity level c while only the two most recent test individuals in scenario e) would have a level of continuity larger than 0. In contrast, all test individuals in each of the scenarios *could* have identical amounts of β-drift relative to the anchor. Hence, unless β-drift is 0, it can be difficult to assess any level of continuity at all based exclusively on the anchor statistic. However, when it can be assumed that one test individual has no β-drift (*i.e.* it has continuity level 1 with the anchor population) the absolute β-drift can be assessed. Since $sf9$ was sampled from the same location as sf12 and both were estimated to be around $7,000$ years old (see table 1 in the Appendix), we assumed that sf12 and sf9 were drawn from the same population and could therefore translate values of the anchor statistic to actual β-drift from the population sf12 and sf9 were sampled from. From this analysis we conclude that although the PWC individuals have about 0.02 of β-drift relative to SHG (represented by sf12 and sf9 in this analysis) and thus that they have ancestry in other populations than SHG that were contemporaneous with SHG. Any admixture event with proportion c with a population that has an additional τ amount of β-drift such that $-ln(ce^{-\tau} + 1 - c) \approx 0.02$ would give this pattern. This is consistent with that they have none of their ancestry tracing back to SHG but to a population that diverged from SHG approximately 0.02 units of drift before the sampled individuals in this study. Alternatively, they may have a significant part of their ancestry in SHG but that they are admixed with populations that diverged considerably more than 0.02 units of drift before the population represented by sf12 and sf9. What is clear is that the FBC individuals have an additional 0.04 amount of β-drift from SHG compared to the PWC individuals.

We have shown, through simulations and an empirical example, that this approach has the power to infer population continuity and estimate proportions of ghost admixture. Although we did not find a suitable empirical example to estimate proportions of ghost admixture it is a UNIQUE(?!) and potentially powerful aspect of our approach as the method does not require any modeling of the source of the admixture by substitute populations.

Finally, although our set-up is closely related to an outgroup-f3 if the outgroup is chosen to be another species (the same species that is used to call ancestral and derived state for the anchor analysis), the two statistics are not identical and our statistic has the advantage of being an estimate of a biologically meaningful parameter (β-drift). As such, our set-up is in our view a more natural way of investigating continuity and relationships among ancient samples.

Acknowledgement

We want to thank Olaf Thalmann, Karl-Heinz Herzig, and Jarosoław Walkowiak for access to unpublished data, Nikola Vukovic for help with graphic design and Imke Lankheet for Latex assistance. This work was supported by the Knut and Alice Wallenberg foundation. The computations and data handling were enabled by resources (SNIC2022/2-11 and p2018003) provided by the Swedish National Infrastructure for Computing (SNIC) at UPPMAX.

References

Allentoft, Morten E et al. (2012). "The half-life of DNA in bone: measuring decay kinetics in 158 dated fossils". In: *Proceedings of the Royal Society B: Biological Sciences* 279.1748, pp. 4724–4733.

Alves, Isabel et al. (2022). "Genetic population structure across Brittany and the downstream Loire basin provides new insights on the demographic history of Western Europe". In: *bioRxiv*. DOI: 10.1101/2022.02.03.478491. eprint: https://www.biorxiv.org/content/early/2022/02/04/2022.02.03.478491.full.pdf. URL: https://www.biorxiv.org/content/early/2022/02/04/2022.02.03.478491.

Arcos, M. C. María C. Ávila and C. Raghavan M. Schlebusch (2023). "Going local with ancient DNA: A review of human histories from regional perspectives". In: *Science* 382.6666, pp. 53–58. DOI: 10.1126/science.adh8140. eprint: https://www.science.org/doi/pdf/10.1126/science.adh8140. URL: https://www.science.org/doi/abs/10.1126/science.adh8140.

Beerli, P. (2004). "Effect of unsampled populations on the estimation of population sizes and migration rates between sampled populations". In: *Mol Ecol* 13, pp. 827–836.

Blank, M., KG. Sjögren, and J. Storå (2020). "Old bones or early graves? Megalithic burial sequences in southern Sweden based on 14C datings". In: *Archaeological and Anthropological Sciences* 12 (89), pp. 1866–9565. DOI: 10.1007/s12520-020-01039-9.

Coutinho, Alexandra et al. (2020). "The Neolithic Pitted Ware culture foragers were culturally but not genetically influenced by the Battle Axe culture herders". In: *American Journal of Physical Anthropology* 172.4, pp. 638–649. DOI: https://doi.org/10.1002/ajpa.24079. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ajpa.24079. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/ajpa.24079.

