# An Extended Admixture Pulse Model Reveals the Limitations to Human–Neandertal Introgression Dating

Leonardo N. M. Iasi,[1] Harald Ringbauer ,[2] and Benjamin M. Peter *,[1]

[1]Department of Evloutionary Genetics, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany
[2]Department of Archaeogenetics, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany

***Corresponding author:** E-mail: benjamin_peter@eva.mpg.de.

**Associate editor:** Daniel Falush

## Abstract

**Neandertal DNA makes up 2–3% of the genomes of all non-African individuals. The patterns of Neandertal ancestry in modern humans have been used to estimate that this is the result of gene flow that occurred during the expansion of modern humans into Eurasia, but the precise dates of this event remain largely unknown. Here, we introduce an extended admixture pulse model that allows joint estimation of the timing and duration of gene flow. This model leads to simple expressions for both the admixture segment distribution and the decay curve of ancestry linkage disequilibrium, and we show that these two statistics are closely related. In simulations, we find that estimates of the mean time of admixture are largely robust to details in gene flow models, but that the duration of the gene flow can only be recovered if gene flow is very recent and the exact recombination map is known. These results imply that gene flow from Neandertals into modern humans could have happened over hundreds of generations. Ancient genomes from the time around the admixture event are thus likely required to resolve the question when, where, and for how long humans and Neandertals interacted.**

***Key words:*** **admixture dating, human–Neandertal admixture, gene flow, extended admixture pulse, Neandertal, recombination clock.**

## Introduction

The sequencing of Neandertal (Green et al. 2010; Prüfer et al. 2013, 2017; Mafessoni et al. 2020) and Denisovan genomes (Reich et al. 2010; Meyer et al. 2012) revealed that modern humans outside of Africa interacted, and received genes from these archaic hominins (Fu et al. 2014, 2015; Sankararaman et al. 2014, 2016; Vernot and Akey 2014; Malaspinas et al. 2016; Vernot et al. 2016). There are two major lines of evidence: First, Neandertals are genome-wide more similar to non-Africans than to Africans (Green et al. 2010). This shift can be explained by 2–4% of admixture from Neandertals into non-Africans (Green et al. 2010; Prüfer et al. 2013). Similarly, East Asians, Southeast Asians, and Papuans are more similar to Denisovans than other human groups, which is likely because of gene flow from Denisovans (Meyer et al. 2012).

As a second line of evidence, all non-Africans carry genomic segments that are very similar to the sequenced archaic genomes. As these putative *admixture segments* are up to several hundred kilobases (kb) long, it is unlikely that they were inherited from a common ancestor that predates the split of modern and archaic humans (Sankararaman et al. 2014; Vernot and Akey 2014). Rather, they entered the modern human populations through later gene flow (Sankararaman et al. 2012, 2014, 2016; Vernot and Akey 2014; Vernot et al. 2016).

However, substantial uncertainty remains about when, where, and over which period of time this gene flow happened. A better understanding of the location and timing of the gene flow would allow us to place constraints on the timing of movements of early humans, and the population genetic consequences of their interactions.

Archeological evidence puts some temporal boundaries on the times when Neandertals and modern humans might have interacted. The earliest currently known modern human remains outside of Africa is dated to around 188 thousand years ago (ka) (Hershkovitz et al. 2018; Stringer and Galway-Witham 2018), and the latest Neandertals are suggested to have lived between 37 and 39 ka old (Higham et al. 2014; Zilhão et al. 2017). Thus, the time window where Neandertals and modern humans might have been in the same area stretches over more than 140,000 years. However, there is less direct evidence of modern humans and Neandertals in the same geographical location at the same time. In Europe, for example, Neandertals and modern humans likely overlapped only for less than 10,000 years (Bard et al. 2020).

### Genetic Dating of Gene Flow

A common approach to learn about admixture dates from genetic data uses a *recombination clock* model: Conceptually, admixture segments are the result of the introduced chromosomes being broken down by recombination.

The first-generation offspring of an archaic and a modern human parent will have one whole chromosome each of either ancestry. Thus, the genomic markers in these individuals are in full ancestry linkage disequilibrium (ALD); all archaic variants are present on one DNA molecule, and all modern human variants on the other one.

If this individual has offspring in a largely modern human population, in each generation meiotic recombination will reshuffle the chromosomes, progressively breaking down the ancestral chromosome down into shorter segments of archaic ancestry (Falush et al. 2003; Gravel 2012; Liang and Nielsen 2014), and ALD similarly decreases with each generation after gene flow (Chakraborty and Weiss 1988; Stephens et al. 1994; Wall 2000).

This inverse relationship between admixture time and either segment length or ALD is commonly used to infer the timing of gene flow (Pool and Nielsen 2009; Moorjani et al. 2011; Pugach et al. 2011, 2018; Gravel 2012; Sankararaman et al. 2012, 2016; Loh et al. 2013; Hellenthal et al. 2014; Liang and Nielsen 2014; Jacobs et al. 2019). Most commonly, it is assumed that gene flow occurs over a very short duration, referred to as an *admixture pulse*, which is typically modelled as a single generation of gene flow (Moorjani et al. 2011). This model has the advantage that both the length distribution of admixture segments and the decay of ALD with distance will follow an exponential distribution, whose parameter is directly informative about the time of gene flow (Pool and Nielsen 2009; Gravel 2012; Liang and Nielsen 2014).

In segment-based approaches, dating starts by identifying all admixture segments, which can be done using a variety of methods (Seguin-Orlando et al. 2014; Sankararaman et al. 2016; Vernot et al. 2016; Racimo et al. 2017; Skov et al. 2018). The length distribution of inferred segments is then used as a summary for dating when gene flow happened.

Alternatively, ALD-based methods use linkage disequilibrium (LD) patterns, without explicitly inferring the segments (Chimusa et al. 2018) (fig. 1*B*). Instead, admixture dates are estimated by fitting a decay curve of pairwise LD as a function of genetic distance, implicitly summing over all compatible segment lengths (Moorjani et al. 2011; Loh et al. 2013).

## Neandertal Gene Flow Estimates

Using this approach, Sankararaman et al. (2012) dated the Neandertal–human admixture pulse to between 37–86 ka. Later, Moorjani et al. (2016) refined this date to 41–54 ka $\mathrm{CI}_{95\%}$ using an updated method, a different marker ascertainment scheme and a refined genetic map for European populations. A date of 50–60 ka was obtained from the analysis of the genome of *Ust'-Ishim*, a 45,000-year-old modern human from western Siberia. The inferred Neandertal segments in *Ust'-Ishim* are substantially longer than those in present-day humans, which makes their detection easier, and adds further evidence that gene flow between Neandertals and modern humans has happened relatively recently before *Ust'-Ishim* lived (Fu et al. 2014). In addition, we have direct evidence of gene flow from early modern humans from Oase (Fu et al. 2015) and Bacho Kiro (Hajdinjak et al. 2021), dated to 40 and 45ky, respectively. In genomes from both sites, segments of recent Neandertal ancestry less than ten generations before hint at admixture histories with late gene flow in Europe.

## Limitations of the Pulse Model

The admixture pulse model assumes that gene flow occurs over a short time period; however, it is currently unclear how long a time could still be consistent with the data. This makes admixture time estimates hard to interpret, as more complex admixture scenarios might be masked, and so gene flow could have happened tens of thousands of years before or after the estimated admixture time.

That admixture histories are often complicated has been shown in the context of Denisovan introgression into modern humans, where at least two distinct admixture events into East Asians and Papuans were proposed (Browning et al. 2018; Jacobs et al. 2019; Choin et al. 2021). Although the length distributions of admixture segments are similar between populations, there are differences in the genomic distribution of admixture segments, and their similarities to the sequenced high-coverage Denisovan (Browning et al. 2018; Massilani et al. 2020). In contrast, all Neandertal admixture segments are most similar to the Vindija Neandertal (Prüfer et al. 2017), but Neandertal ancestry is slightly higher in East Asians than Western Eurasians (Meyer et al. 2012; Wall et al. 2013; Kim and Lohmueller 2015; Vernot and Akey 2015; Villanea and Schraiber 2019).

One way to refine admixture time estimates is to include two or more distinct admixture pulses. The distribution of admixture segment lengths will then be a mixture of the segments introduced from each event. This is especially useful if the events are very distinct in time, for example, if one event is only a few generations back, and the other pulse occurred hundreds of generations ago (Fu et al. 2014, 2015). In this case, the admixture segments will be either very long if they are recent, or much shorter if they are older.

Zhou, Qiu, et al. (2017) extended this model to continuous mixtures, using a polynomial function as a mixture density. However, they found that even for relatively short admixture events, the large number of parameters led to an underestimate of admixture duration (Zhou, Yuan, et al. 2017).

## Extended Pulse Model

One drawback of these approaches is that they introduce a large number of parameters. Even a discrete mixture of two pulses requires at least three parameters (two pulse times and the relative magnitude of the two events) (Pickrell et al. 2014), and the more complex models require regularization schemes for fitting (Ralph and Coop 2013; Zhou, Yuan, et al. 2017).

Here, we propose an *extended admixture pulse* model (fig. 1*A*) to estimate the duration of an admixture event. It only adds one additional parameter, reflecting the duration of gene flow, while retaining much of the mathematical simplicity present in the simple pulse model. The extended pulse model assumes that the migration rate over time is Gamma distributed, so that the length distribution of admixture segments has a closed form (fig. 1*C* and *D*) with two parameters, the mean admixture time and duration.

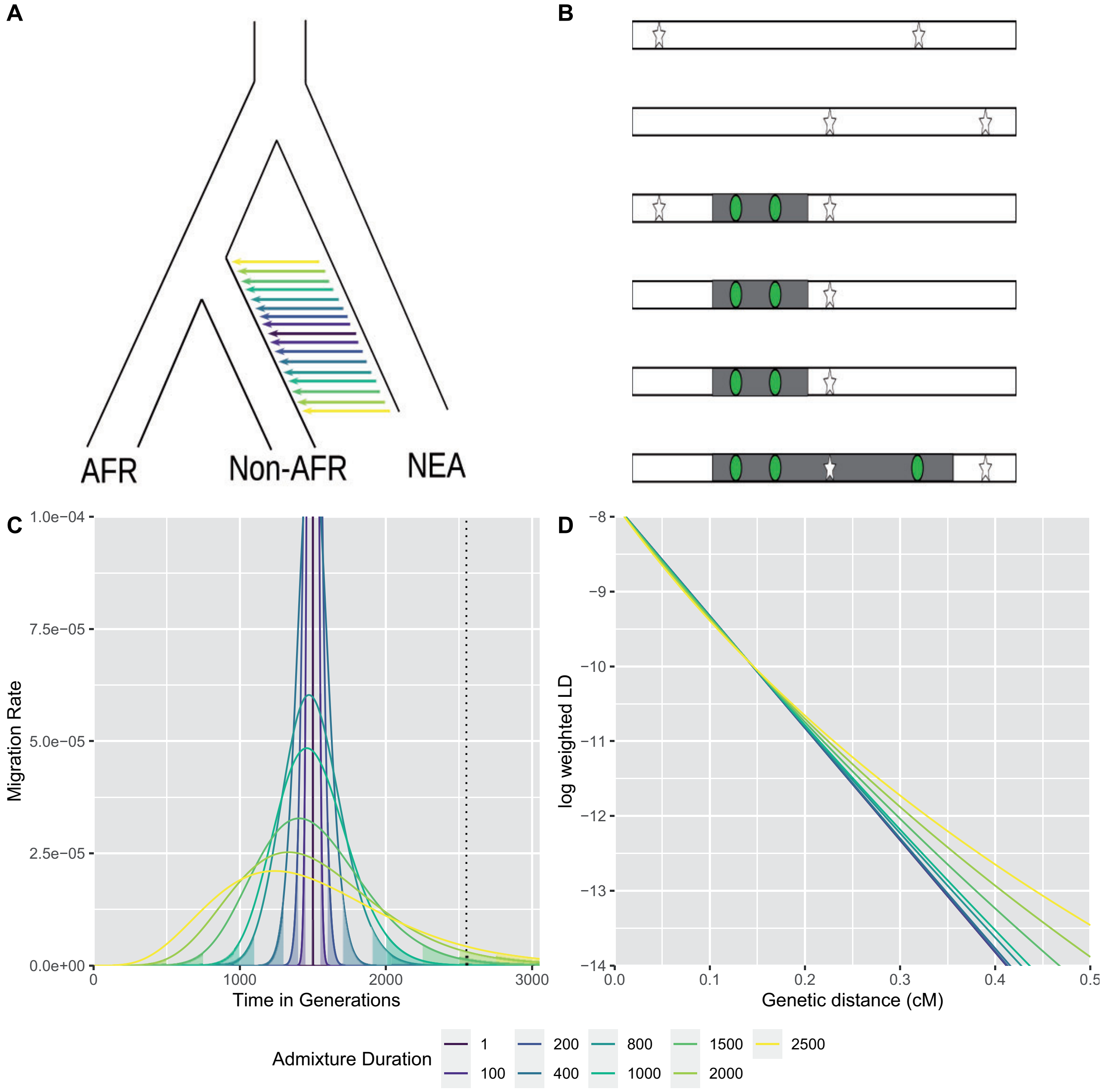

FIG. 1. (*A*) Neandertal introgression into non-Africans with a multitude of potential admixture durations. (*B*) The time and duration of admixture results in different length distributions of introgressed chromosomal segments (gray) containing Neandertal variants (green circles) in high LD to each other compared with the background (human variants white stars). The ALD approach estimates linkage between the introgressed variants (green circles), whereas the haplotype approach tries to estimate the segment directly (gray area). (*C*) Migration rate per generation modeled using the extended pulse model for different admixture durations (colored lines). The filled area under the curve indicates the boundaries of the discrete realization of the duration of gene flow $t_d$. The dotted line indicates the oldest possible time of gene flow (as defined in the simulations). (*D*) The expected LD decay under the extended pulse model.

