Received: 29 September 2020 | Revised: 20 September 2021 | Accepted: 22 September 2021

DOI: 10.1111/1755-0998.13519

RESOURCE ARTICLE

MOLECULAR ECOLOGY RESOURCES WILEY

Estimating the time since admixture from phased and unphased molecular data

Thijs Janzen[1,2] | Verónica Miró Pina[3,4,5]

[1]Groningen Institute for Evolutionary Life Sciences, University of Groningen, Groningen, The Netherlands

[2]Carl von Ossietzky University, Oldenburg, Germany

[3]Instituto de Investigaciones en Matemáticas Aplicadas y Sistemas (IIMAS), Universidad Nacional Autónoma de México (UNAM), México City, México

[4]Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Barcelona, Spain

[5]Universitat Pompeu Fabra (UPF), Barcelona, Spain

Correspondence
Thijs Janzen, Groningen Institute for Evolutionary Life Sciences, University of Groningen, Groningen, The Netherlands.
Email: t.janzen@rug.nl

Funding information
Dirección General de Asuntos del Personal Académico, Universidad Nacional Autónoma de México, Grant/Award Number: 2018

Abstract

After admixture, recombination breaks down genomic blocks of contiguous ancestry. The breakdown of these blocks forms a new "molecular clock" that ticks at a much faster rate than the mutation clock, enabling accurate dating of admixture events in the recent past. However, existing theory on the breakdown of these blocks, or the accumulation of delineations between blocks, so-called "junctions", has mostly been limited to using regularly spaced markers on phased data. Here, we present an extension to the theory of junctions using the ancestral recombination graph that describes the expected number of junctions for any distribution of markers along the genome. Furthermore, we provide a new framework to infer the time since admixture using unphased data. We demonstrate both the phased and unphased methods on simulated data and show that our new extensions have improved accuracy with respect to previous methods, especially for smaller population sizes and more ancient admixture times. Lastly, we demonstrate the applicability of our method on three empirical data sets, including labcrosses of yeast (*Saccharomyces cerevisae*) and two case studies of hybridization in swordtail fish and *Populus* trees.

KEYWORDS
admixture, hybridization, junctions, phasing, recombination

1 | INTRODUCTION

The traditional view where taxa accumulate incompatibilities over time and gradually become reproductively isolated from each other—speciation—has led to insight into the processes generating and maintaining biodiversity (Coyne & Orr, 2004). However, it has become apparent that reticulate evolution is common: taxa do not necessarily only branch, they can also come back together—hybridization—(Abbott et al., 2013). This means that, while the ancestry of nonrecombining DNA can be traced linearly back in time (e.g. mitochondrial lineages), the ancestry of genomes backward in time will branch at admixture events. In plants, it has been known for quite some time that hybridization is a widespread phenomenon that can not only generate viable offspring, but also potentially lead to the formation of new taxa and, ultimately, species (Grant, 1981). It has long been debated whether this process is also common in animals, but over the past few years numerous examples have appeared, including, but not limited to, butterflies (Capblancq et al., 2015; Mavárez et al., 2006), cichlid fishes (Keller et al., 2013; Koblmüller et al., 2007), warblers (Brelsford et al., 2011), fruit flies (Schwarz et al., 2005) and sculpins (Nolte et al., 2005).

Thijs Janzen and Verónica Miró Pina contributed equally to this work.

Understanding the timeline of these hybridization events is paramount in obtaining a full understanding of the process and its impact. A "recombination clock" is particularly useful for studying recent evolutionary dynamics since detectable recombination events are markers of admixture (Baird, 2006), and the statistical association across loci created by admixture (admixture linkage disequilibrium) decays slowly (Baird, 2015). After admixture of two taxa, contiguous genomic blocks are broken down by recombination over time. The delineations between these blocks were termed "junctions" by Fisher (1949, 1954), and inheritance of these junctions is similar to that of point mutations. Further work on the theory of junctions has shown how they accumulate over time for sib–sib mating (Fisher, 1954), self-fertilization (Bennett, 1953), alternate parent–offspring mating (Fisher, 1959; Gale, 1964), a randomly mating population (Baird et al., 2003; Stam, 1980) and for substructured populations (Baird, 1995; Barton, 1983; Chapman & Thompson, 2002, 2003).

So far, applying the theory of junctions has proved difficult, as it requires extensive genotyping of the admixed individuals, but also of the source taxa. With the current decrease in genotyping costs (Muir et al., 2016), such analyses are coming within reach, and frameworks are being developed that assist in inferring local ancestry and detecting junctions, given molecular data of source and admixed taxa (Corbett-Detig & Nielsen, 2017; Guan, 2014; Maples et al., 2013; Medina et al., 2018; Paşaniuc et al., 2009; Svedberg et al., 2021). Nevertheless, molecular data always paints an imperfect image of ancestry along the genome, and inferring the number of junctions in a chromosome remains limited by the number of diagnostic markers available (see Figure 1, first panel). Previous work on the theory of junctions does not take into account the effect of a limited number of genetic markers, and so far this effect had to be corrected using simulations (Buerkle & Rieseberg, 2008; MacLeod et al., 2005). Recent work by Janzen et al. (2018) resolves this issue by extending the theory of junctions with the effect of using a limited number of markers, but they had to assume an evenly spacing of markers. However, molecular markers are rarely evenly spaced. The first result we present here is an extension of the theory of junctions which includes the effect of marker spacing on inferring the number of junctions in a genome and that has the advantage of working well with fewer markers.

Furthermore, most existing theory on the accumulation of junctions is developed for the case where ancestry can be determined within a single chromosome (with the exception of ANCESTRY HMM (Corbett-Detig & Nielsen, 2017)). For diploid species, sequencing data presents itself as the pileup of ancestry across both chromosomes, requiring an additional step to separate the contributions of both chromosomes, called "phasing" (see Figure 1, second and third panel). Phasing methods can be classified into three main categories, including haplotype-resolved genome-sequencing, pedigree-based methods and statistical methods. Haplotype-resolved genome-sequencing methods (reviewed in Snyder et al. (2015)) yield accurate results, but are expensive and require a large number of sequences in order to be able to resolve the required haplotypes (but see (Lutgen et al., 2020)). Pedigree-based phasing methods do not require resolved haplotypes, but instead rely on accurate genotyping of related individuals to resolve the required phase. These methods often yield good results but their application has been limited to humans, where large pedigree data sets are available (Browning & Browning, 2011; Kong et al., 2008; Loh et al., 2016). Statistical methods do not require large pedigree data sets and are based on recombination rate estimates and allele frequencies in a population. While some of these methods make use of a reference genome (e.g. Eagle (Loh et al., 2016b), Beagle (Browning & Browning, 2007) or ShapeIt (O'Connell et al., 2016)), others allow de novo haplotype aware assembly (e.g. POLYTE (Baaijens & Schönhuth, 2019)). Further

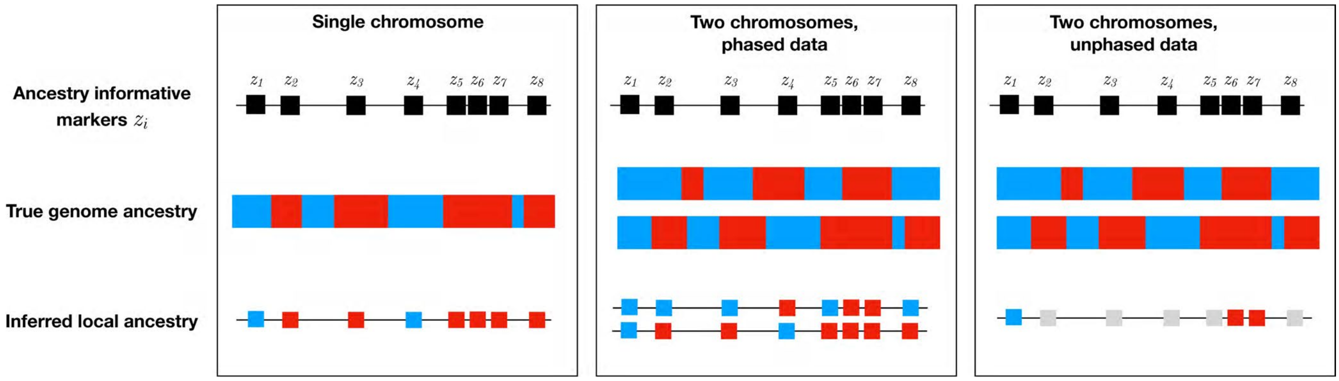

Single chromosome
Two chromosomes, phased data
Two chromosomes, unphased data
Ancestry informative markers z_i
z_1 z_2 z_3 z_4 z_5 z_6 z_7 z_8
True genome ancestry
Inferred local ancestry

FIGURE 1 Visual depiction of the observed data. We show the differences between the type of data generated by the three methods we present in this study. On each panel, the chromosome in the centre is coloured according to ancestry (blue represents source taxa $\mathscr{P}$ and red represents source taxa $\mathscr{Q}$). Above the chromosome are indicated the locations of ancestry informative markers z_i. Resulting inferred ancestry on these markers is shown below, where grey indicates heterozygous ancestry. The first panel represents the one chromosome method. There are seven junctions in the chromosome, but only three are observed in the data due to a limited marker coverage. Notice that a junction in the inferred ancestry corresponds to an odd number of junctions in the "true" genome ancestry and conversely that an even number of junctions in the "true" ancestry yields no inferred junctions. The second and third panels represent the methods that use information from two chromosomes. In the second panel, data are phased whereas in the third panel data are unphased

improvements include the usage of third-generation sequencing (Ebler et al., 2019; Kronenberg et al., 2021; Tangherloni et al., 2019; Tourdot & Zhang, 2021). However, data from hybrid populations are not often available in this form. Across these three groups of methods, the overarching theme is that phasing is often costly and accuracy can be left wanting. Yet, inclusion of information from both chromosomes is expected to improve inference of the onset of admixture considerably and hence expansion of the theory of junctions towards a framework that takes into account data from both chromosomes is warranted.

Here, we provide a framework to estimate the time since admixture using phased or unphased data from two homologous chromosomes, taking into account marker spacing along the chromosome. Our framework is based on modelling the joint genealogy of loci that are located in the same chromosome or in two homologous chromosomes, using the ancestral recombination graph (ARG; Griffths, 1991; Griffths & Marjoram, 1997; Hudson, 1983). It has the advantage of being fast since it relies on mathematical computations and does not require simulations. It has been implemented in the R package "junctions". In comparison with previous methods (ANCESTRY HMM, Corbett-Detig and Nielsen (2017)), our method has the advantage of working well with small population sizes and larger admixture times.

Our study is organized as follows. In Section 2, we introduce our model, which is a simplified version of the ARG, and present three maximum-likelihood methods to infer the time since admixture in hybrid populations: the first method uses information from a single chromosome, the second method uses phased data from two homologous chromosomes, and the third method uses unphased data from two homologous chromosomes. At the end of Section (2.4), we validate our methods using simulations. In Section 3, we provide a detailed comparison to previous methods. Lastly, in Section 4 we apply them to a data set from experimental evolution in yeast and to two case studies of hybridization in swordtail fish and *Populus* trees.

2 | MATERIALS AND METHODS

2.1 | Mathematical model

We assume a diploid population that evolves according to Wright–Fisher dynamics; that is, generations are nonoverlapping, mating is random, and all individuals are hermaphrodites. We only keep track of one chromosome (or one pair of chromosomes), assuming that the accumulation of junctions on different pairs of chromosomes is independent of each other (see Section 2.2.1). We assume that hybridization occurred T generations ago between two source taxa, $\mathscr{P}$ and $\mathscr{Q}$. The proportion of individuals from source $\mathscr{P}$ in the initial generation is p and the proportion of individuals from source $\mathscr{Q}$ is $q = 1 - p$. We define the initial heterogenicity $H_0 := 2pq$ as the probability that, when sampling two individuals from this generation, they are from different source taxa. This also represents the probability that at any given locus, one allele can be traced to source taxon $\mathscr{P}$ and the other allele to $\mathscr{Q}$. Following Janzen et al. (2018), instead of referring to the lesser known term of heterogenicity, we use the term of heterozygosity in the manuscript throughout because these two concepts are mathematically equivalent.

We assume that the length of the chromosome is C Morgan and that there are n molecular markers whose positions are given by $(z_1, \ldots, z_n) \in [0, C]$. For two consecutive markers at sites z_i and z_{i+1}, we define $d_i = z_{i+1} - z_i$, the distance between them in Morgan. We assume that there are enough markers on the chromosome such that the d_i's are small compared to 1. The genealogy of these n (or $2n$ for a diploid genome) loci is given by the ARG, defined in Hudson (1983), Griffths (1991), and Griffths and Marjoram (1997). The ARG is a branching-coalescence process which follows backwards in time the ancestry of loci sampled in the present population (time 0). If two loci belong to the same block at time t, they are identical by descent (IBD) with respect to generation t, that is the sampled loci have been inherited from the same individual living t units of time ago.

Although the ARG for many loci has complicated transition rates and is a computationally intensive model, here we consider only two loci (or two pairs of loci for a diploid genome) at a time. This is sufficient since for our maximum-likelihood approach, we only use the expected number of junctions—and not its variance or higher moments (see, e.g. Equation 3).

We assume that the number of diploid individuals N is large, so that we can neglect some transitions (double coalescences and simultaneous coalescence and recombination) and that $d_i \ll 1$ so that there is no more than one crossover per generation between two molecular markers and the mutation rates are small enough so that we can neglect mutations that happened between the admixture time and the present.

2.2 | Two sites, one chromosome

The aim of this section is to derive a formula for the expected number of observed junctions on one chromosome given the population size N, the distances between the markers $(d_1, \ldots, d_n)$ and the initial heterozygosity $H_0 := 2pq$. We start by considering two consecutive loci z_i and z_{i+1} sampled in the same chromosome in the present population. The ARG for these two sites has two possible states:

- If the loci are in state $(z_i \sim z_{i+1})$ at time t, it means that the two sampled loci have been inherited from *the same individual* living t generations before, that is that they are IBD.
- If the loci are in state $(z_i \nsim z_{i+1})$, it means that the two loci have been inherited from different individuals.

To model the observed junctions, we look at identity-by-descent with respect to time T, when admixture took place. Recall that we only model observed junctions, that is where adjacent markers have alleles corresponding to different source taxa (which corresponds to an odd number of "true" junctions, see 1 and its caption).

