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)

2020-01-16

Abstract

Introduction

Detecting admixture, i.e. gene flow between previously isolated populations can shed light into the complex evolutionary history of populations. The sequencing of the Neandertal [1] and Denisovan [2] genome revealed admixture events between Neandertals and modern humans outside of Africa, as well as an independent admixture event between Denisovans and Asian populations. A question of great interest is when this admixture happened. 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. The reduction in length should be proportional to time since the admixture event. Hence, the length distribution of introgressed chromosomal segments is informative of the time since the admixture event and can be approximated by an exponential process, also known as the recombination clock [3][4]. The challenge here is to correctly infer the length of the introgressed segments. There are two main categories of methods to date admixture [5]. One uses the admixture-induced linkage disequilibrium (ALD) decay. Since introgressed segments entered the admixed population on one chromosome fairly recent, the variants they carry are expected to be in high linkage disequilibrium to each other [6]. Over time the LD between the introgressed variants approaches an equilibrium state, as recombination breaks the introgressed chromosomal segments apart and thereby destroying the association between the variants.The decay of pairwise ALD between variants informative for admixture is approximated by an exponential distribution holding a point estimate for the time since the admixture event [7][8]. This method was used to estimate the Neandertal human admixture to be 47,000–65,000 years ago [9] and the Denisovan human admixture to be between 44,000–54,000 years ago [10]. The other haplotype-based methods are partitioning an individuals genome into admixture segments, thereby directly inferring the

length of introgressed segments [4][11][12]. The partitioning is either based on reconstructing an admixed chromosome from reference chromosomes from the two ancestral populations (ChromoPainter) or on criteria characteristic for admixture, such as high density of SNPs in high LD (S*). Recently, Jacobs et al. 2019 used a combination of haplotype-based methods (ChromoPainter [12], a HMM [13][14]and S* [15]) to inferred Denisovan segments in Papuan individuals and revealed two time separated admixture events 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) [16]. An additional Denisovan admixture event was suggested for East-Asian populations by identifying Denisovan segments, using Sprime, which are more diverged from the Denisovan high-coverage genome then previously found segments [17]. No date could be obtained however for this admixture event [17][16]. The two different methods, the direct measure of the segment length and the indirect ALD based method, without first directly inferring introgressed segments such as the popular ALDER program [8], typically require fairly strong assumptions on the data. 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. 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[18]. Second 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. Uncertainties in the genetic map where found to downward bias the time estimates by missing certain recombination events resulting in assuming introgressed segments longer than they actually are. 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. This however limits the applicability of the method, since it requires a population-specific map as well as pedigree information to estimate the error in the map [9][19][10]. Third, the demography is known. Demographic events such like additional gene flow between the ancestral population and the admixed or deep structure in the ancestral population was previously reported to bias the estimates [9][19][10]. 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 [9][19]. The problem here is that introgressed variants from the ancestral archaic population are ascertained using only

a few sequenced individuals as proxies for the actual introgressing archaic population. These individuals are potentially only distantly related to the actual introgressing population or being older than it, which might result in missing introgressed variants. Finally, the admixture is assumed to happened in only a single generation. This assumption of pulse like events might be violated, since e.g. in the case of archaic and modern human admixture, populations overlapped in some regions for a long period of time [20], potentially resulting in continuous admixture over hundreds of generations.

This study aims to evaluate admixture time estimates, whereby we focus on a well studied example of archaic admixture into modern humans. Those admixture events are long time in the past, with few, potentially distantly related genomes available from the archaics as proxies for the true introgressing population. We are particularly interested in scenarios of long continuous gene flow between archaic and human populations and its signals in the ALD distribution. We introduce a model for continuous admixture, where the migration rate over time is gamma distributed with two parameter, 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 admixtue 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.

Methods

Simulations

Simulations were carried out using the msprime coalescent simulator [21] 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 phase 3 data [22]. Since 3 high coverage Neandertal sequences are available [23][24] we choose to simulate 3 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 2x10-8 per base per generation. The recombination rate was set to 1x10-8 per base pair per generation unless specified otherwise. The demographic parameters are based on previous studies [9][19][25]. In the "simple" model, the effective population size is assumed constant at Ne=10000 for all populations, the split time between modern humans and Neandertals 10000 generations ago and a split time between Africans and non-Africans of 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.

