Evaluating Archaic Admixture Time Estimates

Leonardo Nicola Martin Iasi (Max Planck Institute for Evolutionary Anthropology, MPI EVA), Dr. Benjamin Marco Peter (MPI EVA, benjamin_peter@eva.mpg.de)

2020-02-10

Abstract

Introduction

Detecting admixture, i.e. gene flow between previously isolated populations can shed light into the complex evolutionary history of populations. It is an important factor in specialization events and is found in many sexually reproducing species of plants, invertebrates and vertebrates [1], such as an ancient admixture event in the primate lineage between chimpanzees and bonobos [2]. The sequencing of the Neandertal [3] and Denisovan [4] genome revealed admixture events between Neandertals and modern humans outside of Africa [3,5–9], as well as a second admixture event between Denisovans and Asian populations [4,10–12].

Admixture on the genomic level and basic idea how to date it

On the genomic level, admixture introduces highly divergent chromosomal segments into the admixing population. Over time, recombination between parental chromosomes progressively break the introgressed segments apart, reducing their length in each generation [13]. Under a constant recombination rate, the length of introgressed segments should be inversely proportional to the time since the admixture event. Assuming that recombination events occurs according to a Poisson process, the fragment length distribution can be approximated by an exponential distribution, whose parameter is informative about the number of generations since the fragments entered the admixing population [14–16]. Hence, the length distribution of introgressed chromosomal segments ('recombination clock') has been used to inferred the time since the admixture event [11,14,17–21]

To do this, the model usually assumes a one generation long admixture event followed by a period of isolation [22]. The proportion of admixture of individuals in the admixed population is assumed to be identical, tracts

are rare and inbreeding is not significant, such that admixture segments do not recombine with each other to form longer segments [15]. The effect of drift is assumed to be negligible [19], the introgressed segments act neutral [23] and the recombination rate does not change over time [16].

The two approaches of obtaining the length of introgressed segments: ALD

The challenge is to correctly infer the length of the introgressed segments. There are two main approaches to date admixture [24] (Figure 1 A). The first approach uses the admixture-induced linkage disequilibrium (ALD) decay. Gene flow introduces highly divergent archaic segments into the human background. Hence, variants on these archaic segments are expected to be in high linkage disequilibrium to each other at the time of admixture [25–27]. As recombination breaks the introgressed chromosomal segments apart each generation, the association between variants on these introgressed segments decays as genetic distance increases. In case of a recent admixture event, ALD is found over long genetic distances [28] and is therefore easy distinguishable from short range non-ALD [14]. For ancient admixture however, a pairwise LD between introgressed variants is calculated using an ascertainment scheme to filter for the introgressed variants in the admixed population, since LD form admixture and LD from other processes is in a similar length range [18]. The pairwise LD as a function of genetic distant is then fitted using an exponential distribution holding a point estimate for the time since the admixture event [14,19].

The two approaches of obtaining the length of introgressed segments: searching for introgressd haplotypes

The other approach first tries to infer introgressed haplotypes directly, by partitioning an individuals genome into admixture segments [16,29,30]. Multiple methods are available focusing on different characteristics of introgressed haplotypes, such as LD patterns and density of derived variants, to infer them [11,12,30–33]. In the second step, the length distribution of these directly inferred introgressed haplotypes can be used for dating by fitting an exponential distribution to it.

Dates of archaic introgression obtained so far

The two approaches were used to estimate the Neandertal human admixture to be 37,000–86,000 ya (years ago) [18] as well as 40,510–54,454 ya (95% CI) [34] using modern day genomes and 50,000 to 60,000 ya using an ancient genome [7]. An additional contemporaneous admixture event between Neandertals and East-Asian

populations was suggested to explain the higher amount of Neandertal ancestry in these populations [35,36]. The Denisovan human admixture was dated to be between 44,000–54,000 ya using modern day genomes [11]. An additional Denisovan admixture event was suggested for East-Asian populations by identifying Denisovan segments, which are more diverged from the Denisovan high-coverage genome then previously found segments [37]. Comparing the mean length of the two groups of segments introgressed from divergent Denisovan populations did, however, not reveal significant differences, suggesting a lack of power to distinguish the two events by time [21,37]. Analysing genomes from Papuan individuals revealed two time separated admixture events with Denisovan, one in line with previous estimates at 45.7 kya (95% CI 31.9-60.7 kya) and one exclusive to Papuans dated to be around 29.8 kya (95% CI 14.4-50.4 kya) [21].

What we want to do

Here we want to further investigate archaic admixture dates. Especially focusing on the Neandertal admixture into modern humans. Archaic admixture events are assumed to be long time in the past, with few, potentially distantly related genomes available from the archaics as proxies for the true introgressing population. We want to investigate the effect of the model and data assumptions on the obtained dates. We are particularly interested in modeling long continuous admixture between Neandertals and modern humans instead of a one generation pulse. Thereby we want to resolve what does a time estimate of a pulse mean, is it distinguishable from a continuous admixture distribution over multiple hundred generations. This could hold implications about the contact between Neandertals and modern humans, e.g. was gene flow a single isolated incident or common over the whole time of their co-existence. Neandertals and modern human populations overlapped in some regions for a long period of time [38], potentially resulting in continuous admixture over hundreds of generations. We are especially interested in the extrema of a continuous admixture distribution, since the start and end point can inform us about the first contact between humans and Neandertals and the time of extinction of Neandertals.

Assumptions on the data

Both approaches, the direct measure of the segment length and the indirect ALD based method, without first directly inferring introgressed segments such as the popular ALDER method [19], rely on some fairly strong modelling assumptions [15,16,22]. We want to investigate which of these assumptions can be relaxed without causing a strong bias in the estimate.

Segments are is independent and identically exponentially distributed

First, it is assumed that the length of the introgressed segments is independent and identically exponentially distributed. This means that the number of ancestors per admixed individual is identical, such that the proportion of introgressed segments is equal among the individuals. This however is not true for Neandertal admixture in non-African population, since East-Asian population where fund to carry more Neandertal ancestry than other non-African populations [36]. Segments are independent such that segments of the same ancestor do not recombine to form longer segments. Liang and Nielsen 2014 showed that the violation of the independent and identically exponentially distributed segments length assumption can downward bias admixture time estimates when the true admixture time is very recent, the number of ancestors is small or the admixing population is small. This increases the chance of admixture segments to recombine into a larger segment mimicking a segment length distribution of a more recent admixture [22].

Recombination Map

Second, it is assumed that the recombination rate is known. Since the recombination is used as a clock the correct genetic distance between introgressed segments or variants on the segments has to be known.[this is a repeat] Recombination maps are used to assign the genetic length between introgressed segments/variants. Different maps are available[add refs].

Either focusing on short term information by using parent-offspring data to directly observe recombination events [39,40], searching for admixture tracts breakpoints [41,42] or long term population averages of recombination inferred by LD [43–45]. Different genetic maps are highly correlated on a megabase scale but reveal significant population differences on a finer scale [40,41]. The average length of Neandertal introgressed segments is approximately 50 kb [18], and so fine-scale recombination maps are essential. Furthermore, recombination maps can also vary over time [46]. These uncertainties in the genetic map were found to downward bias the time estimates by missing certain recombination events resulting in assuming introgressed segments longer than they actually are [18]. Sankararaman et al. 2012 suggested a method for accounting for uncertainty in the recombination map by estimating a correction parameter for a given recombination map by comparing the distances between a pair of markers in the map to the number of crossovers that span those markers as observed in a pedigree map. This however limits the applicability of the method, since it requires a population-specific map as well as a pedigree based map to estimate the error in the map [7,11,18].