- If the two loci are in state $(z_i \sim z_{i+1})$ at time T, the ancestor can be from source $\mathscr{P}$ (with probability p) or red (with probability q). In both cases, we observe no junction.
- If the two loci are in state $(z_i \nsim z_{i+1})$ at time T (not IBD), they have two different ancestors at generation T, which can be:
  a. Both blue (with probability p^2) or both red (with probability q^2), in this case there is no junction.
  b. One blue and one red (with probability $2pq$), in this case, there is a junction in the chromosome sampled in the present.

The dynamics of the ARG are controlled by two types of events:

- **Recombination** $(z_i \sim z_{i+1}) \rightarrow (z_i \nsim z_{i+1})$ with probability d_i,
- **Coalescence** $(z_i \nsim z_{i+1}) \rightarrow (z_i \sim z_{i+1})$ with probability $\frac{1}{2N}$.

Since $N \gg 1$ and the $d_i \ll 1$, other events (such as simultaneous coalescence and recombination events) have probabilities that are negligible. This yields the following transition matrix:

$$\overline{M} = \begin{pmatrix} 1-d_i & d_i \\ \frac{1}{2N} & 1-\frac{1}{2N} \end{pmatrix}.$$

Let $\overline{P}_t$ be the probability vector at time t for this Markov chain with two states. $(\overline{P}_t)_1$ is the probability of $(z_i \sim z_{i+1})$ at time t and $(\overline{P}_t)_2$ the probability of $(z_i \nsim z_{i+1})$ at time t. We have $\overline{P}_0 = (1,0)$ (in the present, we sample the two loci in the same individual) and $\overline{P}_t = \overline{P}_0\overline{M}^t$. We denote by $\mathbb{P}(J_T(z_i,z_{i+1}))$ the probability that a junction is observed between z_i and z_{i+1} (conditioning on the fact that hybridization occurred T generations ago). We have

$$\mathbb{P}(J_T(z_i,z_{i+1})) = H_0(\overline{P}_t)_2, \tag{1}$$

which corresponds to the probability that the two loci were carried by different individuals T generations ago and they are from different source taxa (see Figure 2, left panel).

Solving Equation (1) gives

$$\mathbb{P}(J_T(z_i,z_{i+1})) = H_0\frac{2Nd_i}{2Nd_i+1}\left(1-\left(1-d_i-\frac{1}{2N}\right)^T\right). \tag{2}$$

Let $\mathbb{E}(J_T)$ be the expected number of observed junctions on one chromosome, we have

$$\mathbb{E}(J_T) = \sum_{i=1}^{n-1}\mathbb{P}(J_T(z_i,z_{i+1})) = H_0\sum_{i=1}^{n-1}\frac{2Nd_i}{2Nd_i+1}\left(1-\left(1-d_i-\frac{1}{2N}\right)^T\right). \tag{3}$$

2.2.1 | Maximum likelihood

For each chromosome, we can calculate the likelihood of observing the data, where the data are $n-1$ pairs of markers $(z_i \sim z_{i+1})$, which is given by

$$\prod_{i=1}^{n-1}\mathbb{P}(z_i,z_{i+1}),$$

where $\mathbb{P}((z_i,z_{i+1}))$ is the probability of observing the pair of markers (z_i,z_{i+1}) given by:

- $\mathbb{P}((z_i,z_{i+1})) = \mathbb{P}(J_T(z_i,z_{i+1}))$, if there is a junction between the two markers. This is given by Equation (2).
- $\mathbb{P}((z_i,z_{i+1})) = 1-\mathbb{P}(J_T(z_i,z_{i+1}))$, if there is no junction between the two markers.For different chromosomes sampled in the same individual, we can assume independence between chromosomes and the full likelihood of observing the data, given the time since admixture, are given by the product of these likelihoods, for example

$$\mathscr{L} = \prod_{c=1}^{C}\prod_{i=1}^{n_C-1}\mathbb{P}(z_i,z_{i+1}), \tag{4}$$

where C indicates the number of chromosomes, and n_C indicates the number of markers on chromosome C.

When we have data from several individuals (which are not independent, due to coalescence), we first compute the maximum-likelihood estimator of T for each individual (using (4)). Our estimator of the time since admixture is the average of the maximum-likelihood estimator of T obtained for all individuals in the sample. Confidence intervals (CIs) are obtained by bootstrap (reported are the 95% interquartile range of different bootstrap samples). To visualize the dispersion of the data, we also plot the distribution of the maximum-likelihood estimator across individuals.

2.3 | Two sites, two chromosomes

We consider two consecutive loci z_i and z_{i+1}, which are at distance d_i (in Morgan) that we sample in two homologous chromosomes. The ARG for these two sites in two chromosomes has seven states (see Durrett (2008), Chapter 3). To describe them, we borrow the notation from Durrett (2008) and we write (z_iz_{i+1}) to indicate that sites z_i and z_{i+1} are IBD, and notation (z_i) (or (z_{i+1})) to indicate that the ancestor to z_i (or (z_{i+1})) is only ancestor to one of the two sites. The resulting seven states are summarized in Table 1. An example of realization of this process is shown in Figure 2 (right panel).

The initial state is S^1 because in the present time we sample two different loci in two different chromosomes. The transition matrix of the ARG with two loci and a sample size 2 can be approximated, when $N \gg 1$ by

$$M^{(i)} = \begin{pmatrix} 1-\frac{1}{2N}-2d_i & 2d_i & 0 & 0 & 0 & \frac{1}{2N} & 0 \\ \frac{1}{2N} & 1-3\frac{1}{2N}-d_i & d_i & 2\frac{1}{2N} & 0 & 0 & 0 \\ 0 & 2\frac{1}{2N} & 1-4\frac{1}{2N} & 0 & 2\frac{1}{2N} & 0 & 0 \\ 0 & 0 & 0 & 1-\frac{1}{2N}-d_i & d_i & \frac{1}{2N} & 0 \\ 0 & 0 & 0 & 2\frac{1}{2N} & 1-3\frac{1}{2N} & 0 & \frac{1}{2N} \\ 0 & 0 & 0 & 0 & 0 & 1-d_i & d_i \\ 0 & 0 & 0 & 0 & 0 & \frac{1}{2N} & 1-\frac{1}{2N} \end{pmatrix}.$$

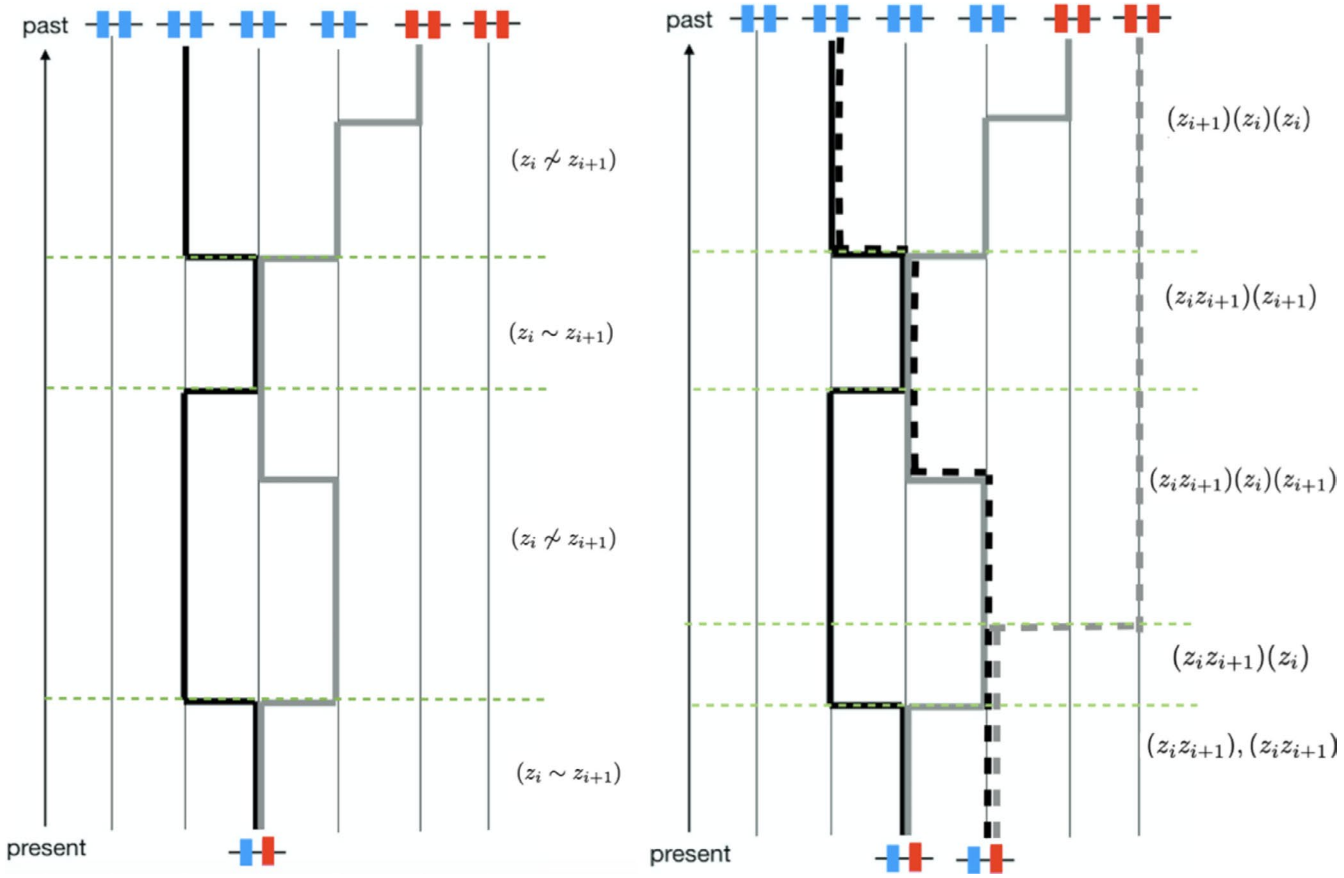


FIGURE 2 The ARG with two markers. Each colour represents one source taxa ($\mathscr{P}$ and $\mathscr{Q}$). The black and grey lines (or dotted lines) represent the ancestral lineage of each marker. In the left panel, we show the ARG for two markers in one chromosome. In the present, there is an observed junction between the two markers. In the past (t generations ago, when hybridization took place), each lineage is carried by a different individual and these two individuals are from different source taxa. The right panel shows the ARG for two markers in two homologous chromosomes

TABLE 1 States of the reduced ARG

State		n_i	n_{i+1}	n_{tot}
S^1	$(z_i z_{i+1}), (z_i z_{i+1})$	2	2	2
S^2	$(z_i z_{i+1})(z_i)(z_{i+1})$	2	2	3
S^3	$(z_i)(z_i)(z_{i+1})(z_{i+1})$	2	2	4
S^4	$(z_i z_{i+1})(z_i)$ or $(z_i z_{i+1})(z_{i+1})$	2 (or 1)	1 (or 2)	2
S^5	$(z_i)(z_{i+1})(z_{i+1})$ or $(z_{i+1})(z_i)(z_i)$	1 (or 2)	2 (or 1)	3
S^6	$(z_i z_{i+1})$	1	1	1
S^7	$(z_i), (z_{i+1})$	1	1	2

Note: n_i (resp. n_{i+1}) denotes the number of ancestors of site z_i (resp. z_{i+1}) and n_{tot} the total number of ancestors to the sample.

All other potential events (e.g. double crossovers or simultaneous crossover and coalescence events) have probabilities that are negligible compared to $1/2N$.

Let $P_t^{(i)}$ be the vector containing the probabilities of observing each of the states (S^1, …, S^7) at time t. $P_t^{(i)}$ satisfies

$$P_t^{(i)} = P_0(M^{(i)})^t,$$

where $P_0 = (1,0,0,0,0,0,0)$, since at time 0 we sample all loci in two homologous chromosomes. This equation can only be solved numerically. Recall that the stationary distribution of this process $P^{(i)}$ satisfies

$$P^{(i)} = P^{(i)} M^{(i)}$$

and has the analytical expression

$$P^{(i)} = \left(0,0,0,0,0,\frac{1}{2d_iN+1},\frac{2d_iN}{2d_iN+1}\right).$$

Thus, for large values of t the system reduces to states S^6 and S^7, which means that each locus has only one ancestor, that is forwards in time the process has reached fixation (at each locus). Recall that state S^6 is the state where there is one ancestor for the sample; thus, we observe no junctions on either chromosome. Furthermore, recall that state S^7 is the state where there are two ancestors, one for the first locus and one for the second locus, and with probability $2pq$ each one of them comes from a different source taxa. This is exactly the probability of observing a junction when $t \to \infty$ for one chromosome (Equation 2). In other words, when t is very large, fixation is reached and the two sampled chromosomes are homozygous so the problem reduces to the single chromosome case.

2.3.1 | Maximum likelihood, phased data

We first consider the case of phased data. Each pair of homologous markers can be in one of four states:

- *PP*, that is, both homologous markers carry the allele from source $\mathscr{P}$,
- *QQ*, that is, both homologous markers carry the allele from source $\mathscr{Q}$,
- *PQ*, that is the marker on the first chromosome carries the allele from source $\mathscr{P}$ and the marker on the second chromosome carries the allele from source $\mathscr{Q}$,
- *QP*, that is the marker on the first chromosome carries the allele from source $\mathscr{Q}$ and the marker on the second chromosome carries the allele from source $\mathscr{P}$.

The data can then be represented as a sequence $(O_i, 1 \le i \le n)$ that takes values in $\{PP, QQ, PQ, QP\}$ such that O_i is the state of the i-th marker. To derive a maximum-likelihood formula for the time since admixture T, we compute the probability of each sequence in $\{PP, QQ, PQ, QP\}^n$ given T, N, C, the distances between the n loci and the initial heterozygosity H_0.