Gene Flow

In the simple model, gene flow is a one generation pulse resulting in an exponentially distributed admixture-induced linkage disequilibrium (ALD) decay curve (Eq. 1), with $\lambda = -dt$ 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,

$$pulse\ ALD \sim exp(\lambda) \tag{1}$$

The migration rate for continuous admixture was modeled as a gamma distribution (Eq. 2)

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

Where k is the shape and theta 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)

$$continuous\ ALD \sim Lomax(k, \theta) \tag{3}$$

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

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

Equation (Eq. 4) shows the relationships between the distribution parameters such that the resulting decay mean length are equal. Here E[x] is mean admixture time in generations and Var[x] is one fourth of the length of admixture time in generations squared, with x as the number of generations of admixture between two populations.

Recombination map

To investigate the effect of a more realistic recombination frequency we simulated using a recombination map. We either used the African-American-Map [26] or the HapMap phase 3 [27] 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 ($1.843 * 10^{-8} \frac{M}{bp}$ AAMap and $1.549 * 10^{-8} \frac{M}{bp}$ HapMap) and we used linear interpolation to calculate the genetic position from the physical position between SNPs. For correcting the genetic distance we interpolated the SNPs genetic distance by the map used for correction.

Inferred demography

To test the impact of demographic history on admixture time estimates, we simulate a more realistic and complex demographic history using estimated effective population sizes and split times. MSMC estimates from YRI as representatives for Africans and CEU for non-Africans from Schiffles & Durbin 2014 [28] were used together with PSMC [29] inferred models from the Vindija33.19 Neandertal. In these models, a mutation rate of 2x10-8 bp/gen and starting effective population size of 1000 is assumed. The estimates were corrected for branch shortening of 1000 generation . To integrate the effective population size at a given time point for modern humans from Schiffels & Durbin 2014 (Figure 4 Excel Table) 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). 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. Two ascertainment schemes were tested to enrich for Neandertal informative sites, which were used previously [9] 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 was calculated using ALDER [8]. A minimal genetic distance d_0 between SNPs is set either to 0.02 cM and 0.05 cM. To obtain the mean time estimates the data is fitted an exponential distribution shown in equation 5, using a non-linear least-square optimization algorithm implemented in R [30]. Where A is the intercept, $\frac{1}{s} = t$ 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 [25]. The duration of continuous admixture is modeled using the Lomax fit shown in Eq. 6. We used the particularization from Kozubowski et al. 2008 [31], here $k = \frac{1}{w}$. The starting value of s 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 [32]. 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).

$$ALD \sim \ A\, e^{-\frac{d}{s}} + c \tag{5}$$

$$ALD \sim \ A \left(\frac{1}{1 + \frac{wd}{s}} \right)^{\frac{1}{w}} + c \tag{6}$$

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

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

$$L_i \sim EXP[\lambda t] \tag{8}$$

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 [18]. 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 [18]. 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).