Demography

A third issue are assumptions about the demography. Demographic processes can create similar genomic signals as admixture. Genomic segments carrying these signals can falsely be included in the estimate of the length distribution of admixed segments, leading to a bias. Admixture segments can also get lost by genetic drift, making it harder to estimate the length distribution. Figure 1 B shows two possible demographies for the Neandertal admixture, a simplified one without no population changes and a more complex one including structured populations, bottlenecks and additional gene flow. The extreme reduction in population sizes of the admixing population can create non-admixture LD [47]. Substructure in an ancestral human population, where human population in Africa are subdivided with a low migration rate, as shown in Figure 1 B, can also introduce divergent chromosomal segment with variants in strong LD, mimicking Archaic introgressed segments [48,49]. Moreover additional gene flow after the Neandertal admixture event between the African and admixed non-African population was shown to potentially influence the detected amount of Neandertal ancestry [50] as well as the time estimates [18].

Ascertainment

An additional bias might be introduced in the ALD based dating method since it requires ascertaining the SNPs for admixture informative sites, especially for archaic admixture [7,18]. The problem here is that introgressed variants from the ancestral archaic population are ascertained using only a few sequenced archaic individuals as proxies for the actual introgressing archaic population. These individuals are potentially only distantly related to the introgressing population, which might result in missing introgressed variants.

Pulse model

Finally, the admixture is assumed to happened in only a single generation [14]. This single pulse is mathematically convenient since all the introgressed segments entering in this single generation can be modeled using a single exponential distribution. This however limits the information gained form the dating since in case of a multi-generation long admixture the single-pulse model results in a mean admixture time between start end end of the gene flow. Ideally one would like to obtain the start and end, hence the duration of admixture between Neandertals and non-African populations. If admixture is happening over n generations, this results in a mixture of n different exponential distribution one for each set of introgressed segments per generation [51]. Previous studies tried to model this, using the ALD, by assuming not a continuous admixture but summarizing it into different independent events. In this case, the decay curve

is comprised out of n different exponential terms, each for one admixture event [51,52]. Methods were introduced to model continuous admixture using a polynomial function with the number of coefficients related to the number of generations of continuous admixture [53]. These models however work best with recent admixture, in the last 100 generations, with a substantial admixture proportion of around 0.3 [53]. This is outside of the range for scenarios of Neandertal admixture, were the admixture proportion is around 0.03 and the timing around 1500 generations ago [5,8,18].

We introduce a model for continuous admixture, where the migration rate over time is Gamma distributed with two parameters, one for the mean time of continuous admixture and one for the duration of the admixture. First, we examining the effect of long continuous admixture on the admixture time estimates in comparison to the effects of the aforementioned model assumptions. Therefore, we quantitatively evaluate the effect of the admixture model, recombination, demography and analysis parameters on the inference. Second, we define the expectation of the resulting segment length distribution for continuous Gama distributed admixture being Lomax distributed, holding a parameter for the duration of admixture. This expectation works for both methods to infer the segments length, either directly or by using the ALD decay. Using this model, we investigate under which scenarios the parameters of the Lomax-distribution can be accurately estimated and for which parameters we can distinguish a pulse-like admixture event from a continuous event. Together this critically evaluates the current state of knowledge about different aspects of archaic admixture time.

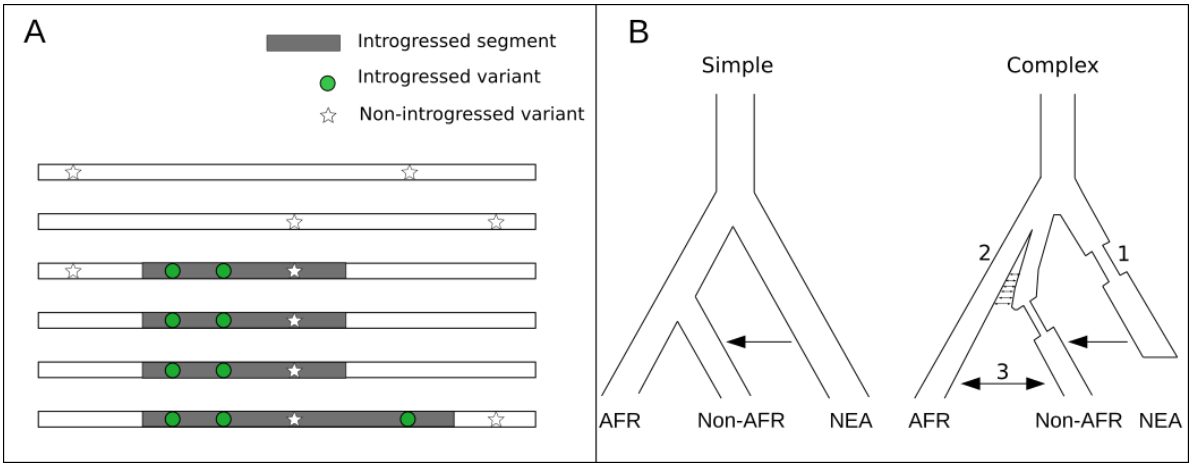

Figure 1: A) Chromosomal sections with introgressed segments (grey). Introgressed variants (green circles) are in high LD compared to the background (stars). The ALD approach estimates linkage between the introgressed variants, wheres the haplotype approach tries to estimate the segment directly. B) Two possible demographies of humans splitting into African population (AFR) and non-African (non-AFR) with admixture from Neandertals (NEA) into non-Africans. The simple demographic model with constant population sizes and the complex with rapid population size changes (1), substructure in Africa, where after an initial earlier split and isolation the structured population exchange migrants till the out-of-Africa event (2) and additional gene flow between Africans and non-Africans after the Neandertal admixture (3).

Methods

We conducted various simulations to assess the effect of continuous admixture compared to a pulse under ideal circumstances. We changed recombination and demographic parameters to simulate more realistic models. We compared the effect of these parameters together with analysis parameters to the effect of continuous admixture on the estimates. After assessing these effects we evaluate the possible parameter range for using a Lomax distribution to fit the ALD decay, enabling to obtain a duration of continuous admixture.

Simulations

We used the msprime coalescent simulator [54] for simulations with sample sizes chosen to reflect presently available data: We simulate 176 diploid African individuals and 170 diploid non-Africans, corresponding to the number of haploid Yoruba (YRI) and Central Europeans from Utah (CEU) sequences in the 1000 Genomes project data [55]. Since three high coverage Neandertal sequences are available [5,9] we choose to

simulate three diploid genomes. For each individual we simulated 20 chromosomes with a length of 150 Mb each. The mutation rate was set for all simulations to $2 * 10^{-8}$ per base per generation. The recombination rate was set to $1 * 10^{-8}$ per base pair per generation unless specified otherwise. The demographic parameters are based on previous studies dating Neandertal admixture [7,18,34]. In the "simple" model (Figure 1 B), the effective population size is assumed constant at Ne=10000 for all populations, the split time between modern humans and Neandertals is 10000 generations ago and a split time between Africans and non-Africans is 2550 generations ago. The migration rate from Neandertals into non-Africans was set to zero before the split from Africans, to ensure no Neandertal ancestry in Africans. Each simulation was repeated 100 times.

Gene Flow

In the admixture pulse model, gene flow is a one generation long resulting in an exponentially distributed admixture-induced linkage disequilibrium (ALD) decay curve (Eq. 1), with λ as the rate parameter of the exponential distribution holding the inverse of time since the admixture event t as a function of genetic distance d,

$$\begin{aligned} x_i &\sim exp(\lambda) \\ \mathbb{E}[x] &= \frac{1}{\lambda} \end{aligned} \tag{1}$$

Continuous gene flow over time was modeled as a gamma distribution (Eq. 2)

$$m_i \sim \Gamma(k+1, \frac{1}{\theta}) \tag{2}$$

Where k is the shape and θ the rate parameter. The parameter values are chosen such that the mean length $\frac{1}{\lambda}$ of the exponentially distributed ALD decay curve resulting from the one generation admixture pulse, is equal to the mean length of a ALD decay curve, as a result of continuous migration with the same total amount of migrants, modeled using a Lomax distribution (Eq. 3)

$$\begin{aligned} x_i &\sim Lomax(k, \theta) \\ \mathbb{E}[x] &= \frac{1}{\lambda} = \frac{\theta}{k} \end{aligned} \tag{3}$$

$$\lambda = \frac{k}{\theta} \tag{4}$$

$$where \qquad k = \mathbb{E}[x]\,\theta \qquad and \qquad \theta = \frac{\mathbb{E}[x]}{Var[x]}$$

Equation (Eq. 4) shows the relationships between the distribution parameters such that the resulting decay mean length are equal. Here x is in generations.