We want to compute the probability of our observations $(O_1, \ldots, O_n)$. These n observations are not independent, as there are nontrivial correlations between loci along the chromosome. However, we can neglect long-range dependencies and assume that O_i only depends on O_{i-1}, that is that the probability of observing $(O_1, \ldots, O_n)$, t units of time after hybridization is

$$\mathbb{P}_t((O_1, \ldots, O_n)) = \mathbb{P}_t(O_1, O_2) \prod_{i=2}^{n-1} \mathbb{P}_t(O_{i+1} | O_i).$$

Recall that ignoring long-range dependencies is a natural approximation and it has been used for example by McVean and Cardin (2005) to define the sequentially Markov coalescent. To compute $\mathbb{P}_t(O_{i+1}|O_i)$, we use the ARG for markers at z_i and z_{i+1} denoted by (Γ_t^i) (and to compute $\mathbb{P}(O_1, O_2)$, we use (Γ_t^1)). For example, we can observe $O_1 = PP$ and $O_2 = QQ$ if:

- $\Gamma_t^1 = S^3$ and the two ancestors for locus 1 are from source taxa $\mathscr{P}$ and the two ancestors for locus 2 from $\mathscr{Q}$, which happens with probability p^2q^2 or,
- $\Gamma_t^1 = S^5$, with probability 1/2 there are two ancestors for locus 1 and one for locus 2 and they are from desired source taxa with probability p^2q. With probability 1/2, there is one ancestor for locus 1 and two for locus 2 and they are from the desired source taxa with probability pq^2 or,
- $\Gamma_t^1 = S^7$ and the ancestor to 1 is from source taxa $\mathscr{P}$ and the ancestor to 2 from $\mathscr{Q}$, which happens with probability pq.

To sum up, when $O_1 = PP$ and $O_2 = QQ$,

$$\mathbb{P}_t(O_1, O_2) = p^2q^2(P_t^{(1)})_3 + \frac{1}{2}(pq^2 + qp^2)(P_t^{(1)})_5 + pq(P_t^{(1)})_7.$$

The probabilities for all combinations of O_1 and O_2 are listed in Figure 3. To compute $\mathbb{P}_t(O_{i+1}|O_i)$ we use Bayes' formula:

$$\mathbb{P}_t(O_{i+1}|O_i) = \frac{\mathbb{P}_t(O_i, O_{i+1})}{\mathbb{P}_t(O_i)},$$

where, using the total probability theorem, $\mathbb{P}_t(O_i)$ can be obtained by summing over the appropriate row in Figure 3. Then, the total probability of observing the data, given N and t, that is

$$\mathbb{P}_t((O_1, \ldots, O_n)) = \mathbb{P}_t(O_1, O_2) \prod_{i=2}^{n-1} \frac{\mathbb{P}_t(O_i, O_{i+1})}{\mathbb{P}_t(O_i)} \tag{5}$$

can be maximized in order to find the maximum-likelihood estimator of t and N. If the data consist of multiple chromosomes from the same individual, we can calculate the joint likelihood of observing the data, given t and N by assuming independence across chromosomes and calculating the likelihood as the product across chromosomes. As in the one chromosome case, we provide estimates of T by averaging across the different individuals of the sample, and the CIs reflect the 95% interquartile range of the T estimates across the different individuals of the sample.

2.3.2 | Maximum likelihood, unphased data

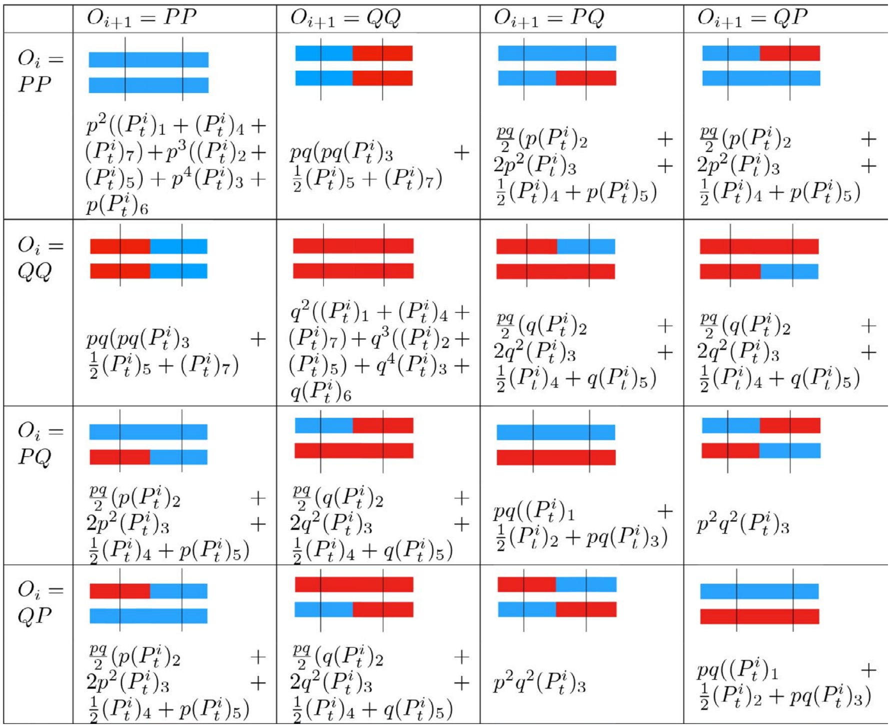
If the data are unphased, we cannot distinguish which allele is in which of the two homologous chromosomes. We can observe one of these three states at each marker:

- *P*, that is we only observe the allele from source $\mathscr{P}$, that is both chromosomes carry the allele from source $\mathscr{P}$,
- *Q*, that is we only observe the allele from source $\mathscr{Q}$.
- *x*, that is we observe both alleles, that is each one of the two homologous chromosomes carries a different allele.

The data can then be represented as a sequence (O_i) of length n that takes values in $\{P,Q,x\}$ such that O_i is the state of the i-th marker. We can perform exactly the same method, as in the previous section, except that now the probabilities of each state are given by Figure 4. Again, if the data consist of multiple chromosomes from the same individual, we can calculate the joint likelihood of observing the data, given t and N by assuming independence across chromosomes and calculating the likelihood as the product across chromosomes.

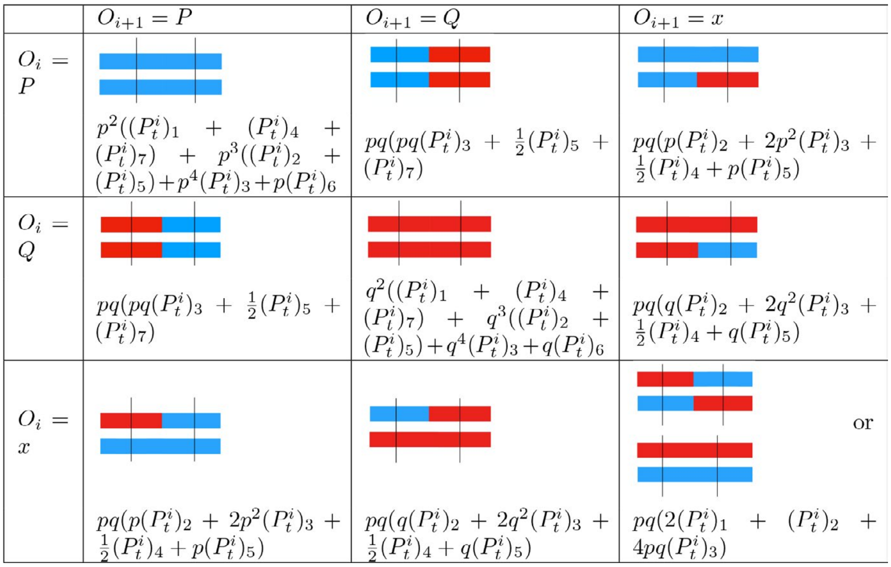
2.4 | Individual-based simulations

To test the validity of our maximum-likelihood approach, we use individual-based simulations, as described in Janzen et al. (2018), that is Wright–Fisher type simulations of randomly mating populations of constant size N, with nonoverlapping generations. We



	$O_{i+1} = PP$	$O_{i+1} = QQ$	$O_{i+1} = PQ$	$O_{i+1} = QP$
$O_i = PP$	$p^2((P_t^i)_1 + (P_t^i)_4 + (P_t^i)_7) + p^3((P_t^i)_2 + (P_t^i)_5) + p^4(P_t^i)_3 + p(P_t^i)_6$	$pq(pq(P_t^i)_3 + \frac{1}{2}(P_t^i)_5 + (P_t^i)_7)$	$\frac{pq}{2}(p(P_t^i)_2 + 2p^2(P_t^i)_3 + \frac{1}{2}(P_t^i)_4 + p(P_t^i)_5)$	$\frac{pq}{2}(p(P_t^i)_2 + 2p^2(P_t^i)_3 + \frac{1}{2}(P_t^i)_4 + p(P_t^i)_5)$
$O_i = QQ$	$pq(pq(P_t^i)_3 + \frac{1}{2}(P_t^i)_5 + (P_t^i)_7)$	$q^2((P_t^i)_1 + (P_t^i)_4 + (P_t^i)_7) + q^3((P_t^i)_2 + (P_t^i)_5) + q^4(P_t^i)_3 + q(P_t^i)_6$	$\frac{pq}{2}(q(P_t^i)_2 + 2q^2(P_t^i)_3 + \frac{1}{2}(P_t^i)_4 + q(P_t^i)_5)$	$\frac{pq}{2}(q(P_t^i)_2 + 2q^2(P_t^i)_3 + \frac{1}{2}(P_t^i)_4 + q(P_t^i)_5)$
$O_i = PQ$	$\frac{pq}{2}(p(P_t^i)_2 + 2p^2(P_t^i)_3 + \frac{1}{2}(P_t^i)_4 + p(P_t^i)_5)$	$\frac{pq}{2}(q(P_t^i)_2 + 2q^2(P_t^i)_3 + \frac{1}{2}(P_t^i)_4 + q(P_t^i)_5)$	$pq((P_t^i)_1 + \frac{1}{2}(P_t^i)_2 + pq(P_t^i)_3)$	$p^2q^2(P_t^i)_3$
$O_i = QP$	$\frac{pq}{2}(p(P_t^i)_2 + 2p^2(P_t^i)_3 + \frac{1}{2}(P_t^i)_4 + p(P_t^i)_5)$	$\frac{pq}{2}(q(P_t^i)_2 + 2q^2(P_t^i)_3 + \frac{1}{2}(P_t^i)_4 + q(P_t^i)_5)$	$p^2q^2(P_t^i)_3$	$pq((P_t^i)_1 + \frac{1}{2}(P_t^i)_2 + pq(P_t^i)_3)$

FIGURE 3 $\mathbb{P}_t(O_i, O_{i+1})$ for phased data. The allele from source taxa $\mathscr{P}$ is represented in blue and the allele from source taxa $\mathscr{Q}$ is represented in red



	$O_{i+1} = P$	$O_{i+1} = Q$	$O_{i+1} = x$
$O_i = P$	$p^2((P_t^i)_1 + (P_t^i)_4 + (P_t^i)_7) + p^3((P_t^i)_2 + (P_t^i)_5) + p^4(P_t^i)_3 + p(P_t^i)_6$	$pq(pq(P_t^i)_3 + \frac{1}{2}(P_t^i)_5 + (P_t^i)_7)$	$pq(p(P_t^i)_2 + 2p^2(P_t^i)_3 + \frac{1}{2}(P_t^i)_4 + p(P_t^i)_5)$
$O_i = Q$	$pq(pq(P_t^i)_3 + \frac{1}{2}(P_t^i)_5 + (P_t^i)_7)$	$q^2((P_t^i)_1 + (P_t^i)_4 + (P_t^i)_7) + q^3((P_t^i)_2 + (P_t^i)_5) + q^4(P_t^i)_3 + q(P_t^i)_6$	$pq(q(P_t^i)_2 + 2q^2(P_t^i)_3 + \frac{1}{2}(P_t^i)_4 + q(P_t^i)_5)$
$O_i = x$	$pq(p(P_t^i)_2 + 2p^2(P_t^i)_3 + \frac{1}{2}(P_t^i)_4 + p(P_t^i)_5)$	$pq(q(P_t^i)_2 + 2q^2(P_t^i)_3 + \frac{1}{2}(P_t^i)_4 + q(P_t^i)_5)$	or $pq(2(P_t^i)_1 + (P_t^i)_2 + 4pq(P_t^i)_3)$

FIGURE 4 $\mathbb{P}_t(O_i, O_{i+1})$ for unphased data. The allele from source taxa $\mathscr{P}$ is represented in blue and the allele from source taxa $\mathscr{Q}$ is represented in red

then recover local ancestry by analysing ancestry at *n* markers whose positions are chosen uniformly at random along the genome.

Unless otherwise specified, we show how time can be accurately inferred for a population of 10,000 individuals, for time points between the first generation and 1000 generations. We use $n = 10{,}000$ markers, which is considered to be sufficient to detect the majority of accumulated junctions (Janzen et al., 2018). We report our findings across 100 replicates, where in each replicate 10 individuals were randomly selected from the admixed population and used to infer *T*. We report the mean and the 95% percentile across these 100 replicates. The number of individuals in the sample for the real data sets analysed is always larger than 10 (see Section 4). We have simulated with three different values of the initial proportion of source taxa $\mathscr{P}$, ($p \in \{0.053, 0.184, 0.5\}$), to vary the initial heterozygosity H_0 in {0.1,0.3,0.5}.

We have first compared the methods we have developed here to previous methods based on the theory of junctions (Figure 5). We observe that, when the number of markers is low, previous methods, that do not take into account marker spacing, tend to underestimate the time since admixture, which is not the case for our methods.

We have then compared the estimations of the time since admixture, using the method for one chromosome and the method for two chromosomes (phased; Figure 6). We observe that using data from the two homologous chromosomes allows to infer the time since admixture more accurately, since it reduces uncertainty.

