Estimating divergence times from DNA sequences

Per Sjödin[*,1], **James McKenna**[*] **and Mattias Jakobsson**[*,1,†]

[*]Human Evolution, Department of Organismal Biology, Uppsala University, Uppsala, Sweden, [†]Science for Life Laboratory, Uppsala University, Uppsala, Sweden

ABSTRACT The patterns of genetic variation within and among individuals and populations can be used to make inferences about the evolutionary forces that generated those patterns. Numerous population genetic approaches have been developed in order to infer evolutionary history. Here, we present the 'Two-Two (TT)' and the 'Two-Two-outgroup (TTo)' methods; two closely related approaches for estimating divergence time based in coalescent theory. They rely on sequence data from two haploid genomes (or a single diploid individual) from each of two populations. Under a simple population-divergence model, we derive the probabilities of the possible sample configurations. These probabilities form a set of equations that can be solved to obtain estimates of the model parameters, including population split-times, directly from the sequence data. This transparent and computationally efficient approach to infer population divergence time makes it possible to estimate time scaled in generations (assuming a mutation rate), and not as a compound parameter of genetic drift. Using simulations under a range of demographic scenarios, we show that the method is relatively robust to migration and that the TTo-method can alleviate biases that can appear from drastic ancestral population size changes. We illustrate the utility of the approaches with some examples, including estimating split times for pairs of human populations as well as providing further evidence for the complex relationship among Neandertals and Denisovans and their ancestors.

KEYWORDS Effective population size; Divergence time; Population divergence; Human evolution

Background

Many population genetic inference approaches compare levels of genetic variation within and across genomes, individuals and/or populations in order to uncover their evolutionary history. A multitude of demographic inference methods have been developed in order to capitalize on the wealth of information that comes with the availability of full genomes from multiple individuals (see Schraiber and Akey 2015, for a review).

The sheer scale and complexity of whole genome data sets poses its own challenge for making inference of population demographic parameters. A common approach for inference has been to compare the observed data, often summarized in some statistic, to simulated data that can be generated under a range of population-genetic models. Building on this idea and combined with a rejection algorithm, Approximate Bayesian Computation (ABC, Beaumont *et al.* 2002; Cornuet *et al.* 2014; Tavaré *et al.* 1997; Pudlo *et al.* 2016) has proven to be one useful tool for both model choice and parameter estimation. However, the problem of choosing which models to test is not trivial for most inference approaches, including ABC, as the set of models to choose from is very large.

In parallel, there have been recent developments in methods that use haplotype information. A challenge for these approaches is how to model the dependence of genealogies along a sequence. One solution has been to approximate the full ancestral recombination graph, a method used in PSMC (Li and Durbin 2011) and similar approaches (Schiffels and Durbin 2014; Terhorst *et al.* 2017; Kelleher *et al.* 2019; Speidel *et al.* 2019; Wang *et al.* 2020).

Another strategy has been to rely on relatively short genetic fragments located sufficiently far away from each other to be able to assume linkage equilibrium between loci, combined with absolute linkage (absence of recombination) within each locus (e.g. Gronau *et al.* 2011). Both these approaches typically lead to set-ups that cannot be solved analytically and often rely on computationally heavy, advanced statistical methods in order

doi: 10.1534/genetics.XXX.XXXXXX
Manuscript compiled: Thursday 15th October, 2020
Human Evolution, Department of Organismal Biology, Uppsala University, Norbyvägen 18C, 75 236 Uppsala, Sweden. per.sjodin@ebc.uu.se, mattias.jakobsson@ebc.uu.se

to estimate parameters (but see Gattepaille *et al.* 2016; Lohse *et al.* 2016). A related strategy is to assume independence among sites, using a composite likelihood framework (Gutenkunst *et al.* 2009; Excoffier *et al.* 2013). From this assumption, the observed variables (*i.e.* frequency spectra) do not depend on the full distribution of genealogical branch lengths, they are functions only of the expected branch lengths (Griffiths and Tavaré 1998; Chen 2012). This observation greatly simplifies the probability computations. To the extent that closed-form solutions can be obtained, the assumption of independence between sites also leads to inference tools that are easier to integrate with other methods, and can provide useful insights into underlying processes (Beichman *et al.* 2017; Terhorst *et al.* 2017). Conversely, a disadvantage of assuming independence between sites is that only information concerning the expected values can be obtained, rather than the full distributions of stochastic variables. For small samples, an alternative is to derive closed form expressions for the probability of observing particular configurations of variants in simple divergence models, including the isolation-with-migration model (Wakeley and Hey 1997; Wilkinson-Herbots 2008; Chen 2012). Lohse *et al.* (2011, 2016) showed that more generally, the probability of observing a particular variant configuration can be obtained from a generating function of genealogical branches. Assuming independence among sequence-blocks, Lohse *et al.* (2011, 2016) outlined an approach for computing the likelihood under various demographic models and sampling schemes.

Regardless of whether independence between sites is assumed or not, all of these methods can be useful for inferring the time of divergence between two populations. Examples of simple and direct methods used to estimate population divergence times include: Green *et al.* (2010); Gutenkunst *et al.* (2009); Wakeley (2009); Schlebusch *et al.* (2012); Theunert and Slatkin (2018). These methods build on the principle of genetic drift accumulating as a function of effective population size and number of generations. Following a population backwards in time, and using that the accumulated drift at generation t is,

$$\sum_{i=1}^{t} \frac{1}{2N(i)},$$

where $N(i)$ is the (effective) number of individuals at generation i, the divergence time is then the number of generations required to generate the estimated drift. Such estimates are typically not dependent on knowing the mutation rate but some assumptions regarding $N(i)$ is required, either by assuming a fixed effective population size or depending on an estimated function of $N(i)$.

Alternatively, one can base the divergence time estimate on an assumed mutation rate (e.g. Wakeley and Hey 1997; Chen 2012; Pickrell *et al.* 2012). By assuming independence among sites, in a two-population divergence model (without migration), the probability of observed sample configurations (summarized as the full SFS, including invariable sites) can be derived analytically. Using a likelihood framework, we can then estimate parameters of interest in the divergence model. Here we present two simple approaches based on picking two gene copies from each of two populations: the 'Two-Two' (TT) method, which was briefly introduced in Schlebusch *et al.* (2017) and the 'Two-Two-outgroup' (TTo) method. These are sufficiently simple to allow for analytical solutions giving closed formulas for the estimates of the model parameters based on the counts of different sample configurations. Specifically, we can directly estimate population divergence time in generations, an estimate which is independent of genetic drift and effects of varying population size since the model is parametrized with both a drift parameter and a time parameter.

Observed data

For the purpose of investigating the demographic relationship between two populations denoted population 1 and population 2, assume that two gene copies have been sampled from each population. For bi-allelic sites, assume that the ancestral (denoted '0') and the derived variant (denoted '1') is known. The number of derived alleles in a sample from population 1 combined with the number of derived alleles in as sample from population 2 is referred to as the joint frequency spectra. Following Chen (2012), a sample configuration of k_1 derived variants in a sample of size n_1 from population 1 and k_2 derived variants in a sample of size n_2 from population 2 can then be denoted as $S_{n_1,n_2}(k_1,k_2)$. In our set up, $n_1 = n_2 = 2$ and k_1, k_2 are either 0, 1 or 2, and there are 9 possible sample configurations, which are presented in Table 1. The observed number of sites with sample configuration $O_{i,j}$ will be denoted by $m_{i,j}$ and the total number of investigated sites by m_{tot}.

Table 1 Notation for the number of sites with 0, 1, or 2 derived variants in the sample from population 1 and the sample from population 2.

	0 in pop. 2	1 in pop. 2	2 in pop. 2
0 in pop. 1	$O_{0,0}$	$O_{0,1}$	$O_{0,2}$
1 in pop. 1	$O_{1,0}$	$O_{1,1}$	$O_{1,2}$
2 in pop. 1	$O_{2,0}$	$O_{2,1}$	$O_{2,2}$

Theory

We study a general population divergence model where the population-branch leading to population 1 and the population-branch leading to population 2 merge (backwards in time) to become the ancestral population. The model makes no assumptions regarding population size and/or population structure changes in the daughter populations. The model assumes no migration between the two daughter populations and that these merge into a panmictic ancestral population.

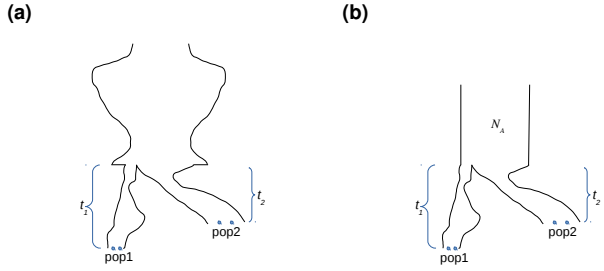

Figure 1 Different assumptions for population divergence models (a) panmictic ancestral population, and (b) constant ancestral population.

We use the following notation, with time measured in number

of generations:

t_1 : time to split for pop. 1,
t_2 : time to split for pop. 2,
α_1 : prob. of two lineages in pop. 1 not coalescing before t_1,
α_2 : prob. of two lineages in pop. 2 not coalescing before t_2,
ν_1 : expected time to coal. in pop. 1 given coal. before t_1,
ν_2 : expected time to coal. in pop. 2 given coal. before t_2.

In addition to the drift parameters α_1 and α_2, the parameters ν_1 and ν_2 are needed because two branches with the same time-length and the same drift can have different distributions of coalescence times. To illustrate, a linearly growing population that starts with size N and ends with size $2N$ will have the same drift as a shrinking population that starts with size $2N$ and ends with size N but they will not have the same distribution of coalescent times within that interval. These parameters also cover cases when the daughter populations are not panmictic. A similar parametrization can be found, for instance, in Rogers and Bohlender (2014).

The composite likelihood assumption of independence between sites implies that the probability of a mutation on a specific branch in a genealogy is the expected length of that branch (given a demographic model) multiplied by the mutation rate. We denote the mutation rate per site and generation by μ, assume independence between sites, and an infinite sites model.

We define the following events for the two sampled lineages for each population:

H_1 : a coalescence in population 1 before t_1; $P(H_1) = 1 - \alpha_1$
H_2 : a coalescence in population 2 before t_2; $P(H_2) = 1 - \alpha_2$.

With $2 \leq k \leq 4$ lineages surviving to enter the ancestral population (depending on whether coalescence events have occurred in the daughter populations), we define A_k to be the number of derived variants in a sample of size k drawn at the split time in the ancestral population and write $a_{ki} = P(A_k = i)$. To illustrate how the probabilities of the sample configurations are derived, we can take an example conditional on no coalescent event in population 1 and a coalescent event in population 2 (the event $\neg H_1 \wedge H_2$). There are then 3 lineages entering the ancestral population. These lineages constitute a sample of size 3 from the ancestral population. Sample configuration $O_{1,0}$ will then be observed with probability $(2/3)a_{31}$ (the ancestral variant has to be assigned to the lineage entering population 2; an event with probability $2/3$) plus the probability that a mutation occurs on either lineage entering population 1 during the time interval t_1. The probability that a mutation hits a branch of length t_1 is μt_1, and the probability that this happens *and* that the derived variant already exists in the ancestral population can be ignored as it requires two mutational events at the same site. Thus, conditional on $\neg H_1 \wedge H_2$, $P(O_{1,0}) = (2/3)a_{31} + 2\mu t_1$. The same reasoning can be applied to derive the conditional probabilities for all 7 (polymorphic) sample configurations and these are shown in Table 2.

Since a sub-sample of size n randomly drawn from a larger sample of size $n + k$ has the same distribution as a sample of size n drawn directly from the population, we can reduce the number of parameters by replacing all a_{ij} with $i < 4$ using a_{ij}-terms with

Table 2 Conditional probabilities.

	$H_1 \wedge H_2$	$H_1 \wedge \neg H_2$	$\neg H_1 \wedge H_2$	$\neg H_1 \wedge \neg H_2$
$O_{1,0}$	$2\mu\nu_1$	$2\mu\nu_1$	$\frac{2a_{31}}{3} + 2\mu t_1$	$\frac{a_{41}}{2} + 2\mu t_1$
$O_{0,1}$	$2\mu\nu_2$	$\frac{2a_{31}}{3} + 2\mu t_2$	$2\mu\nu_2$	$\frac{a_{41}}{2} + 2\mu t_2$
$O_{2,0}$	$\frac{a_{21}}{2} + \mu(t_1 - \nu_1)$	$\frac{a_{31}}{3} + \mu(t_1 - \nu_1)$	$\frac{a_{32}}{3}$	$\frac{a_{42}}{6}$
$O_{0,2}$	$\frac{a_{21}}{2} + \mu(t_2 - \nu_2)$	$\frac{a_{32}}{3}$	$\frac{a_{31}}{3} + \mu(t_2 - \nu_2)$	$\frac{a_{42}}{6}$
$O_{1,1}$	0	0	0	$\frac{2a_{42}}{3}$
$O_{2,1}$	0	$\frac{2a_{32}}{3}$	0	$\frac{a_{43}}{2}$
$O_{1,2}$	0	0	$\frac{2a_{32}}{3}$	$\frac{a_{43}}{2}$

$i = 4$ as follows:

$$a_{21} = P(A_2 = 1) = \sum_{i=0}^{4} P(A_2 = 1|A_4 = i)a_{4i} = \frac{1}{2}a_{41} + \frac{2}{3}a_{42} + \frac{1}{2}a_{43}$$
$$a_{31} = P(A_3 = 1) = \frac{3}{4}a_{41} + \frac{1}{2}a_{42}$$
$$a_{32} = P(A_3 = 2) = \frac{1}{2}a_{42} + \frac{3}{4}a_{43}.$$

These equations together with Table 2 allows us to derive the probabilities for the different sample configurations. For instance:

$$\begin{aligned} P(O_{1,0}) &= (1-\alpha_1)(1-\alpha_2)2\mu\nu_1 + (1-\alpha_1)\alpha_2 2\mu\nu_1 \\ &\quad + \alpha_1(1-\alpha_2)\left(\frac{2}{3}a_{31} + 2\mu t_1\right) \\ &\quad + \alpha_1\alpha_2\left(\frac{1}{2}a_{41} + 2\mu t_1\right) \\ &= 2(1-\alpha_1)\mu\nu_1 + 2\alpha_1\left(\mu t_1 + \frac{1}{4}b_1\right) + \frac{1}{3}\alpha_1(1-\alpha_2)b_2 \end{aligned}$$

where $b_i = a_{4i} = P(A_4 = i)$.

Using the same strategy for the derivation of the other six probabilities, we obtain the probabilities for all seven sample configurations in Table 2. Writing $p_{i,j} = P(O_{i,j})$ for brevity, these are

$$\begin{aligned} p_{1,0} &= 2(1-\alpha_1)\mu\nu_1 + 2\alpha_1\left(\mu t_1 + \frac{1}{4}b_1\right) + \frac{1}{3}\alpha_1(1-\alpha_2)b_2 \\ p_{0,1} &= 2(1-\alpha_2)\mu\nu_2 + 2\alpha_2\left(\mu t_2 + \frac{1}{4}b_1\right) + \frac{1}{3}(1-\alpha_1)\alpha_2 b_2 \\ p_{2,0} &= (1-\alpha_1)\left(\mu t_1 + \frac{1}{4}b_1\right) - (1-\alpha_1)\mu\nu_1 \\ &\quad + \frac{1}{6}(2 - \alpha_1 - \alpha_2 + \alpha_1\alpha_2)b_2 + \frac{1}{4}(1-\alpha_2)b_3 \\ p_{0,2} &= (1-\alpha_2)\left(\mu t_2 + \frac{1}{4}b_1\right) - (1-\alpha_2)\mu\nu_2 \\ &\quad + \frac{1}{6}(2 - \alpha_1 - \alpha_2 + \alpha_1\alpha_2)b_2 + \frac{1}{4}(1-\alpha_1)b_3 \\ p_{1,1} &= \frac{2}{3}\alpha_1\alpha_2 b_2 \\ p_{2,1} &= \frac{1}{3}(1-\alpha_1)\alpha_2 b_2 + \frac{1}{2}\alpha_2 b_3 \\ p_{1,2} &= \frac{1}{3}\alpha_1(1-\alpha_2)b_2 + \frac{1}{2}\alpha_1 b_3 \end{aligned} \quad (1)$$

Furthermore, if we assume a (indefinitely) panmictic ancestral population (figure 1a), we define

$$T_{4i}$$

to be the number of generations a coalescent process that starts with 4 lineages at the (most recent) base of the ancestral population spends with i lineages, so that $T_{mrca} = T_{44} + T_{43} + T_{42}$. Then (see appendix)

$$\begin{aligned} b_1 = P(A_4 = 1) =& \frac{2}{3}\mu E[T_{42}] + 2\mu E[T_{43}] + 4\mu E[T_{44}], \\ b_2 = P(A_4 = 2) =& \frac{2}{3}\mu E[T_{42}] + \mu E[T_{43}], \\ b_3 = P(A_4 = 3) =& \frac{2}{3}\mu E[T_{42}]. \end{aligned}$$

