MOLECULAR ECOLOGY RESOURCES

Supplemental Information for:

Estimating the time since admixture from phased and unphased molecular data

Thijs Janzen and Verónica Miró Pina

Table of Contents:

S1 Appendix: Sensitivity analysis

Using individual based simulations, we test how sensitive our new framework is to variation in the parameters N, p and in the overall rate of recombination C ($C = \sum_{i=1}^{n} d_i$). As in the main text, we simulate with $N = 10,000$, $p = 0.5$ and $C = 1$, we use $n = 10,000$ to mimic coverage found in empirical data (see main text).

1.1 Population size

Fig 1 shows how sensitive inference is if we infer the time since the onset of admixture using a different value of N as that used to simulate the data ($N = 10,000$). If data from a single chromosome is used and if N is much smaller than the population size used in the simulations (e.g. if $N = 1000$), the tie since admixture tends to be overestimated. In other cases, the inferred time is very similar to the simulated time. In particular, if data from two chromosomes is available (phased or unphased), population size is of little impact on the inferred time.

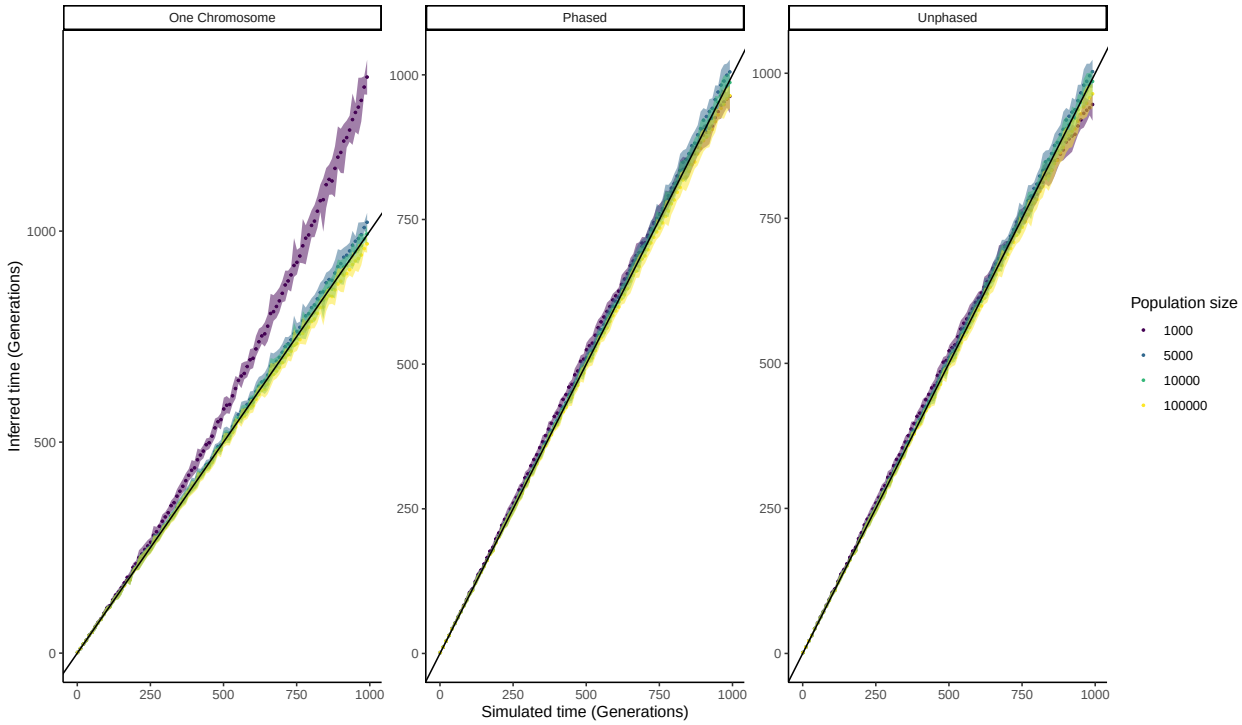

Fig 1. Sensitivity to N. Estimated time since admixture for data simulated with $N = 10,000$, $p = 0.5$, $C = 1$ and $n = 10,000$. The solid black line indicates the simulated = estimated time, colors indicate the inferred time since admixture assuming a different population size than used to simulate. Dots represent the mean estimates across 10 replicates, where for each replicate, the time since admixture was inferred by averaging over 10 individuals, per time point. Shaded areas indicate the 95% interquantile range across these replicates. The three panels represent the three different methods available: using only information from a single chromosome, using phased information from two chromosomes, or using unphased information from two chromosomes.

