bioRxiv preprint doi: https://doi.org/10.1101/2020.09.10.292441; this version posted September 11, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Estimating the time since admixture from phased and unphased molecular data

Thijs Janzen[1,2,†,*], Verónica Miró Pina[3,†]

1 Groningen Institute for Evolutionary Life Sciences, University of Groningen, Box 111039700 CC Groningen, The Netherlands

2 Carl von Ossietzky University, Carl-von-Ossietzky-Str. 9–11, 26111, Oldenburg, Germany

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

† These authors contributed equally to this work. * Corresponding author: t.janzen@rug.nl

Abstract

After admixture, recombination breaks down genomic blocks of contiguous ancestry. The break down 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 break down of these blocks, or the accumulation of delineations between blocks, so called 'junctions', has 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 perform much better than previous methods, especially for more ancient admixture times. Lastly, we demonstrate the applicability of our method on an empirical dataset of labcrosses of yeast (*Saccharomyces cerevisae*) and on two case studies of hybridization in swordtail fish and *Populus* trees.

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.10.292441; this version posted September 11, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Keywords

Admixture, hybridization, recombination, junctions, phasing

1 Introduction

The traditional view where species or lineages accumulate incompatibilities over time and gradually become reproductively isolated from each other has led to insight into the processes generating and maintaining biodiversity (Coyne and Orr, 2004). However, this view has proven to be misleading, and it has become apparent that lineages do not necessarily only branch, but that they can also come back together (Abbott et al., 2013). In plants, it has been known for quite some time that hybridization between lineages can generate not only viable offspring, but also potentially lead to the formation of new lineages, and ultimately, species (Grant, 1981). It has long been debated whether this process could also happen in animals, but over the past few years numerous examples have appeared, including, but not limited to, butterflies (Mavárez et al., 2006; Capblancq et al., 2015), cichlid fishes (Koblmüller et al., 2007; Keller et al., 2013), warblers (Brelsford et al., 2011), fruit flies (Schwarz et al., 2005) and sculpins (Nolte et al., 2005).

Understanding the timeline of these hybridization events is paramount in obtaining a full understanding of the process and its impact. Often, hybridization processes occur fast, on a timescale that is too rapid to accumulate enough mutations, which prevents the use of traditional molecular clocks to infer the onset of hybridization. Instead, recombination processes are sufficiently rapid so as to be used to study the recent evolutionary dynamics of a population. For example, they have been used to infer selective sweeps (Sabeti et al., 2007) or recent demography (Ralph and Coop, 2013; Ringbauer et al., 2017) in human populations. Recombination also leaves a footprint in genomes undergoing hybridization. After admixture of two lineages, 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,

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.10.292441; this version posted September 11, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

1964), a randomly mating population (Stam, 1980; Baird, 1995), and for sub-structured populations (Chapman and Thompson, 2002, 2003).

So far, applying the theory of junctions has shown to be difficult, as it requires extensive genotyping of the admixed lineage, but also of the parental lineages. 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 parental and admixed lineages (Paşaniuc et al., 2009; Maples et al., 2013; Guan, 2014; Corbett-Detig and Nielsen, 2017). 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 Fig 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 (MacLeod et al., 2005; Buerkle and Rieseberg, 2008). 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.

Furthermore, existing theory on the accumulation of junctions is only developed for the case where ancestry can be determined within a single chromosome. 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 Fig 1, second and third panel). Phasing methods can be classified into three main categories. Firstly, direct methods are based on haplotype-resolved genome sequencing (reviewed in Snyder et al. (2015)). These methods yield accurate results, but are expensive and require large amounts of DNA. Recently Lutgen et al. (2020) have shown that linked-read sequencing is efficient enough to provide haplotype resolved sequencing at a population scale at reasonable cost. However, linked-read sequencing is still a fairly new technology, and not yet widespread. Secondly, phasing can be performed using methods based on the analysis of genotypes of closely related individuals. These methods often

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.10.292441; this version posted September 11, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

yield good results but their application has been limited to humans, where large pedigree datasets are available (Browning and Browning, 2011; Loh et al., 2016a; Kong et al., 2008). Lastly, phasing can be performed using statistical methods, based on estimating the recombination rates and allele frequencies in a population. While some algorithms make use of a reference genome (for example Eagle (Loh et al., 2016b), Beagle (Browning and Browning, 2007) or ShapeIt (O'Connell et al., 2016)), others allow *de novo* number of individuals in the sample is small, accuracy is low and only local haplotypes can reliably be inferred (Browning and Browning, 2011; Choi et al., 2018). More recently, statistical methods developed for third generation sequencing data (sometimes combined with Hi-C), do allow to infer long-range haplotypes with good accuracy (Tourdot and Zhang, 2019; Kronenberg et al., 2019; Ebler et al., 2019; Tangherloni et al., 2019). However, data from hybrid populations are not often available in this form. Across these three groups of methods, 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 full 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) (Hudson, 1983; Griffiths, 1991; Griffiths and Marjoram, 1997). 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'.

Our paper is organised as follows. In section 2.1, we introduce our model, which is a simplified version of the ARG. In 2.2 , we present three maximum-likelihood methods to infer the time since admixture in hybrid populations: the first one uses information from a single chromosome and the others use phased or unphased data from two homologous chromosomes. In section 2.3, we validate our methods using simulations. In section 3 we apply them to a dataset from yeast experimental evolution and to two case studies of hybridization in swordtail fish and *Populus* tree.

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.10.292441; this version posted September 11, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

2 Materials and methods

2.1 Mathematical model

We assume a diploid population that evolves according to a Wright-Fisher dynamics, i.e. generations are non-overlapping, 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 on each other. We assume that hybridization occurred at time 0 between two populations, $\mathcal{P}$ and $\mathcal{Q}$. The proportion of individuals from population $\mathcal{P}$ at time 0 is p and the proportion of individuals of type $\mathcal{Q}$ is $q = 1 - p$.

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. The genealogy of these n (or $2n$) loci is given by the Ancestral Recombination Graph (ARG), defined in Hudson (1983); Griffiths (1991); Griffiths and Marjoram (1997). This process is a branching-coalescence process in which loci that belong to the same block at time t are those which were carried by the same ancestor 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) at a time.

We assume that $N \gg 1$ so that we can neglect some transitions (double coalescences and simultaneous coalescence and recombination), $d_i \ll 1$ 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.

2.1.1 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 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}. The ARG for these two sites has two possible states ($z_i \sim z_{i+1}$) (where both loci are carried by the same lineage) and state ($z_i \not\sim z_{i+1}$) (where each locus is carried by a different lineage). The dynamics of this process are controlled by two types of events:

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.10.292441; this version posted September 11, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

- **Recombination** $(z_i \sim z_{i+1}) \to (z_i \not\sim z_{i+1})$ with probability d_i,
- **Coalescence** $(z_i \not\sim z_{i+1}) \to (z_i \sim z_{i+1})$ with probability $\frac{1}{2N}$.

Other events (such as simultaneous coalescence and recombination events) have probabilities that are negligible when N is large. This yields the following transition matrix:

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

Let $\bar{P}_t$ be the probability vector at time t for this Markov chain with two states. $(\bar{P}_t)_1$ is the probability of $(z_i \sim z_{i+1})$ at time t and $(\bar{P}_t)_2$ the probability of $(z_i \not\sim z_{i+1})$ at time t. We have $\bar{P}_0 = (1,0)$ (in the present we sample the two loci in the same individual) and $\bar{P}_t = \bar{P}_0 \bar{M}^t$. 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}, if the hybridization event happened t generations ago. We have

$$\mathbb{P}(J_t(z_i, z_{i+1})) \;=\; H_0(\bar{P}_t)_2, \tag{1}$$

which corresponds to the probability that the two loci were carried by different lineages t generations ago and the two lineages correspond to individuals from different ancestral subpopulations (see Fig 2, left panel).

Solving equation (1) gives

$$\mathbb{P}(J_t(z_i, z_{i+1})) = H_0 \frac{2N}{2N + 1/d_i} \left(1 - \left(1 - d_i - \frac{1}{2N}\right)^t\right).$$

Let $\mathbb{E}(J_t)$ be the expected number of observed junctions, we have

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

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.10.292441; this version posted September 11, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

Table 1. States of the reduced ARG. 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.