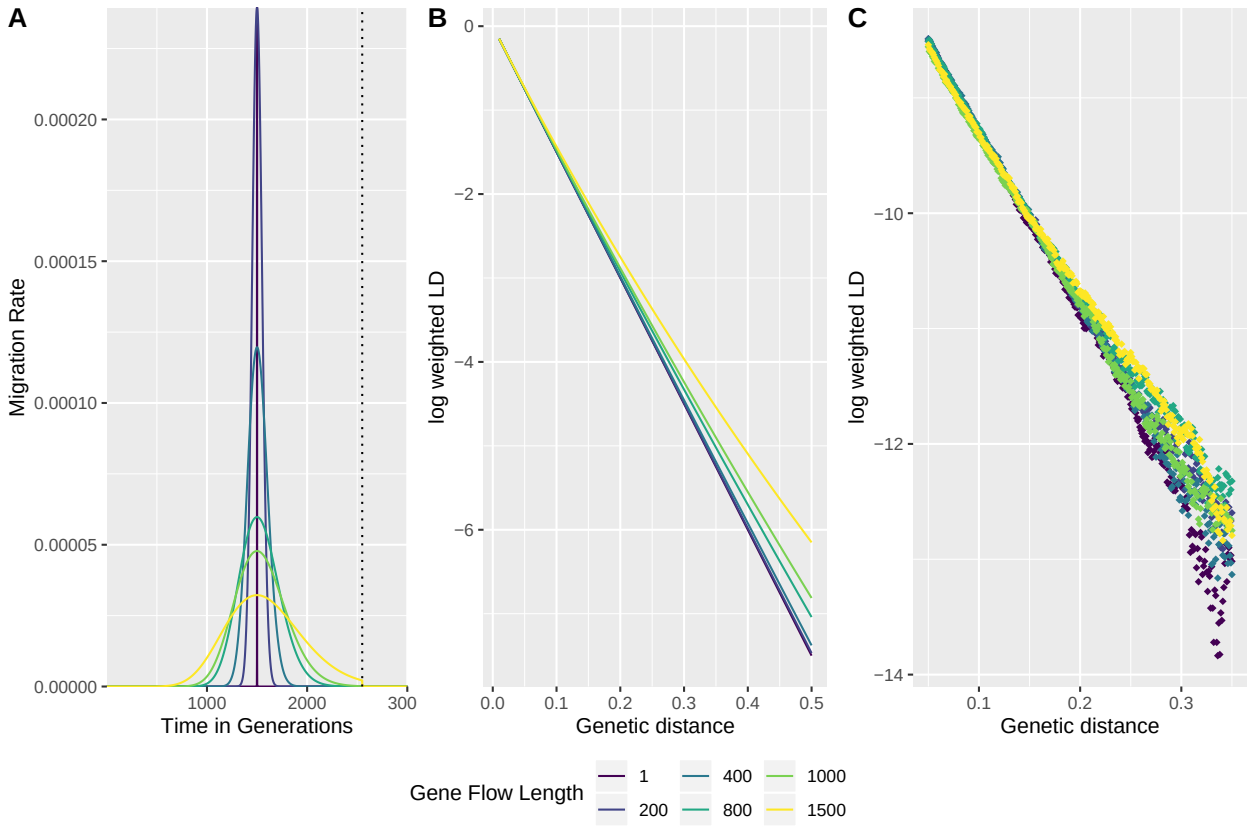

Figure 1: A) Migration rate per generation modeled using a Gamma distribution for different gene flow length, dotted line indicates maximum time of gene flow. B) The expected LD decay modeled as a Lomax distribution for the different length. C) The observed LD decay from msprime simulations

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. 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. Coalescent simulations were carried out simulating 176 Africans, 170 non-African and 3 Neandertal diploid genomes, corresponding to the number of YRI and CEU in the 1000 Genomes project phase 3 and the three available high-coverage Neandertal genomes. Each individual contains 20 chromosomes with a sequence length of 150 Mb, respectively. The total amount of non-symmetric gene flow from Neandertals into non-Africans is set to 3 %. Gene flow is modeled using a Gamma distribution (Eq. 2) holding the

migration rate per generation for different length of continuous gene flow (Figure 1 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 1 C shows the resulting log weighted LD for a pulse admixture and continuous admixture scenarios with differing duration. To estimate admixture times, the weighted LD as a function of genetic distance d is fitted using an exponential distribution holding the time since the admixture event t_{GF}, (Eq. 5). 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 no considerable deviations between the two scenarios and the true admixture time. 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 (Figure 2 C). The slight overestimation is consistent with the findings of previews studies estimating ancient admixture times using modern admixed populations [[9],[19]. 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 2 D. 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.