Conceptually, identifying an extended pulse requires us to establish that the length distribution of admixture segments deviates from an exponential distribution. However, other sources of bias, such as the demography of the admixed population, the accuracy of the recombination map or details in the inference method parameters may also introduce similar biases. Thus, we have to carefully evaluate other potential sources of bias on whether they might lead to confounding signals. (Sankararaman et al. 2012; Fu et al. 2014; Moorjani et al. 2016).

Here, we first define the extended admixture pulse model and derive the resulting segment length and ALD distributions, and introduce inference schemes for either data. We then evaluate under which scenarios these two models can be distinguished. We show that power to distinguish these scenarios is higher for more recent events and longer pulses, but that accurate inference requires high-quality data. Based on these results, we use data from European genomes (1000 Genomes Project Consortium 2015) and find that for the case of Neandertal admixture, a simple pulses cannot be distinguished from continuous admixture over an extended period of time, and the data are consistent with a multitude of durations, up to several tens of thousands of years.

## New Approaches

In this section, we present the mathematical description of the admixture models we use in this paper, and introduce inference algorithms for estimating the admixture time and duration from both segment data and ALD.

### Admixture Models and Inference

We think of admixture as a series of "foreign" chromosomes introduced in a population (for a mechanistic model, see, e.g., Pool and Nielsen [2009]). Throughout, we assume that alleles evolve neutrally, and that recombination is independent of local ancestry. The simple pulse model assumes that all admixture happens in the same generation, (i.e., all chromosomes are introduced to the population at the same time). To extend this model, we allow chromosomes to enter at potentially many different time points, such that the migration rate at time $t$ in the past is given by the function $m(t)$ (Pool and Nielsen, 2009; Ni et al. 2016). For simplicity, we assume that the total amount of introgressed material $\alpha = \int_0^\infty m(t)dt$ is small, so that segments do not interact, but we will discuss violations of this assumptions later. For archaic introgression, $\alpha \approx 0.03$, so this assumption is justified.

Over time, recombination splits up the introgressed genome into smaller pieces, whereas by the neutrality assumption the expected amount of total ancestry remains approximately the same. Thus, if we measure the size of chromosomes in recombination units, a chromosome of size $G$ introduced at time $t$ gives rise to an expected number of $tG$ segments.

### Admixture Segment Lengths

We enumerate the admixture segments in a sample $i = 1 \ldots K$. We denote the length of the $i$-th segment as $L_i$ (measured in Morgan) and the time in the past when segment $i$ entered the population as $T_i$ (measured in generations). We assume that the $L_i$ and $T_i$ are both realizations from more general distributions $L$ and $T$ that reflect the overall segment length and segment age distributions, respectively.

To relate $m(t)$ to $T$, we need to take into account that older fragments had more time to split up (see, e.g., Pool and Nielsen 2009). Hence

$$P(T_i = t) = \frac{tGm(t)}{\int_0^\infty tGm(t)dt}. \qquad (1)$$

The denominator of the right-hand side term in equation (1) is the expected number of admixture segments, $\mathbb{E}[K] = \int_0^\infty tGm(t)dt$.

Given $T_i$, the segment length $L_i$ is exponentially distributed with rate parameter $t$:

$$P(L_i = l | T_i = t) = te^{-tl}. \qquad (2)$$

Integrating over $T$ yields the unconditional distribution of admixture segment lengths:

$$P(L_i = l) = \int_0^\infty P(T_i = t)P(L_i = l|T_i = t)\ dt,$$
$$= \frac{G}{\mathbb{E}[K]} \int_0^\infty t^2 m(t) e^{-tl} dt$$

and we can think of $L$ as an exponential mixture distribution with mixture density proportional to $tm(t)$ (Ralph and Coop 2013; Ni et al. 2016; Zhou, Qiu, et al. 2017).

### Ancestry Linkage Disequilibrium

Alternatively, the impact of gene flow is often characterized using ALD, particularly when accurate identification of archaic segments is difficult. We follow Loh et al. (2013) and note that the ALD from gene flow in a single event at time $t$ generations in the past is

$$D_t = m(1-m)\Delta_x\Delta_y \approx mA, \qquad (4)$$

where $m$ is the fraction of immigrants and $\Delta_x, \Delta_y$ are the differences in allele frequencies between markers in the admixing populations. We assume that terms of the order of $m^2$ can be ignored and that migration is low enough that changes in the allele frequencies in the admixing populations can also be neglected (i.e., $A = \Delta_x\Delta_y$ remains a constant).

At a later generation $s$, the expected LD between two markers a distance $l$ apart is

$$D_s \approx D_t \exp(-l(s-t)), \qquad (5)$$

due to the decay of LD (Sankararaman et al. 2012). If the migration rate $m_f$ is a function of time, we can add up the LD introduced at each time $t$ in the past and approximate $D$ as

$$D_s = A \int_{-\infty}^{s} m_f\Big(t\Big)\exp(-l(s-t))dt. \qquad (6)$$

As we show in the Appendix (Formal Motivation for ALD), equation (6) satisfies the differential equation

$$\frac{dD_s}{ds} = -lD_s + Am_f(s), \qquad (7)$$

where the $-lD_s$-term models the exponential decay of LD due to recombination, and the $Am_f(s)$-term reflects the increase of LD due to admixture (eq. 4).

To connect this equation more directly to the backward-in-time formulation used in the derivation of the admixture segment distribution, we set $s = 0$ and invert the flow of time, such that $m(t) = m_f(-t)$. We obtain

$$D(l) = A \int_0^\infty m(t)\exp(-lt)dt. \qquad (8)$$

Thus, $D$ can be interpreted as the tail function of an exponential mixture with mixture density $m$. Alternatively, the integral in equation (8) is also the (scaled) moment-generating function of $m$ with argument$-l$.

The distribution of admixture segment lengths (eq. 3) and the ALD function (eq. 8) are closely related—in the Appendix (Connection between Admixture Segment Length Distribution and ALD Function) we show that

$$D(l) = \frac{\mathbb{E}(K)}{G} \int_l^\infty P(x)(x - l)\mathrm{d}x \qquad (9)$$

$$P(l) \propto D''(l). \qquad (10)$$

It follows that both functions uniquely determine each other. Consequently, they contain identical information to estimate admixture dates.

Both for the segment and ALD models we use simplifying assumptions that ignore the effects of genetic drift, the recombination between introgressed segments and the replacement of older introgressed material. In the Appendix, we discuss these approximations and show that particularly the replacement of admixed material can be accommodated by replacing $m$ with

$$m_e(t) = m(t)\exp\left[-\int_0^t m(s)\mathrm{d}s\right], \qquad (11)$$

which can be interpreted as the probability of the event that migration happened at time $t$, and no more migration happened later on.

## The Simple Pulse Model

Under the simple pulse model, all fragments enter the population at the same time $t_m$, and $T$ is a constant distribution. We can formalize this model by using a Dirac delta function which integrates to one if the integration interval includes $t_m$ and zero otherwise:

$$m(t) = \alpha\delta_{t_m}(T_i), \qquad (12a)$$

$$P(T_i) = \delta_{t_m}(T_i), \qquad (12b)$$

We obtain the exponential distribution of admixture fragments under this model (Moorjani et al. 2011):

$$P(L_i = l) = t_m \mathrm{e}^{-t_m l} \qquad (13a)$$

$$D(l) \propto \mathrm{e}^{-t_m l}, \qquad (13b)$$

where here and in the remainder of this section we omit the constant term from $D$, which is not relevant for fitting the LD decay. The expected segment length under a simple pulse model is given by

$$\mathbb{E}[L] = \frac{1}{t_m} \qquad (14a)$$

and the variance by

$$\mathrm{Var}[L] = \frac{1}{t_m^2}. \qquad (14b)$$

## The Extended Pulse Model

For the new extended pulse model, we assume that the migration rate $m(t)$ follows a rescaled Gamma distribution so that the total contribution of migrant alleles is $\alpha$. It is convenient to parameterize the migration rate as $\Gamma\left(k, \frac{t_m}{k}\right)$. for $t \geq 0$ and $k \geq 1$.

Using this parameterization, the denominator of equation (1) is $t_m \alpha G$ and

$$P(T_i = t) = \frac{t}{t_m} m(t) \qquad (15a)$$

$$= \frac{1}{\Gamma(k)\left(\frac{t_m}{k}\right)^k} t^{k-1}\mathrm{e}^{-t\frac{k}{t_m}} \qquad (15b)$$

for $t \geq 0$ and $k \geq 2$, which is is the density of a $\Gamma\left(k+1, \frac{t_m}{k}\right)$-distribution with moments

$$\begin{aligned} \mathbb{E}[T] &= \frac{k+1}{k} t_m \\ \mathrm{Var}[T] &= \frac{k+1}{k^2} t_m^2 = \frac{k+1}{k} \left(\frac{t_d}{4}\right)^2. \end{aligned} \qquad (16)$$

Here, we define the admixture duration $t_d = 4 t_m k^{-\frac{1}{2}}$, as a convenient measure for the duration of gene flow. If $k$ is low, then $t_d$ will be large and gene flow extends over many generations. In contrast, if $k$ is large, then $t_d \approx 0$ and we recover the simple pulse model (fig. 1C and D).

The distribution of segment length is calculated by plugging equation (15b) into equation (3) and integrating:

$$P(L = l | k, t_m) = \int_0^\infty \frac{1}{\Gamma(k)\left(\frac{t_m}{l}\right)^k} \ t^{k-1}\mathrm{e}^{-t\frac{k}{t_m}} t \mathrm{e}^{-tl}\mathrm{d}t$$

$$= t_m^{-k} \left(\frac{k+1}{l + \frac{k}{t_m}}\right)^{k+2}.$$

The distribution in equation (17) is known as a *Lomax* or *Pareto-II* distribution, which is a heavier-tailed relative of the Exponential distribution. Under the extended pulse model, the expected segment length will be the same as under the simple pulse model (eq. 14a):

$$\mathbb{E}[L] = \frac{k}{t_m} \frac{1}{(k+1) - 1} = \frac{1}{t_m} \qquad (18)$$

but the variance is larger:

$$\mathrm{Var}[L] = \frac{(k+1)}{(k-1)} \frac{1}{t_m^2}. \qquad (19)$$

We obtain the ALD-function from equation using the moment-generating function of $m(t)$:

$$D_t(l) \propto \left(1 + \frac{t_m l}{k}\right)^{-k}. \qquad (20)$$

### The Constant Migration Model

The simple pulse model can be thought of as the extreme case of the extended pulse model when $k \to \infty$, that is, the pulse gets infinitely short. In the other extreme, the extended pulse model approaches a model of constant migration. In this case, the last migration event at a particular location is exponentially distributed with rate $m$ (eq. 11), which is a model considered by Pool and Nielsen (2009). Setting $t_m = \frac{2}{m}, k = 2$, we obtain

$$m(t) = m\exp(-mt) \sim \Gamma(1, m) \tag{21a}$$

$$P(T_i = t) \sim \Gamma(2, m) \tag{21b}$$

$$P(L_i = l) = \frac{2m^2}{(m + l)^3} \tag{21c}$$

$$D(l) \propto \frac{m}{m + l}. \tag{21d}$$

Equation (21c) differs slightly from equation (6) in Pool and Nielsen (2009) because we approximate the expected number of segments with $n = Kt$, versus theirs $n = 1 + Kt$ (however, they converge to each other for large $Kt$).