2.1.2 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 2 sites in 2 chromosomes has 7 states (see Durrett (2008), Chapter 3). To describe them, we borrow the notation from Durrett (2008) and we write $(z_i z_{i+1})$ to indicate an ancestor that is ancestor to site z_i and z_{i+1}, and notation (z_i) or (z_{i+1}) for an ancestor that is only ancestor to one of the two sites. The resulting 7 states are summarized in Fig 1. An example of realization of this process is shown in Fig 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 2 loci and a sample size 2 can be approximated, when $N \gg 1$ by

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

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

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.10.292441; this version posted September 11, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Let $P_t^{(i)}$ be the vector containing the probabilities of observing each of the states $(S^1, \ldots, S^7)$ 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

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

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 i.e. 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 and 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 ancestral subpopulation. This is exactly the probability of observing a junction when $t \to \infty$ for one chromosome. 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.

2.2 Maximum likelihood estimation

2.2.1 One chromosome case

To infer the admixture time given an observed number of junctions J_{obs}, we have to numerically solve equation (2). The solution of this equation is the maximum likelihood estimator of the time since admixture.

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.10.292441; this version posted September 11, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

2.2.2 Two chromosomes, phased data

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

- PP i.e. both homologous markers carry the allele from parent $\mathcal{P}$,
- QQ i.e. both homologous markers carry the allele from parent $\mathcal{Q}$,
- PQ i.e. the marker on the first chromosome carries the allele from parent $\mathcal{P}$ and the marker on the second chromosome carries the allele from parent $\mathcal{Q}$,
- QP i.e. the marker on the first chromosome carries the allele from parent $\mathcal{Q}$ and the marker on the second chromosome carries the allele from parent $\mathcal{P}$.

The data can then be represented as a sequence $(O_i, 1 \leq i \leq n)$ 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 non-trivial correlations between loci along the chromosome. However, we can neglect long-range dependencies and assume that O_i only depends on O_{i-1}, i.e. 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 subpopulation $\mathcal{P}$ and the two ancestors for locus 2 from $\mathcal{Q}$, which happens with probability p^2q^2 or,

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.10.292441; this version posted September 11, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

- $\Gamma_t^1 = S^5$, with probability 1/2 there are two ancestors for locus 1 and one for locus 2 and they are of from desired subpopulations 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 subpopulations with probability pq^2 or,
- $\Gamma_t^1 = S^7$ and the ancestor to 1 is from subpopulation $\mathcal{P}$ and the ancestor to 2 from $\mathcal{Q}$, 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 Fig 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 Fig 3. Then, the total probability of observing the data, given N and t, i.e.

$$\mathbb{P}_t((O_1, \ldots, O_n)) \;=\; \mathbb{P}_t(O_1, O_2) \prod_{i=2}^{n-1} \frac{\mathbb{P}_t(O_i, O_{i+1})}{\mathbb{P}_t(O_i)} \tag{3}$$

can be maximized in order to find the maximum likelihood estimator of t and N.

2.2.3 Two chromosomes, unphased data

If the data is 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 i.e. we only observe the allele from parent $\mathcal{P}$, i.e. both chromosomes carry the allele from parent $\mathcal{P}$,
- Q i.e. we only observe the allele from parent $\mathcal{Q}$.

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.10.292441; this version posted September 11, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

- x i.e. we observe both alleles, i.e. 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 Fig 4.

2.3 Individual based simulations

To test the validity of our maximum likelihood approach, we use individual based simulations, as described in (Janzen et al., 2018), i.e. Wright-Fisher type simulations of randomly mating populations of constant size N, with non-overlapping generations. We then recover local ancestry by analyzing ancestry at n markers whose positions are chosen uniformly at random along the genome.

As a proof of concept, we show how time can be accurately inferred for a population of $10,000$ individuals, for time points between the first generation and 1,000 generations. We use $n = 10,000$ markers, which should be sufficient to detect the majority of accumulated junctions. We report our findings across 100 replicates, where in each replicate 10 individuals were randomly selected from the population and used to infer the time since the onset of hybridization. We have simulated with three different values of the initial proportion of subpopulation $\mathcal{P}$, ($p \in \{0.053, 0.184, 0.5\}$), to vary the initial heterozygosity H_0 in $\{0.1, 0.3, 0.5\}$.

In Fig 5 we compare the methods we have developed here to previous methods based on the theory of junctions. 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.

In Fig 6 we compare the estimations of the time since admixture, using the method for one chromosome and the method for two chromosomes (phased). We observe that using data from the two homologous chromosomes allows to infer the time since admixture more accurately, since it reduces uncertainty.

In Fig 7 we compare the methods that use phased or unphased information of two homologous chromosomes. We observe that both methods yield very similar results in terms of the relative error.

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.10.292441; this version posted September 11, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 for example (Tourdot and Zhang, 2019; Ebler et al., 2019; Kronenberg et al., 2019)). Switching error rate error has strong effects on the inferred time since admixture, as shown in the bottom panel in Fig 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 be found in S1 Appendix.

3 Results

3.1 Saccharomyces cerevisiae

Experimental evolution provides an important reference point to verify our findings. Here, we re-analyze 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 (Parts et al., 2011; Illingworth et al., 2013). The data consists 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 inferred using a custom procedure, making use of the high levels of homozygosity in the parental

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.10.292441; this version posted September 11, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

lineages. H_0 was 0.5, reflecting a 50/50 contribution of both strains to the first generation. We used three different recombination rate estimates: firstly, we used the linkage map of (Cherry et al., 1997) where the average recombination rate is $1cM/2.7kb$ (1 centimorgan per 2.7 kilobases), secondly, we used the average recombination rate of $1cM/2.2kb$ as inferred in (Mancera et al., 2008) and lastly, we used the average recombination rate of $1cM/5.8kb$ 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 hotspots and coldspots. We assume a large population (N = 100,000), which makes inbreeding effects negligible.

We find that when using the older recombination rate estimates, we consistently underestimate the age of the hybrids (median age using (Mancera et al., 2008): 6.69 generations, median age using (Cherry et al., 1997): 8.45 generations), suggesting that the true recombination rate is slightly lower than assumed. When using the most recent recombination rate estimate (i.e. $1cM/5.8kb$), we slightly overestimate the age (median age estimate: 17.7 generations). Alternatively, we could be overestimating population size, suggesting that perhaps the rate of inbreeding in the experimental design was higher than anticipated. However, this only applies when assuming extremely high degrees of inbreeding, which seems unrealistic.

3.2 Swordtail Fish

Here, we re-analyze 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 et al., 2018, 2014a). 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 (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 favored 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

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.10.292441; this version posted September 11, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

average 38,462 markers per chromosome (95% CI: [18605, 50242]). 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 et al., 2014a): 1830 individuals. We infer the age for each of the 24 linkage groups separately, and analyze 187 individuals from the Tlatemaco population. As a recombination map, we use three approaches. Firstly, we use the average recombination rate of $1cM/378kb$ as used in (Schumer et al., 2014a), which is based on the average genome-wide recombination rate in *Xiphophorus* (Walter et al., 2004). Secondly, we use the average recombination rate of $1cM/500kb$ as reported in (Powell 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 $1cM/485kb$.

When we compare age estimates across chromosomes (see Fig 10 A), we find that chromosomes 17 and 24 are inferred to be much younger, in line with the notion that these chromosomes include large inversions (Schumer et al., 2018), making them unsuitable for admixture analysis. In any subsequent analysis, we have removed these two chromosomes from the dataset. 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 Fig 10 B). We recover a median age of 167 generations (mean: 165, 95% CI: [75, 242]). when using the recombination rate reported in (Schumer et al., 2014b). Using the high density recombination map from (Schumer et al., 2018) we obtain an age estimate of 194 generations (95% CI: [84, 349]), due to the shorter map length. Alternatively, using the most recent recombination rate estimate of $1cM/500kb$ reported in (Powell et al., 2020), we recover a median age of 221 generations (95% CI: [100, 320]).