Writing $\tau_i = \mu E[T_{4i}]$, and replacing the b_i with their respective expression in terms of τ_i, the probabilities for the different sample configurations can be expressed as:

$$\begin{aligned} p_{1,0} =& 2(1-\alpha_1)\mu\nu_1 + 2\alpha_1(\mu t_1 + \tau_4) \\ &+ \frac{1}{9}\alpha_1(4-\alpha_2)(2\tau_2 + 3\tau_3) - \frac{1}{3}\alpha_1\tau_2 \\ p_{0,1} =& 2(1-\alpha_2)\mu\nu_2 + 2\alpha_2(\mu t_2 + \tau_4) \\ &+ \frac{1}{9}(4-\alpha_1)\alpha_2(2\tau_2 + 3\tau_3) - \frac{1}{3}\alpha_2\tau_2 \\ p_{2,0} =& (1-\alpha_1)(\mu t_1 + \tau_4 - \mu\nu_1) \\ &+ \frac{1}{18}(5 - 4\alpha_1 - \alpha_2 + \alpha_1\alpha_2)(2\tau_2 + 3\tau_3) + \frac{1}{6}(\alpha_1 - \alpha_2)\tau_2 \\ p_{0,2} =& (1-\alpha_2)(\mu t_2 + \tau_4 - \mu\nu_2) \\ &+ \frac{1}{18}(5 - \alpha_1 - 4\alpha_2 + \alpha_1\alpha_2)(2\tau_2 + 3\tau_3) + \frac{1}{6}(\alpha_2 - \alpha_1)\tau_2 \\ p_{1,1} =& \frac{2}{9}\alpha_1\alpha_2(2\tau_2 + 3\tau_3) \\ p_{2,1} =& \frac{1}{9}\alpha_2(1-\alpha_1)(2\tau_2 + 3\tau_3) + \frac{1}{3}\alpha_2\tau_2 \\ p_{1,2} =& \frac{1}{9}\alpha_1(1-\alpha_2)(2\tau_2 + 3\tau_3) + \frac{1}{3}\alpha_1\tau_2 \\ p_{0,0} + p_{2,2} =& 1 - \sum_{0<i+j<4} p_{i,j} \end{aligned} \quad (2)$$

These 8 equations point to two challenges: i) it is not possible to completely separate τ_4 from divergence times due to its co-occurrence with μt_1 and μt_2, ii) disregarding τ_4, it is still an under-determined set of equations with 8 parameters but only 7 equations/degrees of freedom ($p_{0,0} + p_{2,2} = 1 - \sum_{0<i+j<4} p_{i,j}$). It can be tempting to reduce the number of parameters by setting $t_1 = t_2$, but because (from equations 1 above)

$$\mu\{t_1 - t_2\} = \frac{1}{2}(p_{1,0} - p_{0,1}) + (p_{2,0} - p_{0,2}) + \frac{1}{2}(p_{2,1} - p_{1,2}),$$

specifying $t_1 = t_2$ would add additional dependence between the equations. Although this will decrease the number of parameters, it also decreases the number of independent equations. Furthermore, allowing for separate divergence times along the two branches is a valuable asset; not only does it allow the framework to be applicable for temporally structured samples, but separate estimates for each branch can be useful more generally. In fact, it turns out that the divergence time estimate based on the population branch represented by a modern-day individual alleviates the potential issue of residual ancient DNA specific properties (DNA degradation, sequencing errors, mapping errors) that could impact divergence time estimates (see below). In contrast, for contemporaneous samples, divergence time estimates should be the same along the two branches (assuming neutrality and the same mutation rate and generation time along the two branches).

The challenges noted above can be dealt with either by assuming a constant ancestral population size (the "TT"-method) or by using an outgroup to increase the number of equations (the "TTo"-method).

Assuming a constant ancestral population size ('TT')

Assuming a constant ancestral population size N_A reduces the number of parameters in the model (figure 1b), so that $E[T_{4k}] = 2N_A/(k(k-1))$ and (with $\theta = \mu N_A$) $\tau_2 = \theta$, $\tau_3 = \theta/3$ and $\theta = \tau_4/6$. Then the probabilities in equations (2) simplify as

$$\begin{aligned} p_{1,0} =& 2\alpha_1 T_1 + 2(1-\alpha_1)V_1 + \frac{1}{3}\alpha_1(4-\alpha_2)\theta \\ p_{0,1} =& 2\alpha_2 T_2 + 2(1-\alpha_2)V_2 + \frac{1}{3}\alpha_2(4-\alpha_1)\theta \\ p_{2,0} =& (1-\alpha_1)(T_1 - V_1) + \frac{1}{6}(6 - 4\alpha_1 - 2\alpha_2 + \alpha_1\alpha_2)\theta \\ p_{0,2} =& (1-\alpha_2)(T_2 - V_2) + \frac{1}{6}(6 - 2\alpha_1 - 4\alpha_2 + \alpha_1\alpha_2)\theta \\ p_{1,1} =& \frac{2}{3}\alpha_1\alpha_2\theta \\ p_{2,1} =& \frac{1}{3}(2-\alpha_1)\alpha_2\theta \\ p_{1,2} =& \frac{1}{3}(2-\alpha_2)\alpha_1\theta \end{aligned}$$

With

$$\begin{aligned} T_1 =& \mu t_1 \\ T_2 =& \mu t_2 \\ V_1 =& \mu\nu_1 \\ V_2 =& \mu\nu_2 \end{aligned}$$

We set $\frac{m_{i,j}}{m_{tot}} = p_{i,j}$ and solve for the parameters (and note that this guarantees that they are also maximum likelihood estimates (Doob 1934; Wald 1949) to obtain:

$$\begin{aligned} \widehat{\alpha_1} &= \frac{2m_{1,1}}{2m_{2,1} + m_{1,1}} \\ \widehat{\alpha_2} &= \frac{2m_{1,1}}{2m_{1,2} + m_{1,1}} \\ \widehat{\theta} &= \frac{1}{m_{tot}}\frac{3}{8}\frac{(2m_{2,1} + m_{1,1})(2m_{1,2} + m_{1,1})}{m_{1,1}} \\ \widehat{T_1} &= \frac{1}{m_{tot}}\left(\frac{m_{1,0}}{2} + m_{2,0} - \frac{(2m_{2,1} + m_{1,1})(6m_{1,2} + m_{1,1})}{8m_{1,1}}\right) \\ \widehat{T_2} &= \frac{1}{m_{tot}}\left(\frac{m_{0,1}}{2} + m_{0,2} - \frac{(6m_{2,1} + m_{1,1})(2m_{1,2} + m_{1,1})}{8m_{1,1}}\right) \\ \widehat{V_1} &= \frac{1}{m_{tot}}\left(\frac{m_{1,0} + m_{1,2}}{2} - m_{1,1}\frac{2m_{2,0} - m_{1,2}}{2m_{2,1} - m_{1,1}}\right) \\ \widehat{V_2} &= \frac{1}{m_{tot}}\left(\frac{m_{0,1} + m_{2,1}}{2} - m_{1,1}\frac{2m_{0,2} - m_{2,1}}{2m_{1,2} - m_{1,1}}\right). \end{aligned} \quad (3)$$

These equations are referred to as the 'TT'-method in Schlebusch *et al.* (2017) where, in order to get the divergence time in years we used $G = 30$ and $\mu = 1.25 \times 10^{-8}$ in

$$\widehat{t_i} = \frac{G}{\mu}\widehat{T_i} \quad (4)$$

where G is the length of a generation.

Note also that if sequencing errors or DNA-degradation mainly result in additional singletons, then errors in the sample from population 1 only affects $m_{1,0}$ and thus only $\widehat{T_1}$ and $\widehat{V_1}$ ($m_{1,0}$ occurs exclusively in the equations for estimating T_1 and V_1).

Adding an outgroup ('TTo')

The equations (1) are useful for data (including SNP-genotype data) where the derived variant at each site has been ascertained in a population that branched off prior to the investigated population split. Such data will ensure that derived variants in the studied sample will be older than the split so that there are no new mutations occurring in the branches. In such a case, μt_1, μt_2, $\mu \nu_1$ and $\mu \nu_2$ can all be set to 0 in the equations 1 above, resulting in a new set of equations (see appendix) that can be solved for the α's to get

$$\widehat{\alpha_1}^* = 2\frac{m^*_{1,0} + m^*_{1,2} + m^*_{1,1}}{2(m^*_{1,0} + 2m^*_{2,0} + m^*_{2,1}) + m^*_{1,1}}$$
$$\widehat{\alpha_2}^* = 2\frac{m^*_{0,1} + m^*_{2,1} + m^*_{1,1}}{2(m^*_{0,1} + 2m^*_{0,2} + m^*_{1,2}) + m^*_{1,1}}.$$

where * indicates that these are the corresponding parameters and sample configuration counts conditional on ascertainment in an outgroup.

With this ascertainment procedure it is important that the population used to ascertain the SNPs represents a true outgroup to our studied populations and that the populations satisfy an assumption of bifurcating topology (or "tree-ness"). To validate such an assumption we can set up tests of tree-ness, since if $\mu t_1 = \mu t_2 = \mu \nu_1 = \mu \nu_2 = 0$, then the test statistics

$$Y_1 = \frac{2m^*_{1,0} + m^*_{1,1}}{2m^*_{0,1} + m^*_{1,1}} - \frac{2m^*_{1,2} + m^*_{1,1}}{2m^*_{2,1} + m^*_{1,1}} \quad (5)$$
$$Y_2 = \frac{(m^*_{1,0} - m^*_{0,1}) + 2(m^*_{2,0} - m^*_{0,2}) + (m^*_{2,1} - m^*_{1,2})}{m^*_{tot}} \quad (6)$$

should be 0 (see appendix, where it is also shown that Y_2 is closely related to the D-statistic, Green *et al.* 2010).

The estimates $\widehat{\alpha_1}^*$ and $\widehat{\alpha_2}^*$ of α_1 and α_2 together with the equations in (2) can furthermore be used to obtain estimates of

$$\begin{aligned}
\widehat{\tau_2}^* &= \frac{1}{m_{tot}}\frac{3}{2}\left(\frac{2m_{2,1} + m_{1,1}}{\widehat{\alpha_2}^*} - \frac{m_{1,1}}{\widehat{\alpha_1}^*\widehat{\alpha_2}^*}\right) \\
&= \frac{1}{m_{tot}}\frac{3}{2}\left(\frac{2m_{1,2} + m_{1,1}}{\widehat{\alpha_1}^*} - \frac{m_{1,1}}{\widehat{\alpha_1}^*\widehat{\alpha_2}^*}\right) \\
\widehat{\tau_3}^* &= \frac{1}{m_{tot}}\left(\frac{5}{2}\frac{m_{1,1}}{\widehat{\alpha_1}^*\widehat{\alpha_2}^*} - \frac{2m_{2,1} + m_{1,1}}{\widehat{\alpha_2}^*}\right) \\
&= \frac{1}{m_{tot}}\left(\frac{5}{2}\frac{m_{1,1}}{\widehat{\alpha_1}^*\widehat{\alpha_2}^*} - \frac{2m_{1,2} + m_{1,1}}{\widehat{\alpha_1}^*}\right) \\
\widehat{B_1}^* &= \frac{1}{m_{tot}}\left(\frac{m_{1,0}}{2} + m_{2,0} + \frac{m_{2,1}}{2} - \frac{5 - \widehat{\alpha_1}^*\widehat{\alpha_2}^*}{\widehat{\alpha_1}^*\widehat{\alpha_2}^*}\frac{m_{1,1}}{4}\right) \\
\widehat{B_2}^* &= \frac{1}{m_{tot}}\left(\frac{m_{0,1}}{2} + m_{0,2} + \frac{m_{1,2}}{2} - \frac{5 - \widehat{\alpha_1}^*\widehat{\alpha_2}^*}{\widehat{\alpha_1}^*\widehat{\alpha_2}^*}\frac{m_{1,1}}{4}\right) \\
\widehat{V_1}^* &= \frac{1}{m_{tot}}\left(\frac{m_{1,0}}{2} - \widehat{\alpha_1}^*\frac{m_{2,0}}{1 - \widehat{\alpha_1}^*} + \frac{m_{1,2}}{2(1 - \widehat{\alpha_1}^*)}\right) \\
\widehat{V_2}^* &= \frac{1}{m_{tot}}\left(\frac{m_{0,1}}{2} - \widehat{\alpha_2}^*\frac{m_{0,2}}{1 - \widehat{\alpha_2}^*} + \frac{m_{2,1}}{2(1 - \widehat{\alpha_2}^*)}\right) \quad (7)
\end{aligned}$$

with

$$B_1 = \mu t_1 + \tau_4$$
$$B_2 = \mu t_2 + \tau_4.$$

Note the two alternative estimates (one using $m_{2,1}$ and one using $m_{1,2}$) for τ_3 and τ_2 and we take the average of these in estimates below.

Based on the obtained estimates of τ_2 and τ_3, we can attempt to approximate τ_4 as a combination of τ_2 and $3\tau_3$. In a constant population, $E[T_{43}]/E[T_{42}] = 1/3$ and $E[T_{44}]/E[T_{43}] = 1/2$ or

$$\frac{\tau_4}{\tau_3} = \frac{E[T_{44}]}{E[T_{43}]} = \frac{3}{2}\frac{E[T_{43}]}{E[T_{42}]} = \frac{3}{2}\frac{\tau_3}{\tau_2}.$$

For this reason we propose to approximate τ_4/τ_3 as $(3/2)x$ where x is the estimated ratio of τ_3/τ_2. This leads to

$$\widehat{\tau_4}^* = \frac{3}{2}\frac{(\widehat{\tau_3}^*)^2}{\widehat{\tau_2}^*}$$

to get

$$\widehat{T_i}^* = \widehat{B_i}^* - \frac{3}{2}\frac{(\widehat{\tau_3}^*)^2}{\widehat{\tau_2}^*}. \quad (8)$$

We refer to this approach to estimate divergence time as "TTo" (as in "TT outgroup").

Picking two gene copies from population 1 and one gene copy from population 2

The method so far described can be seen as an expansion of the simpler case of picking two gene copies from one population, and only one gene copy from the other population. This simpler set-up can be useful, for instance, when dealing with low coverage genome data (e.g. ancient DNA sequence data). With this simpler approach, divergence time estimation needs an outgroup (only assuming a constant population size is not sufficient to solve the equations in this case). This 2 plus 1 approach does, however, provide reliable estimates of branch specific genetic drift (under often reasonable demographic assumptions, see appendix and Wakeley 2009; Schlebusch *et al.* 2012; Skoglund *et al.* 2011).

Simulations and comparison to GPhoCS

The model underlying the TT method assumes a panmictic ancestral population of constant size prior to the split, and no gene-flow between populations after the split. While common to many coalescent-based approaches, such assumptions are rarely realistic for natural populations, and it is increasingly evident that mis-specification of an overly simplistic model may lead to substantially biased parameter estimates (Gronau *et al.* 2011; Mazet *et al.* 2016; Orozco 2016).

Here we investigate the robustness of the TT-method parameter estimation against violation of the basic model assumptions (equations in 3). We compare its performance under these conditions against an alternative method for parameter inference, GPhoCS (Gronau *et al.* 2011). The analytical TT method and the Bayesian inference method GPhoCS are located to some degree at opposite ends of the statistical inference spectrum; instead of relying on independent single bi-allelic sites, GPhoCS assumes complete linkage between individual sites at a genetic locus (typically 10 kb), but independence between these loci. It should

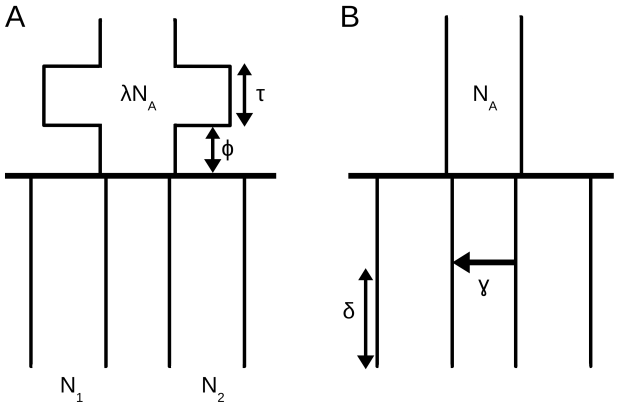

Figure 2 The two general demographic models used to simulate data for testing robustness of TT method, with (A) changes in ancestral population size, and (B) variation in proportion and timing of migration between branches since a population split.

be noted that GPhoCS is capable of estimating parameters under more complicated demographic models than the simple split model we study here. In particular, GPhoCS allows users to specify migration rates and define migration bands between populations, such that it does not share the TT method assumption of no gene-flow occurring between populations after the population split. For this reason, the effect of migration on parameter estimation was investigated only for the TT method.

The software MS (Hudson 2002) was used to generate polymorphic datasets using a standard coalescent algorithm under a variety of demographic scenarios. The effects of changes in ancestral population size (figure 2A) and migration between branches since the split (figure 2B) were investigated. In each model, the ancestral population size, N_A, was fixed at 34,000, corresponding to 17,000 diploid individuals. This value is in line with recent estimates of African ancestral effective population size approximately 1 million years ago (Li and Durbin 2011; Schiffels and Durbin 2014; Schlebusch *et al.* 2017). MS scales time by $4N_e$, and simulations were constructed with true split times of 10,000 and 1500 generations. Assuming a generation time of 30 years this equates to split times of 300,000 and 45,000 years respectively. These were chosen to keep simulations relevant to the findings of previous work where the deepest split among human groups was estimated at >260,000 years (Schlebusch *et al.* 2017), together with more recent divergence events.

MS generates samples assuming $\theta = 4N_1\mu$, where N_1 is the diploid population size of population 1, and μ is the neutral mutation rate. Mutation rates vary across the human genome and estimates vary depending on the method used (Scally and Durbin 2012). Li and Durbin (2011) calculated a human neutral mutation rate of 2.5×10^{-8} per generation (assuming 25 years per generation), whilst recent consensus suggests a lower rate of 1.25×10^{-8} per base pair per generation is more accurate (Moorjani *et al.* 2016). The latter is the mutation rate used across all simulations. Results were filtered such that only those simulations resulting in all sample configurations represented by $> 10,000$ sites were used in subsequent analyses.

The Bayesian inference method GPhoCS is based on likelihood estimation and in order to allow adequate convergence of parameter estimates, a burn-in period of 100,000 iterations was used when applied to MS simulated data.