### Estimation of Admixture Times

For inference, either the admixture segment lengths or ALD can be used. Assuming the admixture segment lengths are known, equation (17) is the likelihood function and can be used for inference. For inference using ALD, we follow Moorjani et al. (2011) and use the decay of ALD with genetic distance as a statistic. Following Moorjani et al. (2016), we add an intercept $A$ and a constant $c$ modeling background LD:

$$\text{ALD} \sim \; Ae^{-t_m \, l} + c \tag{22}$$

$$\text{ALD} \sim \; A\left(1 + \frac{t_m}{k} \, l\right)^{-k} + c. \tag{23}$$

## Results

Here, we investigate under which scenarios we can distinguish the simple and extended pulse models, and when we can infer parameters under either model. We start with an idealized scenario of simulations under the model, and then continue with more realistic coalescent simulations using msprime (Kelleher et al. 2016).

### Power Analysis under the Model

In the easiest case, we assume that segments are known and we simulate directly under the model (eq. 17) and evaluate under which conditions we can tell the two models apart using likelihood-ratio tests on the simulated segments. For this purpose, we compare two scenarios, one where gene flow happened 1,500 generations ago, which reflects Neandertal gene flow inferred from present-day individuals. In the second scenario, which reflects inference from ancient modern human data, the samples are taken 50 generations after gene flow ended. We vary pulse durations from 1 to 2,500 generations, and sample between 100 and 100,000 unique segments. As the simple pulse model is an edge case of the extended pulse model with $k \to \infty$, standard likelihood theory does not apply, and we use empirical significance cutoffs (Kozubowski et al. 2008).

The resulting log-likelihood ratios are given in figure 2. In general, we find that power to distinguish the model increases with pulse duration and the amount of data, and that it is easier to distinguish the models when gene flow had been more recent. For example, with 10,000 unique segments we need an event lasting around 1,000 generations before we are able to confidently distinguish an extended from a simple pulse (fig. 2) using present-day data. In contrast, by sampling closer to the admixture event we are able to distinguish an extended pulse already with a duration of 40–60 generations.

### Population Genetic Model Comparisons

In the previous section, we have shown that we can distinguish long pulses from instantenous gene flow under idealized conditions. As a more realistic scenario, we perform population genetic simulations using msprime (Kelleher et al. 2016). Throughout, we simulate 3% Neandertal admixture into non-Africans using a demographic model of archaic introgression (supplementary fig. 1*B*, Supplementary Material online) with a mean admixture time of 1,500 generations ago and varying durations. We simulate 20 chromosomes of length 150 MB, using either a constant recombination map or the HapMap recombination map (International HapMap Consortium 2007). This results in $\sim 10,000$ introgressed segments. We then perform inference using either the simulated segments, segments inferred from the data (Skov et al. 2018), or ALD calculated using ALDER (Loh et al. 2013). We further vary recombination rate settings as 1) inference and simulation under constant recombination rate (Constant/Constant); 2) simulation using the HapMap genetic map (International HapMap Consortium 2007), and inference using no correction (HapMap/Constant); 3) simulation using HapMap, correction using a different map (HapMap/AAMap) (Hinch et al. 2011); 4) and inference using the same map used for the simulations (HapMap/HapMap).

Using these simulations, we perform model comparisons (fig. 3*A*). For segments, we again use the likelihood-ratio and find that the results for the simulated segments closely match the simulations under the model (fig. 2), showing that our model is a good approximation in the parameter range of interest. In contrast, we find that for inferred segments, results greatly depend on the recombination rate used: For a constant recombination rate, results are similar, but for the HapMap-recombination map, we do not have any power to distinguish these scenarios. As we fit ALD using nonlinear least squares, no formal model-comparison framework exists. Qualitatively, we plot the normalized residual sum-of squares (RSS) and find that they increase with $t_d$ for both recombination scenarios, suggesting that the difference between the two models increase.

Next, we evaluate parameter inference. In figure 3*B*, we present estimates of the mean admixture times, admixture

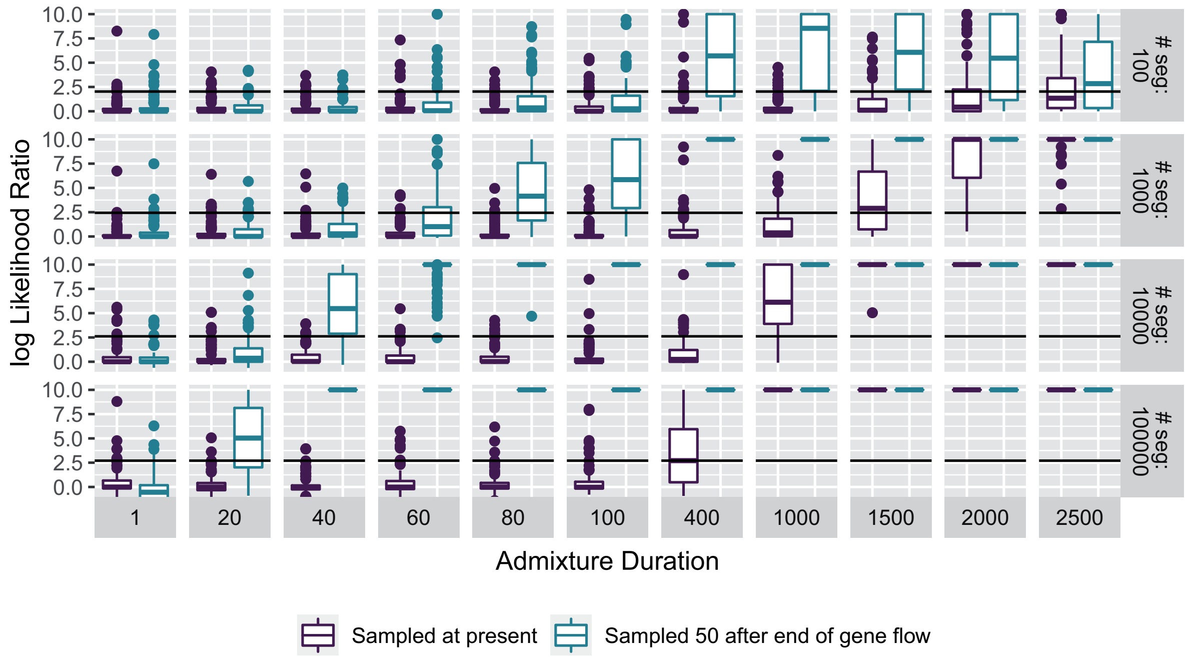

FIG. 2. Model comparisons on perfect data. Segments are either sampled at the present (purple) or 50 generations after the end of gene flow (turquoise). Log-likelihood ratios bigger than 10 are rounded to 10.

duration and the fitted segment and ALD distributions, respectively. We find that the mean admixture times are reasonably accurately estimated in most scenarios, the exception being the inferred segments when using the variable (HapMap) recombination map. The admixture duration estimates are often less accurate, and in most cases has very large variation between simulations.

We detect a slight, but consistent underestimate of the mean admixture times, which increases with $t_d$. For the segments, this underestimate is likely due to the slight downward bias caused by recombination and coalescence between admixed segments (Liang and Nielsen 2014, see also Appendix Genetic Drift and Recombination). For ALD, this bias is much less severe, particularly for inference under the extended pulse model. For scenarios where the recombination map is misspecified, $t_m$ is estimated to be only around half of its true value (supplementary fig. 2, Supplementary Material online). However, we find that in some cases, the extended pulse model provides a better estimate of $t_m$ by estimating the pulses to be extremely long.

In figure 3C, we show examples of the estimated segment length and ALD distributions compared with the simulated data. For these log-plots, the slope of the curve corresponds to the estimate of $t_m$, and the deviation from linearity reflects the duration of gene flow. In all cases, we find that the expected decay is very close to linear, matching our finding that power to differentiate these old events is limited. We find that particularly when using a constant recombination map, all three summaries give a very close fit, and the segment length and ALD-decay distribution closely follow their expectations, which is consistent with the generally good parameter estimates under these conditions. In the case of a variable recombination map, we find that particularly inferred fragments perform poorly, which is reflected by a substantial downward bias of $t_m$ and $t_d$.

## Comparing Effect Sizes for Technical Covariates

As we find that ALD performs as good or better than inferred segments (fig. 3), we focus on ALD for the remainder of this article. Our next goal is to more carefully evaluate the relative importance of common assumptions made in the inference of admixture times, under both the simple and extended pulse models in the ALD framework on the bias and accuracy of estimates of $t_m$ under either model.

In particular, we use a Bayesian generalized linear model (GLM) framework to contrast the effect of extended gene flow on admixture time inference with 1) the effects of a simple/complex demographic history (supplementary fig. 1, Supplementary Material online); 2) recombination map variation; 3) the ALD ascertainment scheme; 4) $d_0$, the minimum genetic distance between variants; and 5) the number of makers used to estimate the ALD curve (see Materials and Methods for details). For each modeling parameter and gene flow model, we use a simple model as the base case, and we study the impact of a more “realistic” alternative model.

In figure 4, we present the estimated effect sizes for these six variables and four key interaction terms. To model bias, we fit a model to the standardized difference between the true and estimated mean admixture time, and to model accuracy, we us the absolute deviation (Materials and Methods, supplementary table 1, supplementary fig. 3, and supplementary table 2, Supplementary Material online). These effect sizes are estimated using simulations under all possible parameter combinations on a scenario with admixture happening 1,500 generations ago (supplementary figs. 4 and 5, Supplementary Material online).