3.3 *Populus* trees

Here, we re-analyze a dataset of *Populus* trees, published in (Suarez-Gonzalez et al., 2016). The dataset focuses on two species of trees, *P. trichocarpa*, found mainly in West-America, in humid, moist conditions, and *P. balsamifera*, which is found in Northern America (e.g. Alaska, Canada) and is more frost tolerant. The two species are thought to have diverged relatively recently, around

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.10.292441; this version posted September 11, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 dataset consists of 32 individuals which are mainly *P. balsamifera*, admixed with *P. trichocarpa* and 36 individuals that are mainly *P. trichocarpa*, admixed with *P. balsamifera*. Three chromosomes of interest were Illumina sequenced, being chromosomes 6, 12 and 15. 68 admixed individuals were included, and unphased data was available for on average 60071 ancestry informative markers per chromosome (95% CI: [28745, 101425]). 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 $1cM/10.4kb$ (Slavov et al., 2012), $1cM/22.2kb$ (Olson et al., 2010) and $1cM/93.3kb$ (Wang et al., 2016). Local ancestry was determined using ANCESTRY HMM (Corbett-Detig and Nielsen, 2017), assuming equal admixture of both parental species. 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. Local ancestry was translated into hard ancestry calls based on fixed thresholds. These thresholds are presented as Phred ancestry scores, which are $-10\log_{10}(p)$, where p indicates the ancestry uncertainty.

Across all Phred Ancestry scores, we find that the time since admixture strongly correlates with the recombination rate used (See Figure 11 A), with a median number of generations since admixture of 6 (95% CI: [3, 15]) when using the highest estimate of recombination ($1cM/10.4kb$ (Slavov et al., 2012)), an intermediate estimate of 12 generations (95% CI: [6, 30]) when using a recombination rate of $1cM/22.2kb$ (Olson et al., 2010) and a much higher age estimate of 48 generations (95% CI: [22, 122]) when using the lowest recombination estimate (Wang et al., 2016). When we correlate the age estimate for a Phred Ancestry score of 30 with the fraction of local ancestry in the sample attributable to *P. trichocarpa*, we find that individuals with intermediate ancestry tend to have a higher estimated age, and that individuals with a genomic ancestry more similar to either of the parental species tend to be younger.

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.10.292441; this version posted September 11, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

4 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 (2)). Second, we have considered the case in which there is sequencing data from two homologous chromosomes. We have developed a maximum likelihood approach that allows to infer the time since admixture, whether the data is phased or unphased. We have used a powerful mathematical model which is the ARG (Hudson, 1983; Griffiths, 1991; Griffiths and Marjoram, 1997). In the one chromosome case, we get an explicit formula for the number of junctions (equation (2)) and in the two chromosomes case, we get a semi-explicit formula for the likelihood of the observations (equation (3)) that can be solved numerically.

We have validated our method using simulations. We have shown that our formula for the number of observed junctions in one chromosome performs better than previous methods that ignore the effect of having a limited number of markers or asume that they are even-spaced (see Fig 5). We expected that using information from two chromosomes would improve accuracy of the estimation considerably, and this is also what we find when using phased data, especially for small population sizes (see Fig 6 and Fig 2 in S2 Appendix. Surprisingly, a similar performance is achieved by the method that uses unphased data (see Fig 7 and Fig 3 in S2 Appendix. The phased and unphased approaches differ mainly in their treatment of markers that are heterozygous for ancestry, and hence we expected mainly differences between these methods to manifest themselves during the initial stages of admixture, when heterozygosity is still high. We did find that there were slight differences during these stages (Fig 7 and Fig 3 in S2 Appendix), but these were neglibile compared to the overall uncertainty.

When we take into account additional errors in ancestry inference due to incorrect phasing, we have shown that our unphased method outperforms the phased method (see Fig 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 (for example in (Choi et al., 2018)). Human data sets are typically of very high quality, and these error rates represent an extremely favourable scenario. In

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.10.292441; this version posted September 11, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

addition, not all data can be phased, for example if no reference haplotypes are available or if the sample sizes are small, which is often the case of data from hybrid species. This makes the unphased method particularly interesting.

In addition, 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). We have found that our method is quite sensitive to H_0 but this parameter can easily be estimated from the proportion of markers that come from each parental population. One advantage of our approach is that age inference is not very sensitive to population size (see Fig 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 Fig 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 analyzing 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 indicative of an insufficient marker number.

The main issue with our method is its sensitivity to the recombination rate. This is shown in Fig 2 of S1 Appendix but also exemplified by the varying results in the empirical datasets, 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. Hence, we advocate for extending research on inferring recombination landscapes. At the same time we realize that inferring local recombination rates is

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.10.292441; this version posted September 11, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

labour intensive, and restrictive for organisms with large generation times (where crossing schemes would take very long to realize). The recombination rate does not only factor in during admixture time inference, but is also typically used to infer local ancestry. Methods such as AncestryHMM (Corbett-Detig and Nielsen, 2017), ELAI (Guan, 2014) and MSG (Andolfatto et al., 2011) use the local recombination landscape to assess the probability of an ancestry switch between neighboring nucleotides. Thus, any variation introduced at the start of the analysis in the recombination landscape, echoes down the analysis pipeline both through impact on local ancestry and on the time since admixture. This further stresses the need for improved methods of inferring the recombination landscape.

To validate our approach we have re-analysed three datasets. The first dataset is from a crossing experiment on *S. cerevisiae*. We found that equation (2) provides a slightly better estimation of the time since admixture than previous methods. However, since the number of markers is very large, we did not expect a major improvement (see Fig 5). In addition, taking into account the marker positions is particularly interesting when a detailed recombination map is available and the recombination rates between each pair of markers are known (here they are assumed to be proportional to the distance in base pairs, which is not necessarily true). Nevertheless, our 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 dataset we re-analyzed is of Swordtail fish (*Xiphophorus*). We infer an admixture time that is older than previous estimates (Schumer et al., 2014b) but that is in line with more recent estimates done by the same authors (M. Schumer, personal communication) using more recent recombination rate estimates (Powell et al., 2020) . The advantage of our method is that it is faster, since it does not rely on simulations. In the original dataset, the authors removed chromosomes 17 and 24 from their analysis because these chromosomes contain large inversions. We also find strongly differing age estimates for these chromosomes, indicating that indeed these chromosomes have not been subject to the same evolutionary history as the others. Again, we find that the results are sensitive to assumptions made regarding the recombination rate.

Finally, we have re-analyzed a dataset on *Populus* trees (Suarez-Gonzalez et al., 2016). We infer

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.10.292441; this version posted September 11, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

an admixture time that is in line with previous findings, but would like to stress that the original analysis did not focus on admixture time, and only used admixture time to infer local ancestry. Furthermore, we find that the time since admixture correlates strongly with the genetic distance to either of the parents, with individuals more closely related to either of the parents inferred to be younger. In the case of incidental hybridization and subsequent backcrossing, we would expect the exact opposite, with individuals more related to the parents to be relatively older. In contrast, the pattern we recover here suggests a hybrid zone between the two parents. However, admixture mapping analyses have shown that perhaps late generation backcrosses have contributed as well to the hybrid population (Suarez-Gonzalez et al., 2018), suggesting perhaps an intermediate form between on the one hand some initial adaptive introgression and back-crossing, and on the other hand the ongoing hybridization across a spatial gradient.

Here we have presented a full 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 is 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 is the best strategy. 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 datasets where phased data is unavailable or impossible to acquire. We have included the derivations and the numerical solution framework in the R package 'junctions'. 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 datasets. We also thank Amaury Lambert

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.10.292441; this version posted September 11, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

who helped building this fruitful collaboration. 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 program.

Data Accessibility

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`. A development version of the package can be found at `https://www.github.com/thijsjanzen/junctions`. All code used in data analysis and visualization for this manuscript has been included in the Supporting Information.

Author contributions

TJ and VMP jointly designed the research. VMP inferred the ARG based mathematics, TJ verified findings using individual based simulations and analyzed the empirical data. TJ and VMP jointly wrote the paper.

References

J. A. Coyne and H. A. Orr. *Speciation*. Sinauer Associates, Inc, 2004.

R. Abbott, D. Albach, S. Ansell, J. W. Arntzen, et al. Hybridization and speciation. *Journal of Evolutionary Biology*, 26:229–246, 2013.