Recombination map

Uncertainties in the recombination map were previously shown to bias admixture time estimates. To investigate the effect of more realistic recombination rate variation we simulated using a recombination map. We either used the African-American-Map [41] or the HapMap phase 3 [56] for simulations under a variable recombination rate, for simplicity, we used the same recombination map (150 Mb of chromosome 1, excluding the first 10 Mb) for all simulated chromosomes. The mean recombination rate was calculated from the 150 Mb map (1.843 $\frac{cM}{Mb}$ AAMap and 1.549 $\frac{cM}{Mb}$ HapMap). To emulate uncertainties in the genetic map we either used the mean recombination rate from the respective map to calculate the genetic position from the physical position for each SNP, used linear interpolation based on the other map (e.g. AAMap used for the msprime simulation and HapMap used to assign genetic distances) or used the same map for simulation and assigning genetic distances.

Inferred demography

Demography such as population size changes are known to influence LD patterns and hence create false admixture signals. To test the impact of demographic history on admixture time estimates, we simulate a more realistic and complex demographic history based on effective population sizes and split times estimated from Neandertal and present-day human genomes. MSMC estimates from YRI as representatives for Africans and CEU for non-Africans from Schiffles & Durbin 2014 [57] were used together with PSMC [58] inferred demographic model for Neandertals based on the Vindija33.19 high-coverage genome [9]. In order to use the effective population size estimates at a given time point for modern humans from Schiffels & Durbin 2014 (Figure 4 Excel Table) for our simulations, we first transformed the time points given in years back to generations by using 30 years for one generation, as assumed in the original study. Second, since the original estimates are based on a different mutation rates ($1.25 * 10^{-8} \frac{bp}{gen}$), we corrected all estimates for

the mutation rate used in the simulations ($2 * 10^{-8} \frac{bp}{gen}$). The split times between Neandertals and modern humans as well as between Africans and non-Africans were kept the same as in the simple simulations (1000 and 2550 generations ago, respectively). The population size of the ancestor of Neandertals and humans before the split was set to 1000. To simulate branch shortening caused by the extinction of Neandertals, Neandertals were sampled 750 generations before the Africans and non-Africans.

Admixture time estimates

Ascertainment scheme

Ascertainment schemes are used to select certain variable positions of interest in a genome. Ascertainment schemes can be used to enrich for Neandertal informative sites in the test population to remove noise and amplify the ALD signal [18]. Two ascertainment schemes were tested to enrich for Neandertal informative sites, which were used previously [18]. The lower-enrichment ascertainment scheme filters for SNPs fixed for the ancestral state in Africans and polymorphic in Neandertals. The higher-enrichment ascertainment scheme restricts the analysis on SNPs fixed for the ancestral state in Africans, polymorphic in Neandertals and polymorphic in non-Africans.

ALD calculation and curve fitting

The pairwise weighted LD between the ascertained SNPs a certain genetic distance d apart was calculated using ALDER [19]. A minimal genetic distance d_0 between SNPs is set either to 0.02 cM and 0.05 cM. This minimal distance cutoff removes extreme short range LD likely confounded by non-ALD, facilitating the fitting procedure. To obtain the mean time estimates the data is fitted with an exponential distribution shown in equation 5, using a non-linear least-square optimization algorithm implemented in R [59]. Where A is the intercept, t the time since the admixture event in generations, d the genetic distance in cM and c is a constant modeling background LD. The model was fitted following Moorjani et al 2016 [34]. The duration of continuous admixture is modeled using the Lomax fit shown in Eq. 6. We used the notation from Kozubowski et al. 2008 [60]. The starting value of t is taken from the exponential fit to ensure convergence of the model. Model validity diagnostics like distribution of residuals, residuals plotted against fitted values and model stability diagnostics like dfbetas were checked for selected simulations and showed no obvious deviations from the model assumptions [61]. To compare the two nested models we used Akaike's information criterion (AIC) measuring the goodness of fit while penalizing the addition of new parameters p and thus controlling for under- and overfitting (Eq. 7).

$$ALD \sim \ A\, e^{-td} + c \tag{5}$$

$$ALD \sim \ A \left(\frac{1}{1 + \frac{1}{k} td} \right)^{k} + c \tag{6}$$

$$AIC = 2p - 2\ln(\hat{L}) \tag{7}$$

Modeling parameter effect sizes

To model and compare parameter effect sizes we simulated 100 replications for each combination of the previously introduced parameters: ascertainment scheme (LES/HES), minimal genetic distance ($d_0 = 0.02\,cM / d_0 = 0.05\,cM$), demography (simple/inferred), recombination rate (constant/variable), gene flow model (pulse/continuous). Results of simulations where the nls-optimization to fit the ALD decay curve did not converge were removed (6 out of 3200). We used a Gaussian linear least-squares model to estimate the effect size of the different parameters. The deviation between the estimated admixture time and the true admixture time, the error in the estimate, was used as the response to the non interacting parameters as model predictors assuming normal distributed and homogeneous residuals. Calculation of variance inflation factors of the predictors where not indicative for colinearity between predictors. Model stability assessment using dfbetas showed stable model predictors and residuals and plotting residuals against fitted values revealed slight deviations from a normal distribution mostly driven by few extreme values of simulations under the variable recombination rate, however, no overall obvious deviation from the assumptions.

Results

Theoretical framework for continuous admixture

First we want to establish a model of continuous admixture and an expectation of the tract length distribution under this model in an ideal-case from perfect data. For this purpose, we assume that the lengths of introgressed tracts are perfectly known. In this case, under some models, the distribution of introgressed tract lengths L_i can be written as

$$x_i \sim exp(\lambda) \tag{8}$$

where t is the time when the fragment entered the population, and λ is a parameter that depends on the model assumptions and recombination rate r [22]. E.g under the SMC, $\lambda = (1-m)r$, and under the SMC' allowing for back coalescence, $\lambda = 2N(1-m)(1-exp^{-t/2N})r$. For Neandertal admixture where m, the admixture fraction, is typically low, the exponential assumption is satisfied [22]. For scenarios where the length of admixture tracts is not exponential, e.g. because admixture is recent or very old, our results do not apply.

It is widely assumed that Neandertal ancestry entered the modern human population over a very short period. As an alternative model, we need to consider t not as a single point in time, but as a random variable itself that follows a mixture distribution $\mathcal{D}_t$. The most widely studied is a small number of discrete "pulses" of admixture, in which case $\mathcal{D}_t$ is categorical. Here, we instead assume a continuous $\mathcal{D}_t$; more precisely we assume $\mathcal{D}_t$ follows a $\Gamma(k+1, \lambda t)$-distribution. This has a number of advantages:

- We just need one additional parameter k, that can be interpreted as the duration of gene flow, instead of the minimal of two additional parameters required for the pulse model.
- The tract length distribution L_i follows a Lomax-distribution, i.e. has the analytically density $Pr(L = l) = \frac{k\lambda t}{(1+\lambda l t)^{k+1}}$
- The cdf is $Pr(L > l) = (1+\lambda l t)^{-k}$
- The mean tract-length is $\frac{1}{\lambda t}$ for all $k > 0$, and undefined otherwise.
- As k approaches infinity, we recover the exponential distribution. Thus, if tract length are directly inferred, one can use a likelihood-ratio test to distinguish continuous from discrete gene flow. As the special case of exponentially lies on the boundary of the parameter space, the test-statistic does not follow a χ^2-distribution (https://pdfs.semanticscholar.org/e6a3/ae271354701ecca576cf94821869f6069e9e.pdf). This is however not possible when tract length are indirectly inferred by ALD.