The effect of varying ancestral population size

In simulating the demographic scenario shown in figure 2A, populations 1 and 2 are constant backwards in time, but not (necessarily) equal in size; each population size is independently drawn from a uniform distribution between 170 and 1,700,000 diploid individuals. A total of 1000 of such demographies were generated. Populations 1 and 2 merge at 10,000 generations to form a single ancestral population of initial size $N_A = 17,000$ individuals for ϕ generations. The ancestral population then changes to λN_A, (drawn uniformly from $[1.7 \times 10^2, 1.7 \times 10^6]$), for τ generations, before returning to N_A. We investigate the impact of that change in ancestral population size (λN_A for τ generations) on TT estimates of population divergence time ($\hat{t}$) and ancestral population size, $\hat{N_A}$, for $\tau = 0, 100, 500, 1,000$ generations and $\phi = 0, 100, 500, 2,000$ generations (figure S1). Figure S1 shows that increasing ancestral population size for τ generations can have the effect of inflating divergence time estimates for the TT method. This behaviour is expected; if we imagine an expansion of the ancestral population to infinite size for τ generations, no coalescence events would occur during that time, and divergence time estimates would be upwardly biased by τ generations. This bias however seems to be relatively minor compared to that arising from severe bottlenecks. For instance, when the true $t = 10,000$, a bottleneck in the ancestral population of 1,500 individuals lasting for 100 generations results in $\hat{t} \approx 9,000$ generations. However, the same severity of bottleneck lasting for 500 generations will result in a greater underestimate of $\hat{t} \approx 5,000$ generations. Similarly, N_A is underestimated when severe bottlenecks occur, though these estimates seem to be more robust than estimates of divergence time (figure S2). When studying more recent splits, (1500 generations), we observe that severe bottlenecks have the potential to result in nonsensical negative split time estimates (figure S3).

Figure 3 shows a comparison between TT method and GPhoCS estimates of population divergence time (t) and ancestral population size (N_A) in cases where the duration of the bottleneck (τ) is fixed at 1,000 generations and true split time is 10,000 generations. Results suggest that both methods react similarly to violations of the assumption of a change in ancestral population size; each being particularly susceptible to bias when severe bottlenecks have occurred. GPhoCS performs somewhat better than the TT method, with severe bottlenecks resulting in less of an underestimate of population divergence time.

An interesting effect appears when a bottleneck of sufficient severity occurs, whereby both methods' $\hat{t}$ estimates begin to rebound towards the true split time of 10,000 generations. Again this behaviour is expected as all lineages will coalesce in a bottleneck of sufficient severity prior to a population divergence event. In this case the bottleneck itself will act as the constant ancestral population size, and as long as it occurs in close proximity to the split, divergence time estimates are not affected much. For the same reason, when a severe bottleneck occurs a long time prior to the split, both methods produce a (slight) overestimate of the true divergence time (figure 3E).

The effect of migration between branches

In simulations based on the demographic scenario shown in figure 2B, a pulse of admixture occurs δ generations ago, with proportion $0 \leq \gamma \leq 1$ of one daughter population made up of migrants from the other daughter population. Thus we examine the effect of increasing proportion of migration occurring at various times between present and the split time. All popula-

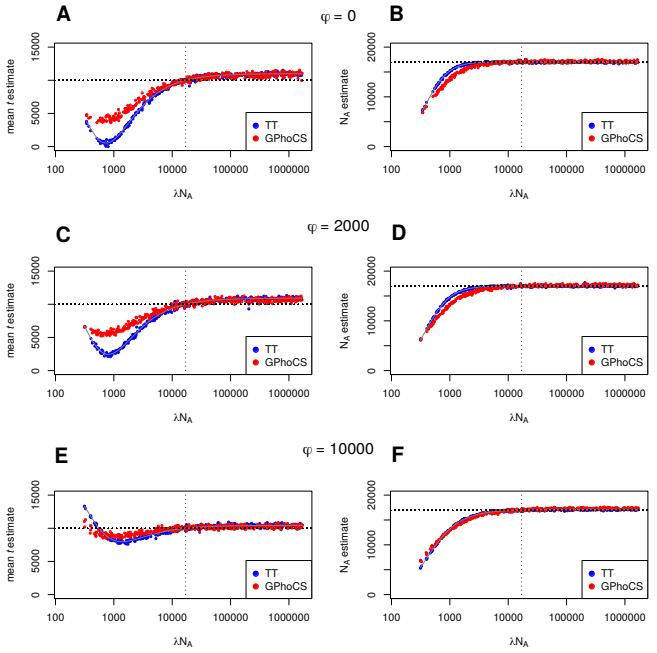

Figure 3 A comparison of the effect of ancestral population size changes on TT method and GPhoCS parameter estimates. The time between population divergence and change in ancestral population size (ϕ) is (A, B) 0, (C, D) 2,000, and (E, F) 10,000 generations. In all cases, the duration of change in ancestral population size (τ) is 1,000 generations and true split time is 10,000 generations.

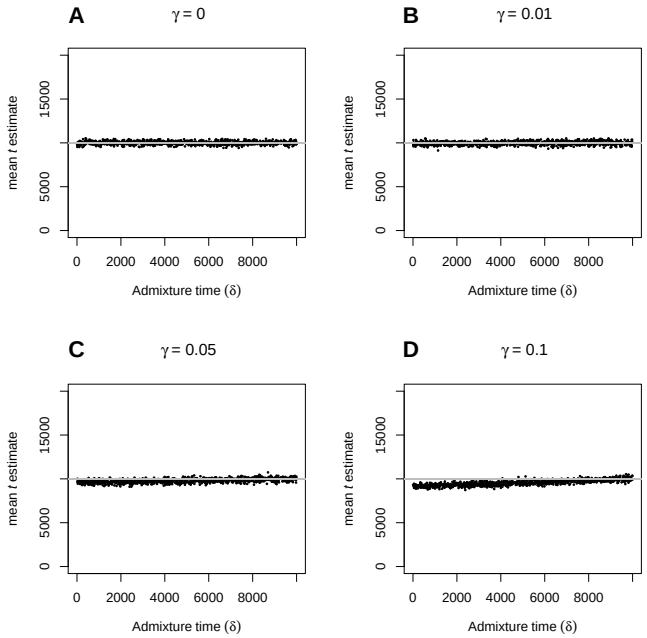

Figure 4 The effect of varying admixture time (δ) on TT split time estimates ($\hat{t}$), when proportion of admixture (γ) is (A) 0, (B) 0.01, (C) 0.05 and (D) 0.1, and true split time is 10,000 generations.

tions are kept fixed and constant at 17,000 diploid individuals ($N_1 = N_2 = N_A$). Figures 4 and 5 show the effect of increasing proportion of migrants on TT divergence time estimates when true split time is 10,000 generations. Divergence times are reliably estimated when the proportion of migrants is below 0.1, and as expected, even at higher proportions the bias decreases the nearer the admixture event is to split time. Note that under this set up, the TT method returns $\hat{t} \approx 3,000$ generations even when the proportion of migrants (γ) is 1 and admixture time (δ) is 0. We would expect in this case a $\hat{t} = 0$, but it seems that this scenario (where all populations are equal in size) is equivalent to a violation of the assumption of a constant ancestral population. As described previously, this has the effect of biasing $\hat{t}$ upwards, and shows that in cases of high proportion of recent admixture, differences between the size of the daughter populations and the ancestral population also has the potential to result in biased $\hat{t}$. Estimates of ancestral population size on the other hand, are only very slightly affected (figures S4 and S5). Very similar results are observed when a more recent true split time of 1500 generations is studied (figures S6 and S7).

While simulations have shown the TT method to be relatively robust to violations of its assumptions in general, it is evident that extensive, recent gene flow between daughter populations or strong, prolonged bottlenecks in the ancestral population have the potential to introduce bias. If however we obtain external estimates of α_1 and α_2 through the outgroup ascertainment procedure outlined above (TTo), we can obtain estimates of divergence time that are much less dependent on assumptions concerning the ancestral population. Figure 6 shows a comparison of TT and TTo method results in scenarios of increasing duration of ancestral population size change. The true values of

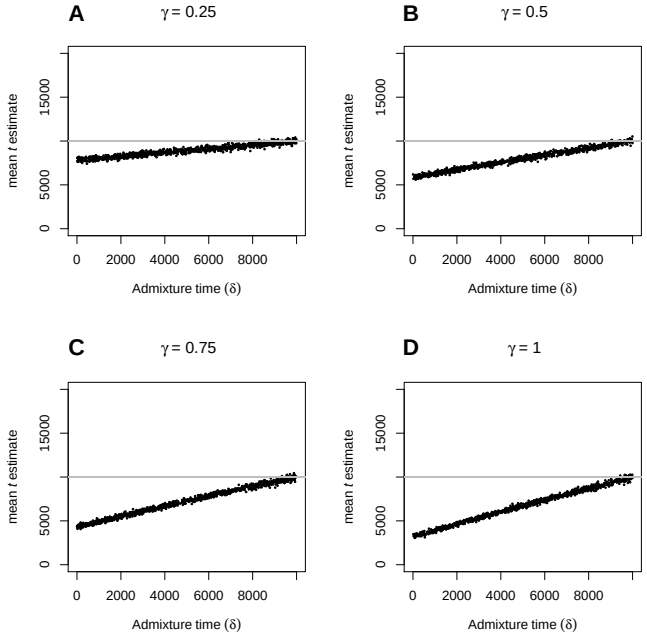

Figure 5 The effect of varying admixture time (δ) on TT split time estimates ($\hat{t}$), when proportion of admixture (γ) is (A) 0.25, (B) 0.5, (C) 0.75 and (D) 1, and true split time is 10,000 generations.

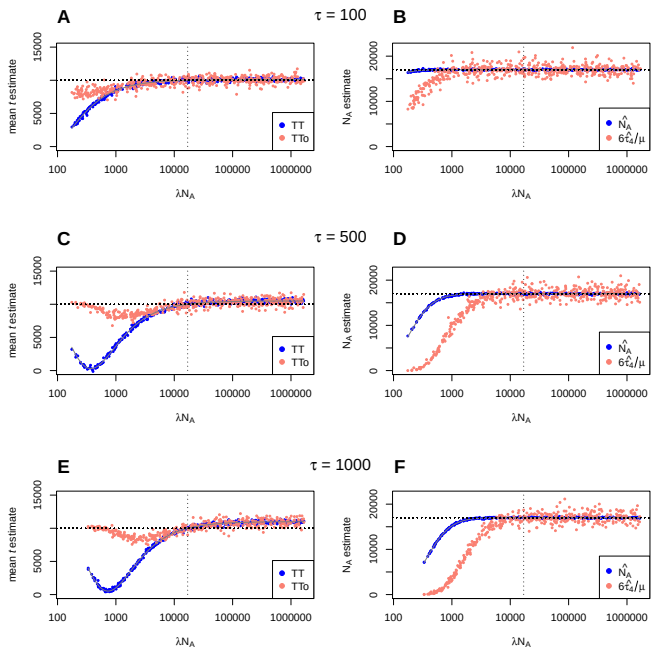

Figure 6 A comparison of TT method estimates of divergence time and ancestral population size with (TTo), and without (TT) using external estimates of drift. The duration of alternative ancestral population size (τ) is (A, B) 100, (C, D) 500, and (E, F) 1,000 generations. In all cases, the change in ancestral population size occurs immediately prior to the split (ϕ=0) and true split time is 10,000 generations.

α_1 and α_2 have been used in equations in 7 to obtain estimates of B_1, B_2, τ_2 and τ_3, that in turn have been used to approximate τ_4 and divergence times following equation 8. These results show that by using external estimates of drift, there is the potential to considerably reduce bias in divergence time estimates when severe bottlenecks have occurred in the ancestral population. Furthermore, figure 6 also compares estimates of N_A, from that of the TT method to one based on the TTo estimate of τ_4 ($6\widehat{\tau}_4/\mu$), which is found to be much more sensitive to ancestral bottlenecks.

Application to data

The TT method requires good quality sequence data, typically high-coverage genome sequence data since diploid genotype calls are utilized, including invariable sites and singletons (in the sample of four chromosomes). The formulas are sufficiently simple to allow for asymptotic confidence intervals based on MLE theory, and one can imagine thinning the genome data to make sites independent of each other (to overcome potential dependence via linkage). However, we chose a more conservative approach of estimating confidence; the weighted block jackknife procedure (Busing *et al.* 1999), which should be more robust to large-scale "outlier" regions driving the signal. We conducted pairwise comparisons among the 11 HGDP individuals and the Denisovan genome and the Altai Neandertal genome from (Meyer *et al.* 2012) and (Prüfer *et al.* 2014) and estimate population divergence times (see appendix for a description of data cleaning and calling of ancestral states). The 11 HGDP individuals include 5 individuals from Africa: one Khoe-San ('San'), one rainforest hunter-gatherer ('Mbuti'), two West-African ('Mandenka' and 'Yoruba') and one East-African ('Dinka'). The other individuals were two Europeans ('French' and 'Sardinian'), two East-Asians ('Han' and 'Dai'), one individual from Oceania ('Papuan') and one individual representing a South-American indigenous population ('Karitiana'). In addition, we used the high coverage ancient southern African hunter-gatherer genome ('Balito Bay A', (Schlebusch *et al.* 2017) as an outgroup for some divergence estimates (see below).

Split model parameter estimates

We refer to split time estimates *in years* in the TT-method as $\widehat{t_i}$ and under the TTo-method as $\widehat{t_i}^*$. These are obtained by setting $G = 30$ and $\mu = 1.25 \times 10^{-8}$ in equation (4) and applying this to the estimates of T_1 and T_2 in equations (3) and equation (8) respectively.

Comparisons are grouped according to the population split they represent. For instance, the comparison between French and San is referred to as the 'Khoe-San split'.

Divergence estimates according to the TT-method

Assuming a constant ancestral population size, N_A, it is possible to estimate N_A, α_1, α_2, ν_1, ν_2 as well as t_1, t_2 without relying on ascertainment procedures. Estimates of α, N_A and ν are shown in supplementary figures S9, S8 and S10 respectively. From figure S10 it is apparent that ν is often poorly estimated and the uncertainty of the estimate appears to be closely linked to the amount of branch-specific genetic drift (figure S11). A closer look at any of the formulas for $p_{i,j}$ reveals that the impact of ν on the probabilities disappears as α approaches 1 (no drift).

Estimates of the ancestral population size remain remarkably constant at around $N_A = 17,000$, regardless of choice of individuals (figure S8).

Values of $\widehat{t}$ are shown in figure 7. To summarize, estimates of the different split times are (in descending order)

- the split between Neanderthal and Denisovans $962 - 979$ kya;
- the split between archaic humans and modern humans $510 - 707$ kya;
- the deepest split among modern human population (between Khoe-San and other human populations) $233 - 266$ kya (see (Schlebusch *et al.* 2017) for the consequence of using the ancient southern African Balito Bay A genome);
- the split between Mbuti and other modern humans (excluding Khoe-San populations) $186 - 220$ kya
- the split between West-Africans and East-Africans $96 - 117$ kya;
- the split between East-Africans and non-African $66 - 82$ kya;
- splits between non-African < 0 ya.

Here, the range for the split between archaic and modern humans takes into account the fact that the archaic genomes are older than 40 ky. There are two obvious odd sets of estimates among these: the negative times for non-Africans, and the deep time between Denisovans and Neandertals contrasted to the younger time between Denisovans/Neandertals and modern humans (note that we assume a constant ancestral population size here). We discuss each of these split time estimates below, but first we revisit the utility of ascertaining variants in an outgroup.

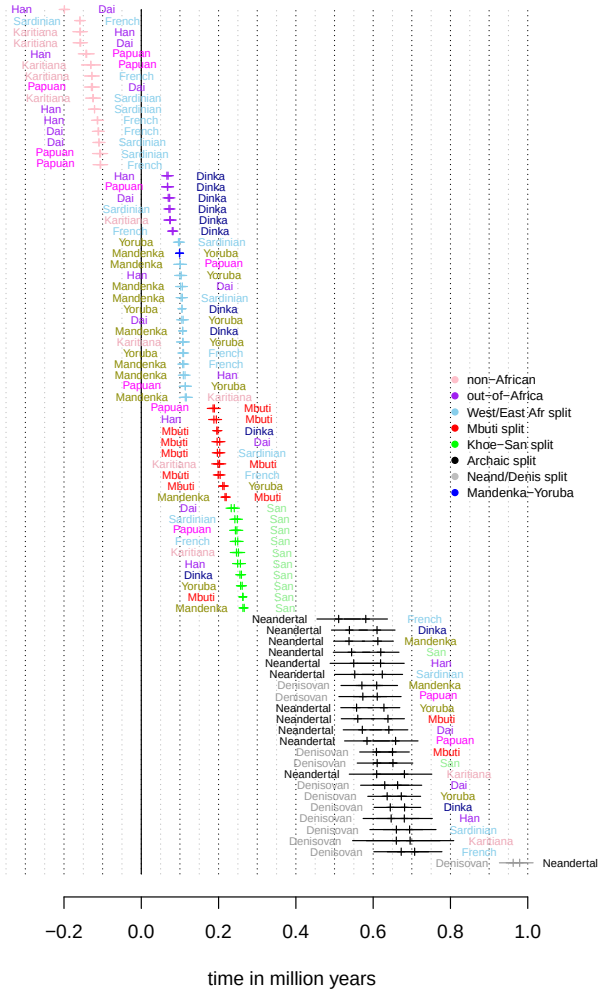

Figure 7 Split time estimates assuming a constant ancestral population and a mutation rate of 1.25×10^{-8} and a generation time of 30 years. Corresponding to the branch specific divergence time, there are two estimates each comparison.