Using unphased data instead of phased data might introduce additional error and in Figure 7, we compare the methods that use

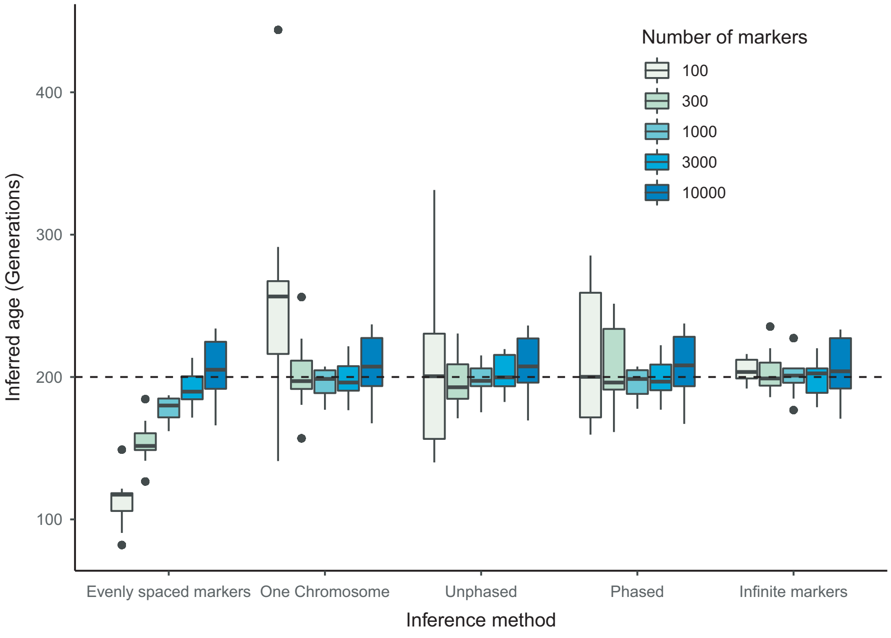


FIGURE 5 Comparison to previous methods. Shown are the median estimates for the time since admixture for 100 replicates, where in each replicate 10 individuals were analysed. Boxplots represent the 95% interquantile range across replicates. The dashed line indicates the simulated time. "Evenly spaced markers" corresponds to the method in Janzen et al. (2018). "Infinite markers" corresponds to an idealized scenario where ancestry is known for every locus in the chromosome and is there to quantify the amount of randomness in the process. The population size was 10,000 individuals, and 10,000 randomly spaced markers were used

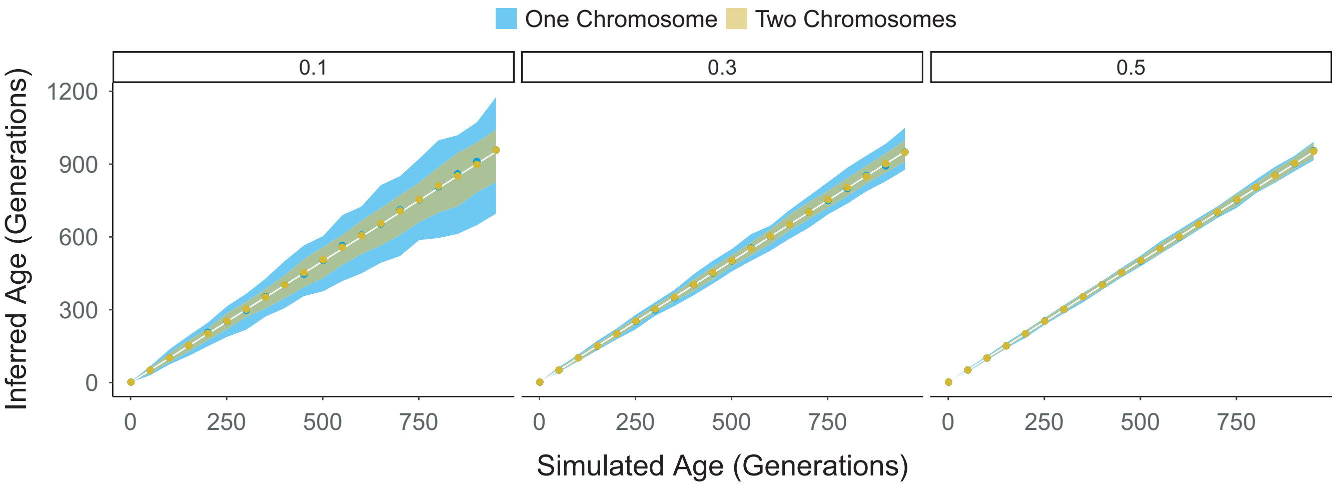


FIGURE 6 Accuracy in age estimate using information from one versus two chromosomes. Inferred time versus simulated time is represented. Shown are the median estimates (dots) for 100 replicates, where in each replicate 10 individuals were analysed. The solid white line indicates the observed is equal to expected line and the shaded area indicates the 95% percentile range across replicates. Shown are results using junction information from one chromosome (blue) and results using phased information from two chromosomes (gold). Numbers above the plots indicate the initial heterozygosity. The population size was 10,000 individuals, and 10,000 randomly spaced markers were used

phased or unphased information of two homologous chromosomes. We observe that both methods yield very similar results in terms of the relative error. This can be due to the fact that homozygous sites have an important contribution to the likelihood and the uncertainty that comes from sites that are of type *x* (in the unphased case) is well managed by our method.

Finally, we explore error in phasing assignment (switching error). We simulate the effect of error in phasing assignment by randomly swapping a fraction of the markers between chromosomes. We explore phasing error in {0.0025,0.005,0.0075,0.01,0.02}. These errors are comparable to the switching error rates reported in the literature. For example, Choi et al. (2018) compared different phasing methods and reported switching error rates between 0.1% and 2%. (Notice that these error rates are for human data where there are good quality references and sample sizes are large). More recent reference-free methods (based on third-generation sequencing techniques) report switching error rates of 1–2% (see, e.g. Ebler et al., 2019; Kronenberg, 2019; Tourdot & Zhang, 2021). Switching error rate error has strong effects on the inferred time since admixture, as shown in the bottom panel in Figure 8. Generally, imposed errors increase the inferred age, by introducing novel junctions due to mis-phased markers.

Another important source of error is the lack of coverage, which would have the effect of reducing the number of markers. An analysis of the sensitivity of our method to reducing the number of markers can

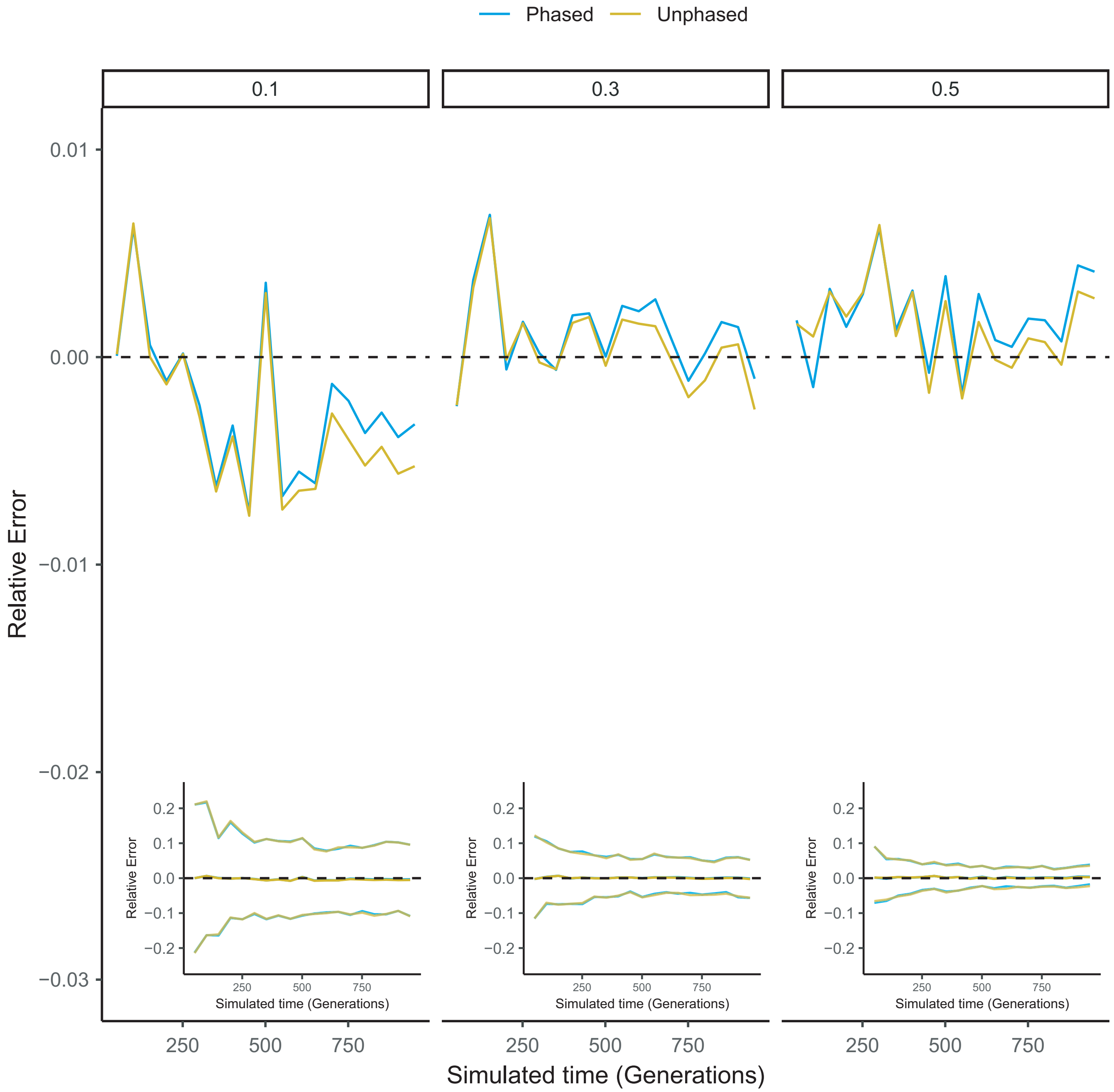


FIGURE 7 Accuracy in age estimate using the unphased framework versus the phased framework. Shown are the median difference across 100 replicates, where for each replicate 100 individuals were analysed. We represent the results for three different initial heterozygozities, as indicated at the top of each plot. The population size was 10,000 individuals, and 10,000 randomly spaced markers were used. The inset plots show the same results, including the 95% percentile, which are far outside the boundaries of the main plot

be found in the S3 Appendix and shows that a reduction of coverage also tends to lead to an overestimation of the time since admixture.

3 | COMPARISON WITH OTHER METHODS

Our framework provides a methodology to infer the time since admixture, provided that local ancestry is known. Previous methods aiding in inferring the time since admixture, such as ELAI (Guan, 2014) and ANCESTRY HMM (Corbett-Detig & Nielsen, 2017) jointly infer the time since admixture and local ancestry, because they use the time since admixture to correct for the impact of recombination on local ancestry estimates. We have chosen to compare our framework with results obtained using ANCESTRY HMM, which we currently consider the most accurate method available to jointly infer local ancestry and time since admixture. We do so in a twofold manner: first, we directly compare the ability of both

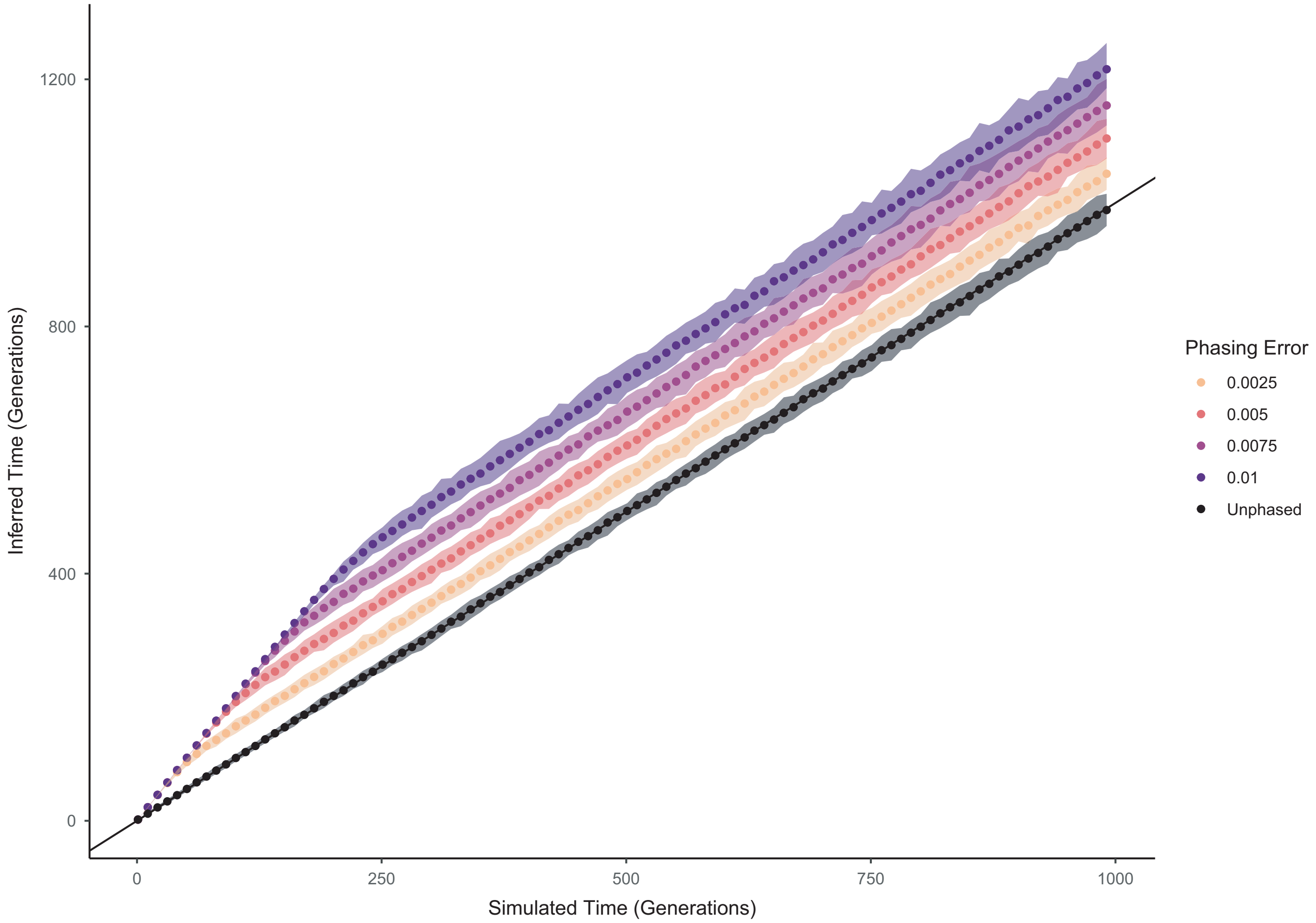


FIGURE 8 Effect of switching error on the estimated time since admixture. Data simulated with $N = 10{,}000$, $p = 0.5$, $C = 1$ and $n = 10{,}000$. The solid black line indicates the simulated = estimated time. Dots indicate the mean estimate across 100 replicates, where in each replicate, 10 individuals were analysed. Coloured area indicates 95% interquantile distribution across replicates. Colours reflect different degrees of phasing error, where a phasing error of 0.01 represents a 1% probability of a SNP being phased incorrectly

methods to infer the time since admixture, provided data with known local ancestry. Second, we provide both packages with imperfect data, where local ancestry is uncertain. Because our framework is not able to infer local ancestry, we directly use the inferred local ancestry of ANCESTRY HMM, which might add an extra layer of error, but is a good reflection of an empirical use case (see also Section 4.3). For both approaches (known and uncertain ancestry), data were simulated using the junctions package, using a population size N of [1000, 10,000] and imposing n markers, where n was [1000, 10,000, 40,000] (in line with the number of markers available in the Swordtail and *Populus* data sets in Section 4). Simulations were run for 10,000 generations, using $H_0 = 0.5$ and C = 1. At each generation, time since admixture was inferred for 10 individuals sampled randomly from the population. Time since admixture was inferred given the superimposed n markers, whose location was drawn randomly in [0, 1]. For each parameter combination, 10 replicate simulations were performed. When the maximum likelihood did not converge to a final value within the given range, the data point was removed from the analysis. This only occurred when $t > N$, and typically only when there were very few ancestry informative markers.