1.2 Recombination rate

Incorrect assessment of recombination rates has drastic effects on the inferred time (Fig 2), regardless on the method used to infer the time since admixture. An underestimation of the amount of recombination dramatically inflates the inferred time since admixture, whereas an overestimate of the amount of recombination leads to an underestimation.

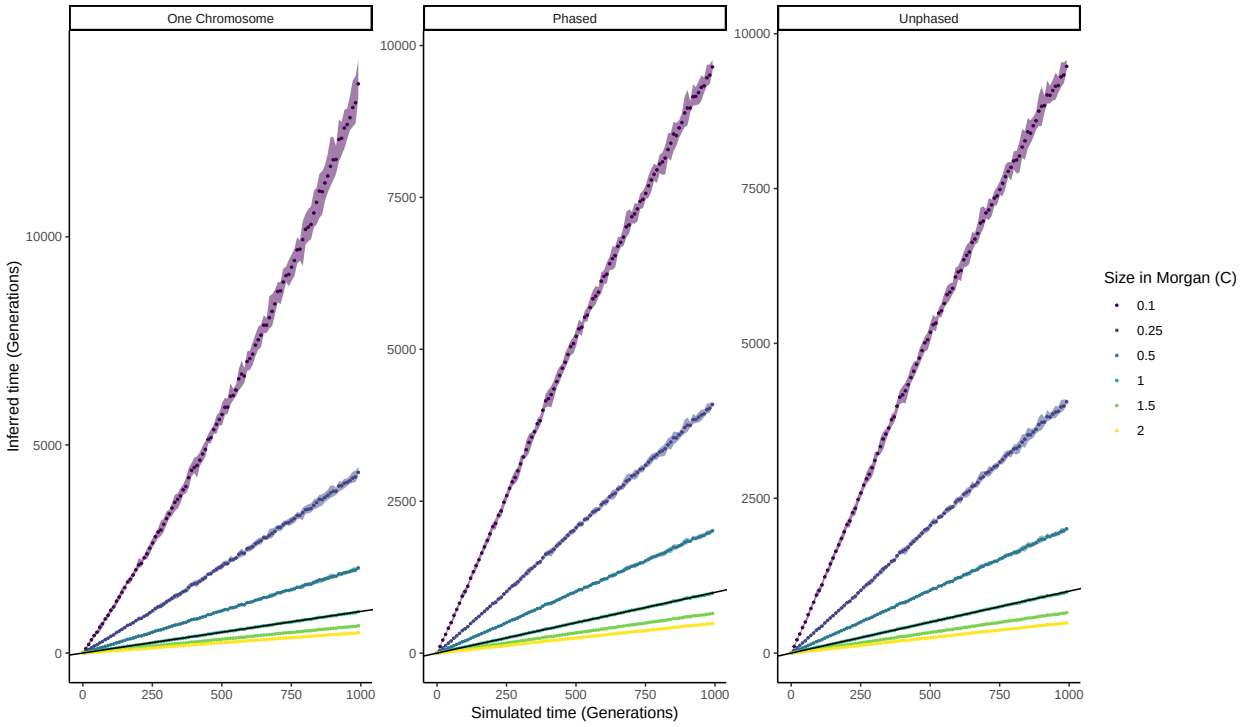

Fig 2. Sensitivity to C. Estimated time since admixture for data simulated with $N = 10,000$, $p = 0.5$, $C = 1$ and $n = 10,000$. The solid black line indicates the simulated = estimated time, colors indicate the inferred time since admixture assuming different recombination rates than used to simulate. Dots represent the mean estimates across 10 replicates, where for each replicate, the time since admixture was inferred by averaging over 10 individuals, per time point. Shaded areas indicate the 95% interquantile range across these replicates. The three panels represent the three different methods available: using only information from a single chromosome, using phased information from two chromosomes, or using unphased information from two chromosomes.