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

J. Mavárez, C. A. Salazar, E. Bermingham, C. Salcedo, et al. Speciation by hybridization in Heliconius butterflies. *Nature*, 441(7095):868–871, 2006.

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

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.10.292441; this version posted September 11, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

S. Koblmüller, N. Duftner, K.M. Sefc, M. Aibara, et al. Reticulate phylogeny of gastropod-shell-breeding cichlids from lake tanganyika–the result of repeated introgressive hybridization. *BMC Evolutionary Biology*, 7(1):7, 2007.

I. Keller, C.E. Wagner, L. Greuter, S. Mwaiko, et al. 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, 2013.

A. Brelsford, Mila B., and D.E. Irwin. Hybrid origin of Audubon's warbler. *Molecular Ecology*, 20: 2380–2389, 2011.

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

A.W. Nolte, J. Freyhof, K.C. Stemshorn, and D. Tautz. 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, Oct 2005.

P. C. Sabeti, P. Varilly, B. Fry, J. Lohmueller, et al. Genome-wide detection and characterization of positive selection in human populations. *Nature*, 449:913–918, 2007.

P. Ralph and G. Coop. The geography of recent genetic ancestry across europe. *PLOS Biology*, 11 (5):1–20, 05 2013. doi: 10.1371/journal.pbio.1001555. URL `https://doi.org/10.1371/journal.pbio.1001555`.

H. Ringbauer, G. Coop, and N.H. Barton. Inferring recent demography from isolation by distance of long shared sequence blocks. *Genetics*, 205(3):1335–1351, 2017. ISSN 0016-6731. doi: 10.1534/genetics.116.196220. URL `https://www.genetics.org/content/205/3/1335`.

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

R. A. Fisher. A fuller theory of "junctions" in inbreeding. *Heredity*, 8:187–197, 1954.

J.H. Bennett. Junctions in inbreeding. *Genetica*, 26(1):392–406, 1953.

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.10.292441; this version posted September 11, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

R. A. Fisher. An algebraically exact examination of junction formation and transmission in parent-offspring inbreeding. *Heredity*, 13:179–186, 1959.

J. Gale. Some applications of the theory of junctions. *Biometrics*, pages 85–117, 1964.

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

S.J.E. Baird. A simulation study of multilocus clines. *Evolution*, 49(6):1038–1045, 1995.

N.H. Chapman and E.A. Thompson. The effect of population history on the lengths of ancestral chromosome segments. *Genetics*, 162(1):449–458, 2002.

N.H. Chapman and E.A. Thompson. A model for the length of tracts of identity by descent in finite random mating populations. *Theoretical Population Biology*, 64(2):141–150, 2003.

P. Muir, S. Li, S.e Lou, D. Wang, et al. The real cost of sequencing: scaling computation to keep pace with data generation. *Genome Biology*, 17(1):53, 2016.

B. Paşaniuc, S. Sankararaman, G. Kimmel, and E. Halperin. Inference of locus-specific ancestry in closely related populations. *Bioinformatics*, 25(12):i213–i221, 2009.

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

Y. Guan. Detecting structure of haplotypes and local ancestry. *Genetics*, 196(3):625–642, 2014.

R Corbett-Detig and R. Nielsen. 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, 2017.

A.K. MacLeod, C.S. Haley, and P. Woolliams, J.A .and Stam. Marker densities and the mapping of ancestral junctions. *Genetical research*, 85(01):69–79, 2005.

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.10.292441; this version posted September 11, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

C. A. Buerkle and L.H. Rieseberg. The rate of genome stabilization in homoploid hybrid species. *Evolution*, 62(2):266–275, 2008.

T. Janzen, A. Nolte, and A. Traulsen. The breakdown of genomic ancestry blocks in hybrid lineages given a finite number of recombination sites. *Evolution*, 72(4):735–750, 2018.

M.W. Snyder, A. Adey, J.O. Kitzman, and J. Shendure. Haplotype-resolved genome sequencing: Experimental methods and applications. *Nature Reviews Genetics*, 16(6):344–358, 5 2015. ISSN 1471-0056. doi: 10.1038/nrg3903.

D. Lutgen, R. Ritter, R.-A. Olsen, H. Schielzeth, et al. Linked-read sequencing enables haplotype-resolved resequencing at population scale. *bioRxiv*, 2020. doi: 10.1101/2020.01.15.907261. URL `https://www.biorxiv.org/content/early/2020/01/15/2020.01.15.907261`.

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

P.R. Loh, P.F. Palamara, and L.P. Alkes. Fast and accurate long-range phasing in a uk biobank cohort. *Nature Genetics*, 48:811–816, 2016a.

A. Kong, G. Masson, M. L. Frigge, A. Gylfason, et al. Detection of sharing by descent, long-range phasing and haplotype imputation. *Nature Genetics*, 40(9):1068–1075, 2008.

P.R. Loh, P.F. Palamara, and A.L. Price. Fast and accurate long-range phasing in a uk biobank cohort. *Nature Genetics*, pages 811–816, 2016b.

S.R. Browning and B.L. Browning. 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):1084–97, 2007.

J. O'Connell, K. Sharp, N. Shrine, L. Wain, I. Hall, M. Tobin, J.F. Zagury, O. Delaneau, and J. Marchini. Haplotype estimation for Biobank-scale data sets. *Nature Genetics*, 48:817–820, 2016.

Y. Choi, A. P. Chan, E. Kirkness, A. Telenti, and N. J. Schork. Comparison of phasing strategies for

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.10.292441; this version posted September 11, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

whole human genomes. *PLOS Genetics*, 14(4):1–26, 04 2018. doi: 10.1371/journal.pgen.1007308. URL `https://doi.org/10.1371/journal.pgen.1007308`.

R.W. Tourdot and C.-Z. Zhang. Whole chromosome haplotype phasing from long-range sequencing. *bioRxiv*, 2019. doi: 10.1101/629337. URL `https://www.biorxiv.org/content/early/2019/05/07/629337`.

Z.N. Kronenberg, A. Rhie, S. Koren, G. .T. Concepcion, and others. Extended haplotype phasing of de novo genome assemblies with falcon-phase. *bioRxiv*, 2019. doi: 10.1101/327064. URL `https://www.biorxiv.org/content/early/2019/04/19/327064`.

J. Ebler, M. Haukness, T. Pesout, T. Marschall, and B. Paten. Haplotype-aware diplotyping from noisy long reads. *Genome Biol.*, 116(20), 2019.

A. Tangherloni, S. Spolaor, L. Rundo, M. S. Nobile, et al. Genhap: a novel computational method based on genetic algorithms for haplotype assembly. *BMC Bioinformatics*, 172(20), 2019.

R.R. Hudson. Properties of the neutral model with intragenic recombination. *Theor. Popul. Biol.*, 23(2):213–201, 1983.

R. C. Griffiths. The two-locus ancestral graph. In I.V. Basawa and R. L. Taylor, editors, *Selected Proceeedings of the Symposium on Applied Probability*, pages 100–117. Institute of Mathematical Statistics, 1991.

R.C. Griffiths and P. Marjoram. An ancestral recombination graph. In P. Donnelly and S. Tavaré, editors, *Progress in Population Genetics and Human Evolution, IMA Volumes in Mathematics and its Applications*, volume 87, pages 257–270. Springer Verlag, 1997.

R. Durrett. *Probability Models for DNA Sequence Evolution*. Springer, 2 edition, 2008.

G.A. McVean and N.J. Cardin. Approximating the coalescent with recombination. *Philos Trans R Soc Lond B Biol Sci*, 1459(360):1387–93, 2005.

L. Parts, F.A. Cubillos, J. Warringer, K. Jain, et al. Revealing the genetic structure of a trait

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.10.292441; this version posted September 11, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

by sequencing a population under selection. *Genome Research*, 21(7):1131–1138, 2011. doi: 10.1101/gr.116731.110. URL `http://genome.cshlp.org/content/21/7/1131.abstract`.

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

J.M. Cherry, C. Ball, S. Weng, G. Juvik, et al. Genetic and physical maps of Saccharomyces cerevisiae. *Nature*, 387(6632 Suppl):67, 1997.

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

M. Schumer, C. Xu, D. L. Powell, A. Durvasula, et al. Natural selection interacts with recombination to shape the evolution of hybrid genomes. *Science*, 360(6389):656–660, 2018.