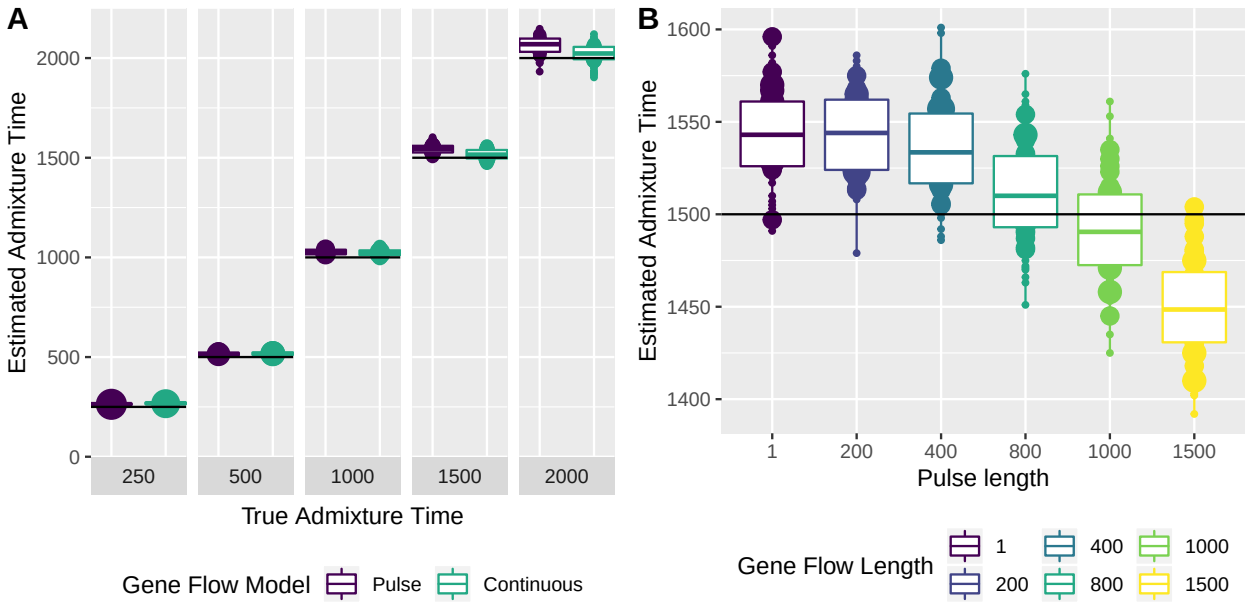

Figure 2: A) Comparison of mean admixture time estimates between pulse and continuous gene flow for different admixture times. The length of continuous gene flow corresponds to 50% of the mean admixture time, black line indicates true mean admixture time. B) Comparison of mean admixture time estimates 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

Next we compared the effect of the two different gene flows with other parameters potentially influencing admixture time estimates. 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 3 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).