3.1 | Known ancestry

Because ANCESTRY HMM requires information on allele frequencies in the source taxa in order to jointly infer local ancestry and time since admixture, input data for ANCESTRY HMM were created assuming both source taxa were fully separated; for example, two alleles were differentially fixed across the source taxa. Results show (Figure 9) that when time is small compared to population size (i.e. when the admixture time is small in “coalescence units”), our method performs as good as ANCESTRY HMM. However, for small population sizes, we observe that, as the number of generations since admixture reaches N, ANCESTRY HMM becomes increasingly incorrect in contrast to our method, which remains accurate. For large values of N, in our simulation, our method performs as good as ANCESTRY HMM. For computational reasons, we could not simulate times exceeding N, but we

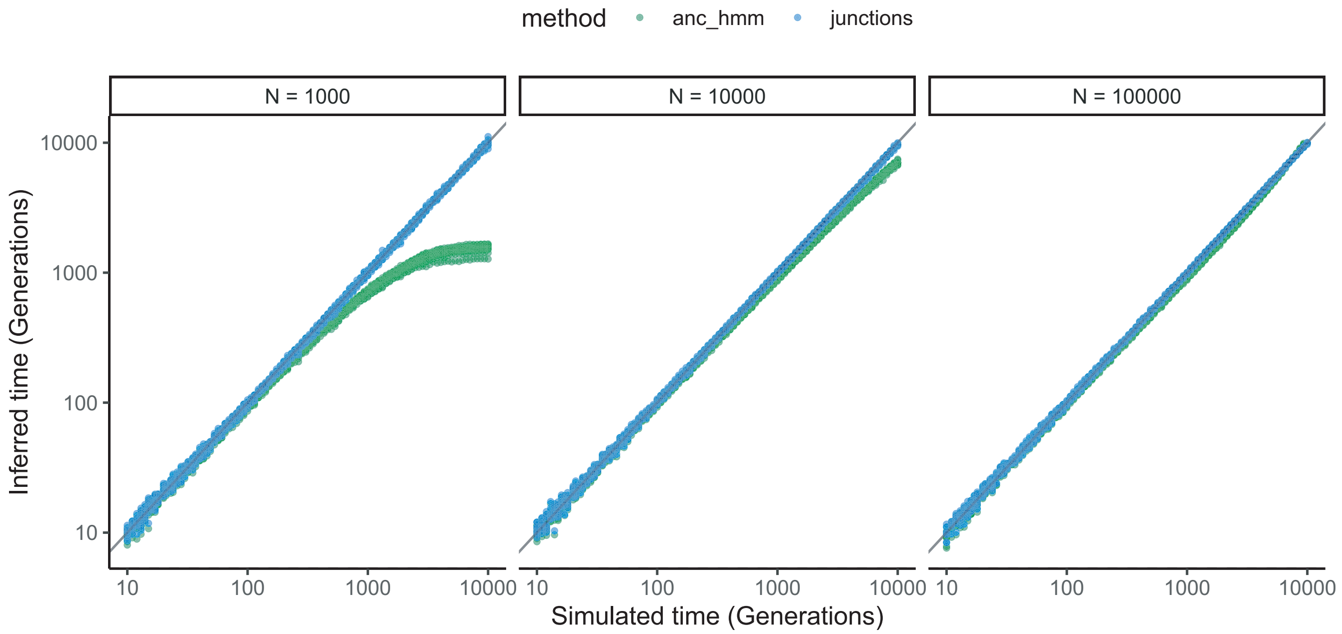


FIGURE 9 Comparison in estimating the time since admixture between ANCESTRY HMM and the method proposed here. The solid black line indicates the simulated = estimated time. Dots indicate the inferred ages, with the green dots representing ages inferred by ANCESTRY HMM, and the blue dots indicate ages inferred by the junctions framework. Age estimates are based on simulated data with known ancestry, using $n = 40{,}000$, $C = 1$, $H_0 = 0.5$

expect the performance of ANCESTRY HMM to decrease when the time since admixture reaches N generations.

3.2 | Uncertain ancestry

To mimic uncertainty in ancestry, we explored different degrees of uncertainty, indicated by different values of the allele frequency differential $d_x = A_x - A_y$, where A_x is the frequency of allele A in population x and A_y is the frequency of allele A is population y, following Shriver et al. (1997). Higher values of d_x indicate higher ancestry information content in the respective marker. Here, we explored values of d_x in [0.5, 0.7, 0.9]. Because our framework does not infer local ancestry, we used the inferred local ancestry by ANCESTRY HMM, using a threshold of including only markers with at least 95% certainty of ancestry. For a large population size ($N = 10{,}000$, S5 Appendix), we observe that ANCESTRY HMM consistently underestimates the true age when the time since admixture becomes large, in line with our findings with known ancestry (Figure 9). In contrast, our method behaves differently, underestimating the true age even further, most likely as a result of a stacking of errors: because our method cannot infer local ancestry, we have used the local ancestry estimates of ANCESTRY HMM. However, when population size is small ($N = 1000$, Figure 10), we observe a striking phenomenon (Figure 10): when the time since admixture becomes large, our framework starts inferring the time correctly again, whereas the age estimates obtained using ANCESTRY HMM remain incorrect and seem to reach a plateau. This effect only occurs when there is a sufficient number of markers of reasonable quality, as we observe that when $n = 1000$ and $d \leq 0.7$, our framework is also unable to accurately infer the time since admixture. We believe that the improved performance when time since admixture becomes large is the result of the interplay between two phenomena. First, while our method can work for any value of T, ANCESTRY HMM typically requires $T \ll log_2(N)$ (this is the same assumption used in SMC' (Liang & Nielsen, 2014)). This could introduce some error in the ancestry assignment and thus lead to the underestimates obtained using ANCESTRY HMM. Second, our method seems to work better with a small number of informative markers. Combined, our method outperforms ANCESTRY HMM in inferring the time since admixture when population size and number of markers are small.

4 | RESULTS

4.1 | Saccharomyces cerevisiae

Experimental evolution provides an important reference point to verify our findings. Here, we reanalyse data from an advanced intercross line (AIL) experiment, where two highly differentiated yeast (*Saccharomyces cerevisiae*) lines were crossed, and the resulting hybrid offspring was outbred for 12 generations in order to obtain maximum genetic diversity (Illingworth et al., 2013; Parts et al., 2011). The data consist of sequencing data for 171 individuals, for all 16 chromosomes. There are on average 3271 ancestry informative markers per chromosome (95% CI: [929, 6284]). Local ancestry was certain in the data, as both source taxa were differentially

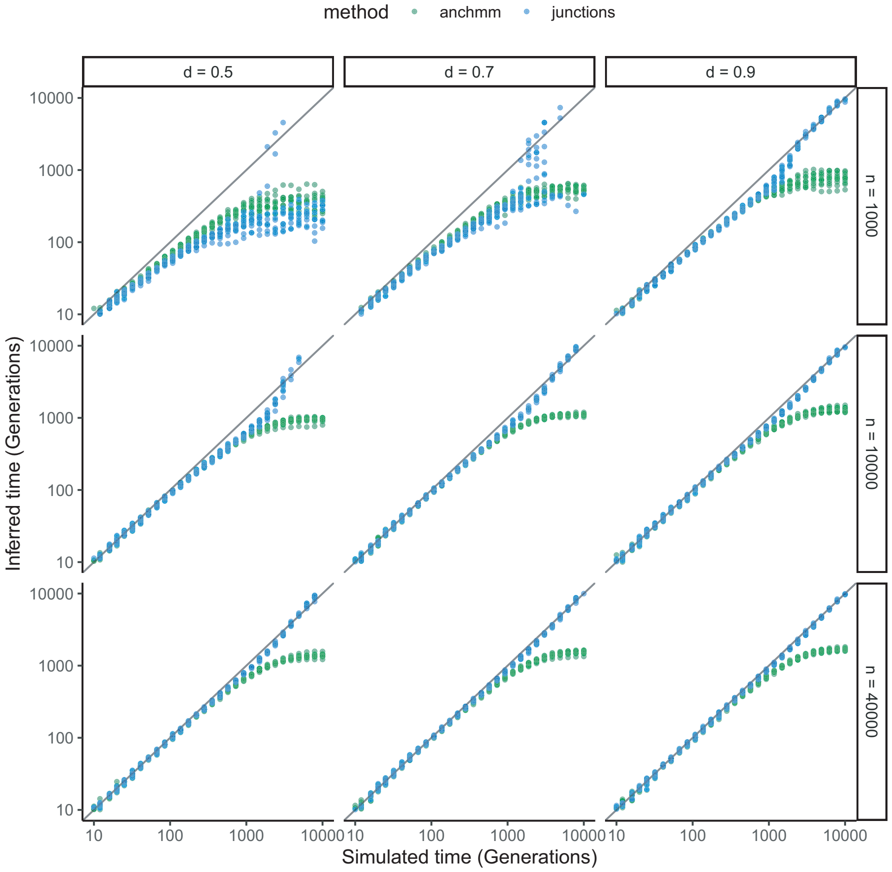


FIGURE 10 Comparison in estimating the time since admixture between ANCESTRY HMM and the method proposed here, for a small population of $N = 1000$. The solid black line indicates the simulated = estimated time. Dots indicate the inferred ages, with the green dots representing ages inferred by ANCESTRY HMM and the blue dots indicate ages inferred by the junctions framework. Missing dots are caused by a lack of convergence of the maximum-likelihood algorithm. Age estimates are based on simulated data with uncertain ancestry, where uncertainty in ancestry is reflected by the allele frequency differential (Shriver et al., 1997). Because the method proposed here does not include ancestry uncertainty, local ancestry as inferred by ANCESTRY HMM was used

homozygous on each ancestry informative SNP. H_0 was 0.5, reflecting a 50/50 contribution of both strains to the first generation. We used three different recombination rate estimates: first, we used the linkage map of Cherry et al. (1997) where the average recombination rate is 1 cM/2.7 kb (1 centi Morgan per 2.7 kilobases); second, we used the average recombination rate of 1 cM/2.2 kb as inferred in Mancera et al. (2008); and lastly, we used the average recombination rate of 1 cM/5.8 kb as inferred for the two-way cross in Illingworth et al. (2013). In the absence of a detailed recombination map, we assume that recombination is constant across the chromosome, ignoring hot spots and cold spots. We assume a large population size ($N = 100{,}000$), reflecting outbreeding.

We find that when using the older recombination rate estimates, we consistently underestimate the age of the hybrids (see Figure 11, panel a). The mean age using the recombination rate from Cherry et al. (1997) corresponds to 9.1 generations (95% CI across all individuals = [6.2, 11.3]) and the mean age using the recombination rate from Mancera et al. (2008) corresponding to 7.0 generations (95% CI: [5.4, 8.6]). These estimates suggest that the true recombination rate is slightly lower than assumed. When using the most recent recombination rate estimate (i.e. 1 cM/5.8 kb, from Illingworth et al. (2013)), we slightly overestimate the age (mean age estimate: 18.7 generations, 95% CI: [13.0, 22.5]). Alternatively, we could be overestimating population size, suggesting that perhaps the rate of inbreeding in the experimental design was higher than anticipated. Reducing the population size tends to increase the time since admixture (S4 Appendix), which could help bring the age estimates using the two oldest recombination rates closer to the true number of generations. However, for the most recent recombination rate estimate, this would only increase the error made. The likelihood profiles (see Figure 11, panel b) suggest that the maximum-likelihood estimates are robust, and thus, it seems more likely that the recombination rate estimates do not accurately reflect the amount of recombination accumulated in the experiment.

4.2 | Swordtail fish

Here, we reanalyse data of hybridizing swordtail fish published in Schumer et al. (2018). Swordtail fish have received considerable attention in the past years, as they have been shown to hybridize readily in nature. We focus here on a hybrid population located in Tlatemaco, Mexico (Schumer, Cui, et al., 2014; Schumer et al., 2018). The population is the result of a hybridization event between *Xiphophorus birchmanni* and *X. malinche*, approximately 100–200 generations ago (Pers. Comm. M. Schumer and (Powell, Moran, et al., 2021; Schumer et al., 2018)). Currently, the hybrid genome consists for 75% of *X. malinche*, suggesting that the initial hybrid swarm was strongly biased towards *X. malinche*, or that strong selection after hybridization has favoured genomic material from *X. malinche*. We use ancestry information provided in the data supplement of Schumer et al. (2018), which contains unphased local ancestry estimates based on multiplexed shotgun genotyping (MSG) results (Andolfatto et al., 2011), with on average