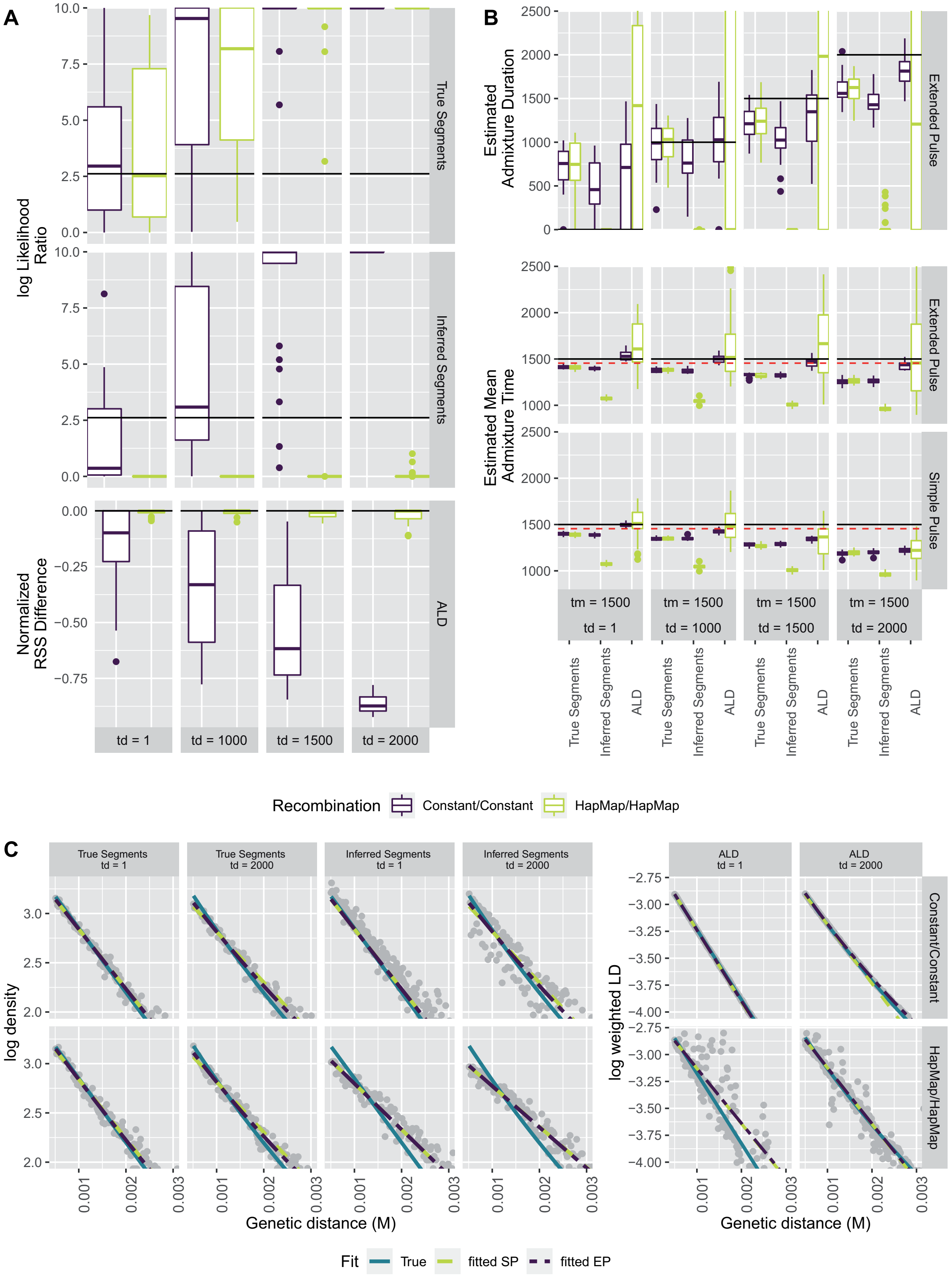

Fig. 3. Model choice, model fit, and parameter estimates. (*A*) Log-likelihood ratios and RSS difference between the simple and extended pulse models for segment data and ALD, respectively. Simulation and inference were done using constant (purple) and an empirical (teal) recombination map. (*B*) Estimates of $t_m$ and $t_d$. Solid black line indicates simulation values, the red dotted line adds a migration corrected $(t_m(1-\alpha))$. (*C*) Model fit for a single simulation in each scenario. Estimated segment density or weighted LD (gray) is compared with the expected (turquoise) and fitted single pulse (yellow) and extended pulse (purple).

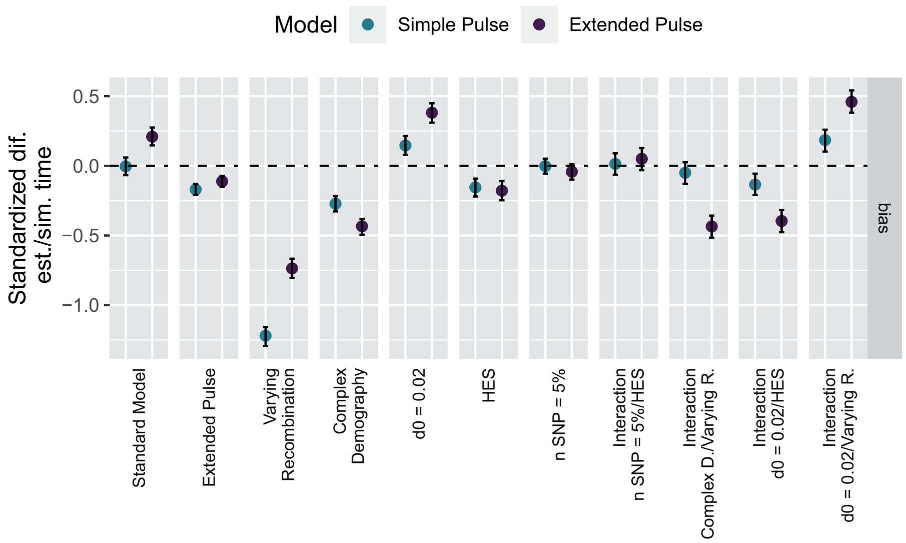

**Fig. 4.** GLM effect sizes for the bias between simulated and estimated mean admixture time and 95% CI for the parameters between the simple and extended pulse models: gene flow (simple/extended), recombination rate (constant/varying), demography (simple/complex), minimal genetic distance (0.02/0.05 cM), SNPs used for ALD calculation (100%/5%), and ascertainment scheme (LES/HES). Estimates are calculated across all possible combinations of parameters. Dotted horizontal line indicates unbiased admixture estimates.

As a baseline, for comparison, we define a standard model using a simple demography (supplementary fig. 1*A*, Supplementary Material online) and a constant recombination rate. This baseline model results in unbiased estimates of $t_m$ under the single pulse model with low deviation of 0.08 (0.02–0.14 $\mathrm{CI}_{95\%}$), and a slight upward bias 0.21 (0.15–0.28) for the extended pulse model.

The effect of simulating an extended-pulse gene flow only results in a slight bias of −0.17; (−0.21 to −0.13) for the simple pulse and no bias for inference under the extended pulse model (−0.11; −0.15 to 0.07). In contrast, uncertainty in the genetic map causes by far the largest downward bias (simple pulse: −1.22, −1.29 to −1.16; extended pulse: −0.74, −0.80 to −0.67) with high deviation in the estimates (supplementary fig. 3, Supplementary Material online). The more complex demography results in an underestimate of $t_m$, presumably because of increased genetic drift, for both the simple pulse (−0.27, −0.33 to 0.22) and extended pulse models (−0.43, −0.49 to −0.38). The remaining parameters largely only have very minor effects, the biggest of which is changing the minimum cutoff from 0.05 to 0.02 cM.

## Application to Neandertal Data

Our next aim is to apply our model on the case of Neandertal gene flow into Eurasians. We estimate the Neandertal admixture pulse from the 1000 Genomes data (1000 Genomes Project Consortium 2015) and three high-coverage Neandertal genomes (Prüfer et al. 2013, 2017; Mafessoni et al. 2020) by fiting pulses with durations ranging from 1 generation up to 2,500 generations to the ALD-decay curve (fig. 5, supplementary table 3, Supplementary Material online). Plotting these best-fit ALD curves shows the extremely slight difference predicted under these drastically different gene flow scenarios (fig. 5*A*). The difference between scenarios becomes more apparent if we log-transform the *y*-axis (fig. 5*B*), where we see that ongoing gene flow results in a heavier tail in the ALD distributions. However, these LD values are very close to zero, and are thus only very noisily estimated.

For short gene flows (less than 1,000 generations), our estimates for $t_m$ are very similar and identical to the simple pulse, at around 1,682 (1,526–1,839 $\mathrm{CI}_{95\%}$) generations. Extremely high values of $t_d$ result in slightly higher values of $t_m$ with overlapping compatibility intervals; but all predict that Neandertals would have survived until 30 ka, for which the archeological evidence is extremely sparse (Hublin 2017). From the RSS, the models perform equally well, with longer extended pulses of gene flow achieving marginally better fits (supplementary table 4, Supplementary Material online). Therefore, we find that all scenarios are compatible with the observed data, and that there is little power to differentiate these cases from genetics alone.

## Sampling Closer to the Admixture Event

Since Neandertal gene flow happened long in the past, much of the signal has been lost, and we have shown that in this scenario, power to distinguish different scenarios is low.

However, we have also shown in figure 2 that inference is easier for more recent gene flow, a case that is relevant for many study systems. We investigate this in a series of simulations where the time between sampling and gene flow is smaller (fig. 6). We use the simple demographic scenario with a constant-sized populations (supplementary fig. 1, Supplementary Material online), and use ALD for inference using the optimized settings for the Neandertal case (ascertainment scheme = LES and d0 = 0.05 cM).

In figure 6*A* and *B*, we show the accuracy of estimating $t_d$ and $t_m$ for increasingly longer pulses, sampled 50 generations after gene flow ended. The corresponding comparison of

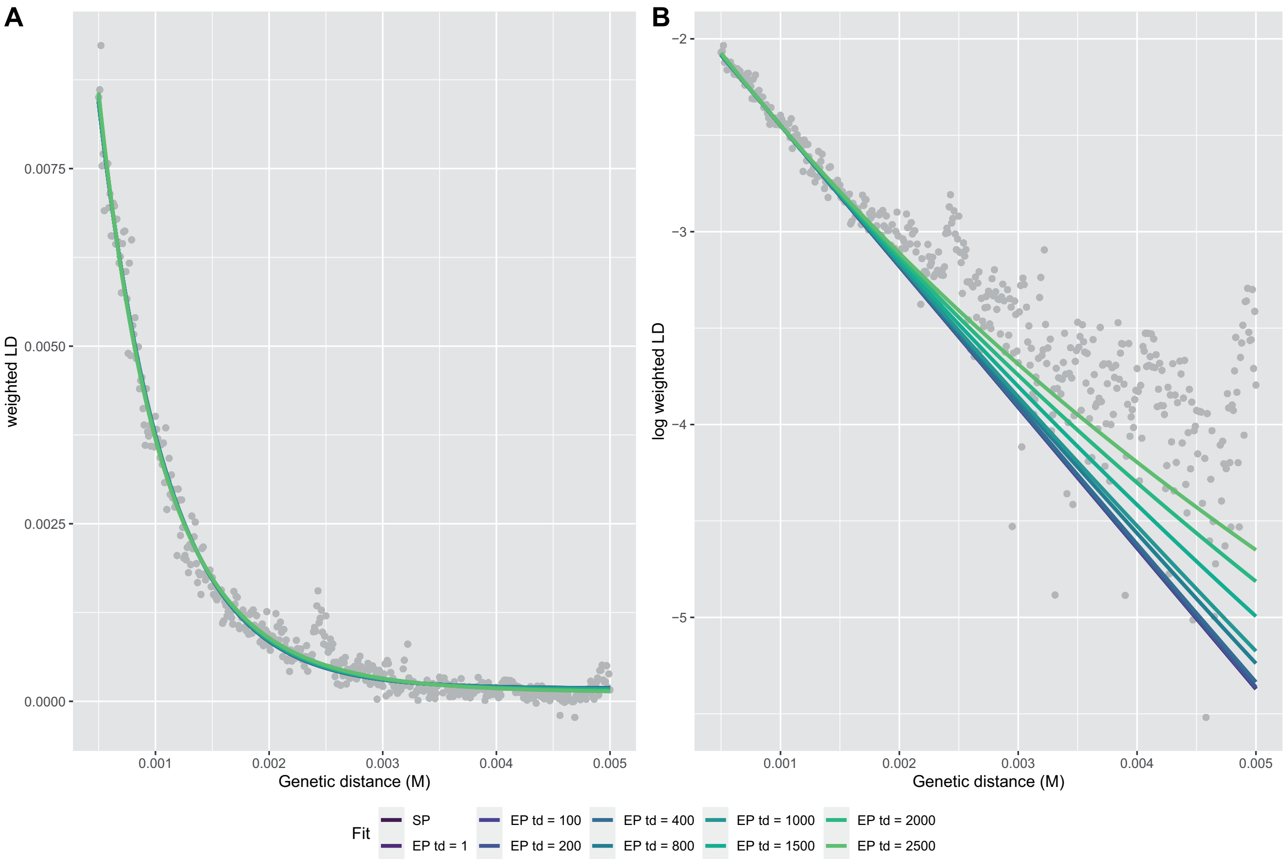

FIG. 5. Neandertal gene flow in modern humans. We fit models with fixed $t_d$ from 1 to 2,500 generations of gene flow to the ALD-curve calculated from CEU individuals on (*A*) natural and (*B*) log-scale.

model fit is depicted in supplementary figure 6, Supplementary Material online. For these cases, we find that inference of $t_m$ under the simple pulse model works well for the shortest pulses but becomes increasingly downward biased as $t_d$ increases. Estimates of $t_m$ are less biased for inference under the extended pulse model, where we get accurate estimates particularly if recombination is constant. In the scenarios with a variable recombination rate, we find that for short, recent pulses, all corrections give good results, but for longer pulses, particularly assuming a constant recombination rate leads to a stronger bias. We also find that we are able to accurately infer the admixture duration, particularly if the recombination rate is constant.

In 6C and D, we keep the pulse duration constant at $t_d = 800$ but move it successively further into the past. For the first two cases of $t_m = 450$ and $t_m = 500$ where the pulse is recent, we again obtain good parameter estimates, but performance deteriorates for $t_m \geq 600$, which also results in low power to distinguish the simple and extended pulse model (supplementary fig. 6, lower panel, Supplementary Material online).