Dabney, Jesse and Matthias Meyer (2019). "Extraction of highly degraded DNA from ancient bones and teeth". In: *Ancient DNA*. Springer, pp. 25–29.

Diego-Ortega-Del, V. and S. Montgomery (2018). "FST between Archaic and Present-Day Samples". In: *bioRxiv*. DOI: 10.1101/362053.

Durvasula, A. and S. Sankararaman (2020). "Recovering signals of ghost archaic introgression in African populations". In: *Science Advances* 6.7. DOI: 10.1126/sciadv.aax5097.

François, O. et al. (2019). "Inference of population genetic structure from temporal samples of DNA". In: *bioRxiv*. DOI: 10.1101/801324.

François, Olivier and Flora Jay (2020). "Factor analysis of ancient population genomic samples". In: *Nature communications* 11.1, pp. 1–11.

Fraser, Magdalena et al. (2018). "New insights on cultural dualism and population structure in the Middle Neolithic Funnel Beaker culture on the island of Gotland". In: *Journal of Archaeological Science: Reports* 17, pp. 325–334.

Green, RE. et al. (2010). "A Draft Sequence of the Neandertal Genome". In: *Science* 328.5979, pp. 710–722. DOI: 10.1126/science.1188021.

Griffiths, RC (2003). "The frequency spectrum of a mutation, and its age, in a general diffusion model". In: *Theor. Popul. Biol.* 64, pp. 241–251.

Günther, T. et al. (Jan. 2018). "Population genomics of Mesolithic Scandinavia: Investigating early postglacial migration routes and high-latitude adaptation". In: *PLOS Biology* 16.1, pp. 1–22. DOI: 10.1371/journal.pbio.2003703. URL: https://doi.org/10.1371/journal.pbio.2003703.

Haak, W. et al. (Mar. 2015). "Massive migration from the steppe was a source for Indo-European languages in Europe". In: *Nature* 522.7555, p. 207. DOI: 10.1038/nature14317.

Kelleher, J., AM. Etheridge, and G. McVean (2016). "Efficient Coalescent Simulation and Genealogical Analysis for Large Sample Sizes". In: *PLoS Comput Biol* 12, e1004842. DOI: 10.1371/journal.pcbi.1004842.

Kimura, M. and T. Ohta (1969). "The average number of generations until fixation of a mutant gene in a population". In: *Genetics* 61, pp. 763–771.

Lawson, DJ., L. van Dorp, and D. Falush (2018). "A tutorial on how not to over-interpret STRUCTURE and ADMIXTURE bar plots". In: *Nat Commun* 9.3258. DOI: 10.1038/s41467-018-05257-7.

Lazaridis, I. et al. (2014). "Ancient human genomes suggest three ancestral populations for present-day Europeans". In: *Nature*, pp. 409–13. DOI: 10.1038/nature13673.

Lazaridis, I. et al. (2016). "Genomic insights into the origin of farming in the ancient Near East". In: *Nature*, pp. 1476–4687. DOI: 10.1038/nature19310.

Li, H. (2011). "A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data". In: *Bioinformatics* 27(21), pp. 2987–2993.

Lillie, Malcolm et al. (2000). *Europe's first farmers*. Cambridge University Press.

Malmström, Helena et al. (2015). "Ancient mitochondrial DNA from the northern fringe of the Neolithic farming expansion in Europe sheds light on the dispersion process". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 370.1660, p. 20130373.

Malmström, H. et al. (2009). "Ancient DNA Reveals Lack of Continuity between Neolithic Hunter-Gatherers and Contemporary Scandinavians". In: *Curr. Biol.* 19.20, pp. 1758–1762.

Malmström, H. et al. (2019). "The genomic ancestry of the Scandinavian Battle Axe Culture people and their relation to the broader Corded Ware horizon". In: *Proceedings of the Royal Society B* 286. DOI: 10.1098/rspb.2019.1528.

Mattila, TM. et al. (2023). In: *Communications Biology* 6. DOI: 10.1038/s42003-023-05131-3.

Mazet, O. et al. (2016). "On the importance of being structured: instantaneous coalescence rates and human evolution–lessons for ancestral population size inference?" In: *Heredity* 116, pp. 362–371. DOI: 10.1038/hdy.2015.104.

McVean, G. (2009). "A genealogical interpretation of principal components analysis". In: *Nature* 5.10, e1000686. DOI: 10.1371/journal.pgen.1000686.