1.3 Initial heterozygosity

Incorrect information about the genetic contribution of one of the two source taxa at the onset of hybridization only leads to an overestimate of the time since admixture for extreme deviations from the value used to simulate the data (e.g. only for $p = 0.01$ or 0.99, whilst the data was simulated with $p = 0.5$), if information from both chromosomes is used (phased or unphased). If only information of a single chromosome is available, incorrect identification of p is more detrimental to the inferred time since admixture, and always leads to an overestimate (Fig 3).

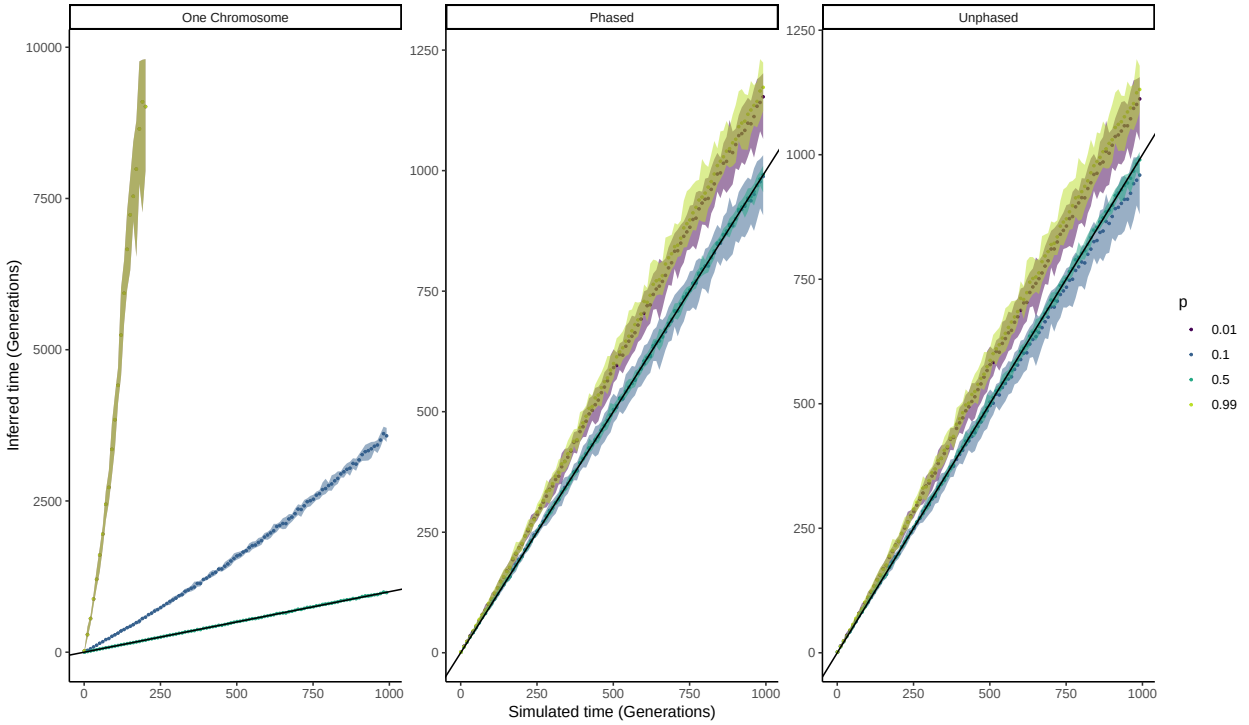

Fig 3. Sensitivity to p. Estimated time since admixture for data simulated with $N = 10,000$, $p = 0.5$, $C = 1$ and $n = 10,000$. The solid black line indicates the simulated = estimated time, colors indicate the inferred time since admixture assuming different contributions of one of the two source taxa (e.g. p) than used to simulate. Dots represent the mean estimates across 10 replicates, where for each replicate, the time since admixture was inferred by averaging over 10 individuals, per time point. Shaded areas indicate the 95% interquantile range across these replicates. The three panels represent the three different methods available: using only information from a single chromosome, using phased information from two chromosomes, or using unphased information from two chromosomes.