## Discussion

In this article, we introduce a new population genetic model for dating extended pulses of gene flow. Our model has just two parameters, that can be interpreted as the mean time and duration of gene flow; and has simple closed form solutions for the segment length and ALD distributions. We show that both the instantaneous pulse and constant migration models are special cases of our model, where the duration is extremely short or long, respectively. We also demonstrate that the segment length distribution and ALD-decay can be directly transformed into each other; in particular, the segment-length distribution is proportional to the second derivative of the ALD-decay curve. This makes our theory and models generally applicable beyond gene flow between Neandertals and humans. In fact, we find that we have little resolution for the parameter settings relevant for archaic gene flow, as the data resulting from simple and extended pulses long in the past are extremely similar. In contrast, we have much more power to estimate the duration of gene flow from events in the recent past, a scenario relevant for many hybridizing species. One limitation of our approach is that we assume that the overall amount of introduced material is low, and that we ignore the effects of genetic drift and selection.

Previous approaches to date Neandertal–human gene flow have focused almost entirely on the mean time of gene flow using a simple pulse model, for which reasonably tight credible intervals can be estimated (Sankararaman et al. 2012; Moorjani et al. 2016). Under this model, the credible intervals of this time are bounds of when gene flow between Neandertals and early modern humans could have happened.

Our estimate of the $t_m$ for Neandertal gene flow of 1,682 generations corresponds to a mean time estimate of 49 ky

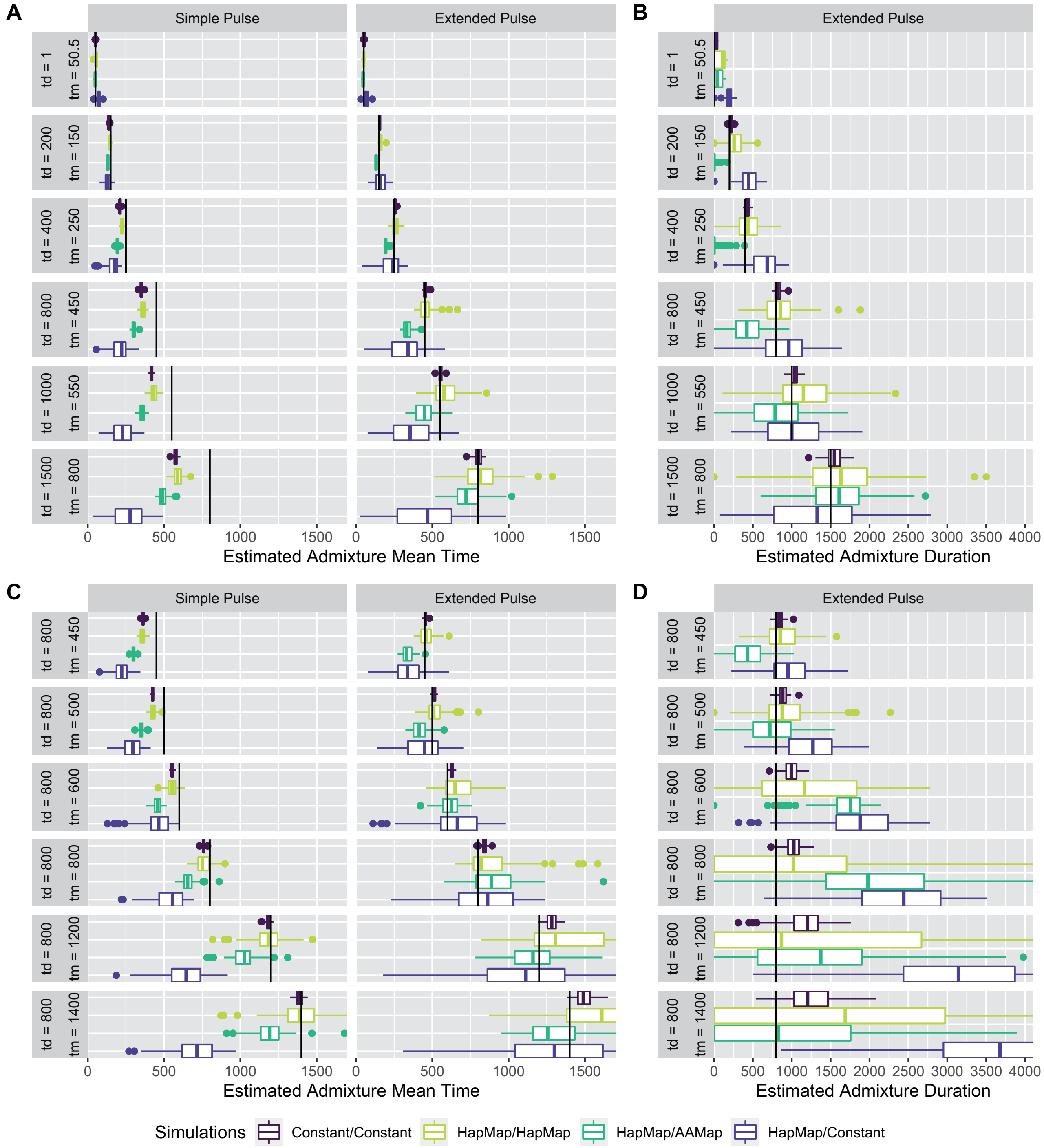

Fig. 6. Parameter estimation using ALD from recent admixture pulses. We show estimates of $t_m$ (panels A, C) and $t_d$ (Panels B and D) for a series of simulations with increasing $t_d$ and $t_m$ such that we sample 50 generations after gene flow (top) and a series of simulations with an increasingly older extended pulse (bottom). All times are given in generations.

(assuming a generation time of 29 years; Moorjani et al. 2016), with bounds of 44–54 ky. This is in almost perfect agreement with the previous result of Moorjani et al. (2016) (41–54 ky), which is based on largely the same method. However, here we show that models of extended gene flow with $t_d$ up to a thousand generations provide very similar fits to the data; and that marginally better fits are achieved with very long gene flows. However, these models all would have Neandertals survive until around 30 ka, whereas archeological evidence for Neandertals surviving beyond 40 ky is increasingly sparse (Hublin 2017), so that these models of extremely long gene flow might be rejected on these grounds.

Our finding that the observed data are compatible with models involving hundreds of generations of gene flow means that while likely substantial amounts of gene flow happened around these mean times, gene flow might have also happened tens of thousands of years before or after. This is of great practical importance, as it makes linking genetic admixture date estimates with biogeographical events much more difficult (Sankararaman et al. 2012; Lazaridis et al. 2016; Douka et al. 2019; Jacobs et al. 2019; Vyas and Mulligan 2019).

The discovery of early modern human genomes dated to 40,000–45,000 ya with very recent Neandertal ancestors less than ten generations ago (Fu et al. 2014; Hajdinjak et al. 2021) illustrates that gene flow likely happened over at least several thousand years. In general, inference based on ancient genomes (Fu et al. 2014, 2015; Moorjani et al. 2016; Hajdinjak et al. 2021) promises to resolve some of these dating issues, as inference is substantially easier when admixture is more recent, as the time difference between gene flow and sampling time is much lower (figs. 2 and 6). However, using these genomes for dating leads to further hurdles, particularly pertaining to the spatial distribution of admixture events; whereas we can assume that the spatial structure present in initial upper paleolithic modern humans is largely homogenized in present-day people, the introgression signals observed in Bacho-Kiro and Oase could be partially private to these populations, and thus these populations may have a different admixture time distribution than present-day people.

The uncertainty over the duration of Neandertal gene flow also has some implications for selection on introgressed Neandertal haplotypes. Neandertal alleles have been suggested to be deleterious in modern human populations due to an increased mutation load (Harris and Nielsen 2016; Juric et al. 2016). Some details of these models may be affected if migration occurred over a longer time. For example, Harris and Nielsen (2016) suggested that an initial pulse of gene flow of up to 10% Neandertal ancestry might be necessary to explain current amounts of Neandertal ancestry, with very high variance in the first few generations after gene flow. More gradual gene flow could mean that such high admixture proportions were never reached, but rather a continuous migration–selection balance process persisted for the contact period, where deleterious Neandertal alleles continually entered the modern human populations, but were selected against immediately. However, in terms of the overall frequencies, there is likely little difference. For example, Juric et al. (2016) showed using a two-locus model that the frequencies of Neandertal haplotypes alone cannot be used to distinguish different admixture histories.

In addition, we find that modeling and method assumptions have an impact on admixture time estimates that are of a similar or larger magnitude than the effect of assuming a one-generation pulse. In particular, recombination rate variation poses a practical limitation to the accuracy of admixture date estimates for old gene flow, and has to be very carefully considered when making inferences about admixture times. A possible reason is that both an extended pulse as well as a nonhomogeneous recombination map lead to an admixture segment distribution that deviates from the expected exponential distribution. Throughout, we measure segment lengths and LD-decay distance in recombination units. Misspecification of the recombination rate will increase the variance in ALD or segment lengths, which might be confounded with a longer admixture pulse (Sankararaman et al. 2012). Therefore, population-specific fine-scale recombination maps are needed for accurate admixture time estimates, at least for admixture that happened more than a thousand generations ago. Estimates of more recent admixture appear to be more robust, perhaps because coarser-scale recombination maps are better estimated, differ less between populations (Hinch et al. 2011) and the error relative to fragment length is substantially lower.

To further refine admixture time estimates, time series data from more admixed early modern human and Neandertal genomes are needed. In particular, measures based on population differentiation (Wall et al. 2013; Browning et al. 2018; Villanea and Schraiber 2019) hold much promise to understand the different events that contributed to archaic ancestry in modern humans. Although Neandertal ancestry in present-day people has been largely homogenized due to the substantial gene flow between populations, samples from both the Neandertal and early modern human populations immediately involved with the gene flow could refine when and where this gene flow happened.

## Materials and Methods

### Power Analysis under the Model

To test the power to distinguish the simple from an extended pulse we simulated 100, 1,000, 10,000, and 100,000 unique times $T_i$ from a Gamma distribution, with shape parameter $k+1$ and scale $k/t_m$, setting $t_m$ to 1,500 generations. Segment lengths $L_i$ are obtained by sampling for each $T_i$ from an exponential distribution with rate parameter $T_i$ for present day samples and $T_i^{(\text{closer})} = T_i - t_m - t_d/2 - 50$ for sampling 50 generations after the end of gene flow. We obtain maximum-likelihood estimates for the simple (Eq. 13a) and extended pulse (Eq. 17) using the optim function implemented in R (R Core Team 2019).

### Coalescent Simulations

We further test our approach on coalescence simulations using msprime (Kelleher et al. 2016). We focus on scenarios mimicking Neandertal admixture and choose sample sizes to reflect those available from the 1000 Genomes data (1000 Genomes Project Consortium 2015). For ALD simulations, we simulate 176 diploid African individuals and 170 diploid non-Africans, corresponding to the number of Yoruba (YRI) and Central Europeans from Utah (CEU). For inference based on segments, we simulated 50 diploid non-Africans. Since three high-coverage Neandertal genomes are available (Prüfer et al. 2013, 2017; Mafessoni et al. 2020), we simulate three diploid Neandertal genomes.

The demographic parameters are based on previous studies dating Neandertal admixture (Sankararaman et al. 2012; Fu et al. 2014; Moorjani et al. 2016; Skov et al. 2018). In the "simple" demographic model (supplementary fig. 1A, Supplementary Material online), the effective population size is assumed constant at $N_e = 10,000$ for all populations, the split time between modern humans and Neandertals is

10,000 generations, and the split between Africans and non-Africans is 2,550 generations. The migration rate from Neandertals into non-Africans was set to zero before the split from Africans, to ensure that there is no Neandertal ancestry in Africans. For a more complex scenario of human population history, we followed Skov et al. (2018) and used a similar demographic model, but only simulated the Europeans. We changed the $N_e$ for the ancestral humans, out-of-Africa bottleneck and ancestral Eurasians to 7,000, 250, and 5,000, respectively. The effective population size for Neandertals was set to 5,000 and the split time of non-Africans is kept the same as in the ALD simulations (2,550 generations ago) (supplementary fig. 1*B*, Supplementary Material online).