Midgley, Magdalena (1992). "TRB culture. The first farmers of the north European plain." In.

Miles, A. and Murillo, R. and Ralph, P. and Harding, N and Pisupati, R. and Rae, S. and Millar, T. (Feb. 19, 2010). *scikit-allel*. Version 1.3.3. DOI: doi.org/10.5281/zenodo.4759368.

Novembre, J. and M. Stephens (2008). "Interpreting principal component analyses of spatial population genetic variation". In: *Nat Genet* 40, pp. 646–649.

Olalde, Iñigo et al. (2018). "The Beaker phenomenon and the genomic transformation of northwest Europe". In: *Nature* 555.7695, pp. 190–196.

Patterson, N. et al. (2012). "Ancient admixture in human history". In: *Genetics* 192, pp. 1065–1093. DOI: 10.1534/genetics.112.145037.

Paulsson, Bettina Schulz (2010). "Scandinavian models: Radiocarbon dates and the origin and spreading of passage graves in Sweden and Denmark". In: *Radiocarbon* 52.3, pp. 1002–1017.

Racimo, F., G. Renaud, and M. Slatkin (2016). "Joint Estimation of Contamination, Error and Demography for Nuclear DNA from Ancient Humans". In: *PLOS Genetics* 12.4, pp. 1–27. DOI: 10.1371/journal.pgen.1005972.

Raghavan, M. et al. (2014). "The genetic prehistory of the New World Arctic". In: *Science* 345.6200, p. 1255832. DOI: 10.1126/science.1255832.

Rasmussen, M. et al. (2015). "The ancestry and affiliations of Kennewick Man". In: *Nature* 523, pp. 455–458. DOI: 10.1038/nature14625.

Reich, D. et al. (2010). "Genetic history of an archaic hominin group from Denisova Cave in Siberia". In: *Nature* 468, pp. 1053–1060. DOI: 10.1038/nature09710.

Rosen, Z. et al. (2018). "Geometry of the Sample Frequency Spectrum and the Perils of Demographic Inference". In: *Genetics* 210.2, pp. 665–682.

Schraiber, JG. (2017). "Assessing the Relationship of Ancient and Modern Populations". In: *Genetics*. DOI: 10.1534/genetics.117.300448.

Silva, Nuno Miguel et al. (2018). "Bayesian estimation of partial population continuity using ancient DNA and spatially explicit simulations". In: *Evolutionary applications* 11.9, pp. 1642–1655.

Sjödin, P., P. Skoglund, and M. Jakobsson (2014). "Assessing the Maximum Contribution from Ancient Populations". In: *Molecular Biology and Evolution* 31.5, pp. 1248–1260. DOI: 10.1093/molbev/msu059.

Skoglund, P. et al. (2012). "Origins and genetic legacy of Neolithic farmers and hunter-gatherers in Europe". In: *Science* 336, pp. 466–469.

Skoglund, P. et al. (2014a). "Investigating population history using temporal genetic differentiation". In: *Mol Biol Evol* 31, pp. 2516–2527. DOI: 10.1093/molbev/msu192.

Skoglund, Pontus et al. (2014b). "Separating endogenous ancient DNA from modern day contamination in a Siberian Neandertal". In: *Proceedings of the National Academy of Sciences* 111.6, pp. 2229–2234.

Skov, Laurits et al. (Sept. 2018). "Detecting archaic introgression using an unadmixed outgroup". In: *PLOS Genetics* 14.9, pp. 1–15. DOI: 10.1371/journal.pgen.1007641. URL: https://doi.org/10.1371/journal.pgen.1007641.

Slatkin, M. (2005). "Seeing ghosts: the effect of unsampled populations on migration rates estimated for sampled populations". In: *Molecular ecology* 14, pp. 67–73.

— (2016). "Statistical methods for analyzing ancient DNA from hominins". In: *Curr Opin Genet Dev* 41, pp. 72–76.

Städler, T. et al. (2009). "The impact of sampling schemes on the site frequency spectrum in nonequilibrium subdivided populations". In: *Genetics* 182, pp. 961–965. DOI: 10.1534/genetics.108.094904.

Tavaré, S. (1984). "Line-of-Descent and Genealogical Processes, and Their Apllications in Population Genetics Models". In: *Theor. Popul. Biol.* 26, pp. 119–164.

Verdu, P. et al. (2014). "Patterns of Admixture and Population Structure in Native Populations of Northwest North America". In: *PLOS Genetics* 10.8, pp. 1–17. DOI: 10.1371/journal.pgen.1004530.