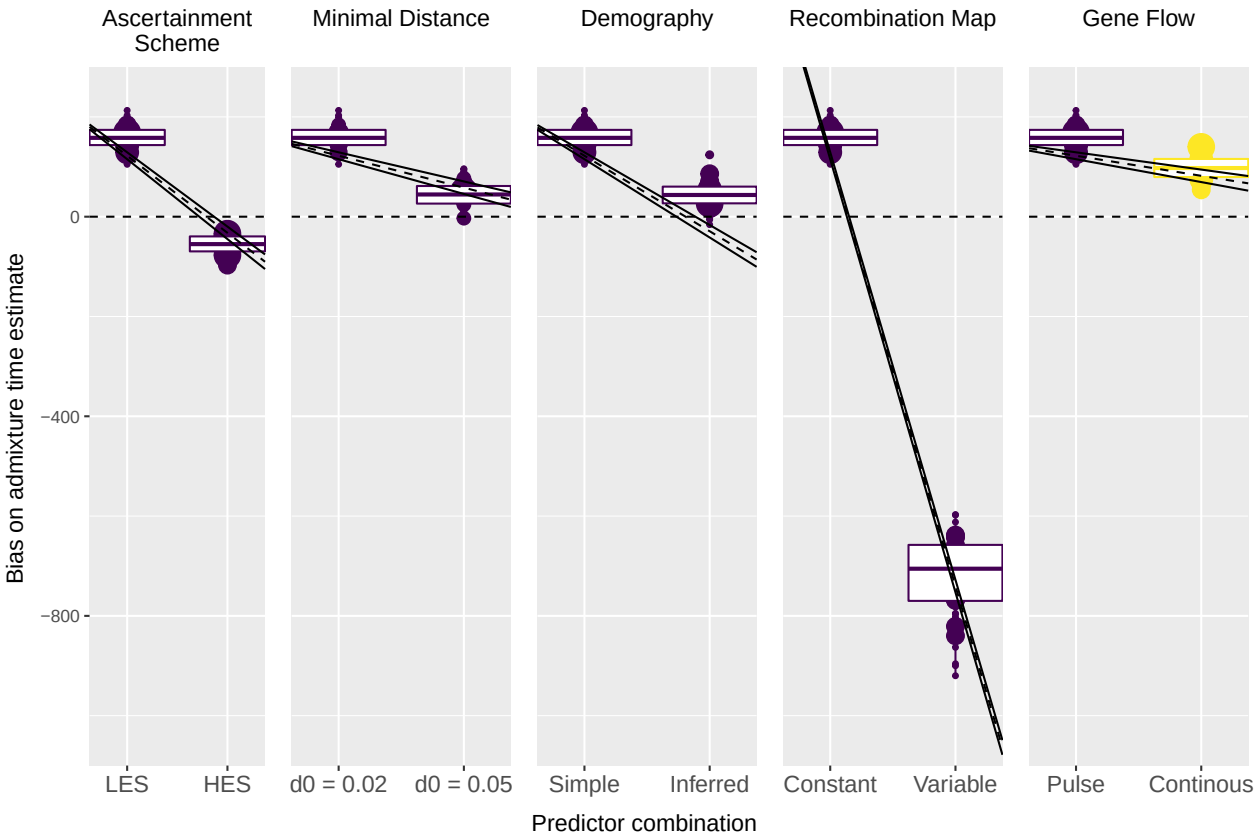

Figure 3: Comparison of the bias on the admixture time estimates between each parameter and the standard model of 100 replications each. Dotted line represents the model prediction with the 2.5 % and 97.5 % compatibility intervals solid lines. Dotted horizontal line indicates unbiased admixture estimates.

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 4 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 4 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 underpowerd. 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.

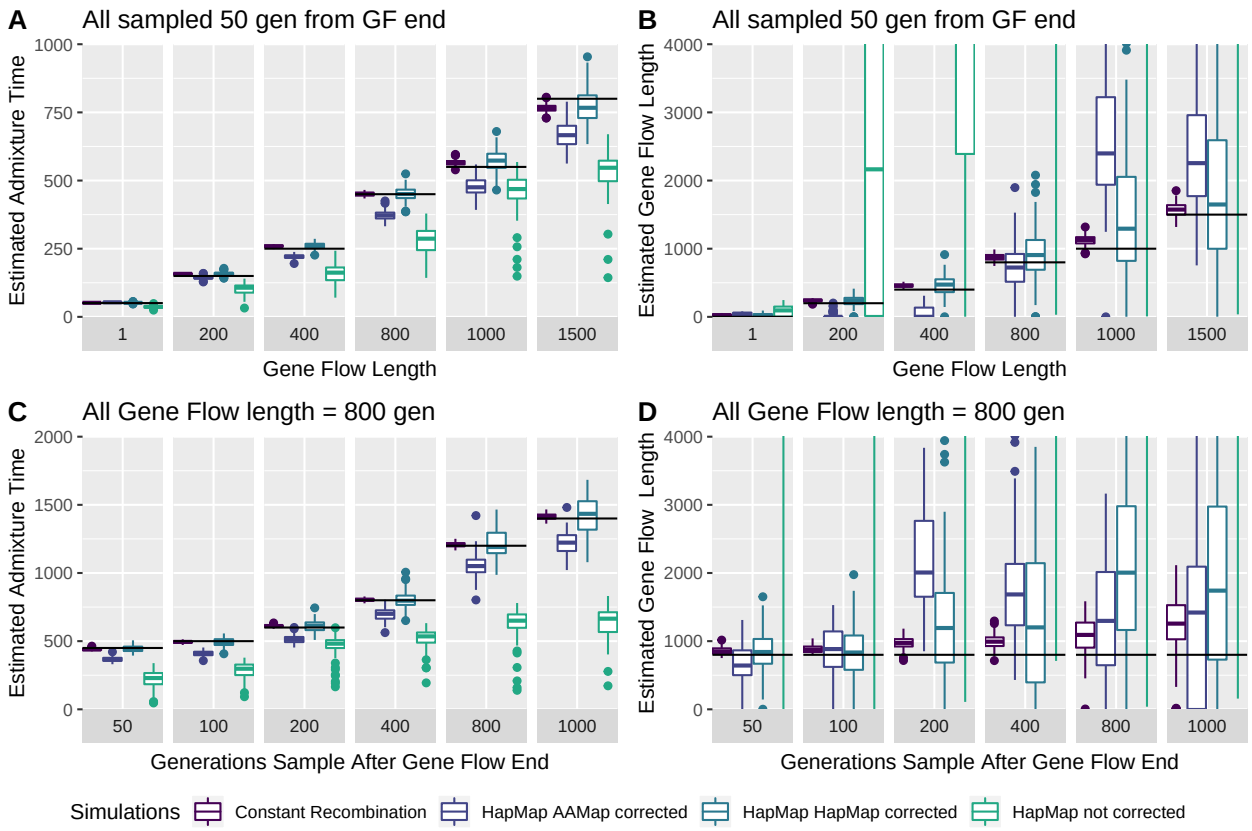

Figure 4: ll

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. 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