For each individual, we simulate 20 chromosomes with a length of 150 Mb each. The mutation rates are set to $2 \times 1 0^{-8}$ and $1.2 \times 10^{-8}$ per base per generation for the "simple" and "complex" models, respectively. The recombination rates are set to $1 \times 10^{-8}$ per base pair per generation for the simple demography and $1.2 \times 10^{-8}$ per base pair per generation for the complex demographic model, unless specified otherwise.

Since inferring archaic segments is slow, we use 25 replicates for scenarios where we compare segment-based and ALD-based inference and use 100 replicates when we only perform ALD-based inference.

## Simulating Admixture

We specify simulations under the extended pulse model using the mean admixture time $t_m$ and the duration $t_d$. We recover the simple pulse model by setting $t_d = 1$, up to errors due to discrete generations. To obtain the migration rates in each generation, we use a discretized version of the migration density (eq. 15b), which we then scale to the approximate amount of Neandertal ancestry in non-Africans ($\alpha = 0.03$).

## Recombination Maps

Uncertainties in the recombination map were previously shown to influence admixture time estimates (Sankararaman et al. 2012, 2016; Fu et al. 2014). To investigate the effect of more realistic recombination rate variation, we perform simulations using empirical recombination maps. For the GLM, we use the African-American map (Hinch et al. 2011) for simulations and for the remaining simulations we use the HapMap phase 3 map (International HapMap Consortium 2007). For simplicity, we use the same recombination map (150 Mb of chromosome 1, excluding the first and last 10 Mb) for all simulated chromosomes. When simulating under an empirical map, with the analysis assuming a constant rate (i.e., no correction), we use the mean recombination rate from the respective map to calculate the genetic distance from the physical distance for each SNP. The mean recombination rate is calculated from the 150-Mb map ($1.017 \frac{\text{cM}}{\text{Mb}}$ AAMap, $0.992 \frac{\text{cM}}{\text{Mb}}$ HapMap). For inference, each segment is either assigned a length based on its physical length ("constant"), the African-American map or HapMap recombination map, depending on the inference scenario.

## Estimating Admixture Time from Simulated Segment Data

For estimating admixture time and duration from introgressed segments, we either used the simulated segment lengths directly or alternatively added an inference step using the HMM from Skov et al. (2018). We only considered inferred segments with an average posterior probability of 0.9 or higher. Furthermore, we use an upper and lower cutoff for inferred segment length of 0.05 and 1.2 cM. We fit the simple (eq. 13a) and extended pulse (eq. 17) using the optim function implemented using R 4.0.3 (method="L-BFGS-B") with lower and upper constrains being 1 and 5,000 for $t_m$ and 2 and $10^{10}$ for $k$, respectively.

## Estimating Admixture Time from ALD Data

### Ascertainment Scheme

Since ALD for ancient admixture events can be quite similar to the genomic background, SNPs need to be ascertained to enrich for Neandertal informative sites in the test population. This removes noise and amplifies the ALD signal (Sankararaman et al. 2012). We evaluate the impact of the ascertainment scheme by contrasting two distinct schemes (Sankararaman et al. 2012; Fu et al. 2014). The lower-enrichment ascertainment scheme (LES) only considers sites that are fixed for the ancestral state in Africans and polymorphic or fixed derived in Neandertals. The higher-enrichment ascertainment scheme (HES) is more restrictive in that it further excludes all sites that are not polymorphic in non-Africans.

### ALD Calculation

The pairwise weighted LD between the ascertained SNPs a certain genetic distance $d$ apart is calculated using ALDER (Loh et al. 2013). A minimal genetic distance $d_0$ between SNPs is set to either 0.02 or 0.05 cM. This minimal distance cutoff removes extremely short-range LD, which might also be due to inheritance of segments from the ancestral population (incomplete lineage sorting ILS) and not gene flow.

### Parameter Estimates

We estimate parameters by fitting the ALD-curve to equations (22) and (23) using a nonlinear least square approach implemented in the nls function in R 4.0.3 (algorithm="port") with lower and upper constrains being 1 and 5,000 for $t_m$ and $1/10^{10}$ and 1/2 for $1/k$, respectively. To achieve better conversion, we prefit the functions using the estimates of the DEoptim optimization (Ardia et al. 2016) as starting parameters for the nls function. To improve estimates for $t_d$, we run the fitting using ten iterations to avoid local optima. We select the estimate with the lowest RSS.

## Modeling Parameter Effect Sizes

To estimate the effect size of the different parameters (eq. 24) we use a Bayesian GLM, where $E$ is the response, and $A, M, D, R, S$, and $G$ are binary predictors.

The model can be written as

$$\begin{gathered}E_i \sim \text{Normal}(\mu_i, \sigma) \\ \mu_i = \alpha + \beta_a A_i + \beta_m M_i + \beta_d D_i + \beta_r R_i + \beta_s S_i + \\ \beta_g G_i + \beta_{sa} S_i A_i + \beta_{ma} M_i A_i + \beta_{dr} D_i R_i + \beta_{mr} M_i R_i \\ \alpha \sim \text{Normal}(0, 2) \\ \beta_a, \beta_m, \beta_d, \beta_r, \beta_s, \beta_g, \beta_{sa}, \beta_{ma}, \beta_{dr}, \beta_{mr} \sim \text{Normal}(0, 2) \\ \sigma \sim \text{Exponential}(1)\end{gathered}, \qquad (24)$$

where the variables define

- ascertainment scheme: $A_i$ = LES/HES
- minimal genetic distance: $M_i$ = 0.02cM/0.05cM
- demography: $D_i$ = simple/complex
- recombination rate: $R_i$ = constant/variable
- n SNPs used: $S_i$ = 100%/5%
- gene flow model: $G_i$ = simple pulse (SP)/extended pulse (EP)

We fit two models, a primary model aimed at investigating the bias of our estimates, and a second model aimed at investigating the deviation. In the first case, the response variable $E_i$ is

$$E_i = \frac{t_{est} - t_{sim}}{\sigma_{t_{est}}},$$

and in the second case we use the absolute error

$$E_i = \frac{|t_{est} - t_{sim}|}{\sigma_{t_{est}}},$$

where $\sigma$ is the standard deviation of $t_{est}$. We also modeled the interaction between number of used SNPs and the ascertainment scheme ($\beta_{sa}$), minimal distance and ascertainment ($\beta_{ma}$), demography and recombination ($\beta_{dr}$), and minimal distance and recombination($\beta_{mr}$).

We perform simulations using all possible parameter combinations. For the effect of the amount of SNPs, that is, accuracy of the ALD estimates, we downsampled the data by randomly choosing 5% of the overall SNPs for ALD calculation. We define a standard model having a constant recombination rate, simple demography and gene flow, LES ascertainment, and $d_0 = 0.05$. The genetic distance is assigned from the physical position using the average recombination rate of the African-American genetic map (i.e., assuming the recombination rate is constant over the simulated chromosome given by this value) for simulations under a variable recombination rate. For each of the possible sets of parameters, we simulate 100 replicates each and fit ALD-decay curves. We excluded a small number of simulations for which either the simple pulse or extended pulse curve could not be estimated (87 out of 6,400).

We assume a Normal likelihood because it is the maximum entropy distribution in our case. We obtained the posterior probability using a Hamiltonian Monte Carlo MCMC algorithm, as implemented in STAN (Carpenter et al. 2017) using an R interface (Stan Development Team 2018; McElreath 2020). The Markov chains converged to the target distribution (Rhat = 1) and efficiently sampled from the posterior (supplementary tables 1 and 2, Supplementary Material online).

### Estimating Neandertal Admixture Time

We estimate the Neandertal admixture time distribution using ALD from the 1000 Genomes data (1000 Genomes Project Consortium 2015), together with the Altai, Vindija, and Chagyrskaya high-coverage Neandertals. We include the 107 unrelated individuals from the YRI as representatives of unadmixed Africans and all CEU as admixed Europeans. We only consider biallelic sites and determine the ancestral allele using the Chimpanzee reference genome (panTro4). We used the CEU-specific fine-scale recombination map (Spence and Song 2019) to convert the physical distance between sites into genetic distance.

## Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

## Acknowledgments

We thank Svante Pääbo, Janet Kelso, Fabrizio Mafessoni, Stéphane Peyrégne, Laurits Skov, Divyaratan Popli, Alba Bossom Mesa, Arev Sümer, and Shai Carmi for helpful comments and discussions. This work was supported by the Max Planck Society and the European Research Council (Grant No.: 694707) to Svante Pääbo.

## Author Contributions

Conceptualization (Design of study): B.M.P.; Software: L.N.M.I.; Methodology—lead: B.M.P.; Methodology—support: L.N.M.I., H.R.; Formal Analysis: L.N.M.I.; Visualization: L.N.M.I.; Data Curation: L.N.M.I.; Writing (original draft preparation)—lead: L.N.M.I.; Writing (original draft preparation)—support: B.M.P.; Writing (review and editing): input from all authors; Supervision: B.M.P.

## Data Availability

The simulation script, the analysis pipeline, and the *Extended Admixture Pulse* model are available on GitHub: https://github.com/LeonardoIasi/Extended_Admixture_Pulse. No new data were generated for this study.

## Appendix: Derivation and Approximation Details

### Formal Motivation for ALD

In the main text, we motivate the ALD decay using the intuitive argument of "adding up" ALD introduced at different generations in the past. Here we present a more formal derivation: In one generation, LD changes due to LD-decay through recombination, and through the introduction of new ALD through migration.

$$\begin{aligned} D(t+1) &= (1-r-m(t))D(t) + \Delta_X \Delta_Y m(t) \\ &\approx (1-r)D(t) + Am(t), \end{aligned}$$

where $A$ is a constant and $r$ the map distance between the pair of markers. The approximation in the second line is valid if we ignore the effect of new introgressed material replacing older introgressed material.

This leads to the following linear differential equation:

$$\frac{dD}{dt} = -rD + Am(t). \qquad \text{(A2)}$$

It is straightforward to verify that equation (6) is a solution to this equation:

$$\begin{aligned} D(t) &= A\int_{-\infty}^{t} m(s)\exp(-r(t-s))ds \\ \frac{dD}{dt} &= \frac{d}{dt}\left[A\int_{-\infty}^{t} m\Big(s\Big)\exp(-r(t-s))ds\right] \\ &= Am(t) + \int_{-\infty}^{t}\left[\frac{d}{dt}Am(s)\exp(-r(t-s))\right]ds \\ &= Am(t) - r\int_{-\infty}^{t} Am(s)\exp(-r(t-s))ds \\ &= Am(t) - rD(t), \end{aligned}$$

where the third line follows from Leibniz' integral rule.

### Replacement during Pulse

The "old"-LD also changes due to addition of new introgressed material, which is accommodated using

$$\frac{dD}{dt} = -(r+m(t))D + Am(t). \qquad \text{(A4)}$$

This equation has solution

$$D(t) = A\int_{-\infty}^{t} \exp[-r(t-s) - M(s,t)]m(s)ds, \qquad \text{(A5)}$$

where

$$M(s,t) = \int_{s}^{t} m(x)dx,$$

which can be interpreted as contributing how much LD has decayed from introgression time $s$ to the observation time $t$ due to replacement from new introgressed material.

This follows from

$$\begin{aligned} D(t) &= A\int_{-\infty}^{t} m(s)\exp\left[-r(t-s) - \int_{s}^{t} m(x)dx\right]ds \\ \frac{dD}{dt} &= \frac{d}{dt}\left[A\int_{-\infty}^{t} m(s)\exp\left(-r(t-s) - \int_{s}^{t} m(x)dx\right)ds\right] \\ &= Am(t) + \int_{-\infty}^{t} Am(s)\left[\frac{d}{dt}\exp(-r(t-s))\exp\left(-\int_{s}^{t} m(x)dx\right)\right]ds \\ &= Am(t) - (r+m(t))\int_{-\infty}^{t} Am(s)\exp(-r(t-s)) \\ &\exp\left(-\int_{s}^{t} m(x)dx\right)ds = Am(t) - (r+m(t))D, \end{aligned}$$

where the derivative can be evaluated using the product rule:

$$\begin{aligned} \frac{d}{dt}\exp(-r(t-s)) &= -r\exp(-r(t-s)) \\ \frac{d}{dt}\exp\left(-\int_{s}^{t} m(x)dx\right) &= -m(t)\exp\left(-\int_{s}^{t} m(x)dx\right). \end{aligned}$$

Changing the flow of time and setting $t=0$ as in the main text, times, this results in

$$D(l) = A\int_{0}^{\infty} \exp[-lrt]\exp[-M_b(t)]m_b(t)dt, \qquad \text{(A7)}$$

where again $m_b(t) = m(-t)$ and

$$M_b(t) = \int_{0}^{t} m_b(x)dx.$$

This motivates the "effective" migration rate

$$m_e(t) = m_b(t)\exp[-M_b(T)];$$

for the case of constant migration, $M_b(t) = mt$ and $m_b(t) = me^{-mt}$, which is an exponential density (Pool and Nielsen 2009).