Using this model we i) examining the effect of continuous admixture on the admixture time estimates calculated using the exponential model assuming a pulse like admixture. ii) comparing this effect to the effects by demography, recombination rate and analysis parameters used for the indirect inference of admixture tract length using ALD, namely the ascertainment scheme and the minimal distance between SNPs. iii) are interested under which conditions the parameters of the Lomax-distribution can be estimated accurately for a scenario of Neandertal admixture and if it is possible to distinguish a pulse-like admixture event (resulting

in exponentially-distributed track lengths) from a continuous event (resulting in Lomax-distributed track lengths).

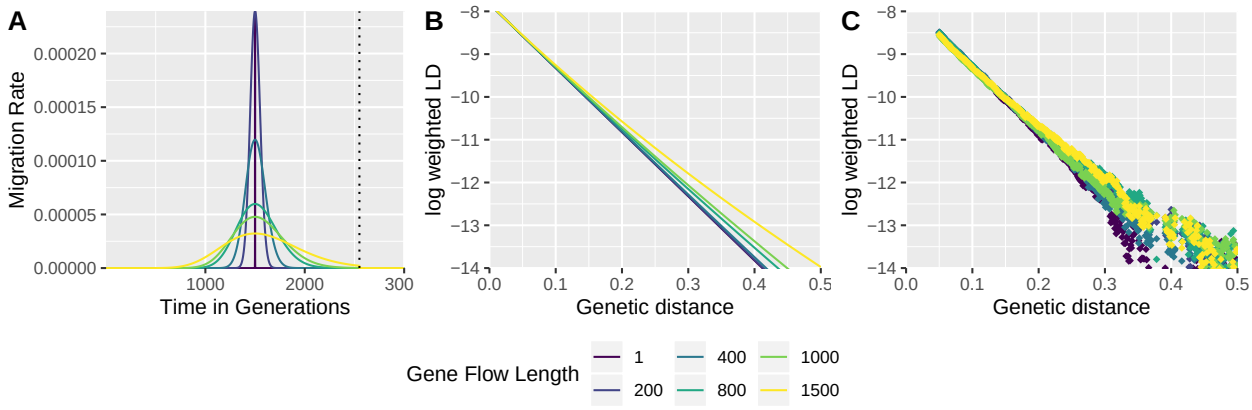

Figure 2: A) Migration rate per generation modeled using a Gamma distribution for different gene flow length, dotted line indicates maximum time of gene flow. B) The expected LD decay modeled as a Lomax distribution for the different length. C) The observed LD decay from msprime simulations

The effect of continuous admixture on the admixture time estimates

To asses the effect of continuous ancient admixture on the admixture time estimates calculated from modern human populations, we compared two models of admixture between Neandertals and non-Africans using coalescent simulations. The simplified model of a pulse like gene flow and a continuous admixture event with several generations of gene flow. The total amount of gene flow m from Neandertals into non-Africans in the two models is equal. Gene flow is modeled using a Gamma distribution (Eq. 2) holding the migration rate per generation for different length of continuous gene flow (Figure 2 A). The shape and scale parameter of the Gamma distribution are chosen such that the resulting weighted LD decay curves as functions of genetic distance share the same mean for a pulse (Eq. 1) and a continuous admixture (Eq. 3). Sites informative for Neandertal introgression into non-Africans where enriched using the lower-enrichment ascertainment scheme filtering for SNPs ancestral in all Africans and polymorphic in Neandertals. The pairwise weighted LD between the ascertained SNPs was computed using the ALDER program. Figure 2 C shows the resulting weighted LD for different length of continuous admixture ranging from a one generation pulse to 1500 generations.

Comparing estimates for different mean admixture times ranging from 250 generations ago to 2000 generations ago, simulated either with a pulse or a continuous admixture with a length of 50 % of the mean

admixture time, reveals minor deviations between the two scenarios and the true admixture time of 3% to 5% for the pulse and 1% to 7% for the continuous (Figure 3 A). Estimates for mean admixture times older than 1000 generations show a slight overestimation, which is lesser for the estimates of simulations under a continuous gene flow . The slight overestimation is consistent with the findings of previous studies estimating ancient admixture times using modern admixed populations [7,18,34]. To further investigate the effect of pulse and continuous gene flow on the admixture time estimates, we compared different durations of continuous gene flow simulated under a fixed mean time of admixture of 1500 generations ago, displayed in Figure 3 B. Estimates between pulse and continuous gene flow start to deviate for 800 generations of continuous admixture, with increasingly lower estimates for simulations under continuous gene flow compared to simulations under a pulse like gene flow per increase of gene flow duration. This bias is likely caused by the differences in LD between sites entered in the tails of the gamma distribution. LD between sites arising from early admixture events, simulated by the right tail of the gamma distribution, is not detected anymore, while LD between sites from late admixture is still present, biasing the estimate towards younger dates. However, deviations in estimates between the two scenarios of ~100 generations in the most extreme case are moderate compared to the mean admixture time of 1500 generations, making admixture time estimates for long continuous gene flow compatible with the used method.

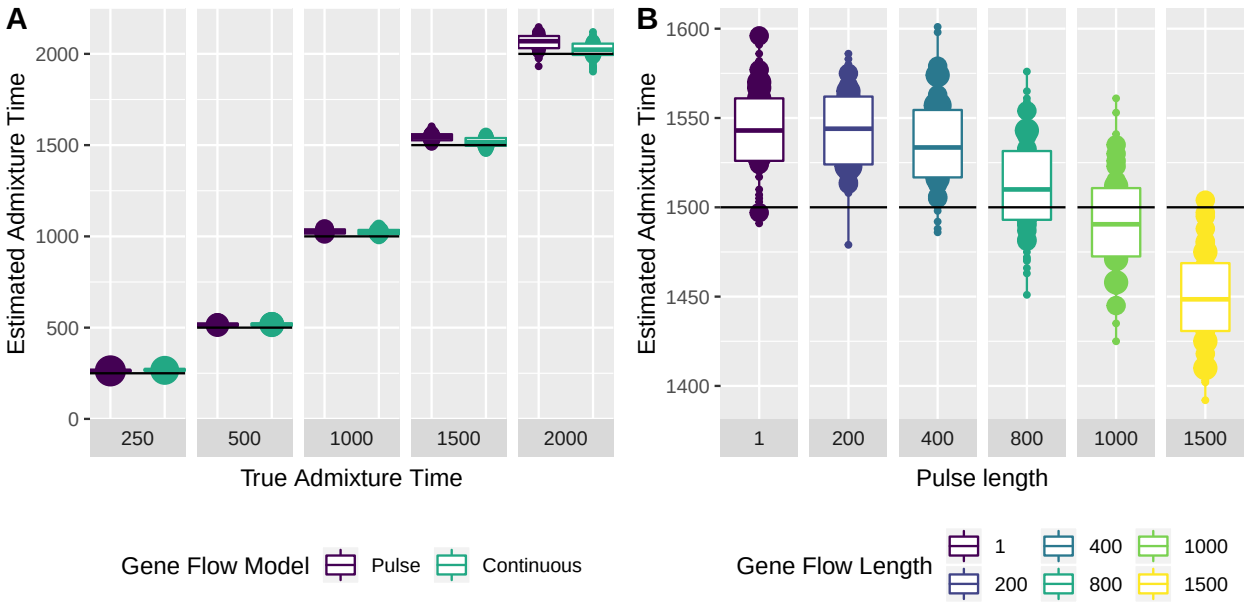

Figure 3: A) Comparison of mean admixture time estimates between pulse and continuous gene flow for different admixture times. The length of continuous gene flow corresponds to 50% of the mean admixture time, black line indicates true mean admixture time. B) Comparison of mean admixture time estimates for simulations with a mean time of admixture of 1500 generations ago, at a varying length of gene flow. Boxplot created from 100 simulation replicates, respectively.