1.4 Number of markers

Having a low number of markers drastically increases the uncertainty of the estimated time since admixture, but only for very low numbers (e.g. <1000 markers per chromosome) (Fig 4). Although the median estimates remain identical to the expected time, the variance becomes very high when the number of markers is very low.

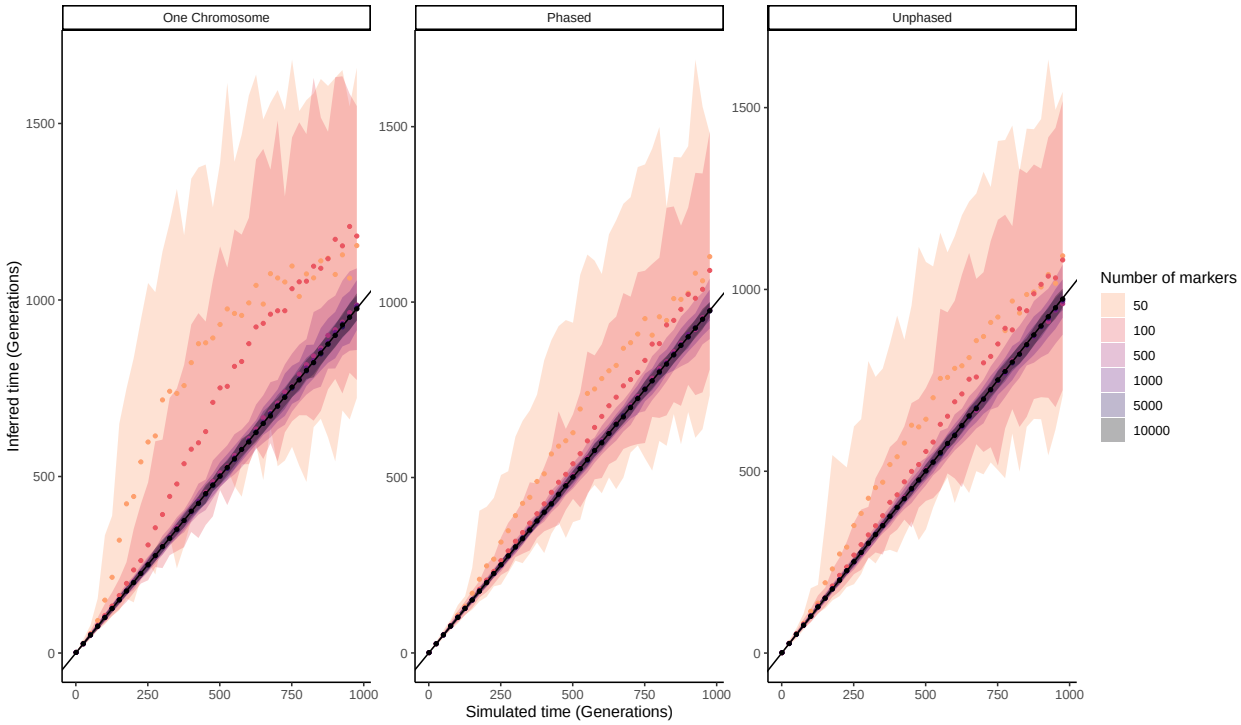

Fig 4. Sensitivity to n. Estimated time since admixture for data simulated with $N = 10,000$, $p = 0.5$ and $C = 1$. The solid black line indicates the simulated = estimated time, colors indicate the inferred time since admixture assuming different numbers of markers. Shown are the mean estimates across 100 replicates, and the 95% envelope across these replicates. The time since admixture was inferred by averaging across 10 individuals per replicate, per time point. The three panels represent the three different methods available: using only information from a single chromosome, using phased information from two chromosomes, or using unphased information from two chromosomes.

S2 Appendix: Small population size

In this appendix, we test the validity of our maximum likelihood approach and its sensitivity to the different parameters, but using a smaller value of the population size in the simulations ($N = 1000$ instead of 10000).