Divergence estimates according to the TTo-method

By comparing two individuals using only those sites where the derived variant was present in an outgroup, it is possible to: i) test whether the outgroup represents a true outgroup, and ii) obtain estimates of α_1 and α_2 that do not rely on assumptions concerning the ancestral population. We utilized the Mbuti, Balito Bay A or Neandertal/Denisovan as outgroups. The estimates of α conditional on the derived variant being present in an outgroup are shown in figure S15. These three options were variably suitable as outgroups depending on the comparison being made. For instance, when comparing an individual from outside Africa to an African individual, Neandertal/Denisovan would not be true outgroups given the archaic admixture shared among non-African individuals (Green *et al.* 2010). This was also visible in the tests based on equations 5 and 6 above (as well as the D-test, see figure S12 in supplementary material). A likely consequence of the documented additional Denisovan ancestry in Papuan (Meyer *et al.* 2012) is that no comparison involving Papuan passed the outgroup tests. Perhaps more surprising, any comparison involving Mbuti failed the tests when Balito Bay A was used as the outgroup. Moreover, both Mbuti and Balito Bay A were expected to be true outgroups for the comparison of Neandertal versus Denisovan, but the test, however, pointed to them not being true outgroups.

Comparisons between estimates of θ (assuming a constant ancestral population size) to estimates of τ_2 and $3\tau_3$ using different outgroups for ascertainment are shown in figures S16, S18 and S17. Since there is presently no suitable outgroup for comparisons between a modern human and one of the two archaic humans – this would require a genome from an archaic human that split off before the Neandertal/Denisovan branch – it was not possible to estimate τ_2 and $3\tau_3$ for such comparisons.

When reliable outgroup ascertained estimates of α_1 and α_2 can be obtained, we estimate τ_2, τ_3, B_1 and B_2 using equations 7 that are used in equation (8) to obtain an estimate of T_i^*. This in turn gives $\widehat{t^*}$ that are shown in figures S16, S17, S18, S19, S20, S21. For the majority of comparisons, such an approach does not yield different estimates compared to assuming a constant ancestral population size. The major exceptions to this are those comparisons involving non-Africans, that shows positive and realistic divergence time estimates using the ascertainment scheme (Figure 8).

Divergence times outside Africa

The divergence time estimates for non-African populations under a constant model ($\widehat{t_i}$) are nonsensical, (negative values). This is likely a consequence of the severe out-of-Africa bottleneck that leads to $\tau_4 = \mu E[T_{44}]$ being much smaller than $\tau_2/6$, which then violates the assumption of a constant N_A ($E[T_{nk}] = 2N_A/k(k-1)$ in a constant population with N_A chromosomes). Estimates based on the three outgroup ascertainment schemes ($\widehat{t_i^*}$) give more reasonable values as shown in figure 8.

Here, the split times estimates are:

- 50 to 75 kya between Europeans and Asians/Americans
- $\sim$ 40 kya between Sardinians and French
- 25 to 30 kya between Dai and Karitiana
- 25 to 30 kya between Dai and Han
- $<$ 25 kya between Han and Karitiana

These estimates are generally consistent with the prevailing

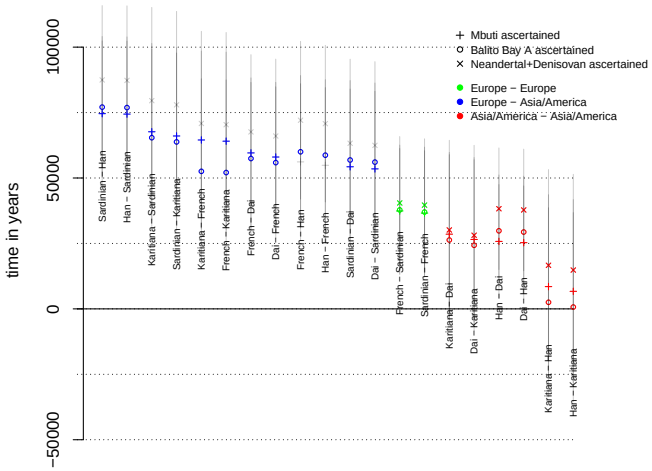

Figure 8 Different estimates of split times using outgroup ascertainment assuming a mutation rate of 1.25×10^{-8} and a generation time of 30 years. Comparisons with Papuans not included as no such comparison passed the outgroup-tests. Three estimates are shown: estimates where outgroup ascertainment is performed in Mbuti (+), in Balito Bay A (○) and in Neandertal/Denisovan (×). Transparent grey represents standard deviation and for comparisons that failed the outgroup tests.

view of the demographic history outside Africa. For instance, the deep split between Sardinians and French may reflect previous findings that while Sardinians trace their ancestry mostly to the early Neolithic farmers, the French are more admixed with European hunter-gatherers and components of the Yamnaya expansion (Skoglund *et al.* 2014; Günther *et al.* 2015; Allentoft *et al.* 2015; Haak *et al.* 2015). To interpret the split times between Han, Dai and Karitiana, it should be noted that Karitiana is best modelled as a combination of three source populations (an ancient Siberian Eurasian source, a north East Asian source and an Australasian source), where the north East Asian contribution is substantially greater than the other two sources combined (Skoglund *et al.* 2015; Raghavan *et al.* 2015). The fact that the Karitiana show a more recent divergence with Han than with Dai likely reflects north East Asians contributing substantially to Native Americans, and that the Dai has a south East Asian component (closer to an Australasians, Skoglund *et al.* 2015; Raghavan *et al.* 2015). This admixture pattern results in shallower divergence between Karitiana and Han, and deeper divergence between Karitiana and Dai and between Han and Dai (see *e.g.* figure 2 in Raghavan *et al.* 2015).

Western vs Eastern Africa and timing of the out-of-Africa event

Assuming a constant ancestral population size, we estimate the split between non-Africans and East-Africans (Dinka) to between 66 and 82 kya. The split between Mandenka and Yoruba is estimated to 100 kya while the split between Western Africans and Eastern Africans (Dinka and non-Africans) is estimated to between 96 and 117 kya.

The estimates based on the three outgroup ascertainment schemes ($\widehat{t_i}^*$) are generally older. Although the demographic history of Western and Eastern Africa appears to be particularly complex (Pickrell *et al.* 2014; Gurdasani *et al.* 2015; Triska *et al.* 2015; Busby *et al.* 2016; Hollfelder *et al.* 2017), and the standard deviation estimates suggest one should not over-interpret the $\widehat{t_i}^*$ values, there are a few interesting tendencies among these estimates (figure 9). First, estimated split times are consistently lower among Yoruba, Mandenka and Dinka than between any of these populations and a non-African population; likely an effect of gene flow among the three African populations (Gurdasani *et al.* 2015; Busby *et al.* 2016; Schlebusch and Jakobsson 2018). Second, estimates between Yoruba and Dinka (or non-Africans) are deeper than split time estimates between Mandenka and Dinka (or non-Africans). This is consistent with some observations suggesting that Mandenka have a greater east African/European ancestry component compared to Yoruba (Gurdasani *et al.* 2015; Patin *et al.* 2017; Schlebusch and Jakobsson 2018). Although Mandenka is more distant geographically from Dinka than Yoruba, there is evidence that historical trading routes along the Sahel belt may have resulted in more gene-flow between East Africans and Mandenka (than with Yoruba) (Triska *et al.* 2015; Černý *et al.* 2018). Third, there is a tendency for split estimates between East Asians (Dai or Han) and West Africans (Yoruba, Mandenka or Dinka) to be deeper than split estimates between Europeans (French or Sardinian) and West Africans. This observation, combined with gene-flow between east and west Africa, is consistent with previous suggestions of migration into East Africa from a European or Middle Eastern source (Llorente *et al.* 2015).

Deepest splits among modern human populations

The split between Khoe-San and other modern human populations are estimated to around 250 kya using the TT-method

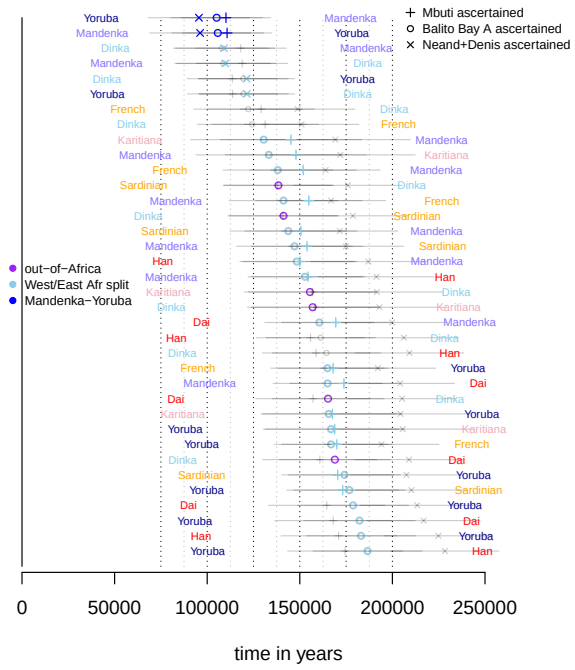

Figure 9 Different estimates of split times using outgroup ascertainment assuming a mutation rate of 1.25×10^{-8} and a generation time of 30 years. Three estimates are shown: estimates where outgroup ascertainment is performed in Mbuti (+), in Balito Bay A (○) and in Neandertal/Denisovan (×). Transparent grey represents standard deviation and for comparisons that failed the outgroup tests.

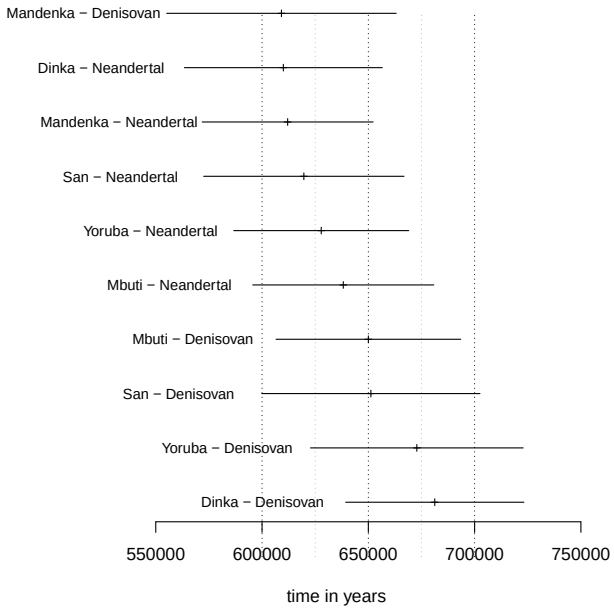

Figure 10 Split time estimates between the five African individuals and the two archaic humans assuming a constant ancestral population. Only estimates based on the non-ancient branch are shown. A mutation rate of 1.25×10^{-8} and a generation time of 30 years is assumed.

(Schlebusch *et al.* 2017). It was further demonstrated that all modern-day Khoe-San groups and individuals, including the HGDP San individuals investigated here, were affected by Eurasian/east African admixture, which in turn impacts estimates of the deepest divergence of modern humans (different methods are differently sensitive to admixture), (Schlebusch *et al.* 2017). This observation became evident in comparisons with an ancient southern African individual (the Balito Bay A boy who lived some 2,000 years ago), closely related to modern-day Khoe-San individuals, but without the Eurasian/east African admixture that post-dated the life-time of the Balito Bay A boy. Population divergence time estimates based on the ancient Balito Bay A boy predates the estimates based on modern-day Khoe-San individuals, and give an estimate that is unaffected by the migration and admixture in the last 2000 years (Schlebusch *et al.* 2017).

Above we showed the effect of assuming a constant ancestral population size, and how violations of this assumption by a bottleneck in the ancestral population can bias divergence time estimates. However, we find no evidence for such a bottleneck in the common ancestral population to all modern humans, and hence

In general, we find very little difference between the TT and the TTo estimates (figure S19). In fact, there is a tendency for $\widehat{t}_i^*$ to be lower than $\widehat{t}_i$ for the Mbuti-split, providing additional support that the Khoe-San split is the deepest split among modern human populations (Schlebusch *et al.* 2012).

Archaic split times

The split between modern humans and both Neandertal and Denisovan is estimated to between 510 amd 707 kya, which is in line with previous such estimates (e.g., Prüfer *et al.* 2014). In fact, restricting the analysis to split times on the non-ancient branch (to alleviate issues with fossil dating and potential excess ancient DNA damage) and only to comparisons between Africans and the two archaic humans, gives a range of estimates from 609 to 681 kya (figure 10). Unfortunately, there is (presently) no suitable outgroup for comparisons between modern humans and archaic humans in order to utilize the outgroup ascertainment approach.

The estimated split between Neandertal and Denisovan is around 970 kya; more than 250 ky older than the split between modern humans and archaic humans. This is likely an artifact of violating the model assumptions; the existence of a more complex demography is indicated by our finding that neither Mbuti nor Balito Bay A were found to be true outgroups to the Neandertal-Denisovan comparison according to our outgroup test and the D-test (figures S14 and S13). Some studies hypothesize that the demographic relationship between Neandertals and Denisovans was governed by meta-population dynamics (Rogers *et al.* 2017). Others suggest complicating demographic factors such as admixture between Denisovans and *Homo erectus* (Prüfer *et al.* 2014) or admixture from the modern human branch into the Altai Neandertal (Kuhlwilm *et al.* 2016).

Conclusion

We present a simple approach to estimate parameters under a comparatively general split model. In particular, no assumptions are needed concerning the population size processes/changes in the daughter populations (i.e., more recent than the split). The underlying model does not include gene-flow between daughter population, however, we can show that moderate violation of this assumption has little impact on the population divergence-time estimates. Assuming a constant ancestral population size, this approach provides an unbiased estimate of divergence time. However, when the ancestral population is not constant, and particularly in the case of severe bottlenecks, divergence time estimates can be biased. Indeed, simulations comparing the TT-method to GPhoCS – an alternative, fundamentally different approach to demographic inference, has shown that the two methods are sensitive to violations of the same assumptions. The reason for this can be intuitively understood in terms of the tMRCA in the ancestral population; most of this time is spent with two lineages and the duration of this is utilized by both methods to estimate the size of the ancestral population. Since, by assumption, the ancestral size is constant, if the time to the first coalescent event in the ancestral population is shorter than expected, (for instance, due to a bottleneck shortly before the divergence), then both methods underestimate the true population divergence time. When such severe bottlenecks have occurred, we have shown that it is possible to reduce much of this bias through the outgroup ascertainment procedure implemented in the TTo-method.

Applying the TT-method to a sample of 11 genomes from the HGDP panel together with the Neandertal and Denisovan genomes, we provide further information on the details of the various splits within the sample and corroborate many previously estimated population divergence times.

Finally, accumulating evidence suggests that human evolution is highly reticulated, and perhaps not well approximated by the sort of bifurcating tree-models studied here (Schlebusch and Jakobsson 2018; Henn *et al.* 2018; Scerri *et al.* 2018; Stringer 2016). Nonetheless, the framework presented here is still a useful tool: different population genetic methods vary in their assumptions and sensitivities to model violations, thus it is important when investigating the complex demographies underlying the evolution of humans to have access to a variety of different methods. The TT-method is relatively robust to model violations and provides a simple and transparent analytic framework that can be compared to, and potentially even integrated with, other, more computationally demanding methods (Beichman *et al.* 2017; Terhorst *et al.* 2017; Wang *et al.* 2020).

Data availability

Genome sequence data from the following publications was extracted and reprocessed in order to avoid mapping and filtering biases: Prüfer *et al.* (2014); Schlebusch *et al.* (2017). Scripts used in simulations and plotting of results, together with open source code for running the TT method is freely available in Python at github.com/jammc313/TT-method.

Literature Cited

Allentoft, M. E., M. Sikora, K.-G. Sjögren, S. Rasmussen, M. Rasmussen, *et al.*, 2015 Population genomics of bronze age eurasia. Nature **522**: 167.

Beaumont, M., A. Zhang, and W. Balding, 2002 Approximate Bayesian computation in population genetics. Genetics **162**: 2025–2035.

Beichman, A. C., T. N. Phung, and K. E. Lohmueller, 2017 Comparison of single genome and allele frequency data reveals discordant demographic histories. G3: Genes, Genomes, Genetics **7**: 3605–3620.

Busby, G. B., G. Band, Q. Si Le, M. Jallow, E. Bougama, *et al.*, 2016 Admixture into and within sub-saharan africa. eLife **5**: e15266.

Busing, F. M. T. A., E. Meijer, and R. V. D. Leeden, 1999 Delete-m jackknife for unequal m. Statistics and Computing **9**: 3–8.

Chen, H., 2012 The joint allele frequency spectrum of multiple populations: A coalescent theory approach. Theoretical Population Biology **81**: 179 – 195.

Cornuet, J., M. Pudlo, P. Veyssier, J. Dehne-Garcia, A. Gautier, *et al.*, 2014 DIYABC v2.0: a software to make approximate Bayesian computation inferences about population history using single nucleotide polymorphism, DNA sequence and microsatellite data. Bioinformatics **30**: 1187–1189, doi:10.1093/bioinformatics/btt763.

Doob, J. L., 1934 Probability and statistics. Transactions of the American Mathematical Society **36**: 759–775.

Excoffier, L., I. Dupanloup, E. Huerta-Sánchez, V. C. Sousa, and M. Foll, 2013 Robust demographic inference from genomic and SNP data. PLoS Genet. **9**: e1003905.

Gattepaille, L., T. Günther, and M. Jakobsson, 2016 Inferring past effective population size from distributions of coalescent times. Genetics **204**: 1191–1206.

Green, R. E., J. Krause, A. W. Briggs, T. Maricic, U. Stenzel, *et al.*, 2010 A Draft Sequence of the Neandertal Genome. Science **328**: 710–722.

Griffiths, R. and S. Tavaré, 1998 The age of a mutation in a general coalescent tree. Communications in Statistics. Stochastic Models **14**: 273–295.

Gronau, I., M. Hubisz, J. Gulko, B. Danko, and C. Siepel, 2011 Bayesian inference of ancient human demography from individual genome sequences. Nature Genetics **43**: 1031–1035, doi:10.1038/ng.937.