Wakeley, J. (2009). *Coalescent Theory: An Introduction*. 1st. Greenswood Village, Colorado: Roberts & Company Publishers.

Wall, Jeffrey D et al. (2019). "Identification of African-specific admixture between modern and archaic humans". In: *The American Journal of Human Genetics* 105.6, pp. 1254–1261.

Wallin, P. and H. Martinsson-Wallin (2016). "Decoding Neolithic Atlantic and Mediterranean Island Ritual". In: Oxbow books. Chap. Collective spaces and material expressions: ritual practice and island identities in Neolithic Gotland.

Wang, C., S. Zollner, and NA. Rosenberg (2012). "A quantitative comparison of the similarity between genes and geography in worldwide human populations". In: *PLoS Genet* 8, e1002886.

Yang, MA. and S. Montgomery (2015). "Using Ancient Samples in Projection Analysis". In: *G3* 6, pp. 99–105.

Appendix

Moments for the population frequency

For brevity, we suppress the dependence of p and τ in the notation and write $E[X_{\downarrow}^{n}]$ for the n'th moment of the population frequency after τ units of genetic

drift *forwards* in time conditional on population frequency p and $E[X_{\uparrow}^n]$ for the n'th moment of the population frequency going back τ units of genetic drift conditional on population frequency p. These moments are derived below.

Moments for $X_{\downarrow}$

To derive the moments of $X_{\downarrow}$, note that given a population frequency x of the derived variant, the probability of obtaining a sample of size n with only the derived allele is x^n. Thus, $X_{\downarrow}^n$ is the probability to pick only the derived variant in a sample of size n, and $E[X_{\downarrow}^n]$ is the expected value of this probability. Conditional on the frequency of the derived variant being p, τ time units ago, the expected probability for such a sample is derived by averaging over the number of ancestors to the sample τ time units ago, where each of those ancestors need to be of the derived type. Hence, $E[X_{\downarrow}^n|X_0 = p]$, or the expected probability to observe only the derived variant in a sample of size n is

$$E[X_{\downarrow}^n|X_0 = p] = \sum_{k=1}^{n} p^k g_{n,k}(\tau) \tag{2}$$

with $g_{n,k}(\tau)$ being the probability of there being k ancestors at time τ to a sample of n gene-copies (a much longer recursive proof using diffusion theory can be obtained from the authors upon request). Specifically

$$g_{n,k}(\tau) = \frac{1}{\binom{k}{2}} \sum_{i=k}^{n} e^{-\binom{i}{2}\tau} \binom{i}{2} \prod_{l=k, l\neq i}^{n} \frac{\binom{l}{2}}{\binom{l}{2} - \binom{i}{2}} \tag{3}$$

for $2 \leq k < n$ with special cases $g_{n,n}(\tau) = e^{-\binom{n}{2}\tau}$ and $g_{n,1}(\tau) = 1 - \sum_{k=2}^{n} g_{n,k}(\tau)$ (Tavaré, 1984; Wakeley, 2009).

Since $g_{1,1}(\tau) = 1$, $g_{2,2}(\tau) = e^{-\tau}$, $g_{2,1}(\tau) = 1 - e^{-\tau}$ we have

$$\begin{aligned}
E[X_{\downarrow}|X_0 = p] &= \sum_{k=1}^{1} p^k g_{1,k}(\tau) = p g_{1,1}(\tau) = p \\
E[X_{\downarrow}^2|X_0 = p] &= \sum_{k=1}^{2} p^k g_{2,k}(\tau) = p g_{2,1}(\tau) + p^2 g_{2,2}(\tau) \\
&= p(1 - e^{-\tau}) + p^2 e^{-\tau} = p - e^{-\tau} p(1 - p)
\end{aligned}$$

Moments for $X_{\uparrow}$

Define $\phi(x, \tau, p)$ to be the probability density of frequency x at time τ, given that $X_A = p$. The density backward in time is, in fact, the same as the forward density conditional on extinction (since by definition the derived variant goes

extinct backwards in time) (Griffiths, 2003). The forward density conditional on the allele ultimately going extinct is (using $*$ to denote conditional on extinction)

$$\phi^*(x,\tau,p) = \phi(x,\tau,p)\frac{u_0(x)}{u_0(p)} \tag{4}$$

where $u_0(z)$ is the probability of an allele at frequency z ultimately being lost. In a neutral model, $u_0(z) = 1 - z$. We get

$$\begin{aligned} E[X_\uparrow^n | X_0 = p] &= \int_0^1 x^n \phi^*(x,\tau,p)dx = \int_0^1 x^n \phi(x,\tau,p)\frac{1-x}{1-p}dx \\ &= \frac{1}{1-p}\left(E[X_\downarrow^n | X_0 = p] - E[X_\downarrow^{n+1} | X_0 = p]\right). \end{aligned}$$