M. Schumer, R. Cui, D. L. Powell, R. Dresner, et al. High-resolution mapping reveals hundreds of genetic incompatibilities in hybridizing fish species. *Elife*, 3:e02535, 2014a.

P. Andolfatto, D. Davison, D. Erezyilmaz, T. T. Hu, et al. Multiplexed shotgun genotyping for rapid and efficient genetic mapping. *Genome research*, 21(4):610–617, 2011.

R.B. Walter, J.D. Rains, J.E. Russell, T.M Guerra, et al. A microsatellite genetic linkage map for xiphophorus. *Genetics*, 168(1):363–372, 2004.

D. L. Powell, M. García-Olazábal, M. Keegan, P. Reilly, et al. Natural hybridization reveals incompatible alleles that cause melanoma in swordtail fish. *Science*, 368(6492):731–736, 2020.

M. Schumer, G.G. Rosenthal, and P. Andolfatto. How common is homoploid hybrid speciation? *Evolution*, 68(6):1553–1560, 2014b.

A. Suarez-Gonzalez, C.A. Hefer, C. Christe, O. Corea, et al. 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, 2016.

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.10.292441; this version posted September 11, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

M.S. Olson, A.L. Robertson, N. Takebayashi, S. Silim, et al. Nucleotide diversity and linkage disequilibrium in balsam poplar (Populus balsamifera). *New Phytologist*, 186(2):526–536, 2010.

G.T. Slavov, S.P. Difazio, J. Martin, W. Schackwitz, et al. Genome resequencing reveals multiscale geographic structure and extensive linkage disequilibrium in the forest tree populus trichocarpa. *New Phytologist*, 196(3):713–725, 2012.

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

Supporting information

S1 Appendix. Sensitivity analysis. Using individual based simulations, we test how sensitive our new framework is to variation of the different parameters.

S2 Appendix. Small population size. We test the validity of our method using a smaller value of the population size ($N = 1000$).

S3 Appendix. Phasing error. We extend the analysis done in the main text to the case of a smaller number of markers.

S4 Appendix. Simulation code. A collection of simulation code, simulation data and visualization scripts, for all figures in the main text and in appendices S1, S2 and S3.

Tables and Figures

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.10.292441; this version posted September 11, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

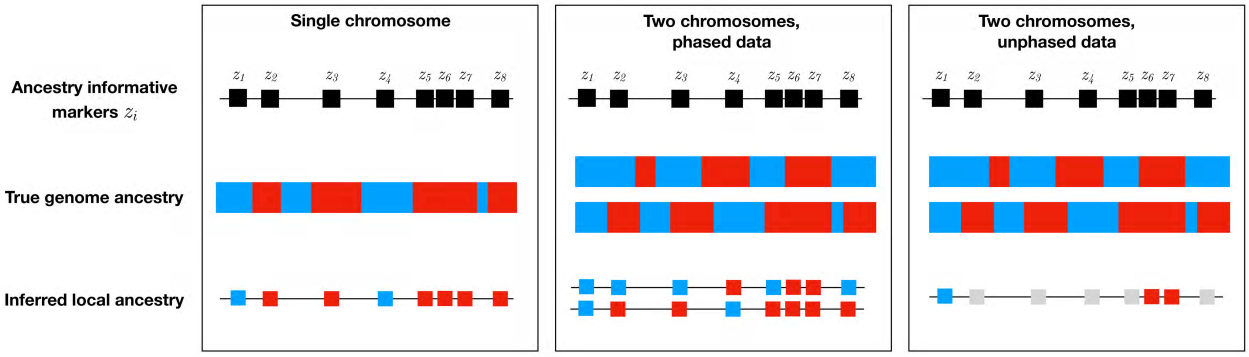


Fig 1. Visual depiction of the observed data. We show the differences between the type of data generated by the three methods we present in this paper. On each panel, the chromosome in the center is colored according to ancestry (blue represents parental population $\mathcal{P}$ and red represents parental population $\mathcal{Q}$). 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 7 junctions in the chromosome, but only 3 are observed in the data due to a limited marker coverage. The second and third panels represent the methods that use information from two chromosomes. In the second panel data is phased whereas in the second panel data is unphased.

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.10.292441; this version posted September 11, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

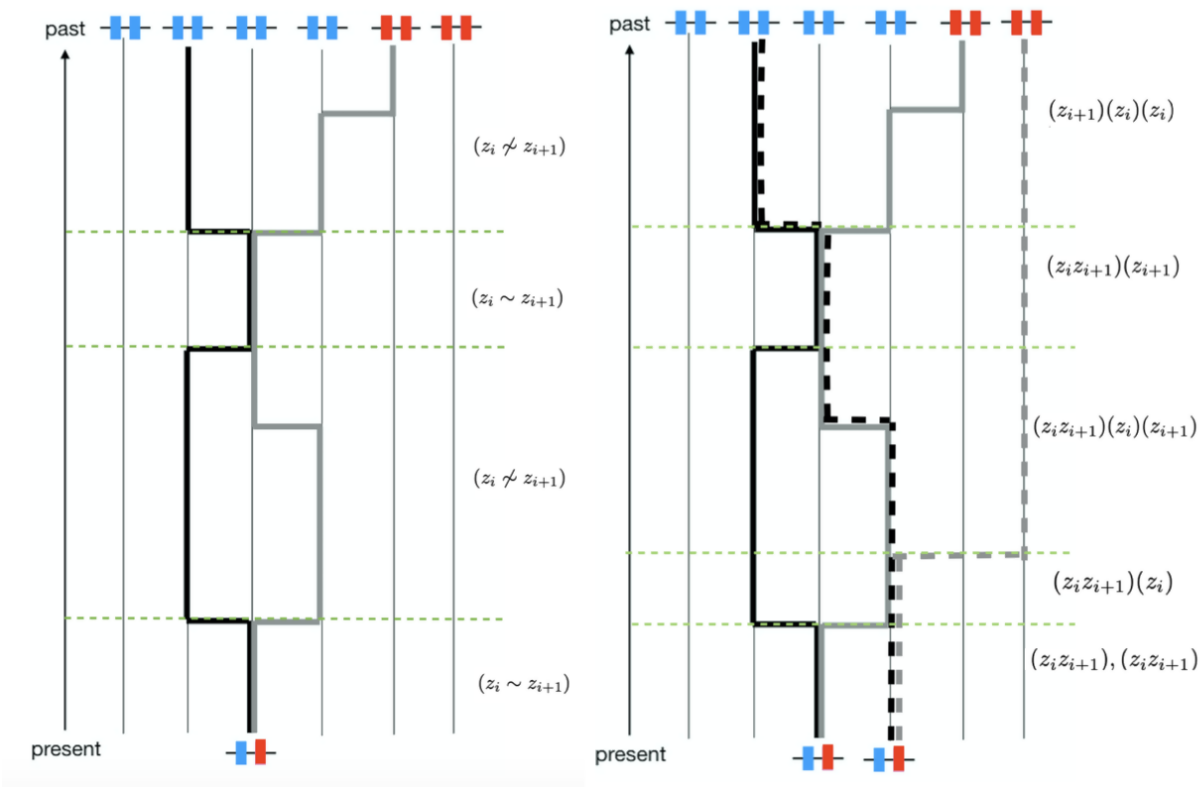


Fig 2. The ARG with two markers. Each color represents one parental population ($\mathcal{P}$ and $\mathcal{Q}$). The black and grey lines (or dotted lines) represent the ancestral lineage of each marker. In the left panel, we show the ARG for two markers in one chromosome. In the present, there is an 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 subpopulations. The right panel shows the ARG for two markers in two homologous chromosomes.