2. 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

3. 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

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

5. 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

6. 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/

7. 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

8. 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

9. 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

10. 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

11. 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

12. 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

13. 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

14. 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

15. 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

16. 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

17. 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

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

19. 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

20. 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

21. 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

22. 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

23. 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

24. 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

25. 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

26. 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

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

28. 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

29. 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

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

31. 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.

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

Supplement

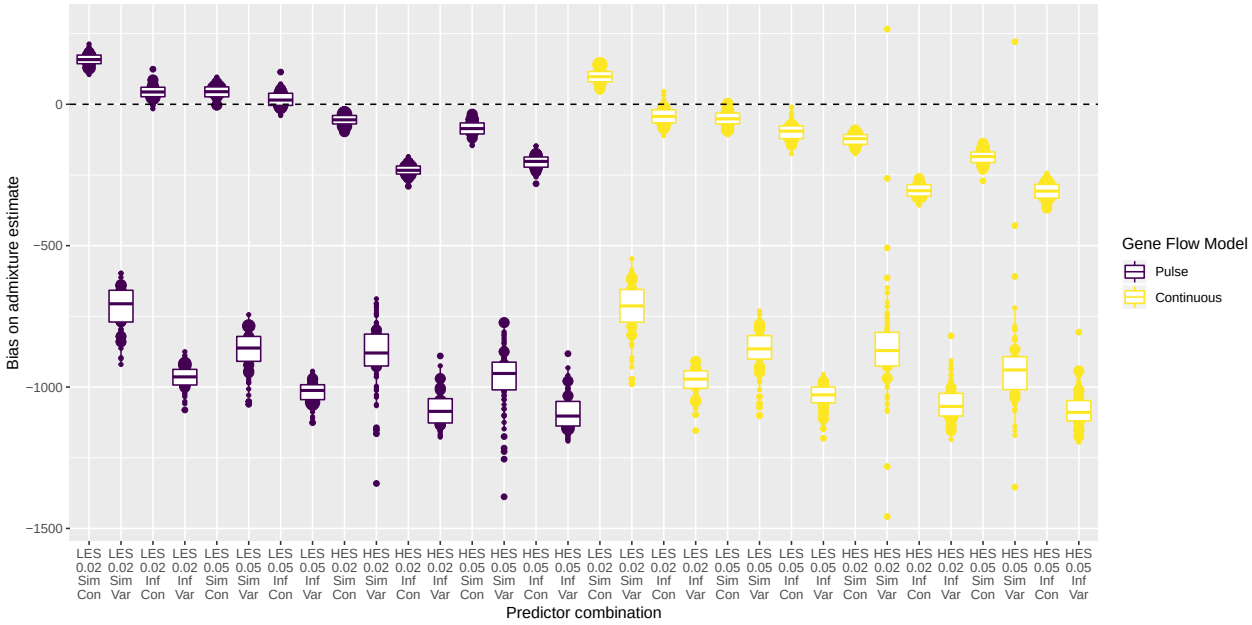

Fig. S 1: Comparison of the bias on the admixture time estimates in generations between all combinations the parameters: ascertainment scheme = LES/HES, d_0 = 0.02/0.05 cM, demography = simple/inferred, recombination = constant/variable and the gene flow model = pulse/continuous. Dotted horizontal line 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)	122.12209	115.22918	129.01499	3.515519	34.73800	0
AscertainmentHES	-154.49308	-160.12111	-148.86505	2.870409	-53.82267	0
min_dist0.05	-63.98735	-69.61538	-58.35932	2.870409	-22.29206	0
DemographyInferred	-151.27818	-156.90621	-145.65014	2.870409	-52.70265	0
Recomb.ratevariable	-862.50113	-868.12917	-856.87310	2.870409	-300.48020	0
GFContinous	-40.21065	-45.83868	-34.58262	2.870409	-14.00868	0