For $n = 1$ we have

$$\begin{aligned} E[X_\uparrow | X_0 = p] &= \frac{1}{1-p}\left(E[X_\downarrow | X_0 = p] - E[X_\downarrow^2 | X_0 = p]\right) \\ &= \frac{1}{1-p}\left(p - p + e^{-\tau}p(1-p)\right) = e^{-\tau}p \end{aligned}$$

Sample metadata

Table 1: Information on the 10 individuals used in this study. mt, mitochondrial; cal, calibrated; BCE, before common era; PWC, Pitted Ware Culture; TRB, Funnel Beaker Culture.

Sample	Context	Origin	Genome coverage	Contamination (%)		Source	Age (cal. years BCE)	Genetic sex
				mt	autosomal			
sf12	Mesolithic	Stora Förvar per Stora Karlsö	62.16	0.015	0.73	(Günther et al., 2018)	7083-6807	XX
sf9	Mesolithic	Stora Förvar, Stora Karlsö	1.15	5.36	0	(Günther et al., 2018)	7350-7038	XX
ajv36	PWC	Ajvide, Gotland	10.96	0.029	0.69	(Coutinho et al., 2020)	3200-2300*	XX
ajv49	PWC	Ajvide, Gotland	20.26	0.049	0.66	Unpublished	3200-2300*	XY
ajv54	PWC	Ajvide, Gotland	17.83	0.024	0.47	(Malmström et al., 2019)	3200-2300*	XY
ajv58	PWC	Ajvide, Gotland	17.21	0.045	0.53	(Skoglund et al., 2014a)	2950-2650	XY
ajv82	PWC	Ajvide, Gotland	20.88	0.185	0.65	(Wallin and Martinsson-Wallin, 2016)	3200-2300*	XY
gok18	TRB	Gökhem, Västergötland	14.64	0.089	0.75	(Paulsson, 2010)	3760-3340†	XY
gok21	TRB	Gökhem, Västergötland	13.14	0.158	1.11	(Paulsson, 2010)	3760-3340†	XY
gok22	TRB	Gökhem, Västergötland	13.90	0.205	1.07	(Paulsson, 2010)	3760-3340†	XY
ros16	TRB	Rössberga, Västergötland	19.53	0.065	0.62	(Blank, Sjögren, and Storå, 2020)	3082-2894	XX

* Radiocarbon dates corrected for marine-reservoir effect
† Archaeological context-dated (Paulsson, 2010)

Data processing

Before calling, a base quality score recalibration (BQSR) step was performed on the 5 terminal bases of each read by reducing the quality of all 5´ T:s and all 3´ A:s to phred-score 2 (♯), using a custom python script. Dummy read groups (RG) were added with Picard v 1.118 (ref), followed by indel realignment using known indel sites in the 1000g phase1 data set (ref) with GATK v. 3.5.0 (ref).

A site in the anchor was chosen as an anchor position if 1) the read coverage was neither within the 5% lower or 5% upper tail of the coverage distribution 2) two variants observed at the site and the minor allele frequency was at least 1/3, 3) one of the variants could be confidently called as the ancestral variant (in our case, we demanded that all three apes had the same variant (without missingness) and that this variant was one of the variants observed in the anchor). Allele counts at anchor positions were parsed using samtools mpileup 1.17 with parameter flags -Q 30 -q 30 (Li, 2011). For site i among the anchor positions, the probability to pick the derived variant and the ancestral variant –$p_d(i)$ and $p_a(i)$ – for a test individual x was calculated. The anchor statistic was then calculate as

$$R_d(A, x) = \frac{\sum_{i=1}^{L} p_d(i)}{\sum_{i=1}^{L} (p_d(i) + p_a(i))}$$

where L is the number of anchor positions.

Anchor statistic and *f*-statistics

All scripts used for generation of simulated datasets and application of anchor test of population continuity from empirical datasets are available at XXX. Calculation of anchor statistics were performed using custom python scripts and functions. The python program scikit-allel (Miles, A. and Murillo, R. and Ralph, P. and Harding, N and Pisupati, R. and Rae, S. and Millar, T., 2010) was used to calculate *f*2-statistics.