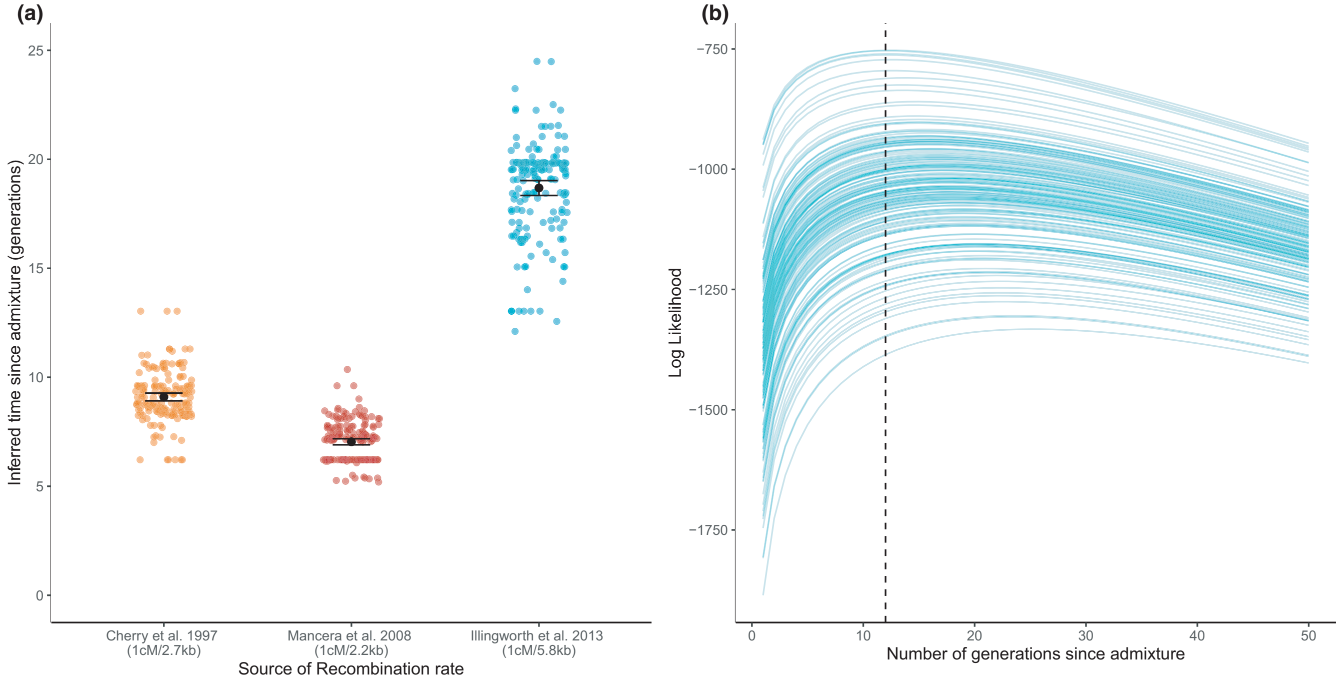


FIGURE 11 Inferred age for F12 hybrid yeast (*Saccharomyces cerevisiae*) individuals. (a) Inferred age for three different recombination rates: 1 cM/2.7 kb (Cherry et al., 1997), 1 cM/2.2 kb (Mancera et al., 2008) and 1 cM/5.8 kb (Illingworth et al., 2013). Shown is the distribution of inferred ages across 171 individuals. Solid black dots indicate the bootstrapped average across all individuals; black error bars indicate the 95% CI of these bootstraps. (b) Loglikelihood profile across all chromosomes, where each line represents one individual. Shown is the loglikelihood profile for the most recent recombination rate estimate, from Illingworth et al. (2013). The vertical dotted line indicates the 12 generations line (e.g. the true age of the F12 hybrid individuals)

38,936 markers per chromosome (95% CI: [22,765, 50,256]). The MSG pipeline provides a posterior probability of observing local ancestry. Following Schumer et al. (2018), we converted local ancestry probabilities of >95% to hard ancestry calls. To obtain age estimates, we use the estimated population size in Schumer, Cui, et al. (2014) of 1830 individuals. We infer the age for 187 individuals from the Tlatemaco population. We use ancestry information from 24 linkage groups, but remove linkage groups 17 and 24, as these are known to include large inversions (Schumer et al., 2018), making them unsuitable for admixture analysis. As a recombination map, we use three approaches. First, we use the average recombination rate of 1 cM/378 kb as used in Schumer, Cui, et al. (2014), which is based on the average genome-wide recombination rate in *Xiphophorus* (Walter et al., 2004). Second, we use the average recombination rate of 1 cM/500 kb as reported in Powell, García-Olazábal, et al. (2020). Lastly, we use the high-density recombination map reconstructed from linkage disequilibrium patterns as presented in Schumer et al. (2018), which represents an average recombination rate of 1 cM/485 kb.

We find that the distribution of ages inferred for individuals from the Tlatemaco population is overall higher than the previously inferred age but still consistent with those estimates (see Figure 12a). We recover a mean age of 163 generations (95% CI: [81, 201]) when using the recombination rate reported in Schumer et al. (2014). Using the high-density recombination map from Schumer et al. (2018), we obtain a mean age estimate of 210 generations (95% CI: [111, 257]), due to the shorter map length. Alternatively, using the most recent recombination rate estimate of 1 cM/500 kb reported by Powell, García-Olazábal, et al. (2020), we recover a mean age of 217 generations (95% CI: [115, 265]). Thus, we generally find that the population is perhaps slightly older than expected, but we would like to emphasize that the likelihood profile across all samples (Figure 12b) is extremely flat, suggesting that the age estimates obtained tend to be uncertain and susceptible to potential inconsistencies in the data (e.g. sequence errors). Exploration of different values for the population size (S4 Appendix) shows that if the true population size is larger than reported by Schumer, Cui, et al. (2014), this would decrease the time since admixture, but only minimally so (by only a few generations).

4.3 | *Populus* trees

Here, we reanalyse a data set of *Populus* trees, published by Suarez-Gonzalez et al. (2016). The data set focuses on two species of trees, *Populus trichocarpa*, found mainly in West America, and *Populus balsamifera*, which is found in Northern America. The two species are thought to have diverged relatively recently, around 760k years ago. Where their ranges meet (around the southern tip of Alaska), the two species hybridize, and a hybrid population has been established. The data set consists of 32 individuals which are mainly *P. balsamifera*, admixed with *P. trichocarpa* and

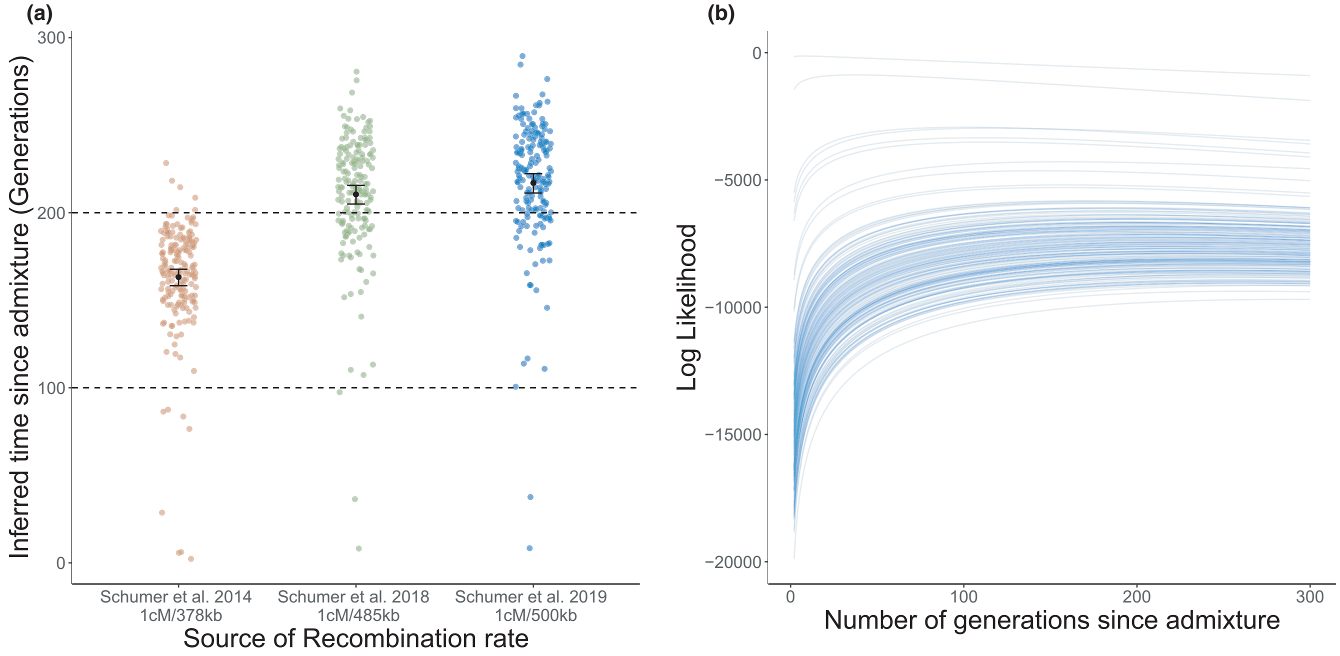


FIGURE 12 Inferred age for hybrid *Swordtail* fish (a) Inferred age for hybrid *Xiphophorus* fish from Tlatemaco (Mexico). Shown is the distribution age inferences across 187 individuals, based on three different recombination maps: 1 cM/370 kb (Schumer, Cui, et al., 2014), 1 cM/485 kb (Schumer et al., 2018) and 1 cM/500 kb (Powell, García-Olazábal, et al., 2020). Dotted lines indicate the hypothesized age limits of the admixed population. Solid black dots indicate the bootstrapped average across all individuals; black error bars indicate the 95% CI of these bootstraps. (b) Loglikelihood profile across all chromosomes, where each line represents one individual. Shown is the loglikelihood profile for the most recent recombination rate estimate (Powell, García-Olazábal, et al., 2020)

36 individuals that are mainly *P. trichocarpa*, admixed with *P. balsamifera*. Three chromosomes of interest (chromosomes 6, 12 and 15) were Illumina-sequenced, and unphased data were available for on average 60,071 ancestry informative markers per chromosome (95% CI: [28,745, 101,425]). Ancestry information in the markers was very high, with on average 40,852 markers with an allele frequency differential of at least 0.7 (Shriver et al., 1997), and 21,009 markers with an allele frequency differential of at least 0.9. We use three different population-level recombination rates recovered from the literature, being $\rho = 0.00219$ (Wang et al., 2016), $\rho = 0.0092$ (Olson et al., 2010) and $\rho = 0.0197$ (Slavov et al., 2012). We converted these population-level recombination rates to individual rates using an effective population size of 5106 individuals, as estimated using phylogenetic methods in Slavov et al. (2012). This yielded three local recombination rates of 1 cM/10.4 kb (Slavov et al., 2012), 1 cM/22.2 kb (Olson et al., 2010) and 1 cM/93.3 kb (Wang et al., 2016). Local ancestry was determined using ANCESTRY HMM (Corbett-Detig & Nielsen, 2017), assuming equal admixture of both source taxa. Because admixture differed strongly across samples, we used the average local ancestry per sample as input for a second run of ANCESTRY HMM in order to obtain accurate local ancestry calls. We converted local ancestry probabilities of >95% to hard ancestry calls. Lastly, we compared the observed variation in ancestry across samples with the expected variation in ancestry for a single panmictic population, as given by equation 8 in Gravel (2012).

We find that the time since admixture strongly correlates with the recombination rate used (see Figure 13a), with a mean number of generations since admixture of 14 (95% CI: [13, 15]) when using the highest estimate of recombination (1 cM/10.4 kb (Slavov et al., 2012)), an intermediate estimate of 30 generations (95% CI: [29, 33]) when using a recombination rate of 1 cM/22.2 kb (Olson et al., 2010) and a much higher age estimate of 123 generations (95% CI: [131, 10]) when using the lowest recombination estimate of 1 cM/93.3 kb (Wang et al., 2016). The likelihood profiles (see Figure 13b) intersect on multiple occasions, suggesting multiple optima. We measure a variation in ancestry across the three chromosomes of 0.042. However, we obtain expected levels of variation of ancestry of 0.0088, 0.0010 and 0.0008 for the three different recombination rate estimates (and their corresponding age estimates). Observed variation in the data is thus much higher than expected from admixture in a single panmictic population.

5 | DISCUSSION

The aim of this article was to improve the estimation of the time since admixture in hybrid populations. To do so, we have extended the theory of junctions in two directions. First, we have derived a formula for the expected number of observed junctions in one chromosome that takes into account the number of markers and their positions (Equation (3)). Second, we have considered

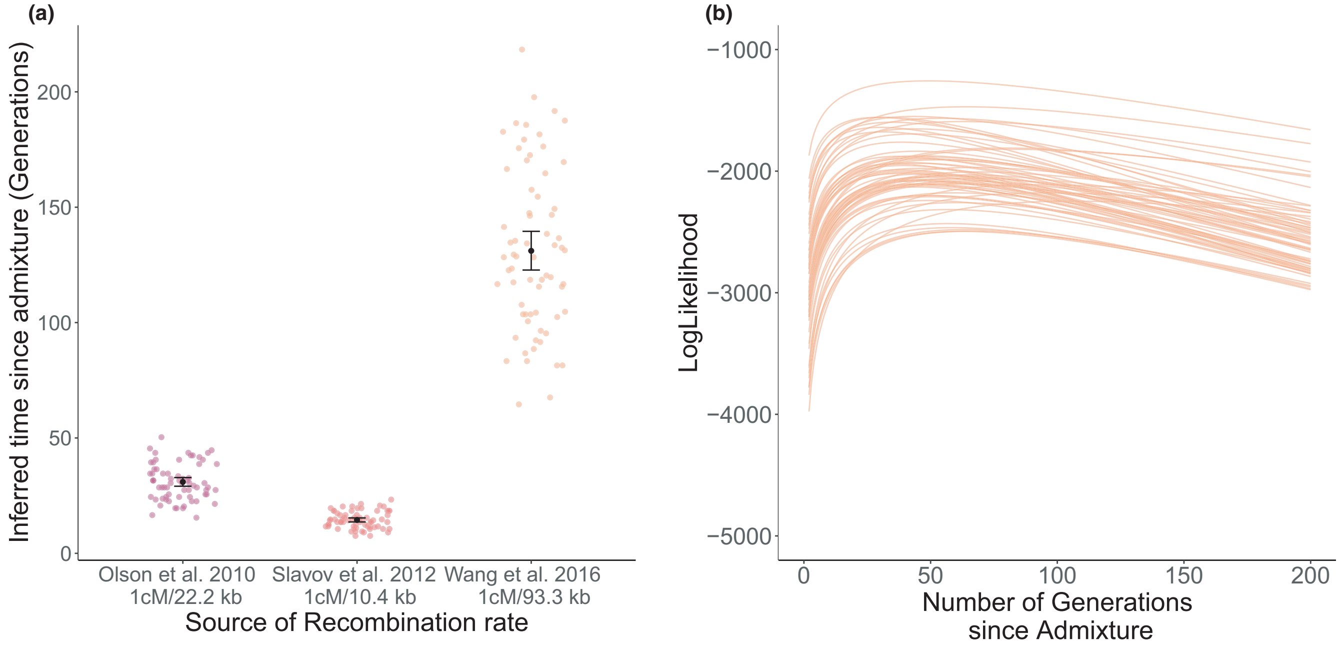


FIGURE 13 Inferred age for hybrid *Populus* trees. (a) Distribution of inferred time since admixture for 68 individuals. Colours indicate different recombination rates used: 1 cM/10.4 kb (Slavov et al., 2012), 1 cM/22.2 kb (Olson et al., 2010) and 1 cM/93.3 kb (Wang et al., 2016). Solid black dots indicate the bootstrapped average across all individuals; black error bars indicate the 95% CI of these bootstraps. (b) Inferred time since admixture, split out across the average frequency of *Populus trichocarpa* in the admixed individual. Loglikelihood profile across all chromosomes, where each line represents one individual. Shown is the loglikelihood profile for the most recent recombination rate estimate (Wang et al., 2016)

the case in which there is sequencing data from two homologous chromosomes. We have developed a maximum-likelihood approach that infers the time since admixture, for both phased and unphased data.