2.5 Validation of the method

In this section, we use the same parameters as in the individual based simulations in the main text but with a smaller population size ($N = 1000$).

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

Fig 6 shows that using the one chromosome method only small hybridization times can accurately be inferred. When the hybridization times are longer than a threshold value (around 5000 generations for $H_0 = 0.5$), the method tends to overestimate the time since admixture. This is due to the fact that there is a plateau in the function giving the number of junctions per unit of time (equation (2) in the main text). This phenomenon is called the maximum packing density of junctions.

However, with the method using information from two homologous chromosome, this problem did not appear and we could always provide an accurate estimation of the time since admixture. The same problem would have appeared if we had performed longer simulations. However, hybridization occurs in short timescales and longer timescales are incompatible with the hypothesis that we can ignore mutation events.

Fig 7 shows a comparison between the two methods that use information from two homologous chromosomes, whether the data is phased or unphased. For short times (below 5000 generations), the phased method provides slightly better results than the unphased method (the relative error is smaller). However, for long times, both methods perform equally well.

2.6 Sensitivity analysis

We also tested if changing the population size to $N = 1000$ changes the sensitivity of our method to the different parameters.

2.6.1 Population size

Varying the population size has some interesting effects (see Fig 1 in S1 Appendix). When using information from a single chromosome we observe that if N is smaller than the population size in the simulations, the time since admixture is over estimated while, if N is larger it is underestimated. When using information from two chromosomes, whether it is phased or unphased, overestimating the population size initially increases the age estimate, although extreme overestimates contrastingly lead to an underestimate.

2.6.2 Recombination rate

If we vary the amount of recombination (Fig 9), we observe that higher levels of recombination (e.g. overestimates of recombination) lead to a younger age. When using information from two chromosomes, this changes the age estimate with a constant amount. But when restricted to a single chromosome, an increased recombination leads to levelling off of the age estimate due to reaching age horizon resulting from the maximum packing density of junctions sooner.

2.6.3 Initial heterozygosity

We find very similar results S1 Appendix. Fig 10 shows that varying the initial heterozygosity has a strong impact in the method that uses one chromosome. However, when using information from the two homologous chromosomes, errors in the estimation of p don't yield such important errors in the estimation of the time since admixture.

2.6.4 Number of markers

With a smaller population size, we are able to simulate to a much larger number of generations and show that the accuracy of inference tends to be retained, provided that the number of markers is large (5,000 - 10,000 markers per chromosome), and that information on both chromosomes is used (Fig 11). When information of only a single chromosome is used, accuracy in admixture time inference is low. When information from both chromosomes is used, we observe that when the admixture time is above a certain threshold, that depends on the number of markers, inference becomes impossible (the 95% enveloppe becomes extremely large and biased towards high values).