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.10.292441; this version posted September 11, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

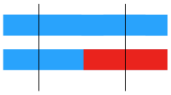
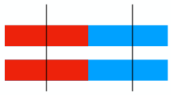
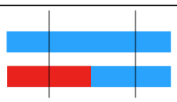
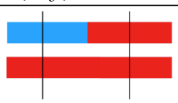
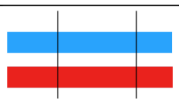
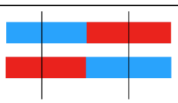
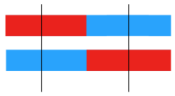
	$O_{i+1} = PP$	$O_{i+1} = QQ$	$O_{i+1} = PQ$	$O_{i+1} = QP$
$O_i = PP$	$p^2((P_t^i)_1 + (P_t^i)_4 + (P_t^i)_7) + p^3((P_t^i)_2 + (P_t^i)_5) + p^4(P_t^i)_3 + p(P_t^i)_6$	$pq(pq(P_t^i)_3 + \frac{1}{2}(P_t^i)_5 + (P_t^i)_7)$	 $\frac{pq}{2}(p(P_t^i)_2 + 2p^2(P_t^i)_3 + \frac{1}{2}(P_t^i)_4 + p(P_t^i)_5)$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

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

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.10.292441; this version posted September 11, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

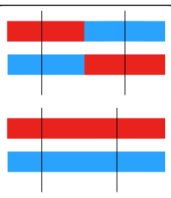
	$O_{i+1} = P$	$O_{i+1} = Q$	$O_{i+1} = x$
$O_i = P$	$p^2((P_t^i)_1 + (P_t^i)_4 + (P_t^i)_7) + p^3((P_t^i)_2 + (P_t^i)_5) + p^4(P_t^i)_3 + p(P_t^i)_6$	$pq(pq(P_t^i)_3 + \frac{1}{2}(P_t^i)_5 + (P_t^i)_7)$	$pq(p(P_t^i)_2 + 2p^2(P_t^i)_3 + \frac{1}{2}(P_t^i)_4 + p(P_t^i)_5)$
$O_i = Q$	$pq(pq(P_t^i)_3 + \frac{1}{2}(P_t^i)_5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

Fig 4. $\mathbb{P}_t(O_i, O_{i+1})$ **for unphased data.** The allele from parent $\mathcal{P}$ is represented in blue and the allele from parent $\mathcal{Q}$ is represented in red.

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.10.292441; this version posted September 11, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

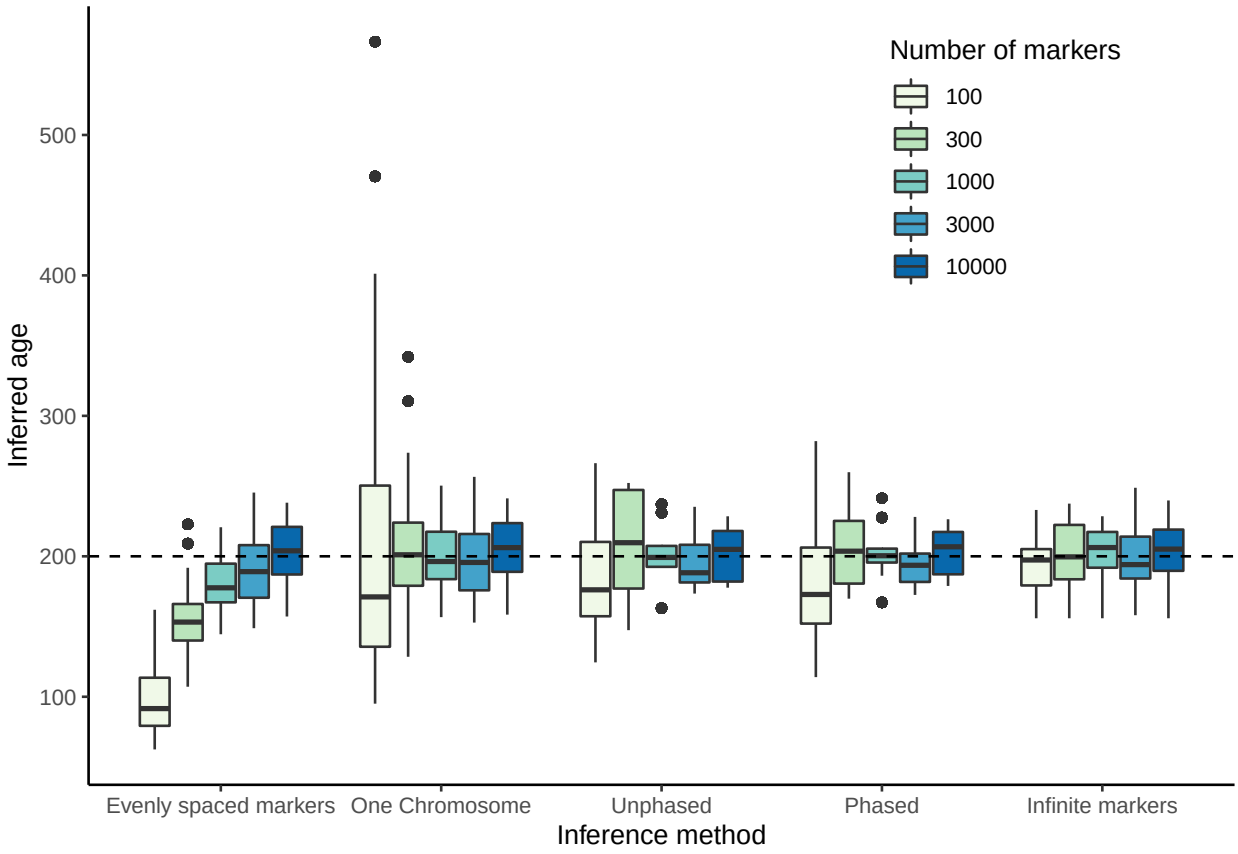


Fig 5. Comparison to previous methods. Shown are the median estimates for the time since admixture (dots) for 100 replicates, where in each replicate 10 individuals were analyzed. The dashed line indicates the simulated time. 'Evenly spaced markers' corresponds to the method in (Janzen et al., 2018). '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.

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.10.292441; this version posted September 11, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

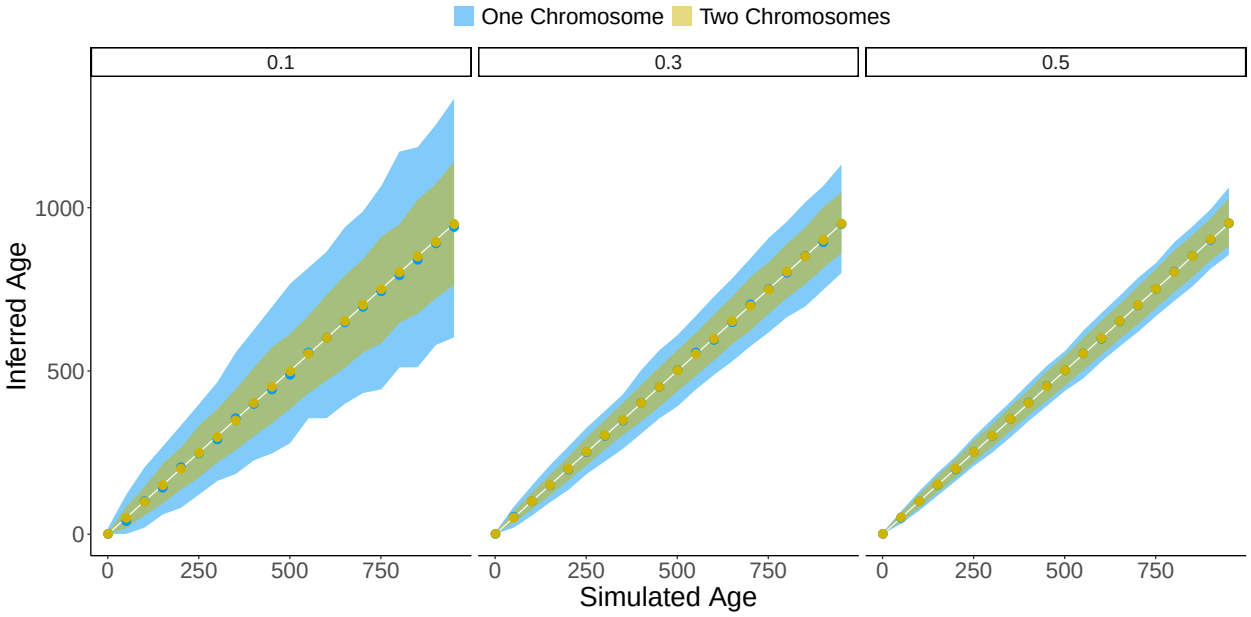


Fig 6. Accuracy in age estimate using information from one versus two chromosomes. Inferred time versus simulated time is represented. Shown are the median estimates (dots) for 100 replicates, where in each replicate 10 individuals were analyzed. The solid white line indicates the observed is equal to expected line and the shaded area indicates the 95% percentile. Shown are results using junction information from one chromosome (blue) and results using 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.