Comparing effect sizes

Having established that the duration of continuous admixture under ideal circumstances is only marginally influential on admixture time estimates, we deploy more realistic simulation scenarios. Specifically, we investigated the influence of two analysis parameter by changing the ascertainment scheme to a higher-enrichment scheme requiring additional to the LES non-Africans to be polymorphic at a SNP, and applying a larger minimal distance of 0.05 cM between SNPs used for the fitting of the exponential distribution. Additionally, we changed the demographic model from a simple to a complex one based on inferred effective population sizes and time points taken from MSMC estimates for Africans and non-Africans and PSMC estimates for the simulated Neandertals. Split times between the population where unchanged and Neandertals were sampled 750 generations before the modern humans to simulate branch shortening. We further applied the African-American-Map to simulate varying recombination rates across the 150 Mb long chromosomes, instead of a constant rate. We simulated every combination of these parameter

sets resulting in 32 different sets with 100 replications each (Supp. Fig. 1). A Gaussian linear model was applied to estimate effect sizes of the four predictors being ascertainment scheme, minimal distance, demography and recombination on the bias of admixture estimates (Supplement Table 1). Figure 4 shows the comparison of the bias on admixture time estimates between the previously used model (ascertainment = LES, $d_0 = 0.02cM$, demography = simple and recombination = constant) further refereed to as the standard model and a model with one of the four parameter changed, respectively and the corresponding model prediction. The previously observed overestimation of the standard model was estimated to be 122 (+7/-7) generations. Every parameter change results in lower estimates compared to the standard model, with the biggest difference between a constant and a varying recombination (-863 +5/-6 generations) and the smallest differences between a pulse and continuous gene flow model (-40 +5/-6 generations). Most accurate estimates are achieved using the LES ascertainment scheme in combination with a minimal distance of 0.05 cM. These analysis parameter combination shows more robustness towards changes in the demography (Supplement Figure 1).

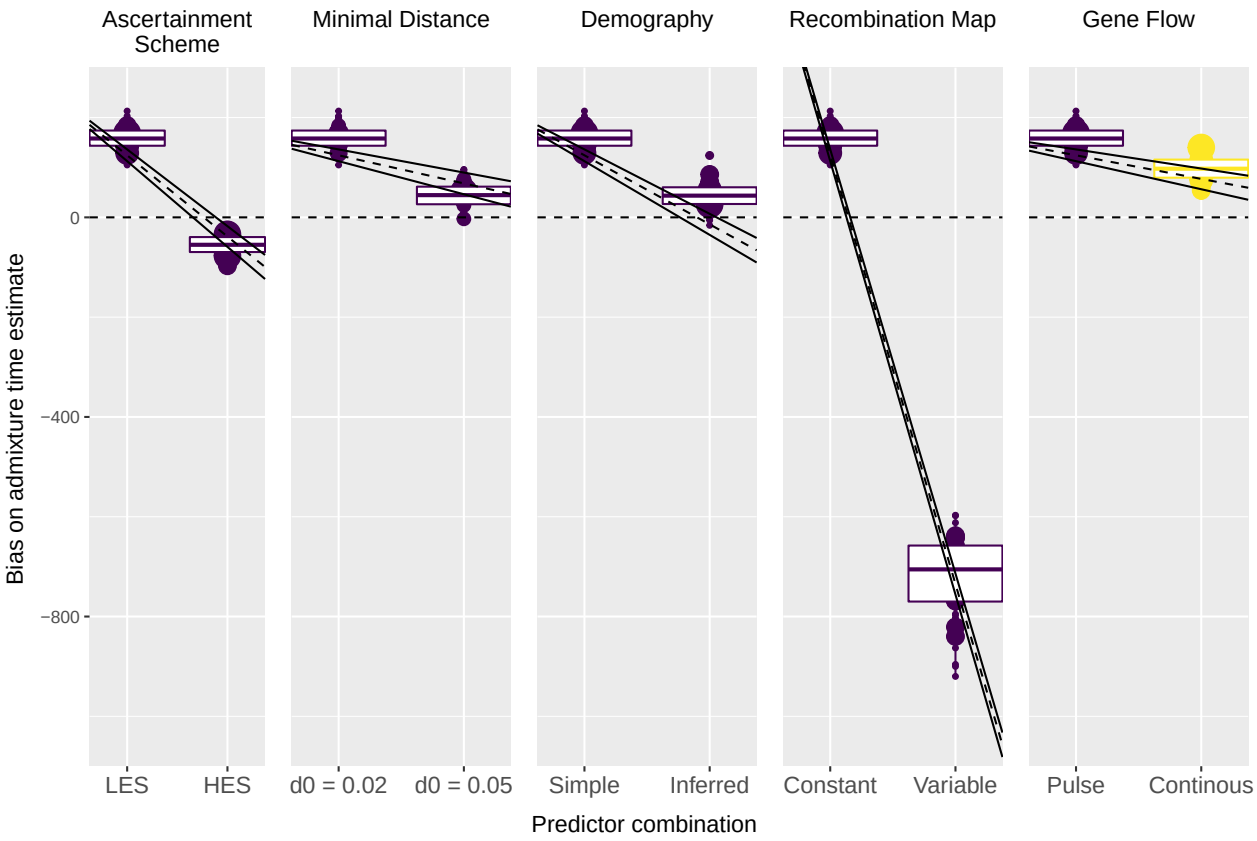

Figure 4: Comparison of the bias on the admixture time estimates between each parameter and the standard model of 100 replications each. Dotted line represents the model prediction with the 2.5 % and 97.5 % compatibility intervals solid lines. Dotted horizontal line indicates unbiased admixture estimates.

Estimating the Lomax-parameters under different conditions

We investigated the effect of continuous admixture in comparison to other parameters on the admixture time estimates of a model only considering a single generation pulse. The continuous admixture revealed to have the smallest effect, whereas the recombination rate shows a sever underestimation of the admixture time. Building on the previous result we want to find out the conditions to retrieve the Lomax parameters i.e. the duration of gene flow. We simulated an admixture scenario under a simple demography with varying duration of continuous admixture and sampled the population at different time points since the end of the gen flow. Doing so allows us to examine how close one has to sample form the end of a continuous admixture to still accurately estimate its duration. Moreover, since the recombination map is highly influential we test the duration estimates under a constant and variable recombination using the HapMap genetic map, whereby we

correct the genetic distance by: assuming a constant rate, using the AAMap or the HapMap itself. We used the LES ascertainment scheme in combination with a minimal distance of 0.05 cM. To fit the Lomax we can take advantage of the fact that the bias of the mean time estimates using the simplified one generation pulse model is pretty accurate. We can thus estimate lambda first by fitting an exponential and in a second step estimating k using a Lomax distribution with a starting parameter for lambda received form the exponential fit. Figure 5 A and C show the mean time estimates received from the Exponential fit. A shows estimates for different durations of continuous admixture with all populations sampled 50 generations after the end of gene flow. C shows estimates for a 800 generation long continuous admixture sampled after different times after the end of the gene flow. In both scenarios, for simulations under a constant recombination the mean time can be estimated confidently with only slight underestimation for long continuous gene flow of 1500 generations. Estimates for simulation under a recombination map only assuming a constant rate when calculating the LD results in severely underestimation of the mean time estimates. Using the AAMap yields results closer to the true value and less downwards biased, however only using the exact same genetic map gives unbiased results. Figure 5 B and D shows the corresponding duration estimates by the Lomax fit. Accurate estimates can be obtained throughout the different gene flow length under a constant recombination rate, when sampled recent from the end of the continuous admixture. The further away one samples form the end of the admixture event the less accurate the estimates. All simulations using a recombination map show a much higher variance in the estimates, especially for a gene flow length longer then 800 generations or sampled later then 50 generations away from the end of gene flow the estimates are not reliably. If no precise genetic map is used to inferred the genetic distance between SNPs, no accurate duration estimates can be obtained. The mean duration over all replicates for the simulation corrected by the exact same map seems relatively unbiased for a recent gene flow with a duration shorter 1000 generations. Under a realistic scenario with a varying recombination rate, the duration can only be accurately estimated with a highly precise map for recent continuous admixture events not longer then 400 generations. With regard to scenarios of Neandertal admixture, accurately estimating the duration of possible continuous admixture from present day human genomes even under a constant rate is probably is underpowered. The time since the end of the gene flow when using present day human genome is to far away such that the signal of continuous gene flow, even under simple simulation scenarios using a constant recombination rate, is not strong enough.