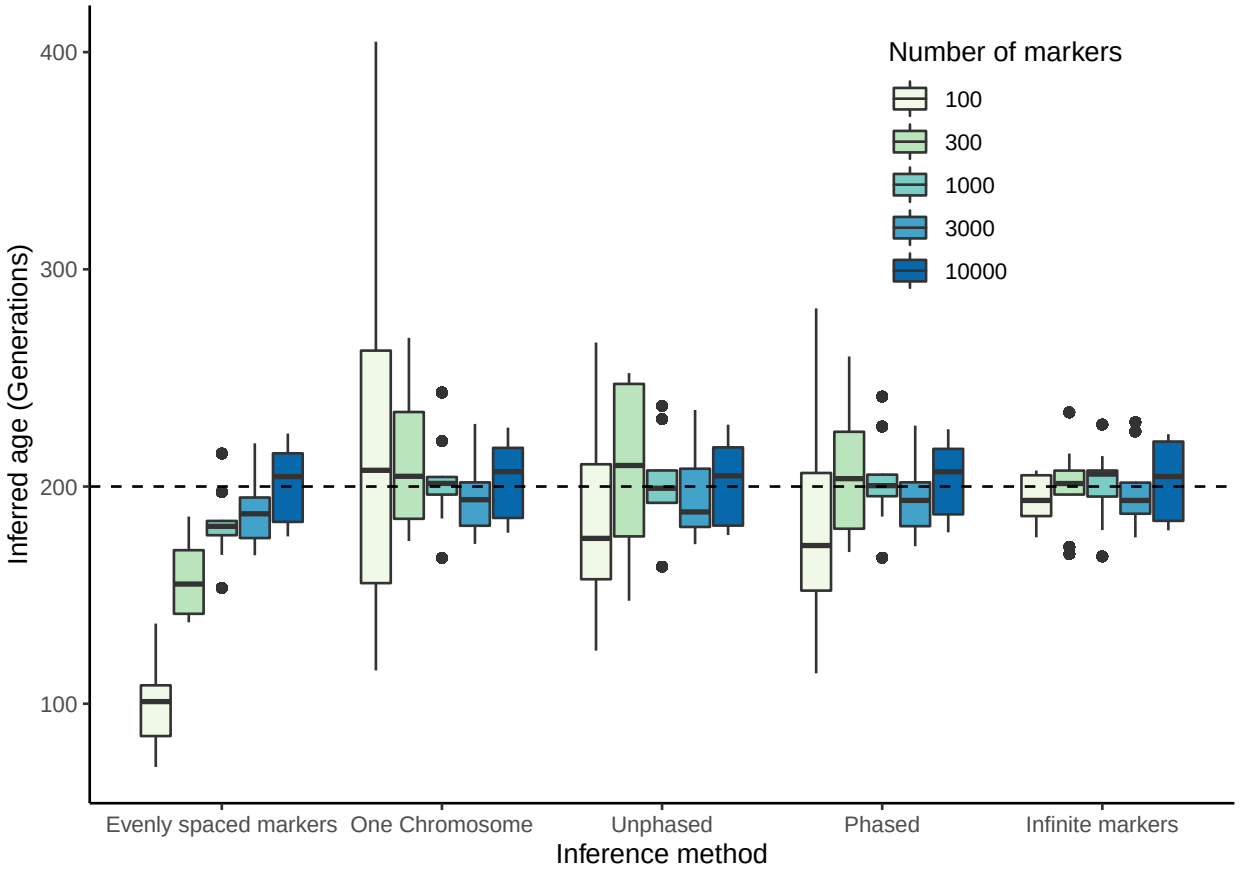

Fig 5. Inferred time using different methods. Shown are the mean estimates for 100 replicates, where in each replicate 10 individuals were analyzed. Boxplots represent the 95% interquantile range across replicates. The dashed line indicates the simulated time. 'Evenly spaced markers' corresponds to the 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 1,000 individuals, and 10,000 randomly spaced markers were used.