### Connection between Admixture Segment Length Distribution and ALD Function

Here, we describe how the admixture segment length distribution $P(l)$ and the ALD function $D(l)$ are interconnected (under the assumptions of this work) and in fact uniquely determine each other as claimed in the main text (eqs. 9 and 10).

Following the models outlined in the main text, throughout we assume that admixture is rare and that consequently admixture segments do not interact. Using these assumptions, we first describe how the admixture segment length distribution uniquely determines the ALD function, and second how the ALD function in reverse uniquely determines the admixture segment length distribution.

### From the Admixture Segment Length Distribution to the ALD Function

Without loss of generality, we assume that derived allele frequencies are 0 in the target and 1 in the admixture source population, otherwise the resulting ALD curve can be reweighted with a constant factor $A$ as in equation (4).

We start by noting that two-point ALD between two loci can be written as:

$$D = x_{11} - p \cdot q,$$

where $x_{11}$ denotes the frequency of 11 haplotypes and $p$ and $q$ the allele frequencies at these two loci. Under the assumption that introgressed segments do not interact with each other, genome-wide excess 11 haplotypes beyond random association ($p \cdot q$ for each pair of loci) originate from pairs of markers on the same introgressed segments. To get the genome-wide average $D$, we therefore have to sum over the contribution of 11 haplotypes from introgressed segments of all lengths.

For all pairs of markers a map distance $l$ apart, only segments of length $x > l$ contribute pairs of markers at distance $l$. Let us first describe a single introgressed segment of length $x$. In the limit of long chromosome (of length $G$) and of high marker density, a fraction $(x - l)/G$ of all pairs of markers at distance $l$ fall both onto this segment. Then, denoting the expected number of segments of length $x$ as $E(x)$, we sum over all segment lengths $x > l$ to get:

$$D(l) = \frac{1}{G}\int_l^\infty E(x)(x - l)\mathrm{d}x. \tag{A8}$$

The expected number of segments of a given length $E(x)$ can be directly derived from the segment length distribution via $\mathbb{E}[K]$, the total number of all introgressed segments:

$$E(x) = P(x)\mathbb{E}[K]. \tag{A9}$$

Plugging equation (A9) into equation (A8) yields:

$$D(l) = \frac{\mathbb{E}[K]}{G}\int_l^\infty P(x)(x - l)\mathrm{d}x. \tag{A10}$$

Equation (A10) now allows one to directly calculate $D(l)$ from $P(l)$, which shows that $P(l)$ uniquely determines $D(l)$.

### From the ALD Function to the Admixture Segment Length Distribution

To derive the inverse relationship, we start by differentiating $D(l)$ twice with respect to l ($\frac{\mathrm{d}}{\mathrm{d}l}$). Using equation (A10) and the Leibniz integral rule yields:

$$D'(l) = -\frac{\mathbb{E}[K]}{G}\int_l^\infty P(x)\mathrm{d}x$$
$$D''(l) = \frac{\mathbb{E}[K]}{G}P(l).$$

Simple rearrangement and plugging in the reweighting of LD with $A$ (that we omitted above) yield:

$$P(l) = \frac{1}{A}\frac{G}{\mathbb{E}[K]}D''(l). \tag{A11}$$

Thus, $D(l)$ uniquely determines $P(l)$.

### For the Continuous Admixture Model

For the concrete case of the continuous admixture model, we derived explicit formulas for both the ALD function and the admixture segment length distribution (eqs. 3 and 8). We can validate the above derived functional relationships (eq. A10) directly and show that the more general derivation above holds for these specific formulas central for this work.

To check the relationships, we start with reiterating the explicit formulas (eqs. 3 and 8):

$$P(l) = \frac{G}{\mathbb{E}[K]}\int_0^\infty t^2 m(t)\exp(-lt)\mathrm{d}t \tag{A12}$$

$$D(l) = A\int_0^\infty m\left(t\right)\exp(-lt)\mathrm{d}t. \tag{A13}$$

First, plugging equation (A12) into the functional relationship equation (A10) yields:

$$\begin{aligned}\frac{\mathbb{E}[K]}{G}\int_l^\infty P(x)(x - l)\mathrm{d}x &= \int_l^\infty \mathrm{d}x \int_0^\infty dt\ t^2 m(t)\exp(-xt)(x - l)\\ &= \int_0^\infty \mathrm{d}t\ t^2 m(t)\int_l^\infty \mathrm{d}x\exp(-xt)(x - l)\end{aligned}$$

$$\frac{1}{t^2}\exp(-lt) = \int_0^\infty m(t)\exp(-lt)\mathrm{d}t.$$

If we reweight the right-hand side with the allele-frequency differences A, it becomes D(l) from equation (A13), which finishes the validation that the first functional relationship equation (A10) holds.

To verify the second functional relationship (eq. A11), we differentiate $D(l)$ twice and multiply with $\frac{1}{A}\frac{G}{\mathbb{E}[K]}$:

$$\begin{aligned}\frac{1}{A}\frac{G}{\mathbb{E}[K]}D(l) &= \frac{G}{\mathbb{E}[K]}\int_0^\infty m(t)\frac{\mathrm{d}^2}{\mathrm{d}l^2}\exp(-lt)\mathrm{d}t\\ &= \frac{G}{\mathbb{E}[K]}\int_0^\infty t^2 m(t)\exp(-lt)\mathrm{d}t = P(l).\end{aligned}$$

## Genetic Drift and Recombination

Our model assumes that Neandertal segments in the human population always recombine with non-Neandertal haplotypes, and that the effect of genetic drift can be neglected. In this appendix, we discuss some possible extensions of the model to incorporate aspects of genetic drift and recombination between admixture segments.

### Single Pulse Theory of Recombination between Fragments

Under our modeling assumptions, we show that $\mathbb{E}L_i = t_m^{-1}$, but this ignores recombination between introgressed material. Under a single pulse, Liang and Nielsen (2014) showed that under the SMC model (McVean and Cardin 2005) the expectation is reduced as

$$\mathbb{E}L_i = \left[t_m(1-\alpha)\right]^{-1}, \qquad \text{(A14)}$$

and under the SMC' model (Marjoram and Wall 2006) this is

$$\mathbb{E}L_i = \left[2N(1-\alpha)\left(1-\exp\left(-\frac{t_m}{2N}\right)\right)\right]^{-1}.$$

Using a Taylor expansion around $N \rightarrow \infty$ and ignoring terms of the order $\frac{1}{N^2}$, this can be approximated as

$$\mathbb{E}L_i \approx \left[t_m(1-\alpha)\left(1-\frac{t_m}{4N}\right)\right]^{-1}, \qquad \text{(A15)}$$

which makes the similarity to the SMC model more apparent.

This is compared with $\mathbb{E}L_i = \frac{1}{t_m}$ as we obtained from equation (14a). The $(1-\alpha)$-term models recombination between adjacent introgressed segments; and the $\left(1-\frac{t_m}{4N}\right)$ can be thought of reflecting genetic drift.

The justification for both of these formulas is that they are geometric mixtures of exponential distributions (Liang and Nielsen 2014), which are themselves exponential. Under the extended pulse model, the segment length distribution is no longer exponential, so the segments may have a more complicated mixture distribution.

For the case of Neandertal admixture, assuming gene flow happened over a short duration, these equations can be used to estimate the error made from ignoring drift and recombination between Neandertal segments. As $\alpha \approx 0.03$, $t_m \approx 1,600$, $N \approx 10,000$, and so the expected combined error of these two terms is on the order of 10%.

## Effect of Reduced “Effective” Recombination and Coalescence

In the ALD framework, we can take further complications into account. For example, we can motivate

- an “effective” recombination rate $r(t) = r(1-\alpha(t))$ that takes into account as the admixture fraction increases, some recombination events will be between introgressed material, and we denote the total amount of introgressed material by time $t$ as $\alpha(t)$.
- the allele frequencies in the admixing populations may change, so that we replace the constant $A = \Delta_x\Delta_y$ by $A(t) = \Delta_x(t)\Delta_y(t)$.
- genetic drift will fix some haplotypes, which then can no longer decay. This happens at rate $\frac{1}{2N(t)}$.

Taken together, the analogous equation is

$$\frac{\mathrm{d}D}{\mathrm{d}t} = -\left[r(1-\alpha(t)) + m(t) + \frac{1}{2N(t)}\right]D + A(t)m(t). \qquad \text{(A16)}$$

This equation is still a first-order nonhomogeneous linear differential equation, so the solution will have the same form

$$D(t) = \int_{-\infty}^{t} \exp[-F(s,t)]m(s)A(s)\mathrm{d}s, \qquad \text{(A17)}$$

where

$$\begin{aligned} F(s,t) &= \int_s^t \left[r(1-\alpha(x)) - m(x) - \frac{1}{2N(x)}\right]\mathrm{d}x \\ &= r(t-s) - r\int_s^t \alpha(x)\mathrm{d}x + \int_s^t m(x)\mathrm{d}x + \int_s^t \frac{1}{2N(x)}\mathrm{d}x, \end{aligned}$$

For example, if we assume $N$, $A(t)$, and $\alpha$ are all constant, and migration is a simple pulse at $t_m$

$$\begin{aligned} &D_t(l) = A\int_0^t \delta_{t_m}(s)\exp\left(-\frac{s}{2N}\right)\exp(-l(1-\alpha)s)\mathrm{d}s \\ &D_t''(l) \propto \exp(lt_m(1-\alpha)). \end{aligned}$$

## References

1000 Genomes Project Consortium. 2015. A global reference for human genetic variation. *Nature* 526:68.

Ardia D, Mullen KM, Peterson BG, Ulrich J. DEoptim: Differential Evolution in R. 2016. Available from: https://CRAN.R-project.org/package=DEoptim.

Bard E, Heaton TJ, Talamo S, Kromer B, Reimer RW, Reimer PJ. 2020. Extended dilation of the radiocarbon time scale between 40,000 and 48,000 y BP and the overlap between Neanderthals and Homo sapiens. *Proc Natl Acad Sci U S A.* 117(35):21005–21007.

Browning SR, Browning BL, Zhou Y, Tucci S, Akey JM. 2018. Analysis of human sequence data reveals two pulses of archaic Denisovan admixture. *Cell* 173(1):53–61.e9.

Carpenter B, Gelman A, Hoffman MD, Lee D, Goodrich B, Betancourt M, Brubaker M, Guo J, Li P, Riddell A. 2017. Stan: a probabilistic programming language. *J Stat Soft.* 76(1):1–32.

Chakraborty R, Weiss KM. 1988. Admixture as a tool for finding linked genes and detecting that difference from allelic association between loci. *Proc Natl Acad Sci U S A.* 85(23):9119–9123.

Chimusa ER, Defo J, Thami PK, Awany D, Mulisa DD, Allali I, Ghazal H, Moussa A, Mazandu GK. 2018. Dating admixture events is unsolved problem in multi-way admixed populations. *Brief Bioinformatics.* 21:144–155. doi: 10.1093/bib/bby112.

Choin J, Mendoza-Revilla J, Arauna LR, Cuadros-Espinoza S, Cassar O, Larena M, Ko AM-S, Harmant C, Laurent R, Verdu P, et al. 2021. Genomic insights into population history and biological adaptation in Oceania. *Nature* 592(7855):583–589.

Douka K, Slon V, Jacobs Z, Ramsey CB, Shunkov MV, Derevianko AP, Mafessoni F, Kozlikin MB, Li B, Grün R, et al. 2019. Age estimates for hominin fossils and the onset of the Upper Palaeolithic at Denisova Cave. *Nature* 565(7741):640–644.

Falush D, Stephens M, Pritchard JK. 2003. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* 164(4):1567–1587.