To distinguish a pulse like admixture event resulting in an exponentially distributed ALD decay from a Lomax distributed ALD decay associated with a continuous admixture, we compared the two model via goodness-of-fit to the ALD data. Therefore we used Akaike's information criterion comparing the exponential model nested in the Lomax. We expect that, since the AIC is penalizing extra parameter, that a Lomax fit

to a pulse like simulation scenario should be rejected, wheres for all the other scenarios the Lomax would be preferred. However, for the simplified simulation scenarios under a constant recombination rate were we obtained the best results the Lomax displays a lower AIC in 59 fits out of 100 replications. Model comparison for scenarios of true continuous admixture always rested in the Lomax having a lower AIC. The AIC reject the null hypothesis to often and thus seem to be a to liberal test.

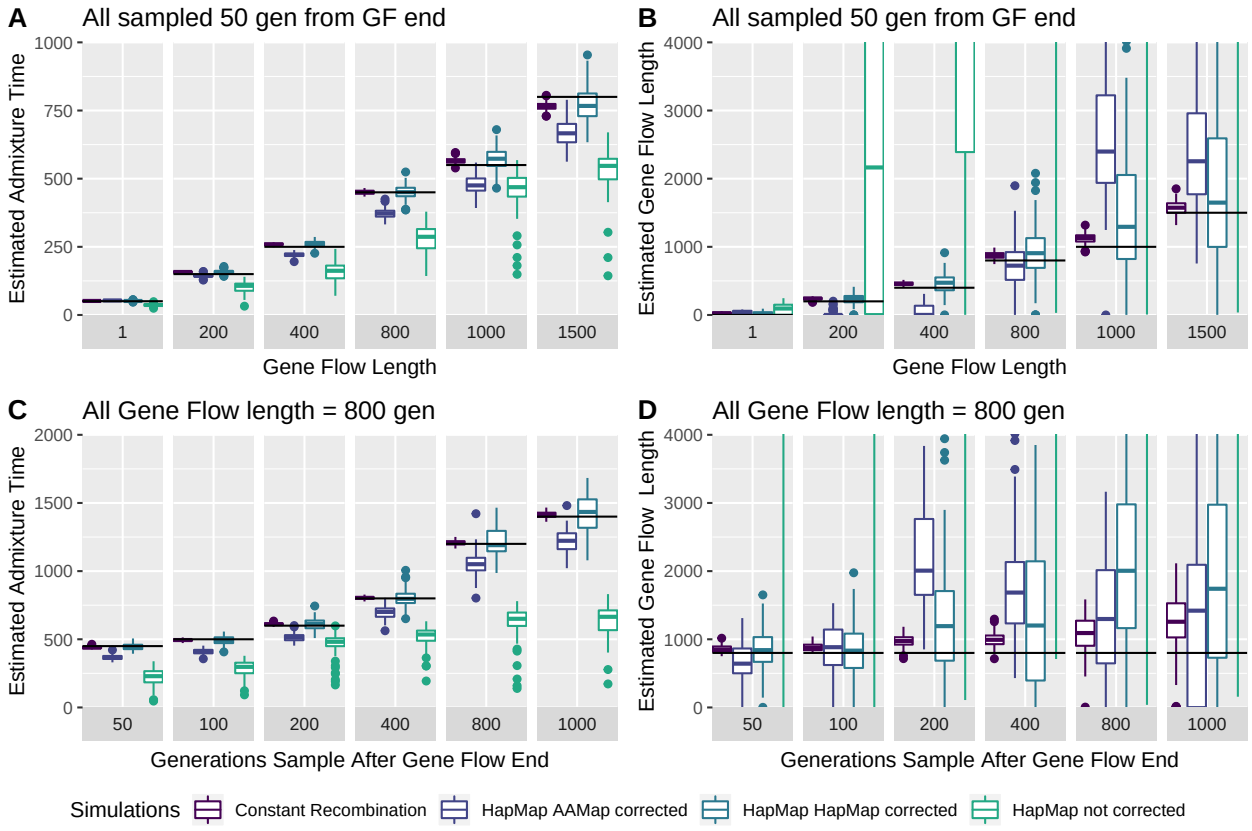

Figure 5: ll

Discussion

Conclusions:

- The mean time of gene flow can be reliably estimated with a pulse model even if the gene flow is actually continuous.
- Technical model variables (ascertainment scheme, minimal distance, demography and recombination map) outweigh the effect of continuous gene flow.

- The Lomax model is under powered to inferring continuous admixture for parameters relevant for Neandertal admixture.
- Estimating the the mean time and especially the duration of gene flow is only possible with a highly precise genetic map.
- The signal of continuous gene flow is hard to detect if the gene flow is not long enough or to far in the past.
- Reliably distinguishing the models is difficult.

Refereces

1. Abbott R, Albach D, Ansell S, Arntzen JW, Baird SJE, Bierne N, et al. Hybridization and speciation. Journal of Evolutionary Biology. 2013;26: 229–246. doi:10.1111/j.1420-9101.2012.02599.x

2. Manuel M de, Kuhlwilm M, Frandsen P, Sousa VC, Desai T, Prado-Martinez J, et al. Chimpanzee genomic diversity reveals ancient admixture with bonobos. Science. 2016;354: 477–481. doi:10.1126/science.aag2602

3. Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, et al. A Draft Sequence of the Neandertal Genome. Science. 2010;328: 710. doi:10.1126/science.1188021

4. Reich D, Green RE, Kircher M, Krause J, Patterson N, Durand EY, et al. Genetic history of an archaic hominin group from Denisova Cave in Siberia. Nature. 2010;468: 1053. Available: https://doi.org/10.1038/nature09710

5. Prüfer K, Racimo F, Patterson N, Jay F, Sankararaman S, Sawyer S, et al. The complete genome sequence of a Neanderthal from the Altai Mountains. Nature. 2013;505: 43. Available: https://doi.org/10.1038/nature12886

6. Fu Q, Hajdinjak M, Moldovan OT, Constantin S, Mallick S, Skoglund P, et al. An early modern human from romania with a recent neanderthal ancestor. Nature. 2015;524: 216–219. doi:10.1038/nature14558

7. Fu Q, Li H, Moorjani P, Jay F, Slepchenko SM, Bondarev AA, et al. Genome sequence of a 45,000-year-old modern human from western Siberia. Nature. 2014;514: 445. Available: https://doi.org/10.1038/nature13810

8. Sankararaman S, Mallick S, Dannemann M, Prüfer K, Kelso J, Pääbo S, et al. The genomic landscape

of neanderthal ancestry in present-day humans. Nature. 2014;507: 354–357. doi:10.1038/nature12961