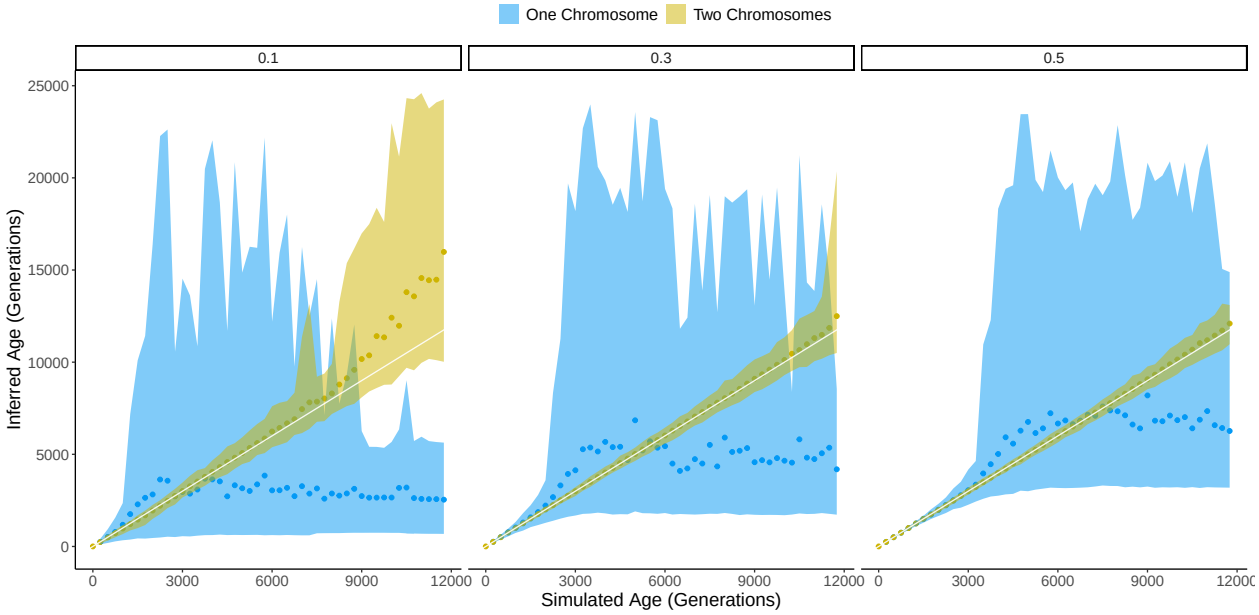

Fig 6. Inferred time versus simulated time using one or two chromosomes. Shown are the mean estimates (dots) across 100 replicates, where per replicate 10 individuals were analyzed. The solid white line indicates the observed is equal to expected line and the shaded area indicates the 95% percentile range across the 100 replicates. 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 1,000 individuals, and 10,000 randomly spaced markers were used.

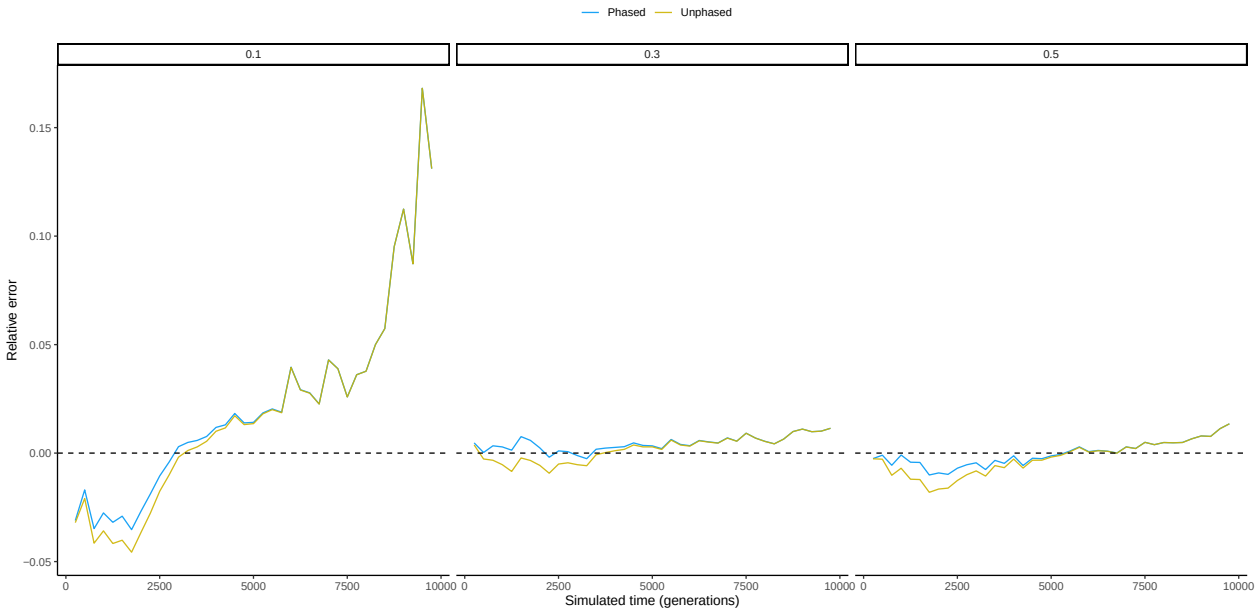

Fig 7. Accuracy in age estimate using the unphased framework versus the phased framework. Shown is the mean difference across 100 replicates where for each replicate, 10 individuals were analyzed. Shown are results for three different initial heterozygozities, indicated at the top of each plot. The population size was 1,000 individuals, and 10,000 randomly spaced markers were used.