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.10.292441; this version posted September 11, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

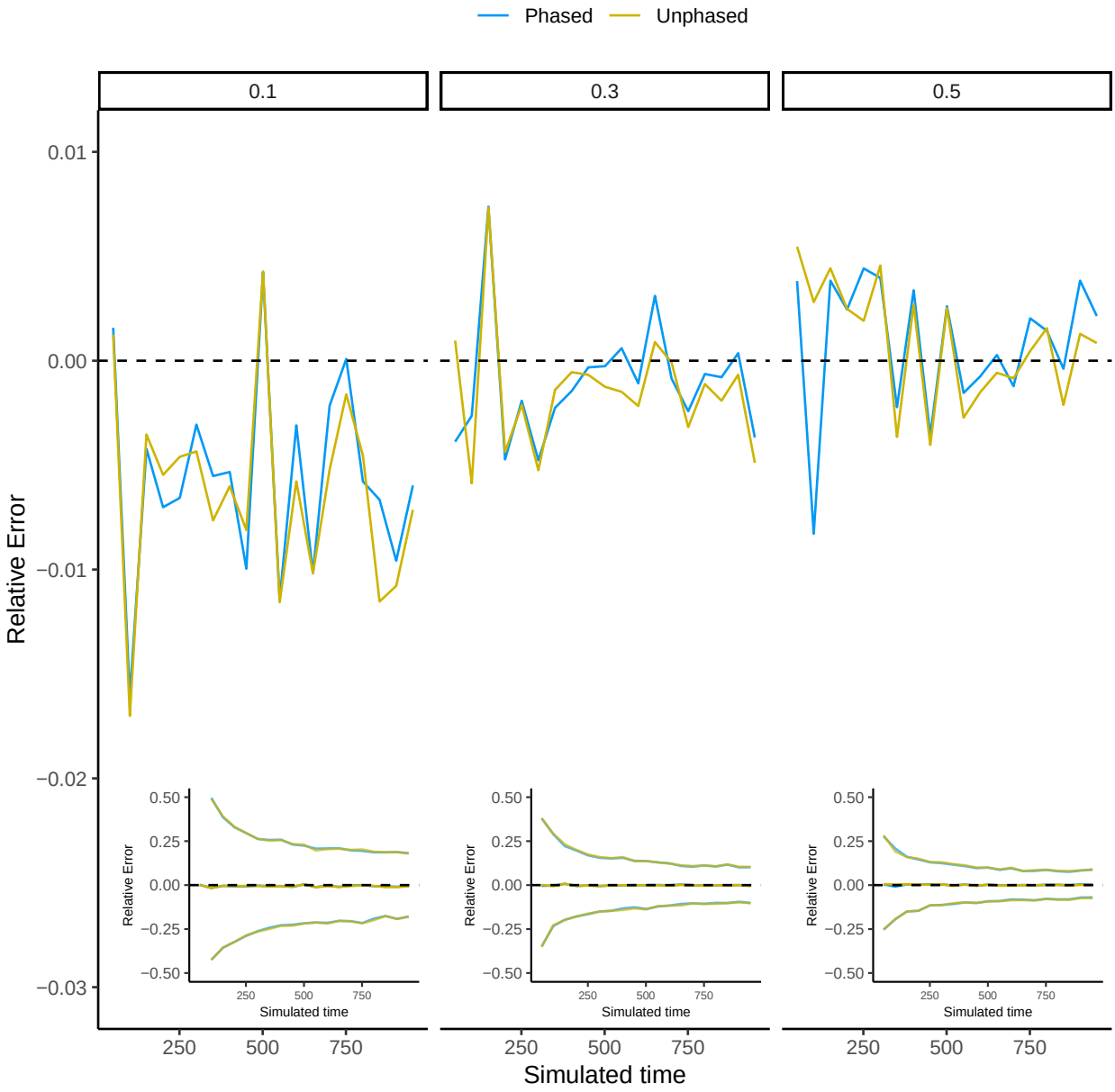


Fig 7. Accuracy in age estimate using the unphased framework versus the phased framework. Shown are the median difference across 100 replicates. We represent the results for three different initial heterozygozities, as indicated at the top of each plot. The population size was 10,000 individuals, and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

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.10.292441; this version posted September 11, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

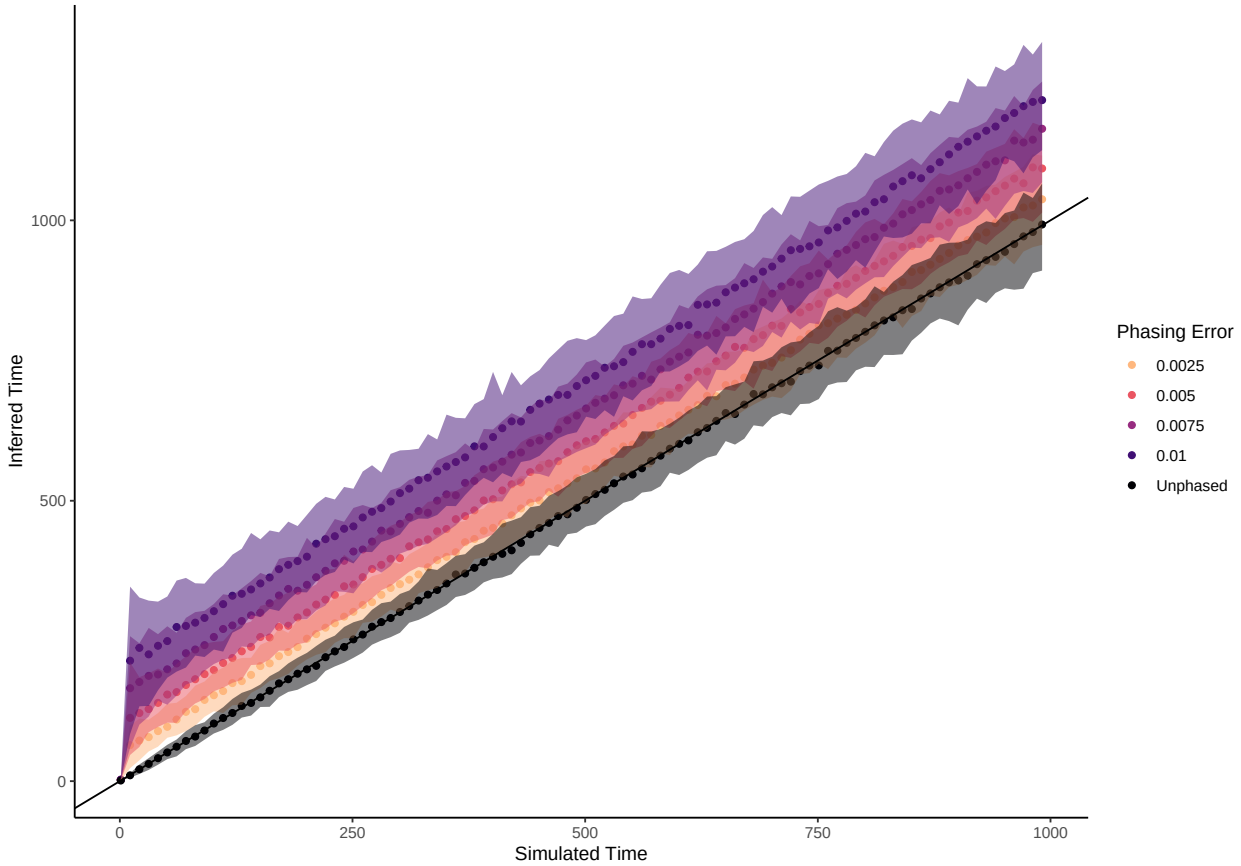


Fig 8. Effect of switching error on the estimated time since admixture. Data simulated with $N = 10,000$, $p = 0.5$, $C = 1$ and $n = 10,000$. The solid black line indicates the simulated = estimated time. Dots indicate the median inferred age and the colored area indicates the 95% confidence interval (CI) envelope. Colors reflect different degrees of phasing error, where a phasing error of 0.01 represents a 1% probability of a SNP being phased incorrectly.