Fu Q, Hajdinjak M, Moldovan OT, Constantin S, Mallick S, Skoglund P, Patterson N, Rohland N, Lazaridis I, Nickel B, et al. 2015. An early modern human from Romania with a recent Neanderthal ancestor. *Nature* 524(7564):216–219.

Fu Q, Li H, Moorjani P, Jay F, Slepchenko SM, Bondarev AA, Johnson PLF, Aximu-Petri A, Prüfer K, de Filippo C, et al. 2014. Genome sequence of a 45,000-year-old modern human from western Siberia. *Nature* 514(7523):445–449.

Gravel S. 2012. Population genetics models of local ancestry. *Genetics* 191(2):607–619.

Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, Patterson N, Li H, Zhai W, Fritz MH-Y, et al. 2010. A draft sequence of the neandertal genome. *Science* 328(5979):710–722.

Hajdinjak M, Mafessoni F, Skov L, Vernot B, Hübner A, Fu Q, Essel E, Nagel S, Nickel B, Richter J, et al. 2021. Initial upper palaeolithic humans in Europe had recent Neanderthal ancestry. *Nature* 592(7853):253–257.

Harris K, Nielsen R. 2016. The genetic cost of Neanderthal introgression. *Genetics* 203(2):881–891.

Hellenthal G, Busby GBJ, Band G, Wilson JF, Capelli C, Falush D, Myers S. 2014. A genetic atlas of human admixture history. *Science* 343(6172):747–751.

Hershkovitz I, Weber GW, Quam R, Duval M, Grün R, Kinsley L, Ayalon A, Bar-Matthews M, Valladas H, Mercier N, et al. 2018. The earliest modern humans outside Africa. *Science* 359(6374):456–459.

Higham T, Douka K, Wood R, Ramsey CB, Brock F, Basell L, Camps M, Arrizabalaga A, Baena J, Barroso-Ruíz C, et al. 2014. The timing and spatiotemporal patterning of Neanderthal disappearance. *Nature* 512(7514):306–309.

Hinch AG, Tandon A, Patterson N, Song Y, Rohland N, Palmer CD, Chen GK, Wang K, Buxbaum SG, Akylbekova EL, et al. 2011. The landscape of recombination in African Americans. *Nature* 476:170.

Hublin J-J. 2017. The last Neanderthal. *Proc Natl Acad Sci U S A.* 114(40):10520–10522.

International HapMap Consortium. 2007. A second generation human haplotype map of over 3.1 million SNPs. *Nature* 449(7164):851–861.

Jacobs GS, Hudjashov G, Saag L, Kusuma P, Darusallam CC, Lawson DJ, Mondal M, Pagani L, Ricaut F-X, Stoneking M, et al. 2019. Multiple deeply divergent Denisovan ancestries in Papuans. *Cell* 177(4):1010–1021.e32.

Juric I, Aeschbacher S, Coop G. 2016. The strength of selection against Neanderthal introgression. *PLoS Genet.* 12(11):e1006340.

Kelleher J, Etheridge AM, McVean G. 2016. Efficient coalescent simulation and genealogical analysis for large sample sizes. *PLoS Comput Biol.* 12(5):e1004842.

Kim B, Lohmueller K. 2015. Selection and reduced population size cannot explain higher amounts of Neandertal ancestry in East Asian than in European human populations. *Am J Hum Genet.* 96(3):454–461.

Kozubowski TJ, Panorska AK, Qeadan F, Gershunov A, Rominger D. 2008. Testing exponentiality versus pareto distribution via likelihood ratio. *Commun Stat Simul Comput.* 38(1):118–139.

Lazaridis I, Nadel D, Rollefson G, Merrett DC, Rohland N, Mallick S, Fernandes D, Novak M, Gamarra B, Sirak K, et al. 2016. Genomic insights into the origin of farming in the ancient Near East. *Nature* 536(7617):419–424.

Liang M, Nielsen R. 2014. The lengths of admixture tracts. *Genetics* 197(3):953–967.

Loh P-R, Lipson M, Patterson N, Moorjani P, Pickrell JK, Reich D, Berger B. 2013. Inferring admixture histories of human populations using linkage disequilibrium. *Genetics* 193(4):1233–1254.

Mafessoni F, Grote S, Filippo CD, Slon V, Kolobova KA, Viola B, Markin SV, Chintalapati M, Peyrégne S, Skov L, et al. 2020. A high-coverage Neandertal genome from Chagyrskaya Cave. *Proc Natl Acad Sci U S A.* 117(26):15132–15136.

Malaspinas A-S, Westaway MC, Muller C, Sousa VC, Lao O, Alves I, Bergström A, Athanasiadis G, Cheng JY, Crawford JE, et al. 2016. A genomic history of Aboriginal Australia. *Nature* 538(7624):207–214.

Marjoram P, Wall JD. 2006. Fast" coalescent" simulation. *BMC Genet.* 7(1):16.

Massilani D, Skov L, Hajdinjak M, Gunchinsuren B, Tseveendorj D, Yi S, Lee J, Nagel S, Nickel B, Devièse T, et al. 2020. Denisovan ancestry and population history of early East Asians. *Science* 370(6516):579–583.

McElreath R. 2020. Statistical rethinking: a Bayesian course with examples in R and STAN. Boca Raton (FL): CRC Press. ISBN 978-0-429-63914-2. Google-Books-ID: 6H_WDwAAQBAJ.

McVean GAT, Cardin NJ. 2005. Approximating the coalescent with recombination. *Philos Trans R Soc Lond B Biol Sci.* 360(1459):1387–1393.

Meyer M, Kircher M, Gansauge M-T, Li H, Racimo F, Mallick S, Schraiber JG, Jay F, Prüfer K, de Filippo C, et al. 2012. A high-coverage genome sequence from an archaic Denisovan individual. *Science* 338(6104):222–226.

Moorjani P, Patterson N, Hirschhorn JN, Keinan A, Hao L, Atzmon G, Burns E, Ostrer H, Price AL, Reich D. 2011. The history of African gene flow into Southern Europeans, Levantines, and Jews. *PLoS Genet.* 7(4):e1001373.

Moorjani P, Sankararaman S, Fu Q, Przeworski M, Patterson N, Reich D. 2016. A genetic method for dating ancient genomes provides a direct estimate of human generation interval in the last 45,000 years. *Proc Natl Acad Sci U S A.* 113(20):5652–5657.

Ni X, Yang X, Guo W, Yuan K, Zhou Y, Ma Z, Xu S. 2016. Length distribution of ancestral tracks under a general admixture model and its applications in population history inference. *Sci Rep.* 6(1):20048.

Pickrell JK, Patterson N, Loh P-R, Lipson M, Berger B, Stoneking M, Pakendorf B, Reich D. 2014. Ancient west Eurasian ancestry in southern and eastern Africa. *Proc Natl Acad Sci U S A.* 111(7):2632–2637.

Pool JE, Nielsen R. 2009. Inference of historical changes in migration rate from the lengths of migrant tracts. *Genetics* 181(2):711–719.

Prüfer K, de Filippo C, Grote S, Mafessoni F, Korlević P, Hajdinjak M, Vernot B, Skov L, Hsieh P, Peyrégne S, et al. 2017. A high-coverage Neandertal genome from Vindija Cave in Croatia. *Science* 358(6363):655–658.

Prüfer K, Racimo F, Patterson N, Jay F, Sankararaman S, Sawyer S, Heinze A, Renaud G, Sudmant PH, de Filippo C, et al. 2013. The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature* 505(7481):43–49.

Pugach I, Duggan AT, Merriwether DA, Friedlaender FR, Friedlaender JS, Stoneking M. 2018. The gateway from Near into Remote Oceania: new insights from genome-wide data. *Mol Biol Evol*. 35(4):871–886.

Pugach I, Matveyev R, Wollstein A, Kayser M, Stoneking M. 2011. Dating the age of admixture via wavelet transform analysis of genome-wide data. *Genome Biol*. 12(2):R19.

R Core Team. 2019. R: a language and environment for statistical computing. Vienna (Austria): R Foundation for Statistical Computing. Available from: https://www.R-project.org/.

Racimo F, Marnetto D, Huerta-Sánchez E. 2017. Signatures of archaic adaptive introgression in present-day human populations. *Mol Biol Evol*. 34(2):296–317.

Ralph P, Coop G. 2013. The geography of recent genetic ancestry across Europe. *PLoS Biol*. 11(5):e1001555.

Reich D, Green RE, Kircher M, Krause J, Patterson N, Durand EY, Viola B, Briggs AW, Stenzel U, Johnson PLF, et al. 2010. Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature* 468(7327):1053–1060.

Sankararaman S, Mallick S, Dannemann M, Prüfer K, Kelso J, Pääbo S, Patterson N, Reich D. 2014. The genomic landscape of Neanderthal ancestry in present-day humans. *Nature* 507(7492):354–357.

Sankararaman S, Mallick S, Patterson N, Reich D. 2016. The combined landscape of Denisovan and Neanderthal ancestry in present-day humans. *Curr Biol*. 26(9):1241–1247.

Sankararaman S, Patterson N, Li H, Pääbo S, Reich D. 2012. The date of interbreeding between Neandertals and modern humans. *PLoS Genet*. 8(10):e1002947.

Seguin-Orlando A, Korneliussen TS, Sikora M, Malaspinas A-S, Manica A, Moltke I, Albrechtsen A, Ko A, Margaryan A, Moiseyev V, et al. 2014. Paleogenomics. Genomic structure in Europeans dating back at least 36,200 years. *Science* 346(6213):1113–1118.

Skov L, Hui R, Shchur V, Hobolth A, Scally A, Schierup MH, Durbin R. 2018. Detecting archaic introgression using an unadmixed outgroup. *PLoS Genet*. 14(9):e1007641.

Spence JP, Song YS. 2019. Inference and analysis of population-specific fine-scale recombination maps across 26 diverse human populations. *Sci Adv*. 5(10):eaaw9206.

Stan Development Team. 2018. RStan: the R interface to Stan. Available from: https://mc-stan.org/rstan/authors.html.

Stephens JC, Briscoe D, O'Brien SJ. 1994. Mapping by admixture linkage disequilibrium in human populations: limits and guidelines. *Am J Hum Genet*. 55(4):809–824.

Stringer C, Galway-Witham J. 2018. When did modern humans leave Africa? *Science* 359(6374):389–390.

Vernot B, Akey J. 2015. Complex history of admixture between modern humans and Neandertals. *Am J Hum Genet*. 96(3):448–453.

Vernot B, Akey JM. 2014. Resurrecting surviving Neandertal lineages from modern human genomes. *Science* 343(6174):1017–1021.

Vernot B, Tucci S, Kelso J, Schraiber JG, Wolf AB, Gittelman RM, Dannemann M, Grote S, McCoy RC, Norton H, et al. 2016. Excavating Neandertal and Denisovan DNA from the genomes of Melanesian individuals. *Science* 352(6282):235–239.

Villanea FA, Schraiber JG. 2019. Multiple episodes of interbreeding between Neanderthal and modern humans. *Nat Ecol Evol*. 3(1):39–44.

Vyas DN, Mulligan CJ. 2019. Analyses of Neanderthal introgression suggest that Levantine and southern Arabian populations have a shared population history. *Am J Phys Anthropol*. 169(2):227–239.

Wall JD. 2000. Detecting ancient admixture in humans using sequence polymorphism data. *Genetics* 154(3):1271–1279.

Wall JD, Yang MA, Jay F, Kim SK, Durand EY, Stevison LS, Gignoux C, Woerner A, Hammer MF, Slatkin M. 2013. Higher levels of neanderthal ancestry in East Asians than in Europeans. *Genetics* 194(1):199–209.

Zhou Y, Qiu H, Xu S. 2017. Modeling continuous admixture using admixture-induced linkage disequilibrium. *Sci Rep*. 7(1):43054–43010.

Zhou Y, Yuan K, Yu Y, Ni X, Xie P, Xing EP, Xu S. 2017. Inference of multiple-wave population admixture by modeling decay of linkage disequilibrium with polynomial functions. *Heredity (Edinb)*. 118(5):503–510.

Zilhão J, Anesin D, Aubry T, Badal E, Cabanes D, Kehl M, Klasen N, Lucena A, Martín-Lerma I, Martínez S, et al. 2017. Precise dating of the Middle-to-Upper Paleolithic transition in Murcia (Spain) supports late Neandertal persistence in Iberia. *Heliyon* 3(11):e00435.