9. Prüfer K, Filippo C de, Grote S, Mafessoni F, Korlević P, Hajdinjak M, et al. A high-coverage Neandertal genome from Vindija Cave in Croatia. Science. 2017;358: 655. doi:10.1126/science.aao1887

10. Meyer M, Kircher M, Gansauge M-T, Li H, Racimo F, Mallick S, et al. A high-coverage genome sequence from an archaic denisovan individual. Science. 2012;338: 222–226. doi:10.1126/science.1224344

11. Sankararaman S, Mallick S, Patterson N, Reich D. The Combined Landscape of Denisovan and Neanderthal Ancestry in Present-Day Humans. Current Biology. 2016;26: 1241–1247. doi:10.1016/j.cub.2016.03.037

12. Vernot B, Tucci S, Kelso J, Schraiber JG, Wolf AB, Gittelman RM, et al. Excavating Neandertal and Denisovan DNA from the genomes of Melanesian individuals. Science. 2016;352: 235–239. doi:10.1126/science.aad9416

13. Falush D, Stephens M, Pritchard JK. Inference of population structure using multilocus genotype data: Linked loci and correlated allele frequencies. Genetics. 2003;164: 1567–1587. Available: https://www.genetics.org/content/164/4/1567

14. Moorjani P, Patterson N, Hirschhorn JN, Keinan A, Hao L, Atzmon G, et al. The History of African Gene Flow into Southern Europeans, Levantines, and Jews. PLOS Genetics. 2011;7: e1001373. doi:10.1371/journal.pgen.1001373

15. Pool JE, Nielsen R. Inference of Historical Changes in Migration Rate From the Lengths of Migrant Tracts. Genetics. 2009;181: 711–719. doi:10.1534/genetics.108.098095

16. Gravel S. Population Genetics Models of Local Ancestry. Genetics. 2012;191: 607–619. doi:10.1534/genetics.112.139808

17. Pugach I, Matveyev R, Wollstein A, Kayser M, Stoneking M. Dating the age of admixture via wavelet transform analysis of genome-wide data. Genome Biology. 2011;12: R19. doi:10.1186/gb-2011-12-2-r19

18. Sankararaman S, Patterson N, Li H, Pääbo S, Reich D. The Date of Interbreeding between Neandertals and Modern Humans. PLOS Genetics. 2012;8: e1002947. doi:10.1371/journal.pgen.1002947

19. Loh P-R, Lipson M, Patterson N, Moorjani P, Pickrell JK, Reich D, et al. Inferring Admixture Histories of Human Populations Using Linkage Disequilibrium. Genetics. 2013;193: 1233. doi:10.1534/genetics.112.147330

20. Pugach I, Duggan AT, Merriwether DA, Friedlaender FR, Friedlaender JS, Stoneking M. The Gateway from Near into Remote Oceania: New Insights from Genome-Wide Data. Molecular Biology and Evolution.

2018;35: 871–886. doi:10.1093/molbev/msx333

21. Jacobs GS, Hudjashov G, Saag L, Kusuma P, Darusallam CC, Lawson DJ, et al. Multiple Deeply Divergent Denisovan Ancestries in Papuans. Cell. 2019;177: 1010–1021.e32. doi:10.1016/j.cell.2019.02.035

22. Liang M, Nielsen R. The Lengths of Admixture Tracts. Genetics. 2014;197: 953–967. doi:10.1534/genetics.114.162362

23. Shchur V, Svedberg J, Medina P, Corbett-Detig R, Nielsen R. On the distribution of tract lengths during adaptive introgression. bioRxiv. 2019; 724815. doi:10.1101/724815

24. Chimusa ER, Defo J, Thami PK, Awany D, Mulisa DD, Allali I, et al. Dating admixture events is unsolved problem in multi-way admixed populations. Briefings in Bioinformatics. [cited 30 Jul 2019]. doi:10.1093/bib/bby112

25. Chakraborty R, Weiss KM. Admixture as a tool for finding linked genes and detecting that difference from allelic association between loci. PNAS. 1988;85: 9119–9123. doi:10.1073/pnas.85.23.9119

26. Stephens JC, Briscoe D, O'Brien SJ. Mapping by admixture linkage disequilibrium in human populations: Limits and guidelines. Am J Hum Genet. 1994;55: 809–824.

27. Wall JD. Detecting ancient admixture in humans using sequence polymorphism data. Genetics. 2000;154: 1271–1279. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1460992/

28. Patterson N, Hattangadi N, Lane B, Lohmueller KE, Hafler DA, Oksenberg JR, et al. Methods for high-density admixture mapping of disease genes. The American Journal of Human Genetics. 2004;74: 979–1000. doi:10.1086/420871

29. Price AL, Tandon A, Patterson N, Barnes KC, Rafaels N, Ruczinski I, et al. Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. PLOS Genetics. 2009;5: e1000519. doi:10.1371/journal.pgen.1000519

30. Lawson DJ, Hellenthal G, Myers S, Falush D. Inference of Population Structure using Dense Haplotype Data. PLOS Genetics. 2012;8: e1002453. doi:10.1371/journal.pgen.1002453

31. Racimo F, Marnetto D, Huerta-Sánchez E. Signatures of Archaic Adaptive Introgression in Present-Day Human Populations. Molecular Biology and Evolution. 2017;34: 296–317. doi:10.1093/molbev/msw216

32. Seguin-Orlando A, Korneliussen TS, Sikora M, Malaspinas A-S, Manica A, Moltke I, et al. Paleogenomics. Genomic structure in Europeans dating back at least 36,200 years. Science (New York, NY). 2014;346: 1113–1118. doi:10.1126/science.aaa0114

33. Skov L, Hui R, Shchur V, Hobolth A, Scally A, Schierup MH, et al. Detecting archaic introgression

using an unadmixed outgroup. PLOS Genetics. 2018;14: e1007641. doi:10.1371/journal.pgen.1007641

34. Moorjani P, Sankararaman S, Fu Q, Przeworski M, Patterson N, Reich D. A genetic method for dating ancient genomes provides a direct estimate of human generation interval in the last 45,000 years. Proceedings of the National Academy of Sciences. 2016;113: 5652. doi:10.1073/pnas.1514696113

35. Kim BY, Lohmueller KE. Selection and reduced population size cannot explain higher amounts of neandertal ancestry in east asian than in european human populations. Am J Hum Genet. 2015;96: 454–461. doi:10.1016/j.ajhg.2014.12.029

36. Vernot B, Akey JM. Complex history of admixture between modern humans and neandertals. Am J Hum Genet. 2015;96: 448–453. doi:10.1016/j.ajhg.2015.01.006

37. Browning SR, Browning BL, Zhou Y, Tucci S, Akey JM. Analysis of Human Sequence Data Reveals Two Pulses of Archaic Denisovan Admixture. Cell. 2018;173: 53–61.e9. doi:10.1016/j.cell.2018.02.031

38. Higham T, Douka K, Wood R, Ramsey CB, Brock F, Basell L, et al. The timing and spatiotemporal patterning of Neanderthal disappearance. Nature. 2014;512: 306–309. doi:10.1038/nature13621

39. Coop G, Wen X, Ober C, Pritchard JK, Przeworski M. High-resolution mapping of crossovers reveals extensive variation in fine-scale recombination patterns among humans. Science. 2008;319: 1395–1398. doi:10.1126/science.1151851

40. Kong A, Thorleifsson G, Gudbjartsson DF, Masson G, Sigurdsson A, Jonasdottir A, et al. Fine-scale recombination rate differences between sexes, populations and individuals. Nature. 2010;467: 1099–1103. doi:10.1038/nature09525

41. Hinch AG, Tandon A, Patterson N, Song Y, Rohland N, Palmer CD, et al. The landscape of recombination in african americans. Nature. 2011;476: 170. Available: https://doi.org/10.1038/nature10336