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.10.292441; this version posted September 11, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

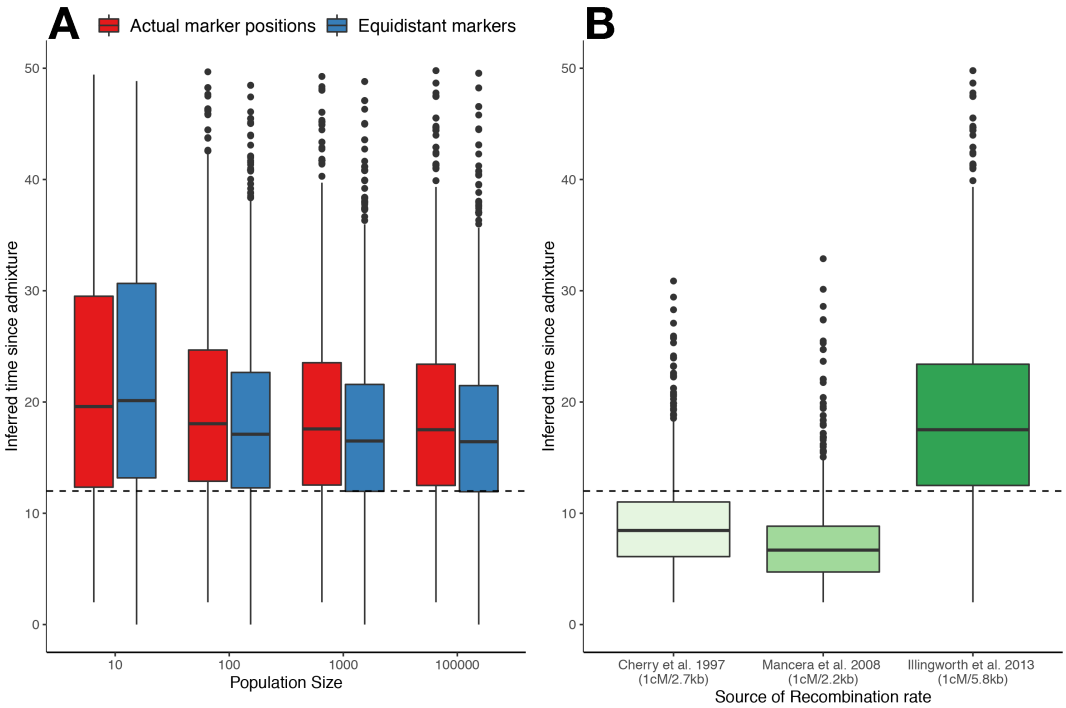


Fig 9. Inferred age for F12 Hybrid Yeast (*Saccharomyces cerevisiae*) individuals. Shown are estimates across all 16 chromosomes. The dotted line indicates the 12 generations used to generate the hybrid individuals. (A) Results using either equidistant markers ((Janzen et al., 2018)) or using actual marker positions (this paper). Shown are results using average recombination rate of $1cM/5.8kb$ as inferred in (Illingworth et al., 2013). (B) Inferred age for different recombination rates, assuming a population size of 100,000 individuals.

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.10.292441; this version posted September 11, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

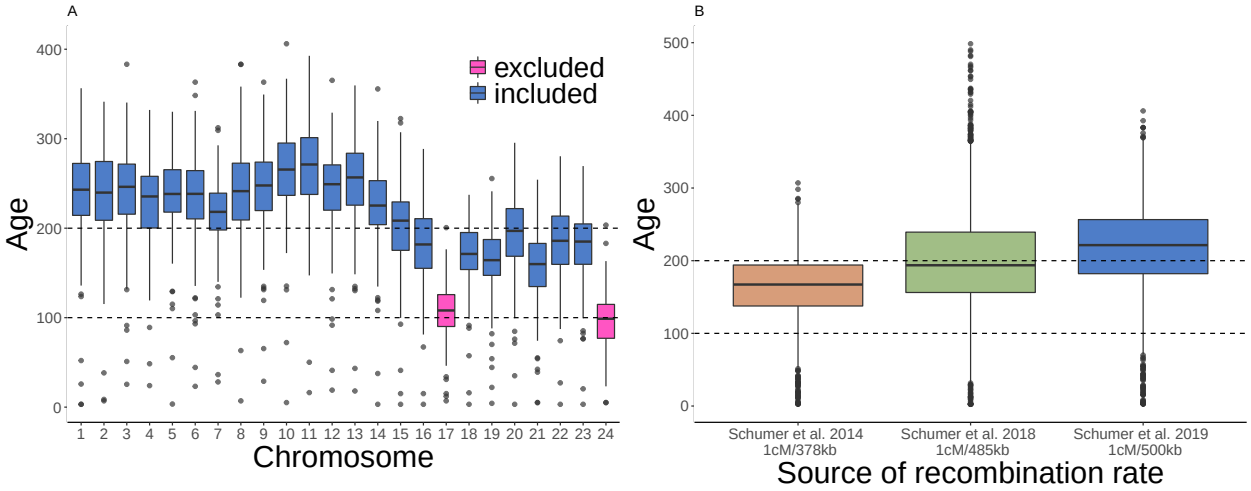


Fig 10. Inferred age for hybrid *Xiphophorus* fish from Tlatemaco (Mexico). (A):Results for each chromosome, where two chromosomes with large inversions are indicated in pink (these were excluded from the subsequent analysis). Shown are inferred ages using the recombination map from (Schumer et al., 2014a). (B) combined results (excluding chromosomes 17 and 24). The dashed line indicates the previously estimated age, based on the decay of linkage disequilibrium (56 generations). Shown are age inferences based on different recombination maps.

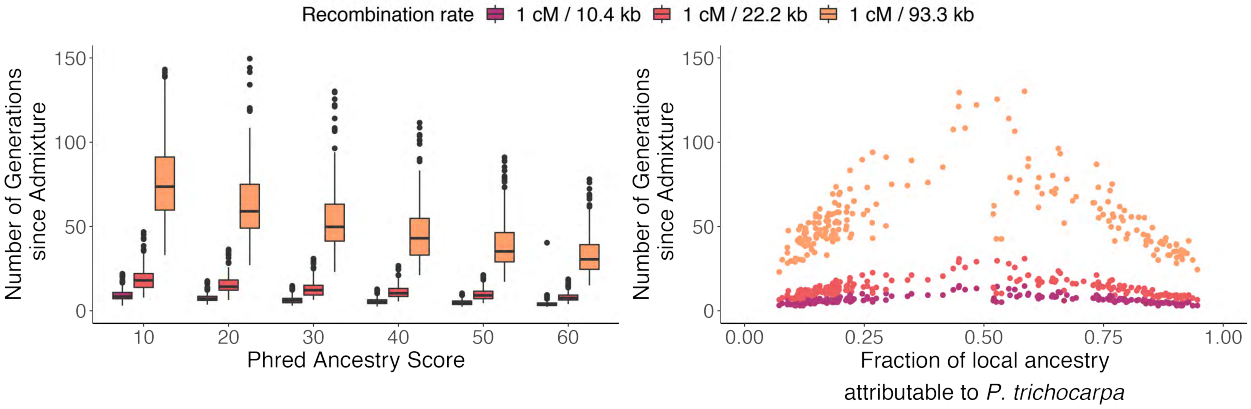


Fig 11. Inferred age for hybrid *Populus* trees. (A) .pdf Inferred time since admixture for all individuals, split out per Phred Ancestry score, where ancestry phred scores indicate the local ancestry uncertainty allowed for inclusion of markers. Colors indicate different recombination rates used: $1cM/10.4kb$ (Slavov et al., 2012), $1cM/22.2kb$ (Olson et al., 2010) and $1cM/93.3kb$ (Wang et al., 2016). (B) Inferred time since admixture for a Phred Ancestry score of 30, split out across the average frequency of *P. trichocarpa* in the admixed individual.