Günther, T., C. Valdiosera, H. Malmström, I. Ureña, R. Rodriguez-Varela, *et al.*, 2015 Ancient genomes link early farmers from atapuerca in spain to modern-day basques. Proceedings of the National Academy of Sciences **112**: 11917–11922.

Gurdasani, D., T. Carstensen, F. Tekola-Ayele, L. Pagani, I. Tachmazidou, *et al.*, 2015 The african genome variation project shapes medical genetics in africa. Nature **517**: 327–332.

Gutenkunst, R. N., R. D. Hernandez, S. H. Williamson, and C. D. Bustamante, 2009 Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. PLoS Genet **5**: e1000695.

Haak, W., I. Lazaridis, N. Patterson, N. Rohland, S. Mallick, *et al.*, 2015 Massive migration from the steppe was a source for indo-european languages in europe. Nature **522**: 207.

Henn, B. M., T. E. Steele, and T. D. Weaver, 2018 Clarifying distinct models of modern human origins in africa. Current Opinion in Genetics & Development **53**: 148 – 156, Genetics of Human Origins.

Hollfelder, N., C. M. Schlebusch, T. Günther, H. Babiker, H. Y. Hassan, *et al.*, 2017 Northeast african genomic variation shaped by the continuity of indigenous groups and eurasian migrations. PLOS Genetics **13**: 1–17.

Hudson, R., 2002 Generating samples under a Wright–Fisher neutral model of genetic variation. Bioinformatics **18**: 337–338, doi:10.1093/bioinformatics/18.2.337.

Kelleher, J., Y. Wong, A. W. Wohns, C. Fadil, P. K. Albers, *et al.*, 2019 Inferring whole-genome histories in large population datasets. Nature Genetics **51**: 1330–1338, doi.org/10.1038/s41588-019-0483-y.

Kuhlwilm, M., I. Gronau, M. J. Hubisz, C. de Filippo, J. Prado-Martinez, *et al.*, 2016 Ancient gene flow from early modern humans into eastern neanderthals. Nature **530**: 429–433.

Li, H. and R. Durbin, 2011 Inference of human population history from individual whole-genome sequences. Nature **475**: 493–496, doi:10.1038/nature10231.

Llorente, M. G., E. R. Jones, A. Eriksson, V. Siska, K. W. Arthur, *et al.*, 2015 Ancient ethiopian genome reveals extensive eurasian admixture in eastern africa. Science **350**: 820–822.

Lohse, K., M. Chmelik, S. H. Martin, and N. H. Barton, 2016 Efficient Strategies for Calculating Blockwise Likelihoods Under the Coalescent. Genetics **202**: 775+.

Lohse, K., R. J. Harrison, and N. H. Barton, 2011 A General Method for Calculating Likelihoods Under the Coalescent Process. Genetics **189**: 977–U398.

Mazet, O., W. Rodriguez, S. Grusea, S. Boitard, and L. Chikhi, 2016 On the importance of being structured: instantaneous coalescence rates and human evolution–lessons for ancestral population size inference? Heredity **116**: 362–371, doi:10.1038/hdy.2015.104.

Meyer, M., M. Kircher, M.-T. Gansauge, H. Li, F. Racimo, *et al.*, 2012 A high-coverage genome sequence from an archaic denisovan individual. Science **338**: 222–226.

Moorjani, P., Z. Gao, and M. Przeworski, 2016 Human Germline Mutation and the Erratic Evolutionary Clock. PLoS Biol **14**: e2000744, doi:10.1371/journal.pbio.2000744.

Orozco, P., 2016 The devil is in the details: the effect of population structure on demographic inference. Heredity **116**: 349–350, doi:10.1038/hdy.2016.9.

Patin, E., M. Lopez, R. Grollemund, P. Verdu, C. Harmant, *et al.*, 2017 Dispersals and genetic adaptation of bantu-speaking populations in africa and north america. Science **356**: 543–546.

Pickrell, J. K., N. Patterson, C. Barbieri, F. Berthold, L. Gerlach, *et al.*, 2012 The genetic prehistory of southern Africa. Nature Communications **3**: 1143.

Pickrell, J. K., N. Patterson, P.-R. Loh, M. Lipson, B. Berger, *et al.*, 2014 Ancient west eurasian ancestry in southern and eastern africa. Proceedings of the National Academy of Sciences **111**: 2632–2637.

Prüfer, K., F. Racimo, N. Patterson, F. Jay, S. Sankararaman, *et al.*, 2014 The complete genome sequence of a neanderthal from the altai mountains. Nature **505**: 43–49.

Pudlo, P., J. Marin, A. Estoup, J. Cornuet, M. Gautier, *et al.*, 2016 Reliable abc model choice via random forests. Bioinformatics **32**: 859–866.

Raghavan, M., M. Steinrücken, K. Harris, S. Schiffels, S. Rasmussen, *et al.*, 2015 Genomic evidence for the pleistocene and recent population history of native americans. Science **349**.

Rogers, A. and R. Bohlender, 2014 Bias in estimators of archaic admixture. Theoretical Population Biology **100**: 63–78, doi.org/10.1016/j.tpb.2014.12.006.

Rogers, A. R., R. J. Bohlender, and C. D. Huff, 2017 Early history of neanderthals and denisovans. Proceedings of the National Academy of Sciences **114**: 9859–9863.

Scally, A. and R. Durbin, 2012 Revising the human mutation rate: implications for understanding human evolution. Nature Reviews Genetics **13**: 745–753.

Scerri, E. M., M. G. Thomas, A. Manica, P. Gunz, J. T. Stock, *et al.*, 2018 Did our species evolve in subdivided populations across africa, and why does it matter? Trends in Ecology & Evolution **33**: 582 – 594.

Schiffels, S. and R. Durbin, 2014 Inferring human population size and separation history from multiple genome sequences. Nature Genetics **46**: 919–925, doi:10.1038/ng.3015.

Schlebusch, C. M. and M. Jakobsson, 2018 Tales of human migration, admixture, and selection in africa. Annual Review of Genomics and Human Genetics **19**: null, PMID: 29727585.

Schlebusch, C. M., H. Malmström, T. Günther, P. Sjödin, A. Coutinho, *et al.*, 2017 Southern african ancient genomes estimate modern human divergence to 350,000 to 260,000 years ago. Science **358**: 652–655.

Schlebusch, C. M., P. Skoglund, P. Sjödin, L. M. Gattepaille, D. Hernandez, *et al.*, 2012 Genomic variation in seven khoesan groups reveals adaptation and complex african history. Science **338**: 374–379.

Schraiber, J. G. and J. M. Akey, 2015 Methods and models for unravelling human evolutionary history. Nature Reviews Genetics **16**: 727–740.

Skoglund, P., A. Götherström, and M. Jakobsson, 2011 Estimation of population divergence times from non-overlapping genomic sequences: Examples from dogs and wolves. Molecular Biology and Evolution **28**: 1505–1517.

Skoglund, P., S. Mallick, M. C. Bortolini, N. Chennagiri, T. Hünemeier, *et al.*, 2015 Genetic evidence for two founding populations of the americas. Nature **525**.

Skoglund, P., P. Sjödin, T. Skoglund, M. Lascoux, and M. Jakobsson, 2014 Investigating population history using temporal genetic differentiation. Mol Biol Evol **31**: 2516–2527, doi:10.1093/molbev/msu192.

Slatkin, M., 1996 Gene genealogies within mutant allelic classes. Genetics **143**: 579–587.

Speidel, L., M. Forest, S. Shi, and S. Myers, 2019 A method for genome-wide genealogy estimation for thousands of samples. Nature Genetics **51**: 1321–1329.

Stringer, C., 2016 The origin and evolution of *Homo sapiens*. Philosophical Transactions of the Royal Society B: Biological Sciences **371**: 20150237.

Tavaré, S., D. J. Balding, R. C. Griffiths, and P. Donnelly, 1997 Inferring coalescence times from DNA sequence data. Genetics **145**: 505–518.

Terhorst, J., J. A. Kamm, and Y. S. Song, 2017 Robust and scalable inference of population history from hundreds of unphased whole-genomes. Nature genetics **49**: 303–309.

Theunert, C. and M. Slatkin, 2018 Estimation of population divergence times from snp data and a test for treeness. bioRxiv .

Triska, P., P. Soares, E. Patin, V. Fernandes, V. Cerny, *et al.*, 2015 Extensive Admixture and Selective Pressure Across the Sahel Belt. Genome Biology and Evolution **7**: 3484–3495.

Černý, V., I. Kulichová, E. S. Poloni, J. M. Nunes, L. Pereira, *et al.*, 2018 Genetic history of the african sahelian populations. HLA **91**: 153–166.

Wakeley, J., 2009 *Coalescent Theory: An Introduction*. Roberts & Company Publishers, Greenswood Village, Colorado, first edition.

Wakeley, J. and J. Hey, 1997 Estimating ancestral population parameters. Genetics **145**: 847–855.

Wald, A., 1949 Note on the consistency of the maximum likelihood estimate. The Annals of Mathematical Statistics **20**: 595–601.

Wang, K., I. Mathieson, J. O'Connell, and S. Schiffels, 2020 Tracking human population structure through time from whole genome sequences. PLOS Genetics **16**: 1–24.

Wilkinson-Herbots, H. M., 2008 The distribution of the coalescence time and the number of pairwise nucleotide differences in the "isolation with migration" model. Theoretical Population Biology **73**: 277–288.

Appendix

Ascertained data

Conditional on the derived variant being present in a true outgroup (in a population that has branched off before the investigated split), there are no mutations within the branches so that μt_1, μt_2, $\mu \nu_1$ and $\mu \nu_2$ can all be set to 0 in the equations 1 above. Thus we are left with

$$\begin{aligned}
p^*_{1,0} =& \frac{1}{2}\alpha_1 b^*_1 + \frac{1}{3}\alpha_1(1-\alpha_2)b^*_2 \\
p^*_{0,1} =& \frac{1}{2}\alpha_2 b^*_1 + \frac{1}{3}(1-\alpha_1)\alpha_2 b^*_2 \\
p^*_{2,0} =& \frac{1}{4}(1-\alpha_1)b^*_1 + \frac{1}{6}(2-\alpha_1-\alpha_2+\alpha_1\alpha_2)b^*_2 + \frac{1}{4}(1-\alpha_2)b^*_3 \\
p^*_{0,2} =& \frac{1}{4}(1-\alpha_2)b^*_1 + \frac{1}{6}(2-\alpha_1-\alpha_2+\alpha_1\alpha_2)b^*_2 + \frac{1}{4}(1-\alpha_1)b^*_3 \\
p^*_{1,1} =& \frac{2}{3}\alpha_1\alpha_2 b^*_2 \\
p^*_{2,1} =& \frac{1}{3}(1-\alpha_1)\alpha_2 b^*_2 + \frac{1}{2}\alpha_2 b^*_3 \\
p^*_{1,2} =& \frac{1}{3}\alpha_1(1-\alpha_2)b^*_2 + \frac{1}{2}\alpha_1 b^*_3.
\end{aligned}$$

Under these equations,

$$\begin{aligned}
\frac{2p^*_{1,0}+p^*_{1,1}}{2p^*_{0,1}+p^*_{1,1}} =& \frac{2p^*_{1,2}+p^*_{1,1}}{2p^*_{2,1}+p^*_{1,1}} \\
p^*_{1,0}+2p^*_{2,0}+p^*_{2,1} =& p^*_{0,1}+2p^*_{0,2}+p^*_{1,2}
\end{aligned}$$

leading to the test statistics Y_1 and Y_2 in equations 5 and 6 above.

The second outgroup test (Y_2) is similar to the D-test since the probability to draw the ancestral (single) allele from population 1 and the derived allele from population 2 is

$$p^*_{0,2} + \frac{1}{2}(p^*_{0,1}+p^*_{1,2}) + \frac{1}{4}p^*_{1,1}$$

and the probability to draw the derived allele from population 1 and the ancestral allele from population 2 is

$$p^*_{2,0} + \frac{1}{2}(p^*_{1,0}+p^*_{2,1}) + \frac{1}{4}p^*_{1,1}.$$

Similarly, for sites where the derived variant is observed in a sample of size 1 in the outgroup (note that our conditioning is different)

$$\begin{aligned}
D =& \frac{\left(m^{**}_{0,2} + \frac{1}{2}\left(m^{**}_{0,1}+m^{**}_{1,2}\right) + \frac{1}{4}m^{**}_{1,1}\right) - \left(m^{**}_{2,0} + \frac{1}{2}\left(m^{**}_{1,0}+m^{**}_{2,1}\right) + \frac{1}{4}m^{**}_{1,1}\right)}{\left(m^{**}_{0,2} + \frac{1}{2}\left(m^{**}_{0,1}+m^{**}_{1,2}\right) + \frac{1}{4}m^{**}_{1,1}\right) + \left(m^{**}_{2,0} + \frac{1}{2}\left(m^{**}_{1,0}+m^{**}_{2,1}\right) + \frac{1}{4}m^{**}_{1,1}\right)} \\
=& 2\frac{(m^{**}_{0,1}-m^{**}_{1,0}) + 2(m^{**}_{0,2}-m^{**}_{2,0}) + (m^{**}_{1,2}-m^{**}_{2,1})}{m^{**}_{1,1} + 2(m^{**}_{0,1}+m^{**}_{1,0}) + 4(m^{**}_{0,2}+m^{**}_{2,0}) + 2(m^{**}_{1,2}+m^{**}_{2,1})}
\end{aligned}$$

where '**' indicates sites where the derived variant is observed in a sample of size 1 in the outgroup (our conditioning is different). The nominator is very similar to the nominator in the second outgroup test (Y_2 in equation 6 above).

Assuming an ancestral population with no structure backwards in time

Define

$$T_{ki}$$

to be the number of generations a coalescent process that starts with k lineages at the (most recent) base of the ancestral population spends with i lineages (so that $T_{mrca} = \sum_{i=k}^{2} T_{ki}$). The probability that there are k derived variants in a sample of size n given that a mutation occurred when there were i lineages is (Slatkin 1996).

$$P(A_n = k | \text{mutation during } T_{n,i}) = \frac{\binom{n-k-1}{i-2}}{\binom{n-1}{i-1}}$$

implying that

$$\begin{aligned} P(A_n = k) &= \sum_{i=n}^{2} P(A_n = k|\text{mutation during } T_{n,i})P(\text{mutation during } T_{n,i}) \\ &= \sum_{i=n}^{2} \frac{\binom{n-k-1}{i-2}}{\binom{n-1}{i-1}} \mu i E[T_{n,i}] \end{aligned}$$

We get

$$\begin{aligned} b_1 = P(A_4 = 1) &= \frac{2}{3}\mu E[T_{42}] + 2\mu E[T_{43}] + 4\mu E[T_{44}] \\ b_2 = P(A_4 = 2) &= \frac{2}{3}\mu E[T_{42}] + \mu E[T_{43}] \\ b_3 = P(A_4 = 3) &= \frac{2}{3}\mu E[T_{42}] \end{aligned}$$

Picking two gene copies from population 1 and one gene copy from population 2

Here, instead of two gene copies from both populations, the method assumes two sampled gene copies from population 1 and one sampled gene copy from population 2. The possible sample configurations are then:

Table 3 Number of derived in the two samples.

	0 in pop2	1 in pop2
0 in pop1	$O_{0,0}$	$O_{0,1}$
1 in pop1	$O_{1,0}$	$O_{1,1}$
2 in pop1	$O_{2,0}$	$O_{2,1}$

The observed number of sites with sample configuration $O_{i,j}$ will be denoted by $m_{i,j}$ and the total number of sites by m_{tot}. Assume independence between sites, an infinite sites model and a split model with no migration where the two daughter populations merge into a panmictic ancestral population and define the event

$$H: \text{ a coalescent event in population 1 before } t_1 \ (P(H) = 1 - \alpha)$$

Also define A_k to be the number of derived variants in a (hypothetical) sample of size k drawn at the split time in the ancestral population. Writing $a_{ki} = P(A_k = i)$, the conditional probabilities for sample configurations $O_{1,0}, \cdots, O_{1,1}$ are as in table 4:

Table 4 Conditional probabilities.

	H	$\neg H$
$O_{1,0}$	$2\mu\nu$	$\frac{2}{3}a_{31} + 2\mu t_1$
$O_{0,1}$	$\frac{1}{2}a_{21} + \mu t_2$	$\frac{1}{3}a_{31} + \mu t_2$
$O_{2,0}$	$\frac{1}{2}a_{21} + \mu(t_1 - \nu)$	$\frac{1}{3}a_{32}$
$O_{1,1}$	0	$\frac{2}{3}a_{32}$

Since

$$a_{21} = P(A_2 = 1) = \frac{2}{3}a_{31} + \frac{2}{3}a_{32}$$

we write $b_i = a_{3i} = P(A_3 = i)$ to get

$$\begin{aligned} p_{1,0} =& 2(1-\alpha)\mu\nu + 2\alpha\left(\mu t_1 + \frac{1}{3}b_1\right) \\ p_{0,1} =& \frac{1}{3}(1-\alpha)b_2 + \mu t_2 + \frac{1}{3}b_1 \\ p_{2,0} =& (1-\alpha)\left(\mu t_1 + \frac{1}{3}b_1\right) - (1-\alpha)\mu\nu + \frac{1}{3}b_2 \\ p_{1,1} =& \frac{2}{3}\alpha b_2 \\ p_{0,0} + p_{2,1} =& 1 - \sum_{0<i+j<3} p_{i,j} \end{aligned} \tag{9}$$

where $p_{i,j} = P(O_{i,j})$.