42. Wegmann D, Kessner DE, Veeramah KR, Mathias RA, Nicolae DL, Yanek LR, et al. Recombination rates in admixed individuals identified by ancestry-based inference. Nat Genet. 2011;43: 847–853. doi:10.1038/ng.894

43. McVean GAT, Myers SR, Hunt S, Deloukas P, Bentley DR, Donnelly P. The fine-scale structure of recombination rate variation in the human genome. Science. 2004;304: 581–584. doi:10.1126/science.1092500

44. Myers S, Bottolo L, Freeman C, McVean G, Donnelly P. A fine-scale map of recombination rates and hotspots across the human genome. Science. 2005;310: 321–324. doi:10.1126/science.1117196

45. A second generation human haplotype map of over 3.1 million SNPs. Nature. 2007;449: 851–861.

doi:10.1038/nature06258

46. Auton A, Fledel-Alon A, Pfeifer S, Venn O, Ségurel L, Street T, et al. A fine-scale chimpanzee genetic map from population sequencing. Science. 2012;336: 193. doi:10.1126/science.1216872

47. Schaper E, Eriksson A, Rafajlovic M, Sagitov S, Mehlig B. Linkage disequilibrium under recurrent bottlenecks. Genetics. 2012;190: 217–229. doi:10.1534/genetics.111.134437

48. Nei M, Li W-H. Linkage disequilibrium in subdivided populations. Genetics. 1973;75: 213–219. Available: https://www.genetics.org/content/75/1/213

49. Pfaff CL, Parra EJ, Bonilla C, Hiester K, McKeigue PM, Kamboh MI, et al. Population structure in admixed populations: Effect of admixture dynamics on the pattern of linkage disequilibrium. Am J Hum Genet. 2001;68: 198–207. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1234913/

50. Petr M, Pääbo S, Kelso J, Vernot B. Limits of long-term selection against neandertal introgression. PNAS. 2019;116: 1639–1644. doi:10.1073/pnas.1814338116

51. Pickrell JK, Patterson N, Loh P-R, Lipson M, Berger B, Stoneking M, et al. Ancient west eurasian ancestry in southern and eastern africa. PNAS. 2014;111: 2632–2637. doi:10.1073/pnas.1313787111

52. Zhou Y, Yuan K, Yu Y, Ni X, Xie P, Xing EP, et al. Inference of multiple-wave population admixture by modeling decay of linkage disequilibrium with polynomial functions. Heredity. 2017;118: 503–510. doi:10.1038/hdy.2017.5

53. Zhou Y, Qiu H, Xu S. Modeling continuous admixture using admixture-induced linkage disequilibrium. Sci Rep. 2017;7: 1–10. doi:10.1038/srep43054

54. Kelleher J, Etheridge AM, McVean G. Efficient Coalescent Simulation and Genealogical Analysis for Large Sample Sizes. PLOS Computational Biology. 2016;12: e1004842. doi:10.1371/journal.pcbi.1004842

55. The 1000 Genomes Project Consortium, Auton A, Abecasis GR, Altshuler (Co-Chair) DM, Durbin (Co-Chair) RM, Abecasis GR, et al. A global reference for human genetic variation. Nature. 2015;526: 68. Available: https://doi.org/10.1038/nature15393

56. A second generation human haplotype map of over 3.1 million SNPs. Nature. 2007;449: 851–861. doi:10.1038/nature06258

57. Schiffels S, Durbin R. Inferring human population size and separation history from multiple genome sequences. Nature Genetics. 2014;46: 919. Available: https://doi.org/10.1038/ng.3015

58. Li H, Durbin R. Inference of human population history from individual whole-genome sequences. Nature.

2011;475: 493. Available: https://doi.org/10.1038/nature10231

59. R Core Team. R: A language and environment for statistical computing. 2019. Available: https://www.R-project.org/

60. Kozubowski TJ, Panorska AK, Qeadan F, Gershunov A, Rominger D. Testing exponentiality versus pareto distribution via likelihood ratio. Communications in Statistics - Simulation and Computation. 2008;38: 118–139.

61. Quinn G, Keough M. Experimental Design and Data Analysis For Biologists. 2002. doi:10.1017/CBO9780511806384

Supplement

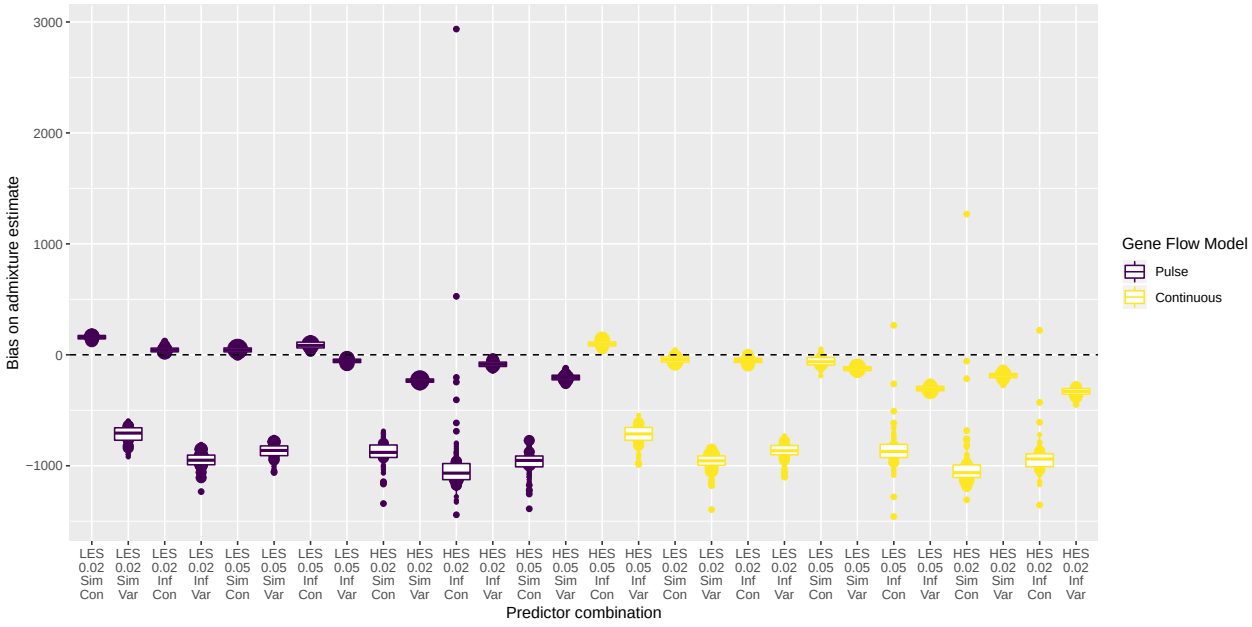

Fig. S 1: Comparison of the bias on the admixture time estimates in generations between all combinations the parameters: ascertainment scheme = LES/HES, d_0 = 0.02/0.05 cM, demography = simple/inferred, recombination = constant/variable and the gene flow model = pulse/continuous. Dotted horizontal line indicates unbiased admixture estimates.

Tab. S 1: Estimates for the predictor from the Gaussian linear model with 2.5/95.5 % compatibility intervals in generations, standard error, t and p-values.

	org	2.5 %	97.5 %	Std. Error	t value	Pr(>|t|)
(Intercept)	124.35797	112.57232	136.14362	6.010919	20.68868	0
AscertainmentHES	-162.79214	-171.75061	-153.83366	4.569002	-35.62969	0
min_dist0.05	-56.32267	-66.26121	-46.38414	5.068852	-11.11152	0
DemographyInferred	-138.38616	-147.68281	-129.08951	4.741477	-29.18630	0
Recomb.ratevariable	-859.26142	-868.55807	-849.96477	4.741477	-181.22230	0
GFContinous	-47.28775	-56.24623	-38.32928	4.569002	-10.34969	0