First, we have used simulations to validate the accuracy of our method. Results from our simulations show that our method for a single chromosome performs better than previous methods that ignore the effect of having a limited number of markers or assume that the markers are even-spaced (see Figure 5). Furthermore, we find that using information from two chromosomes improves accuracy considerably, as expected. This effect is even stronger for small population sizes (see figures 6 and 2 in S2 Appendix).

Surprisingly, the method based on unphased data performs similarly to our method based on phased data (see figures 7 and 3 in S2 Appendix). The phased and unphased approaches differ in their treatment of markers that are heterozygous for ancestry, and hence, we expected differences between these methods to manifest themselves primarily during the initial stages of admixture, when heterozygosity is still high. We did find that there were slight differences during these stages (figures 7 and 3 in S2 Appendix), but these were negligible compared to the overall uncertainty. Furthermore, these errors were much smaller than errors accumulated due to incorrect phasing (see Figure 8). Our findings here are conservative, as we show that the unphased method performs better even for small error rates, comparable to error rates for human data (e.g. in Choi et al., 2018). Human data sets are typically of very high quality, and these error rates represent an extremely favourable scenario. Thus, given the impact of error rates incurred during phasing, we recommend using our unphased framework if possible, to obtain more accurate time estimates.

Apart from sensitivity to phasing error, we have tested the sensitivity of our method to different parameters such as the number of markers n, the population size N, the initial heterozygosity H_0 and the total recombination rate C (see S1 Appendix). Our method seems to be quite sensitive to H_0 but this parameter can easily be estimated from the proportion of markers that come from each source taxa. One advantage of our approach is that age inference is not very sensitive to population size (see figure 1 in S1 Appendix), which was not true for previous methods that rely on a good estimation of N (see Janzen et al. (2018)). Our method is not very sensitive either to the number of markers (see figure 4 in S1 Appendix), provided that it is above a certain threshold. Janzen et al. (2018) inferred that when using regularly spaced markers and information for a single chromosome, the number of markers typically needs to be an order of magnitude larger than $\frac{1}{2}Ct$, where t is the admixture time and C the total amount of recombination. We find similar results when using information from a single chromosome with arbitrarily spaced markers or information from both chromosomes (see S1 Appendix). When analysing empirical data, it is often impossible to know *a priori* whether the number of ancestry informative markers is much larger than the admixture time. However, our simulation results indicate that when the number of markers is too small, variation in the age estimate across different chromosomes tends to increase. Thus, large variation in the estimate of admixture time, or inferred admixture times that tend to extremely large values, are potentially indicative of an insufficient marker number.

The main issue with our method is its sensitivity to the recombination rate. This is shown in figure 2 of S1 Appendix but also exemplified by the varying results in the empirical data sets, dependent on our assumptions about recombination rates. However, it should be noted that this issue is not novel to our approach, but is a general issue with the theory of junctions. Apart from sensitivity to the average recombination rate, local hot spots or cold spots of recombination could potentially also influence admixture time estimates.

Furthermore, the recombination rate factors into inference of local ancestry. Our methodology assumes local ancestry to be known and relies on upstream inference of local ancestry. However, methods to infer local ancestry such as MSG (Andolfatto et al., 2011), ELAI (Guan, 2014) and ANCESTRY HMM (Corbett-Detig & Nielsen, 2017) also use recombination in their calculations to infer local ancestry. Thus, by using these local ancestry estimates, our methodology reuses the recombination rate (both for local ancestry inference, and subsequent age estimation). Any errors in the recombination rate estimate might then propagate through the pipeline and affect age estimates using our framework. Using simulations including ancestry uncertainty, we have looked into this effect of stacking errors and have found that our method performs similarly to ANCESTRY HMM, even if we use the local ancestry inferred by ANCESTRY HMM. Furthermore, our method outperforms ANCESTRY HMM when the number of generations since admixture is small compared to the population size (i.e. when time in "coalescence units" is small, where one coalescence unit corresponds to $2N_e$ generations, which reflects the number of allele copies in a diploid population of effective size N_e, which is reflected by our parameter of population size N). ANCESTRY HMM assumes $T \ll \log_2(N)$ (this is the same assumption used in SMC' (Liang & Nielsen, 2014)), while our framework does not need to make these assumptions to obtain age estimates. However, when marker numbers dwindle (S5 Appendix), our framework can no longer correct for this approximation, and performance of our framework drops below inference using only ANCESTRY HMM. Yet, in the empirical data sets we have analysed, we have generally found marker densities for which these limits were not reached and with full genome sequencing now within reach for many researchers, we expect this not to be an issue.

To validate our approach, we have reanalysed three data sets. The first data set is from a crossing experiment with *S. cerevisiae*. Here, we applied the single chromosome equations, and estimates of the time since the onset of admixture line up well with the experimental design, although assumptions regarding the recombination rate remain of strong influence on the admixture time estimates.

The second data set we reanalysed is of Swordtail fish (*Xiphophorus*). We infer an admixture time that is older than previous estimates (Schumer, Rosenthal, et al., 2014) but that is in line with more recent estimates done by the same authors (M. Schumer, personal communication (Powell, Moran, et al., 2021)) using more recent recombination rate estimates (Powell, García-Olazábal, et al., 2020). However, likelihood profiles are fairly flat, indicating that our maximum-likelihood estimates are sensitive to changes in the data, further reflected by the sensitivity of the results to assumptions made regarding the recombination rate. Other plausible explanations are recombination rate variation along the genome or a more complex evolutionary history that is not well modelled by a single admixture event.

Finally, we have reanalysed a data set on *Populus* trees (Suarez-Gonzalez et al., 2016). We infer an admixture time that is in line with previous findings, but the original analysis did not focus on admixture time and only used admixture time to infer local ancestry. However, we find that the time since admixture correlates strongly with the genetic distance to either of the source taxa, with individuals more closely related to the source taxa inferred to be younger. This suggests that the data set does not consist of a sample from a single, admixed, population and that this invalidates our analysis. Furthermore, variation in ancestry is much higher than expected from the admixture time alone (Gravel, 2012), again indicating that the analysed population is most likely not a single admixing population. Lastly, admixture mapping analyses have shown that perhaps late generation backcrosses have contributed as well to the hybrid population (Suarez-Gonzalez et al., 2018), suggesting an intermediate form between on the one hand adaptive introgression and backcrossing and on the other hand ongoing hybridization across a spatial gradient. Across these results, it is clear that our assumption of a single admixed population is violated, and hence it seems likely that our age estimates are incorrect and that our framework is not a good fit for this empirical data set. Yet, we believe that it provides an interesting example on how our methodology can be applied.

Summarizing, we have presented a framework to estimate the time since admixture using phased or unphased data from two homologous chromosomes, taking into account marker spacing along the chromosome. We have shown that using data from two chromosomes improves the estimations of the admixture time compared to the method that uses only one chromosome. This is true whether the data are phased or unphased. In addition, we have shown, using simulations, that applying the phased or the unphased method yields very similar results. However, given that even small (unavoidable) phasing errors produce overestimates in the time since admixture, we suggest that, in most cases, using unphased data are the best strategy. In comparison with previous methods, such as ANCESTRY HMM (Corbett-Detig & Nielsen, 2017), our method performs better when the population size is small or the time since admixture is long. With our new framework, we hope to have opened new avenues towards inferring the time since admixture in admixed populations, and primarily hope to have brought this analysis within reach also for data sets where phased data are unavailable or impossible to acquire. Furthermore, we would like to emphasize that our method also works for a relatively small number of SNPs, which opens up avenues towards analysis of closely related taxa, where the number of ancestry informative markers might be low. We have included the derivations and the numerical solution framework in the R package "junctions" (Janzen, 2021). By providing the code in an easy to use package, we hope to lower the threshold for other users to apply the theory of junctions to their model system.

ACKNOWLEDGEMENTS

We thank Gianni Litti, Molly Schumer, Adriana Suarez-Gonzalez and Quentin Cronk for their help and enthusiasm with respect to our re-analysis of their data sets. We also thank Amaury Lambert who helped building this fruitful collaboration and Arno Siri-Jégousse for a careful reading of a previous version of this manuscript. We thank Stuart Baird and two anonymous referees for constructive comments on the manuscript. TJ thanks the Carl von Ossietzky University of Oldenburg for use of their computer cluster CARL and the Center for Information Technology of the University of Groningen for their support and access to the Peregrine high performance computing cluster. VMP was funded by the DGAPA-UNAM postdoctoral programme.

AUTHOR CONTRIBUTIONS

T.J. and V.M.P. jointly designed the research. V.M.P. inferred the ARG-based mathematics; T.J. verified findings using individual-based simulations and analysed the empirical data. T.J. and V.M.P. jointly wrote the manuscript.

OPEN RESEARCH BADGES

This article has earned an Open Data badge for making publicly available the digitally-shareable data necessary to reproduce the reported results. The data is available at https://doi.org/10.5061/dryad.xwdbrv1c5.

DATA AVAILABILITY STATEMENT

We have included the derivations and the numerical solution framework in the R package "junctions," which can be found on CRAN on https://CRAN.R-project.org/package=junctions. All code used in data analysis and visualization for this manuscript can be downloaded from data dryad, from: https://doi.org/10.5061/dryad.xwdbrv1c5.

ORCID

Thijs Janzen https://orcid.org/0000-0002-4162-1140

REFERENCES

Abbott, R., Albach, D., Ansell, S., Arntzen, J. W., Baird, S. J. E., Bierne, N., Boughman, J., Brelsford, A., Buerkle, C. A., Buggs, R., Butlin, R. K., Dieckmann, U., Eroukhmanoff, F., Grill, A., Cahan, S. H., Hermansen, J. S., Hewitt, G., Hudson, A. G., Jiggins, C., ... Zinner, D. (2013). Hybridization and speciation. *Journal of Evolutionary Biology*, *26*, 229–246. https://doi.org/10.1111/j.1420-9101.2012.02599.x

Andolfatto, P., Davison, D., Erezyilmaz, D., Hu, T. T., Mast, J., Sunayama-Morita, T., & Stern, D. L. (2011). Multiplexed shotgun genotyping for rapid and efficient genetic mapping. *Genome Research*, *21*(4), 610–617. https://doi.org/10.1101/gr.115402.110

Baaijens, J. A., & Schönhuth, A. (2019). Overlap graph-based generation of haplotigs for diploids and poly-ploids. *Bioinformatics*, *35*(21), 4281–4289. https://doi.org/10.1093/bioinformatics/btz255

Baird, S. J. E. (1995). A simulation study of multilocus clines. *Evolution*, *49*(6), 1038–1045. https://doi.org/10.1111/j.1558-5646.1995.tb04431.x

Baird, S. J. E. (2006). Phylogenetics - Fisher's markers of admixture. *Heredity*, *97*(2), 81–83. https://doi.org/10.1038/sj.hdy.6800850

Baird, S. J. E. (2015). Exploring linkage disequilibrium. *Molecular Ecology Resources*, *15*(5), 1017–1019. https://doi.org/10.1111/1755-0998.12424

Baird, S. J. E., Barton, N. H., & Etheridge, A. M. (2003). The distribution of surviving blocks of an ancestral genome. *Theoretical Population Biology*, *64*(4), 451–471. https://doi.org/10.1016/S0040-5809(03)00098-4

Barton, N. H. (1983). Multilocus clines. *Evolution*, *37*(3), 454–471. https://doi.org/10.1111/j.1558-5646.1983.tb05563.x

Bennett, J. H. (1953). Junctions in inbreeding. *Genetica*, *26*(1), 392–406. https://doi.org/10.1007/BF01690623

Brelsford, A., Mila, B., & Irwin, D. E. (2011). Hybrid origin of Audubon's warbler. *Molecular Ecology*, *20*, 2380–2389. https://doi.org/10.1111/j.1365-294X.2011.05055.x

Browning, S. R., & Browning, B. L. (2007). Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *American Journal of Human Genetics*, *81*(5), 1084–1097. https://doi.org/10.1086/521987

Browning, S. P., & Browning, B. L. (2011). Haplotype phasing: Existing methods and new developments. *Nature Reviews Genetics*, *12*, 703–714.

Buerkle, C. A., & Rieseberg, L. H. (2008). The rate of genome stabilization in homoploid hybrid species. *Evolution*, *62*(2), 266–275. https://doi.org/10.1111/j.1558-5646.2007.00267.x

Capblancq, T., Després, L., Rioux, D., & Mavárez, J. (2015). Hybridization promotes speciation in coenonympha butteries. *Molecular Ecology*, *24*(24), 6209–6222.

Chapman, N. H., & Thompson, E. A. (2002). The effect of population history on the lengths of ancestral chromosome segments. *Genetics*, *162*(1), 449–458. https://doi.org/10.1093/genetics/162.1.449

Chapman, N. H., & Thompson, E. A. (2003). A model for the length of tracts of identity by descent in finite random mating populations. *Theoretical Population Biology*, *64*(2), 141–150. https://doi.org/10.1016/S0040-5809(03)00071-6

Cherry, J. M., Ball, C., Weng, S., Juvik, G., Schmidt, R., Adler, C., Dunn, B., Dwight, S., Riles, L., Mortimer, R. K., & Botstein, D. (1997). Genetic and physical maps of *Saccharomyces cerevisiae*. *Nature*, *387*(S6632), 67–73. https://doi.org/10.1038/387s067

Choi, Y., Chan, A. P., Kirkness, E., Telenti, A., & Schork, N. J. (2018). Comparison of phasing strategies for whole human genomes. *PLOS Genetics*, *14*(4), 1–26. https://doi.org/10.1371/journal.pgen.1007308

Corbett-Detig, R., & Nielsen, R. (2017). A hidden Markov model approach for simultaneously estimating local ancestry and admixture time using next generation sequence data in samples of arbitrary ploidy. *PLoS Genetics*, *13*(1), e1006529. https://doi.org/10.1371/journal.pgen.1006529

Coyne, J. A., & Orr, H. A. (2004). *Speciation*. Sinauer Associates, Inc.