Similar to the case when two gene copies are picked from each sub-population: 1) it is not possible to completely separate b_1 from divergence times due to the co-occurrence of b_1 with μt_1 and μt_2, 2) disregarding b_1, this is an underdetermined set of equations with 5 parameters but only 4 equations/degrees of freedom ($p_{0,0} + p_{2,1} = 1 - \sum_{0<i+j<4} p_{i,j}$). Setting $t_1 = t_2$ does not help since

$$p_{0,1} - p_{2,0} + \frac{1}{2}(p_{1,1} - p_{1,0}) = \mu(t_2 - t_1)$$

and thus reduces the number of independent equations. Assuming the model in figure 1b, does not reduce the number of parameters and does not help in this case.
If ascertainment is done in an outgroup

$$\begin{aligned} p^*_{1,0} =& \frac{2}{3}\alpha b^*_1 \\ p^*_{0,1} =& \frac{1}{3}(1-\alpha)b^*_2 + \frac{1}{3}b^*_1 \\ p^*_{2,0} =& \frac{1}{3}(1-\alpha)b^*_1 + \frac{1}{3}b^*_2 \\ p^*_{1,1} =& \frac{2}{3}\alpha b^*_2 \end{aligned}$$

where '*' indicates that these are the corresponding conditional parameters and probabilities. This can be solved to yield two estimates of α

$$\begin{aligned} \widehat{\alpha}^* =& \frac{m^*_{1,0} + m^*_{1,1}}{2m^*_{0,1} + m^*_{1,1}} \\ \widehat{\alpha}^* =& \frac{m^*_{1,0} + m^*_{1,1}}{2m^*_{2,0} + m^*_{1,0}} \end{aligned}$$

and these two estimates can be compared to create the tests

$$\begin{aligned} \frac{m^*_{1,0} + m^*_{1,1}}{2m^*_{0,1} + m^*_{1,1}} - \frac{m^*_{1,0} + m^*_{1,1}}{2m^*_{2,0} + m^*_{1,0}} &= 0 \\ 2m^*_{0,1} - 2m^*_{2,0} + m^*_{1,1} - m^*_{1,0} &= 0 \end{aligned}$$

of that ascertainment was performed in a true outgroup.

Assuming the model in figure 1a, then $b_2 = \mu E[T_{32}]$ and $b_1 = 3\mu E[T_{33}] + \mu E[T_{32}]$ and

$$\begin{aligned} p_{1,0} =& 2\alpha\left(\mu t_1 + \tau_3\right) + 2(1-\alpha)\mu\nu + \frac{2}{3}\alpha\tau_2 \\ p_{0,1} =& \mu t_2 + \tau_3 + \frac{1}{3}(2-\alpha)\tau_2 \\ p_{2,0} =& (1-\alpha)\left(\mu t_1 + \tau_3\right) - (1-\alpha)\mu\nu + \frac{1}{3}(2-\alpha)\tau_2 \\ p_{1,1} =& \frac{2}{3}\alpha\tau_2 \end{aligned}$$

with $\tau_3 = \mu E[T_{33}]$ and $\tau_2 = \mu E[T_{32}]$. If α is given by $\widehat{\alpha}^*$, we solve to get

$$\begin{aligned}
\widehat{\tau_2} &= \frac{3}{2\widehat{\alpha}^*}\frac{m_{1,1}}{m_{tot}} \\
\widehat{\mu t_1 + \tau_3} &= \frac{1}{m_{tot}}\left(m_{1,0}\frac{1}{2} + m_{2,0} - m_{1,1}\frac{1}{\widehat{\alpha}^*}\right) \\
\widehat{\mu t_2 + \tau_3} &= \frac{1}{m_{tot}}\left(m_{0,1} - m_{1,1}\frac{2-\widehat{\alpha}^*}{2\widehat{\alpha}^*}\right) \\
\widehat{\mu\nu} &= \frac{1}{m_{tot}}\left(m_{1,0}\frac{1}{2} - m_{2,0}\frac{\widehat{\alpha}^*}{1-\widehat{\alpha}^*} + m_{1,1}\frac{1}{2(1-\widehat{\alpha}^*)}\right)
\end{aligned}$$

In (Schlebusch *et al.* 2012) and (Skoglund *et al.* 2011),

$$\frac{3}{2}\frac{m_{1,1}}{m_{2,0}+m_{1,1}}$$

is used to estimate α. A comparison to the framework presented here gives

$$\frac{3}{2}\frac{m_{1,1}}{m_{2,0}+m_{1,1}} = \frac{3}{2}\frac{p_{1,1}}{p_{2,0}+p_{1,1}} = \alpha\frac{b_2}{(1-\alpha)\mu(t_1-\nu) + \frac{1}{3}(1-\alpha)b_1 + \frac{1}{3}(1+2\alpha)b_2}$$

which is approximately α for α close to 1 or if $b_1 \approx 2b_2$ and either $b_2 >> (1-\alpha)\mu(t_1-\nu)$ or $t_1 - \nu \approx 0$.

Assuming a constant ancestral population size (figure 1b) implies $b_1 = 2\mu NA = 2b_2$ and

$$\frac{3}{2}\frac{p_{1,1}}{p_{2,0}+p_{1,1}} = \alpha\frac{N_A}{N_A + (1-\alpha)(t_1-\nu)}$$

which is approximately α for α close to 1, for $N_A >> (1-\alpha)(t_1-\nu)$ and for $t_1 - \nu \approx 0$.

If ascertainment is performed in an outgroup,

$$\frac{3}{2}\frac{m^*_{1,1}}{m^*_{2,0}+m^*_{1,1}} = \frac{3}{2}\frac{p^*_{1,1}}{p^*_{2,0}+p^*_{1,1}} = \alpha\frac{3b^*_2}{(1-\alpha)b^*_1 + (1+2\alpha)b^*_2}$$

which is approximately α for α close to 1 or if $b^*_1 \approx 2b^*_2$.

Supplemental Material

Figures for simulation studies

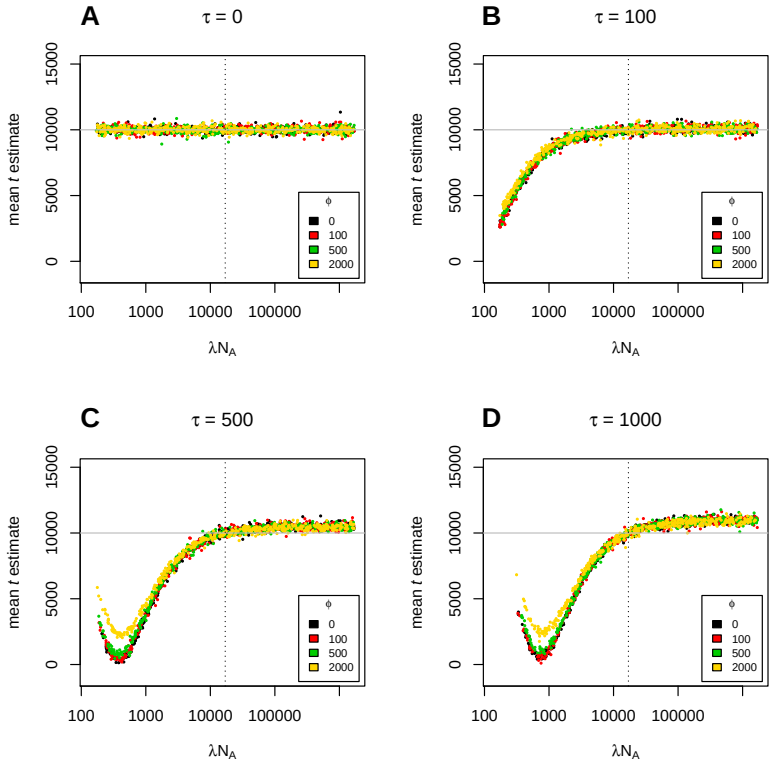

Figure S1 The effect of ancestral population size change, λN_A, on TT method estimates of divergence time, t, with a true split time of 10,000 generations and ancestral population size change lasting (A) 0, (B) 100, (C) 500, and (D) 1000 generations.

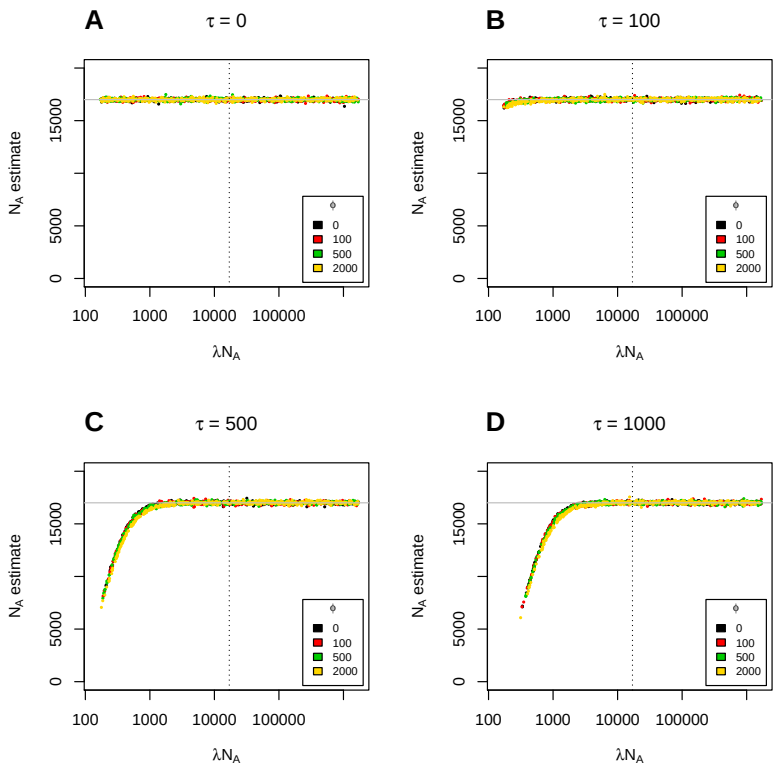

Figure S2 The effect of ancestral population size change, λN_A, on TT method estimates of ancestral population size, N_A, with a true split time of 10,000 generations and ancestral population size change lasting (A) 0, (B) 100, (C) 500, and (D) 1000 generations.

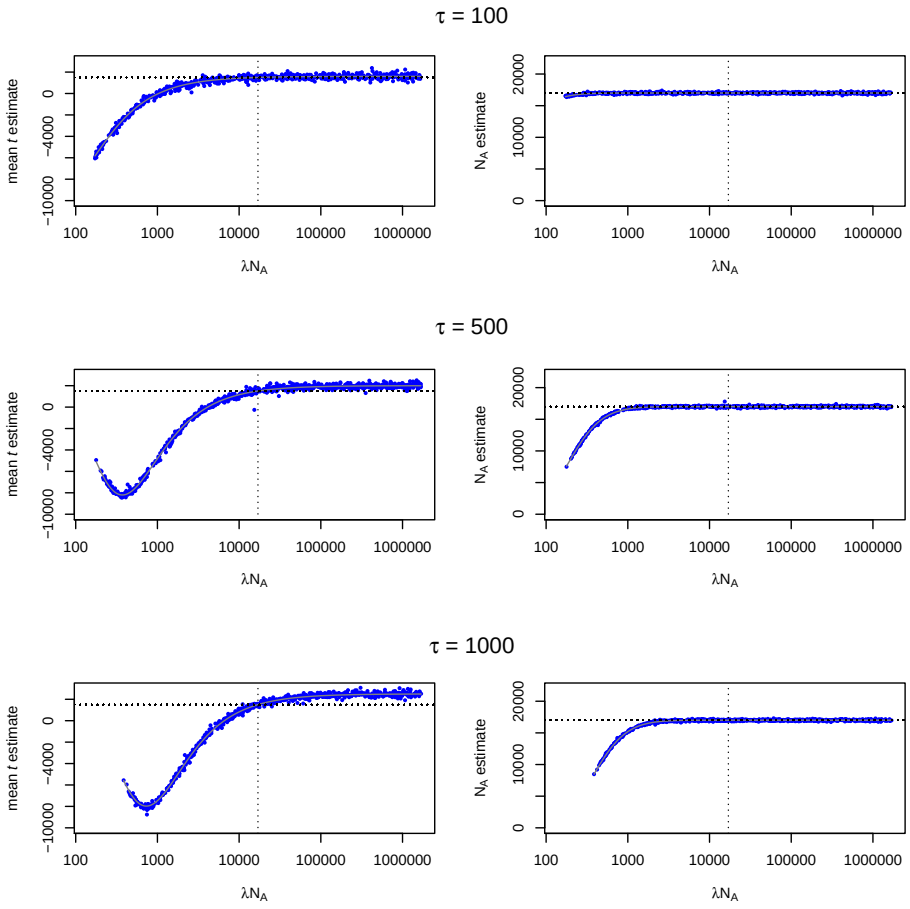

Figure S3 TT method estimates of divergence time and ancestral population size with changes to ancestral population size of duration (τ) (A, B) 100, (C, D) 500, and (E, F) 1,000 generations. In all cases, the change in ancestral population size occurs immediately prior to the split (ϕ=0) and the true split time is 1500 generations.

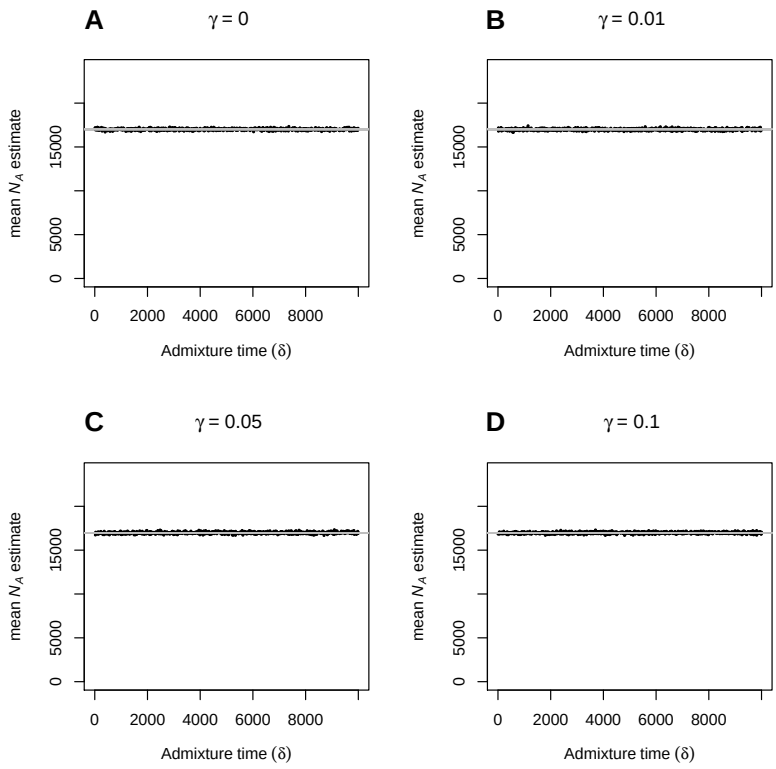

Figure S4 The effect of varying admixture time (δ) on TT method estimates of ancestral population size (N_A), with migration proportions (γ) of (A) 0, (B) 0.01, (C) 0.05 and (D) 0.1 when true ancestral population size is 17,000 and true split time is 10,000 generations.

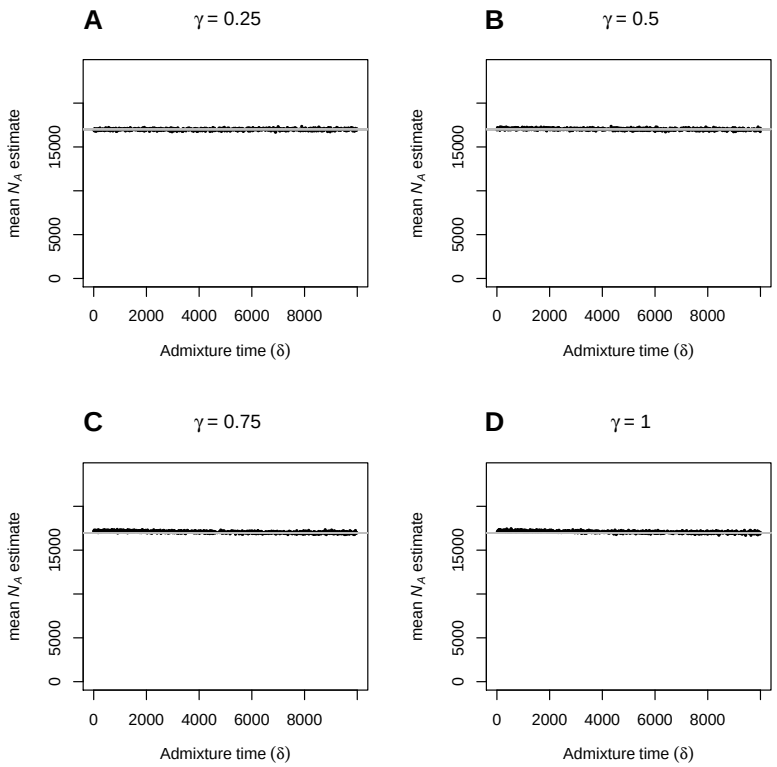

Figure S5 The effect of varying admixture time (δ) on TT method estimates of ancestral population size (N_A), with migration proportions (γ) of (A) 0.25, (B) 0.5, (C) 0.75 and (D) 1 when true ancestral population size is 17,000 and true split time is 10,000 generations.

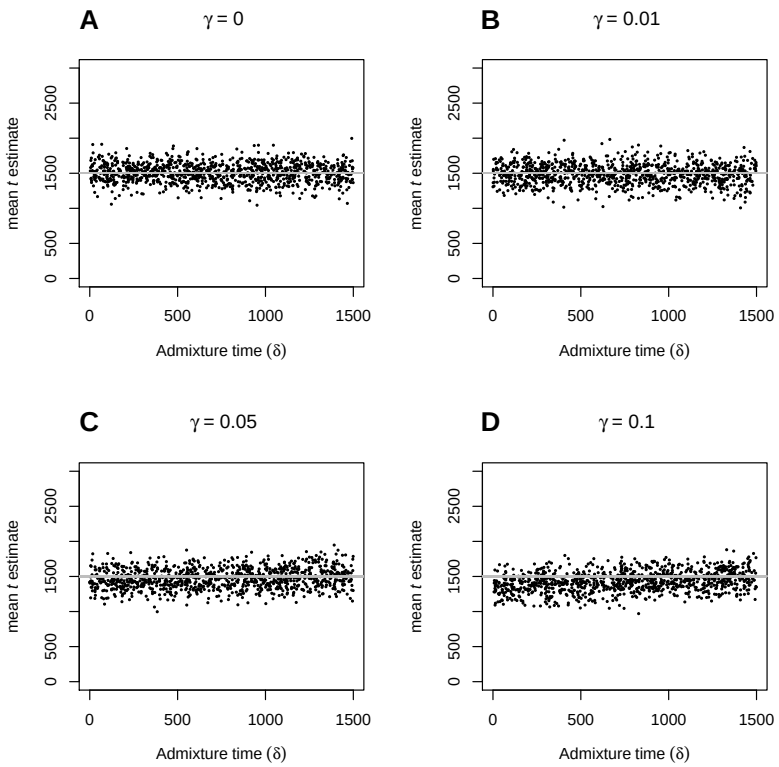

Figure S6 The effect of varying admixture time (δ) on TT split time estimates ($\hat{t}$), when the proportion of admixture (γ) is (A) 0, (B) 0.01, (C) 0.05 and (D) 0.1, and true split time is 1500 generations.

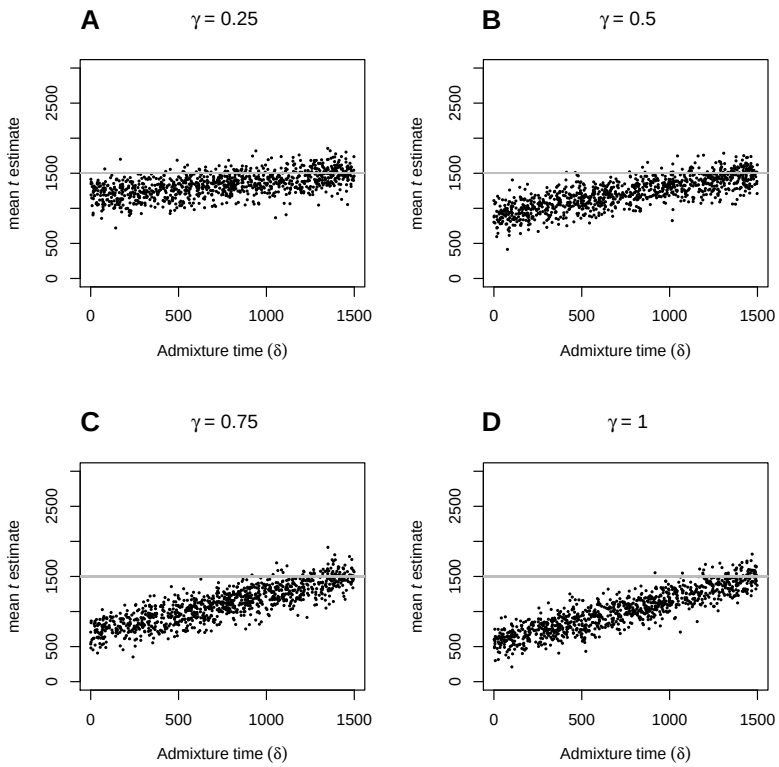

Figure S7 The effect of varying admixture time (δ) on TT split time estimates ($\hat{t}$), when the proportion of admixture (γ) is (A) 0.25, (B) 0.5, (C) 0.75 and (D) 1, and true split time is 1500 generations.

Figures for application to data

TT-estimates under a constant ancestral population

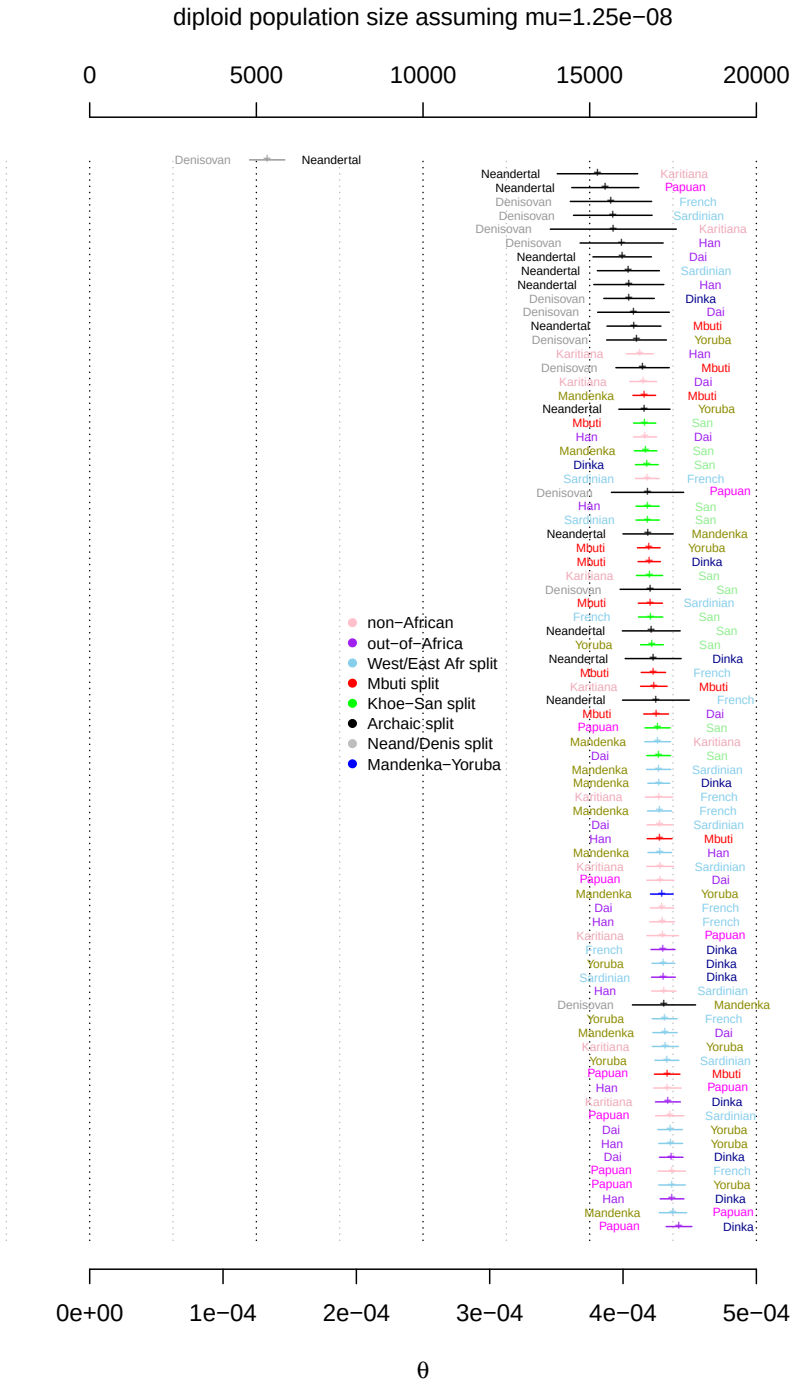

Figure S8 Estimates of $\theta = \mu N_A$ and the corresponding diploid ancestral population size ($N_A/2$) assuming that it is constant and a mutation rate of 1.25×10^{-8}.

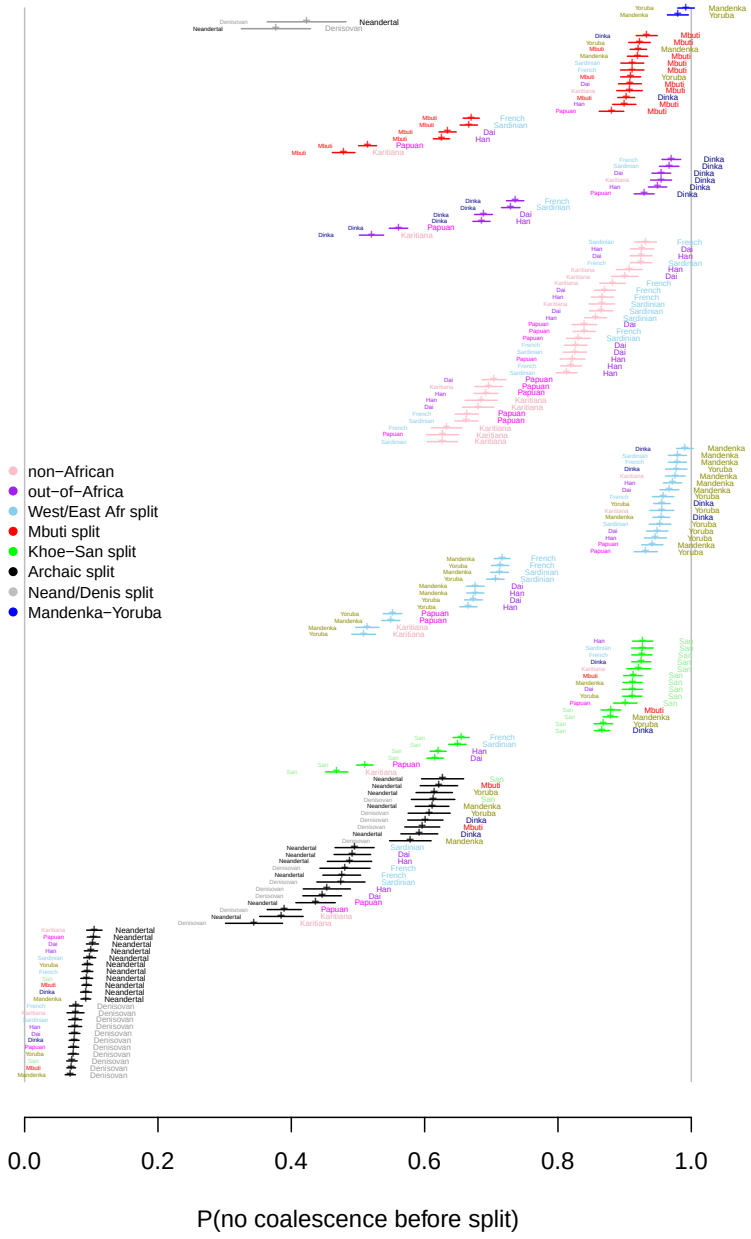

Figure S9 Estimates of α assuming a constant ancestral population.

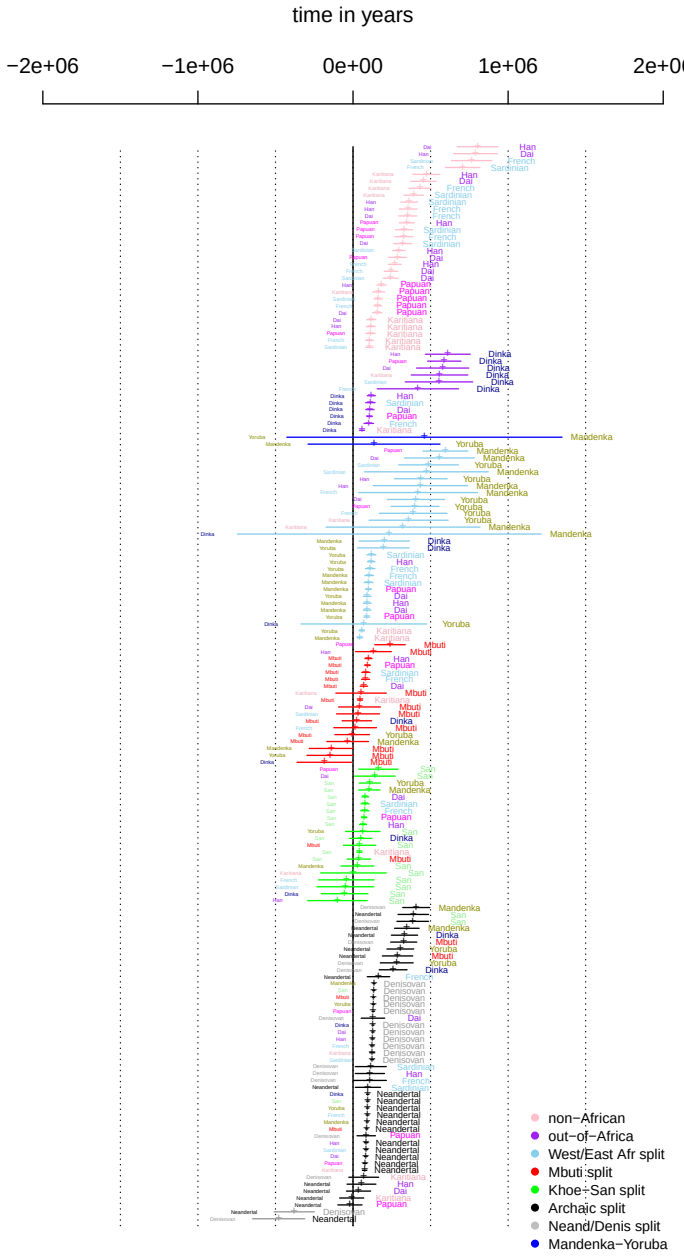

Figure S10 Estimates of expected coalescence times in years given coalescence before split assuming a constant ancestral population. A mutation rate of 1.25×10^{-8} and a generation time of 30 years is assumed.

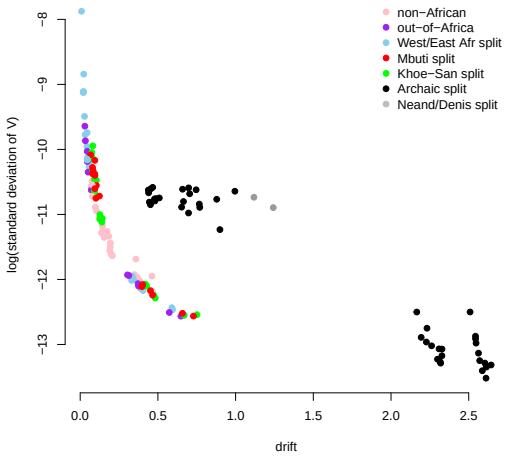

Figure S11 Drift ($-ln(\alpha)$) vs estimated standard devation of $V = \mu\nu$.

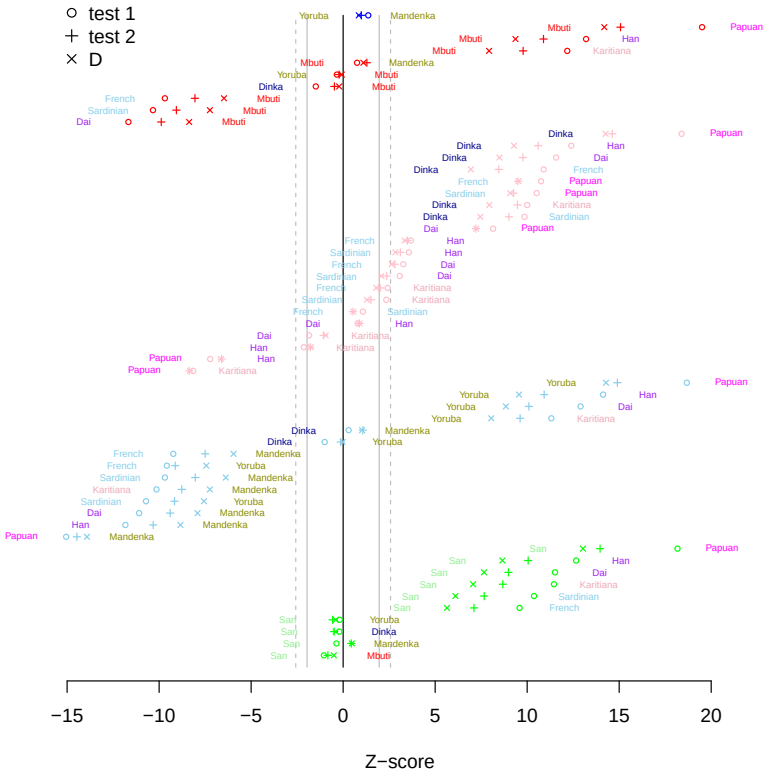

Figure S12 Tests based on equation 5 (○) and equation 6 (+) when data is conditional on the derived variant being present in either the Neandertal or Denisovan genome. D-tests (scaled by its estimated standard deviation, ×) with D('left population', 'right population', Neandertal+Denisovan, outgroup) are also shown. Solid grey lines shows cut-offs for rejecting null-model at the 0.05 level and the dashed lines at the 0.01 level.

TT-estimates using outgroup ascertainment

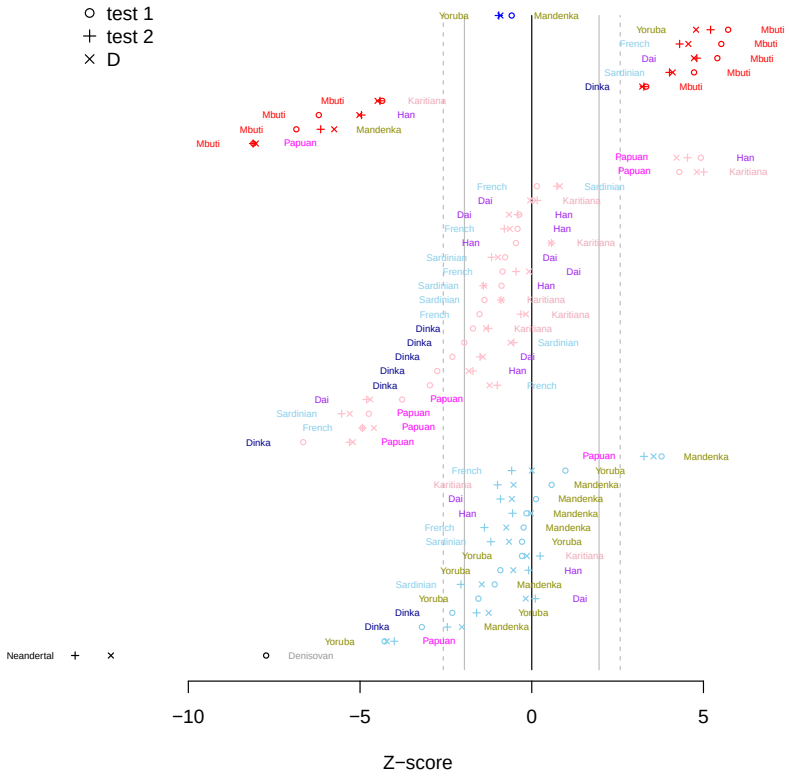

Figure S13 Tests based on equation 5 (○) and equation 6 (+) when data is conditional on the derived variant being present in Balito Bay A. D-tests (scaled by its estimated standard deviation, ×) with D('left population', 'right population', Balito Bay A, outgroup) are also shown. Solid grey lines shows cut-offs for rejecting null-model at the 0.05 level and the dashed lines at the 0.01 level.