Durrett, R. (2008). *Probability models for DNA sequence evolution*, 2nd ed. Springer.

Ebler, J., Haukness, M., Pesout, T., Marschall, T., & Paten, B. (2019). Haplotype-aware diplotyping from noisy long reads. *Genome Biology*, *20*(1), 116. https://doi.org/10.1186/s13059-019-1709-0

Fisher, R. A. (1949). *The Theory of Inbreeding*. Oliver and Boyd.

Fisher, R. A. (1954). A fuller theory of "junctions" in inbreeding. *Heredity*, *8*, 187–197. https://doi.org/10.1038/hdy.1954.17

Fisher, R. A. (1959). An algebraically exact examination of junction formation and transmission in parent-offspring inbreeding. *Heredity*, *13*, 179–186. https://doi.org/10.1038/hdy.1959.21

Gale, J. (1964). Some applications of the theory of junctions. *Biometrics*, *20*, 85–117. https://doi.org/10.2307/2527619

Grant, V. (1981). *Plant speciation*. Columbia University Press.

Gravel, S. (2012). Population genetics models of local ancestry. *Genetics*, *191*(2), 607–619. https://doi.org/10.1534/genetics.112.139808

Griffths, R. C. (1991). The two-locus ancestral graph. In I. V. Basawa, & R. L. Taylor (Eds.), *Selected proceedings of the symposium on applied probability* (pp. 100–117). Institute of Mathematical Statistics.

Griffths, R. C., & Marjoram, P. (1997). An ancestral recombination graph. In P. Donnelly, & S. Tavaré (Eds.), *Progress in population genetics and human evolution*, *IMA volumes in mathematics and its applications*, Vol. *87* (pp. 257–270). Springer Verlag.

Guan, Y. (2014). Detecting structure of haplotypes and local ancestry. *Genetics*, *196*(3), 625–642. https://doi.org/10.1534/genetics.113.160697

Hudson, R. R. (1983). Properties of the neutral model with intragenic recombination. *Theoretical Population Biology*, *23*(2), 213–201.

Illingworth, C. J. R., Parts, L., Bergström, A., Liti, G., & Mustonen, V. (2013). Inferring genome-wide recombination landscapes from advanced intercross lines: application to yeast crosses. *PLoS One*, *8*(5), e62266.

Janzen, T. (2021). *junctions: The Breakdown of Genomic Ancestry Blocks in Hybrid Lineages*. https//github.com/thijsjanzen/junctions. R package version 2.0.0.

Janzen, T., Nolte, A., & Traulsen, A. (2018). The breakdown of genomic ancestry blocks in hybrid lineages given a finite number of recombination sites. *Evolution*, *72*(4), 735–750. https://doi.org/10.1111/evo.13436

Keller, I., Wagner, C. E., Greuter, L., Mwaiko, S. et al (2013). Population genomic signatures of divergent adaptation, gene flow and hybrid speciation in the rapid radiation of lake Victoria cichlid fishes. *Molecular Ecology*, *22*(11), 2848–2863.

Koblmüller, S., Duftner, N., Sefc, K. M., Aibara, M., Stipacek, M., Blanc, M., Egger, B., & Sturmbauer, C. (2007). Reticulate phylogeny of gastropod-shell-breeding cichlids from lake Tanganyika—The result of repeated introgressive hybridization. *BMC Evolutionary Biology*, *7*(1), 7. https://doi.org/10.1186/1471-2148-7-7

Kong, A., Masson, G., Frigge, M. L., Gylfason, A., Zusmanovich, P., Thorleifsson, G., Olason, P. I., Ingason, A., Steinberg, S., Rafnar, T., Sulem, P., Mouy, M., Jonsson, F., Thorsteinsdottir, U., Gudbjartsson, D. F., Stefansson, H., & Stefansson, K. (2008). Detection of sharing by descent, long-range phasing and haplotype imputation. *Nature Genetics*, *40*(9), 1068–1075. https://doi.org/10.1038/ng.216

Kronenberg, Z. N., Rhie, A., Koren, S., Concepcion, G. T., Peluso, P., Munson, K. M., Hiendleder, S., Fedrigo, O., Jarvis, E. D., Phillippy, A. M., Eichler, E. E., Williams, J. L., Smith, T. P. L., Hall, R. J., Sullivan, S. T., & Kingan, S. B. (2021). Extended haplotype-phasing of de novo genome assemblies using Hi-C. *Nature Communications*, *12*, 1935. https://doi.org/10.1038/s41467-020-20536-y

Liang, M., & Nielsen, R. (2014). The lengths of admixture tracts. *Genetics*, *197*(3), 953–967. https://doi.org/10.1534/genetics.114.162362

Loh, P. R., Palamara, P. F., & Alkes, L. P. (2016). Fast and accurate long-range phasing in a UK biobank cohort. *Nature Genetics*, *48*, 811–816. https://doi.org/10.1038/ng.3571

Loh, P.-R., Palamara, P. F., & Price, A. L. (2016). Fast and accurate long-range phasing in a UK Biobank cohort. *Nature Genetics*, *48*(7), 811–816. https://doi.org/10.1038/ng.3571

Lutgen, D., Ritter, R., Olsen, R.-A., Schielzeth, H., Gruselius, J., Ewels, P., Garcia, J. T., Shirihai, H., Schweizer, M., Suh, A., & Burri, R. (2020). Linked-read sequencing enables haplotype-resolved resequencing at population scale. *Molecular Ecology Resources*, *20*, 1311–1322. https://doi.org/10.1111/1755-0998.13192

MacLeod, A. K., Haley, C. S., Woolliams, P., & Stam, J. A. (2005). Marker densities and the mapping of ancestral junctions. *Genetical Research*, *85*(01), *69–79*. https://doi.org/10.1017/S0016672305007329

Mancera, E., Bourgon, R., Brozzi, A., Huber, W., & Steinmetz, L. M. (2008). High-resolution mapping of meiotic crossovers and non-crossovers in yeast. *Nature*, *454*(7203), 479–485. https://doi.org/10.1038/nature07135

Maples, B. K., Gravel, S., Kenny, E. E., & Bustamante, C. D. (2013). Rfmix: A discriminative modeling approach for rapid and robust local-ancestry inference. *The American Journal of Human Genetics*, *93*(2), 278–288.

Mavárez, J., Salazar, C. A., Bermingham, E., Salcedo, C., Jiggins, C. D., & Linares, M. (2006). Speciation by hybridization in Heliconius butteries. *Nature*, *441*(7095), 868–871.

McVean, G. A., & Cardin, N. J. (2005). Approximating the coalescent with recombination. *Philosophical Transactions of the Royal Society of London. Series B*, *Biological Sciences*, *1459*(360), 1387–1393. https://doi.org/10.1098/rstb.2005.1673

Medina, P., Thornlow, B., Nielsen, R., & Corbett-Detig, R. (2018). Estimating the timing of multiple admixture pulses during local ancestry inference. *Genetics*, *210*(3), 1089–1107. https://doi.org/10.1534/genetics.118.301411

Muir, P., Li, S., Lou, S., Wang, D., Spakowicz, D. J., Salichos, L., Zhang, J., Weinstock, G. M., Isaacs, F., Rozowsky, J., & Gerstein, M. (2016) The real cost of sequencing: Scaling computation to keep pace with data generation. *Genome Biology*, *17*(1), 53.

Nolte, A. W., Freyhof, J., Stemshorn, K. C., & Tautz, D. (2005). An invasive lineage of sculpins, Cottus sp. (Pisces, Teleostei) in the rhine with new habitat adaptations has originated from hybridization between old phylogeographic groups. *Proceedings of the Royal Society B*, *272*, 2379–2387.

O'Connell, J., Sharp, K., Shrine, N., Wain, L., Hall, I., Tobin, M., Zagury, J. F., Delaneau, O., & Marchini, J. (2016). Haplotype estimation for Biobank-scale data sets. *Nature Genetics*, *48*, 817–820. https://doi.org/10.1038/ng.3583

Olson, M. S., Robertson, A. L., Takebayashi, N., Silim, S., Schroeder, W. R., & Tiffin, P. (2010). Nucleotide diversity and linkage disequilibrium in balsam poplar (*Populus balsamifera*). *New Phytologist*, *186*(2), 526–536.

Parts, L., Cubillos, F. A., Warringer, J., Jain, K., Salinas, F., Bumpstead, S. J., Molin, M., Zia, A., Simpson, J. T., Quail, M. A., Moses, A., Louis, E. J., Durbin, R., & Liti, G. (2011). Revealing the genetic structure of a trait by sequencing a population under selection. *Genome Research*, *21*(7), 1131–1138. https://doi.org/10.1101/gr.116731.110

Paşaniuc, B., Sankararaman, S., Kimmel, G., & Halperin, E. (2009). Inference of locus-specific ancestry in closely related populations. *Bioinformatics*, *25*(12), i213–i221. https://doi.org/10.1093/bioinformatics/btp197

Powell, D. L., García-Olazábal, M., Keegan, M., Reilly, P., Du, K., Díaz-Loyo, A. P., Banerjee, S., Blakkan, D., Reich, D., Andolfatto, P., Rosenthal, G. G., Schartl, M., & Schumer, M. (2020). Natural hybridization reveals incompatible alleles that cause melanoma in swordtail fish. *Science*, *368*(6492), 731–736. https://doi.org/10.1126/science.aba5216

Powell, D. L., Moran, B., Kim, B., Banerjee, S. M., Aguillon, S. M., Fascinetto-Zago, P., Langdon, Q., & Schumer, M. (2021). Two new hybrid zones expand the swordtail hybridization model system. *Evolution*, *75*, 2524–2539. https://doi.org/10.1111/evo.14337

Schumer, M., Cui, R., Powell, D. L., Dresner, R., Rosenthal, G. G., & Andolfatto, P. (2014). High-resolution mapping reveals hundreds of genetic incompatibilities in hybridizing fish species. *Elife*, *3*, e02535. https://doi.org/10.7554/eLife.02535

Schumer, M., Rosenthal, G. G., & Andolfatto, P. (2014). How common is homoploid hybrid speciation? *Evolution*, *68*(6), 1553–1560. https://doi.org/10.1111/evo.12399

Schumer, M., Xu, C., Powell, D. L., Durvasula, A., Skov, L., Holland, C., Blazier, J. C., Sankararaman, S., Andolfatto, P., Rosenthal, G. G., & Przeworski, M. (2018). Natural selection interacts with recombination to shape the evolution of hybrid genomes. *Science*, *360*(6389), 656–660. https://doi.org/10.1126/science.aar3684

Schwarz, D., Matta, B. M., Shakir-Botteri, N. L., & McPheron, B. A. (2005). Host shift to an invasive plant triggers rapid animal hybrid speciation. *Nature*, *436*(7050), 546–549. https://doi.org/10.1038/nature03800

Shriver, M. D., Smith, M. W., Jin, L., Marcini, A., Akey, J. M., Deka, R., & Ferrell, R. E. (1997). Ethnic-affiliation estimation by use of population-specific DNA markers. *American Journal of Human Genetics*, *60*(4), 957–964.

Slavov, G. T., Difazio, S. P., Martin, J., Schackwitz, W., Muchero, W., Rodgers-Melnick, E., Lipphardt, M. F., Pennacchio, C. P., Hellsten, U., Pennacchio, L. A., Gunter, L. E., Ranjan, P., Vining, K., Pomraning, K. R., Wilhelm, L. J., Pellegrini, M., Mockler, T. C., Freitag, M., Geraldes, A., … Tuskan, G. A. (2012). Genome resequencing reveals multiscale geographic structure and extensive linkage disequilibrium in the forest tree *Populus trichocarpa*. *New Phytologist*, *196*(3), 713–725.

Snyder, M. W., Adey, A., Kitzman, J. O., & Shendure, J. (2015). Haplotype-resolved genome sequencing: Experimental methods and applications. *Nature Reviews Genetics*, *16*(6), 344–358.

Stam, P. (1980). The distribution of the fraction of the genome identical by descent in finite random mating populations. *Genetical Research*, *35*(2), 131–155. https://doi.org/10.1017/S0016672300014002

Suarez-Gonzalez, A., Hefer, C. A., Christe, C., Corea, O., Lexer, C., Cronk, Q. C., & Douglas, C. J. (2016). Genomic and functional approaches reveal a case of adaptive introgression from *Populus balsamifera* (balsam poplar) in *P. trichocarpa* (black cottonwood). *Molecular Ecology*, *25*(11), 2427–2442.

Suarez-Gonzalez, A., Hefer, C. A., Lexer, C., Douglas, C. J., & Cronk, Q. C. B. (2018). Introgression from *Populus balsamifera* underlies adaptively significant variation and range boundaries in *P. trichocarpa*. *New Phytologist*, *217*(1), 416–427.

Svedberg, J., Shchur, V., Reinman, S., Nielsen, R., & Corbett-Detig, R. (2021). Inferring adaptive introgression using Hidden Markov models. *Molecular Biology and Evolution*, *38*(5), 2152–2165. https://doi.org/10.1093/molbev/msab014

Tangherloni, A., Spolaor, S., Rundo, L., Nobile, M. S., Cazzaniga, P., Mauri, G., Liò, P., Merelli, I., & Besozzi, D. (2019). Genhap: A novel computational method based on genetic algorithms for haplotype assembly. *BMC Bioinformatics*, *172*(20). https://doi.org/10.1186/s12859-019-2691-y

Tourdot, R. W., & Zhang, C.-Z. (2021). Whole chromosome haplotype phasing from long-range sequencing. *Genome Biology*, *22*, 139. https://doi.org/10.1186/s13059-021-02330-1

Walter, R. B., Rains, J. D., Russell, J. E., Guerra, T. M., Daniels, C., Johnston, D. A., Kumar, J., Wheeler, A., Kelnar, K., Khanolkar, V. A., Williams, E. L., Hornecker, J. L., Hollek, L., Mamerow, M. M., Pedroza, A., & Kazianis, S. (2004). A microsatellite genetic linkage map for xiphophorus. *Genetics*, *168*(1), 363–372.

Wang, J., Street, N. R., Scofield, D. G., & Ingvarsson, P. K. (2016). Natural selection and recombination rate variation shape nucleotide polymorphism across the genomes of three related populus species. *Genetics*, *202*(3), 1185–1200.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Janzen, T., & Miró Pina, V. (2022). Estimating the time since admixture from phased and unphased molecular data. *Molecular Ecology Resources*, 22, 908–926. https://doi.org/10.1111/1755-0998.13519