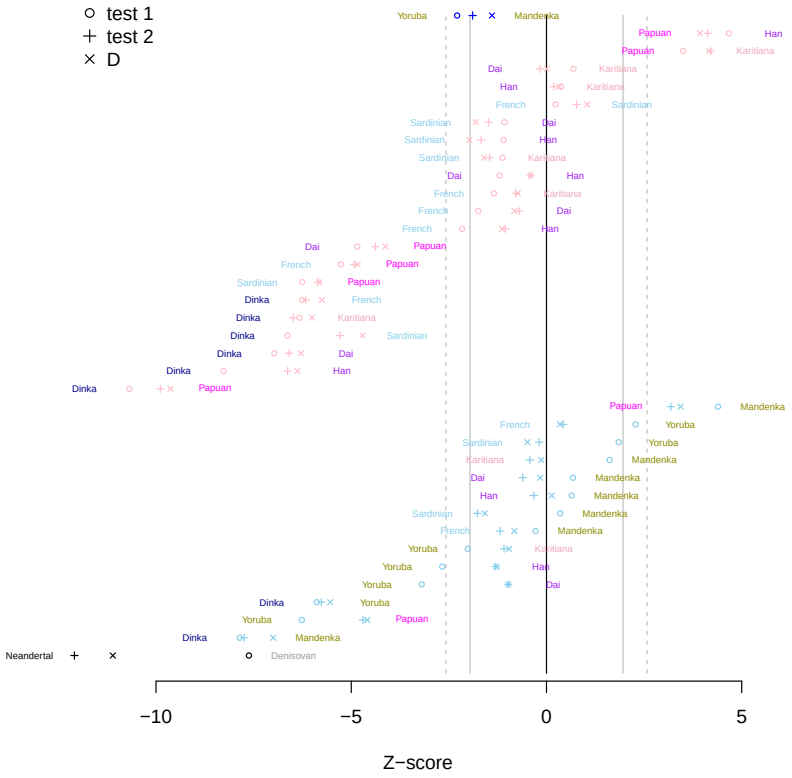

Figure S14 Tests based on equation 5 (○) and equation 6 (+) when data is conditional on the derived variant being present in Mbuti. D-tests (scaled by its estimated standard deviation, ×) with D('left population', 'right population', Mbuti, outgroup) are also shown. Solid grey lines shows cut-offs for rejecting null-model at the 0.05 level and the dashed lines at the 0.01 level.

Figure S15 α-estimates conditional on the derived variant being present in the outgroup. 'Cross' show estimates when the derived variant is present in Mbuti, 'circle' show estimates when the derived variant is present in Neandertal or Denisovan, 'plus' show estimates when the derived variant is present in Balito Bay A. Black was used for estimates that did not fail the outgroup tests while transparent grey was used for comparisons that failed at least one of the two outgroup tests.

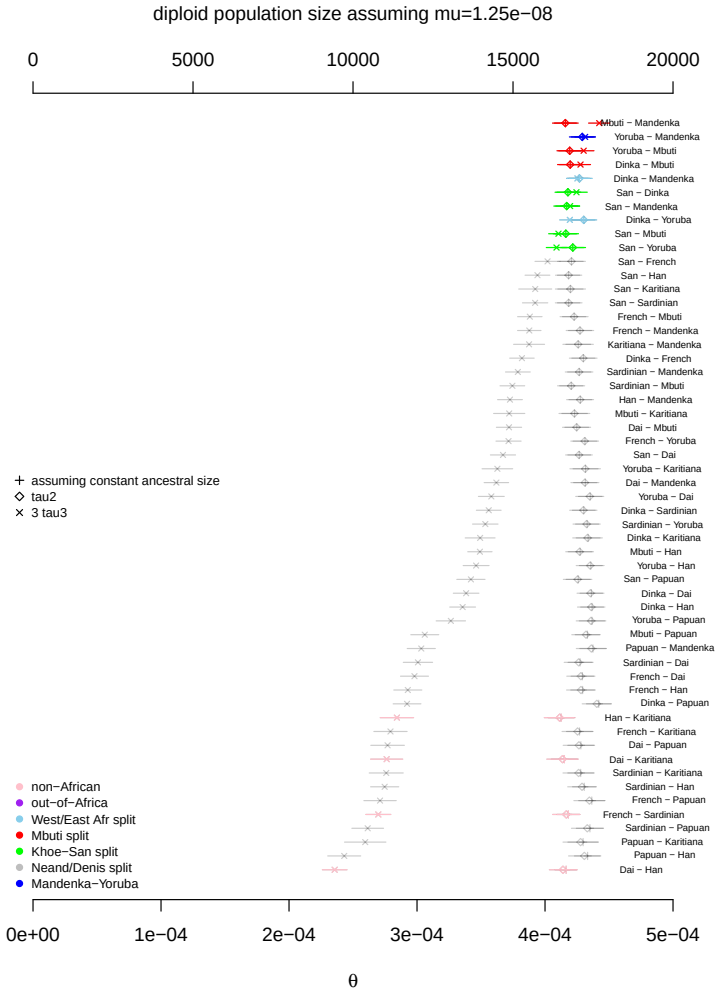

Figure S16 Different estimates of the ancestral population size. Here Neandertal+Denisovan was used as outgroup to estimate τ_2 and τ_3. Transparent grey was used for comparisons that failed at least one of the two outgroup tests.

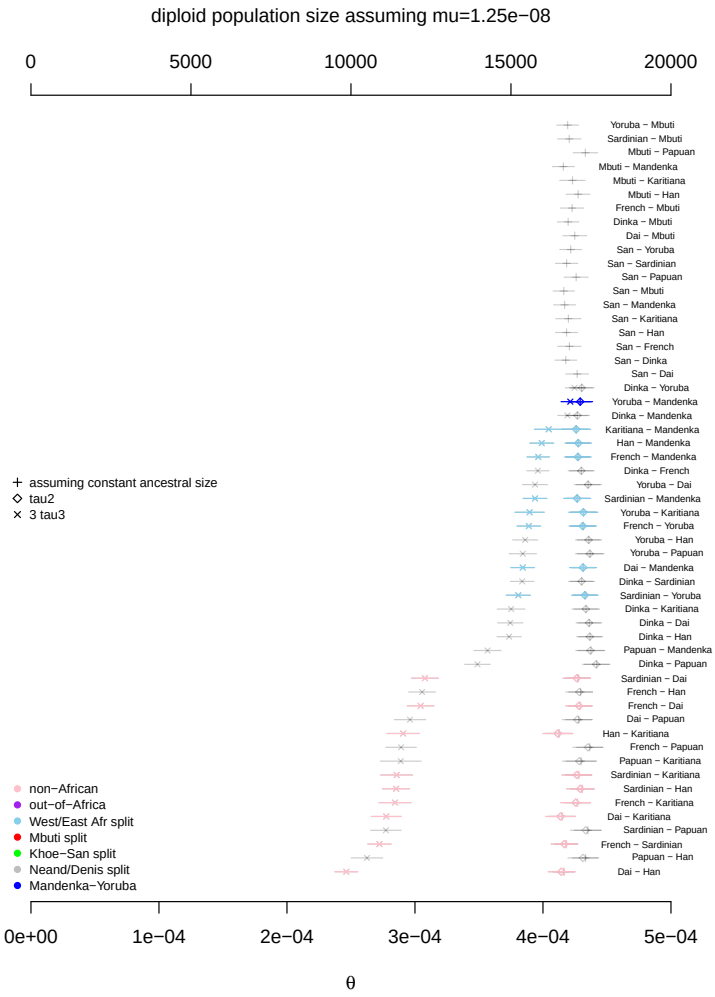

Figure S17 Different estimates of the ancestral population size. Here Mbuti was used as outgroup to estimate τ_2 and τ_3. Transparent grey was used for comparisons that failed at least one of the two outgroup tests.

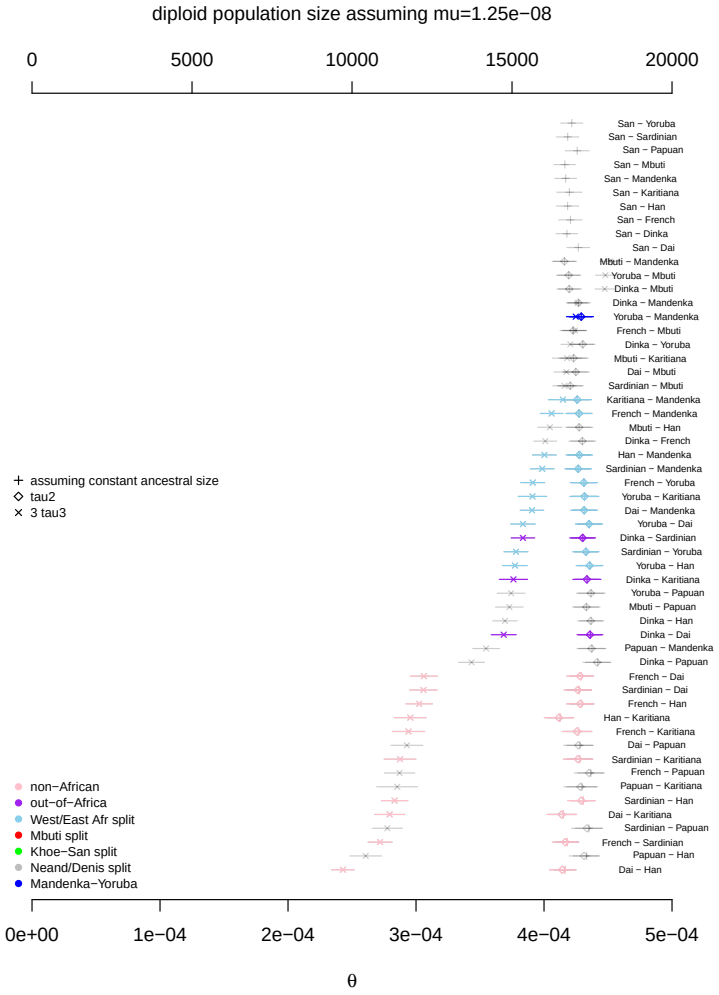

Figure S18 Different estimates of the ancestral population size. Here Balito Bay A was used as outgroup to estimate τ_2 and τ_3. Transparent grey was used for comparisons that failed at least one of the two outgroup tests.

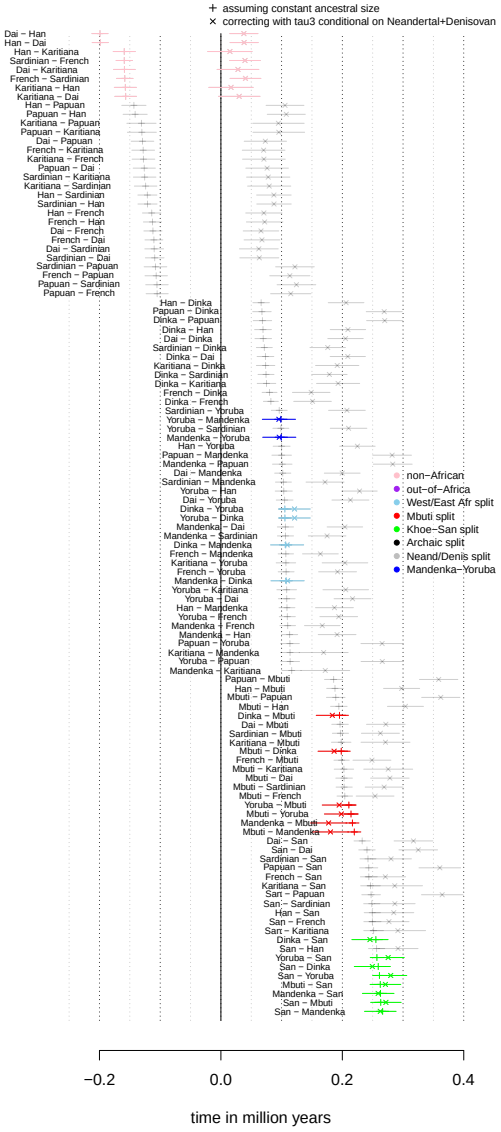

Figure S19 Different estimates of split times assuming a mutation rate of 1.25×10^{-8} and a generation time of 30 years. Two estimates are shown: estimates assuming a constant ancestral (+) and estimates relying on an external estimate of α (×) obtained by outgroup ascertainment in Neandertal+Denisovan. Transparent grey was used for comparisons that failed at least one of the two outgroup tests.

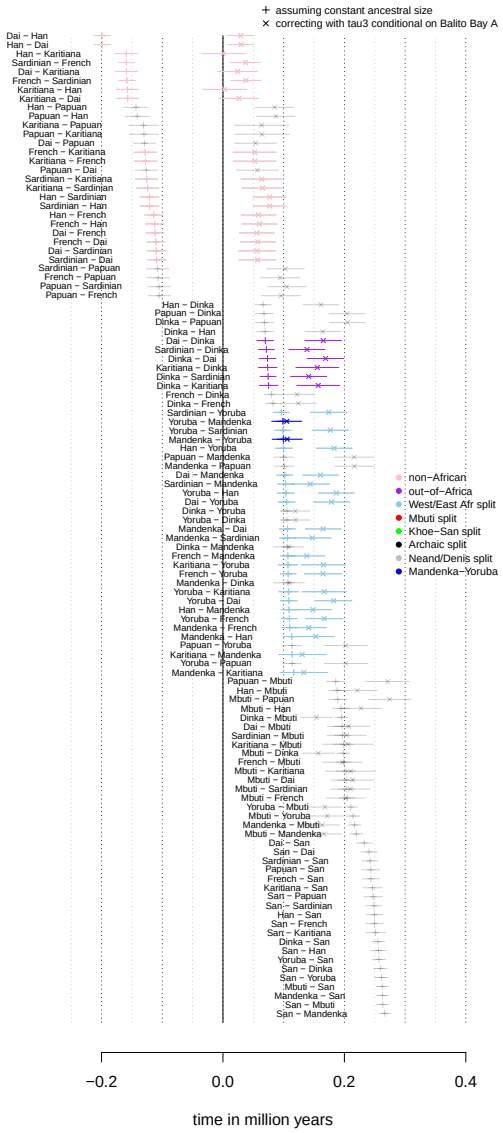

Figure S20 Different estimates of split times assuming a mutation rate of 1.25×10^{-8} and a generation time of 30 years. Two estimates are shown: estimates assuming a constant ancestral (+) and estimates relying on an external estimate of α (×) obtained by outgroup ascertainment in Balito Bay A (an ancient Khoe-San genome). Transparent grey was used for comparisons that failed at least one of the two outgroup tests.

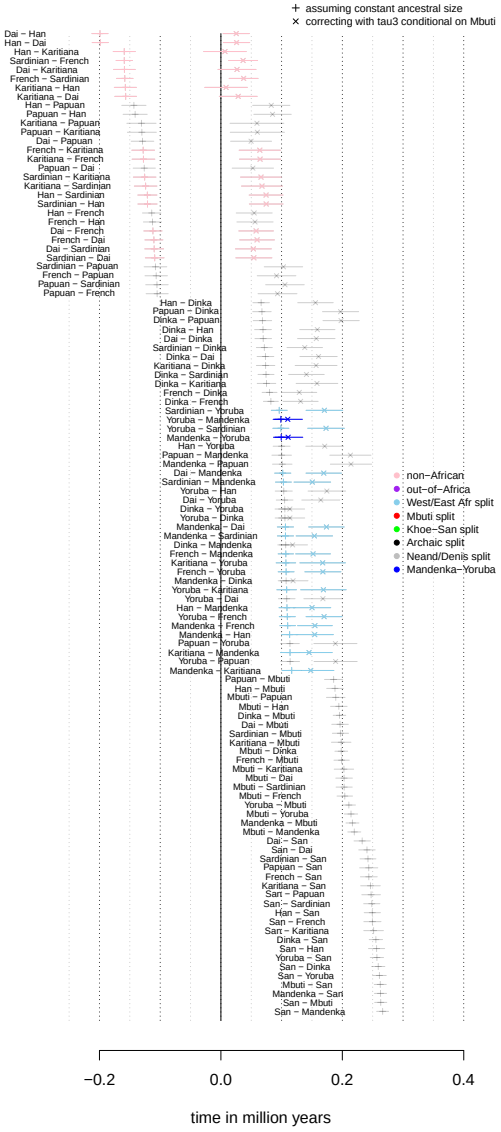

Figure S21 Different estimates of split times assuming a mutation rate of 1.25×10^{-8} and a generation time of 30 years. Two estimates are shown: estimates assuming a constant ancestral (+) and estimates relying on an external estimate of α (×) obtained by outgroup ascertainment in Mbuti. Transparent grey was used for comparisons that failed at least one of the two outgroup tests.