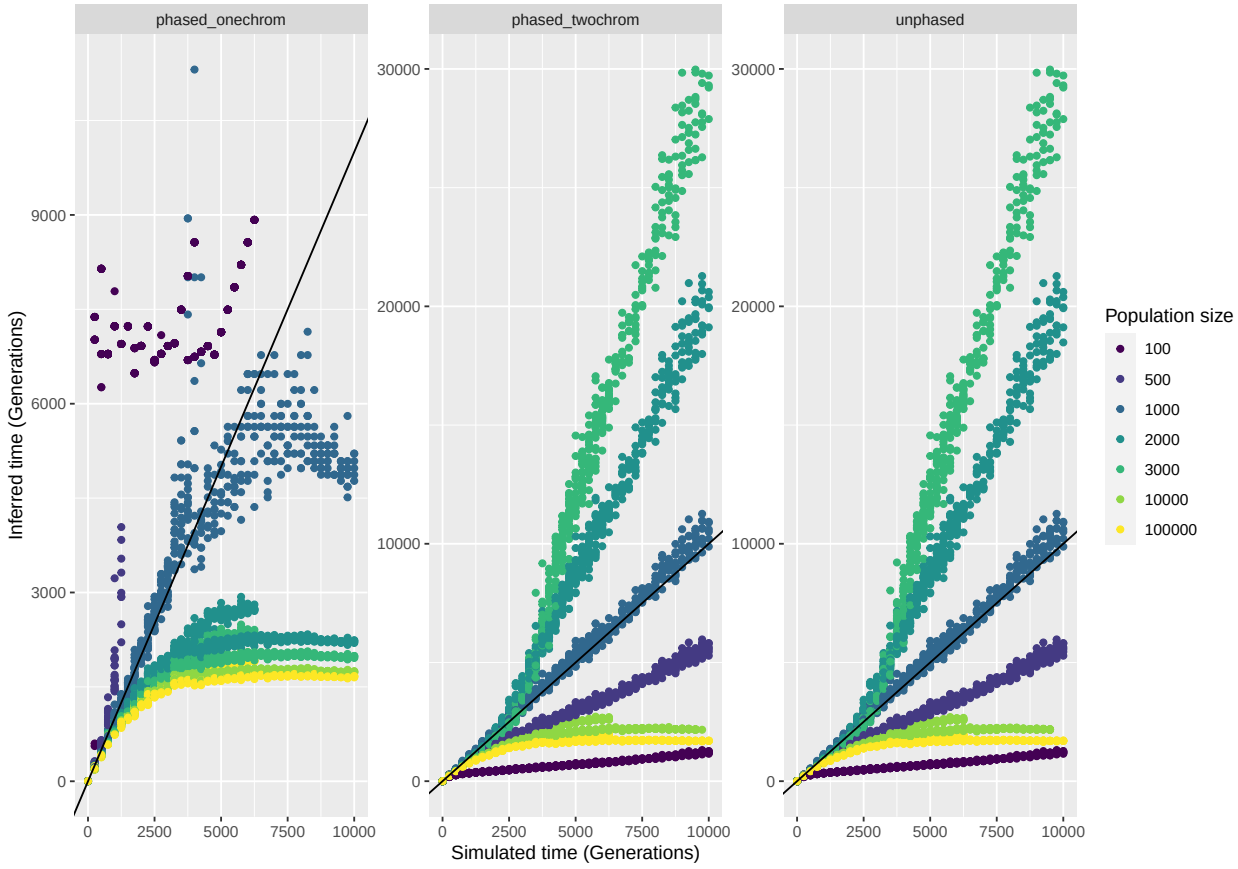

Fig 8. Sensitivity to N Estimated time since admixture for data simulated with $N = 1000$, $p = 0.5$ and $C = 1$. The solid black line indicates the simulated = estimated time, colors indicate the inferred time since admixture assuming a different value of N than used to simulate. Dots indicate the mean time estimate across 10 replicates, where for each replicate, the time since admixture was inferred by averaging over 10 individuals, per time point. Shaded areas indicate the 95% interquantile range across these replicates. The three panels represent the three different methods available: using only information from a single chromosome, using phased information from two chromosomes, or using unphased information from two chromosomes.

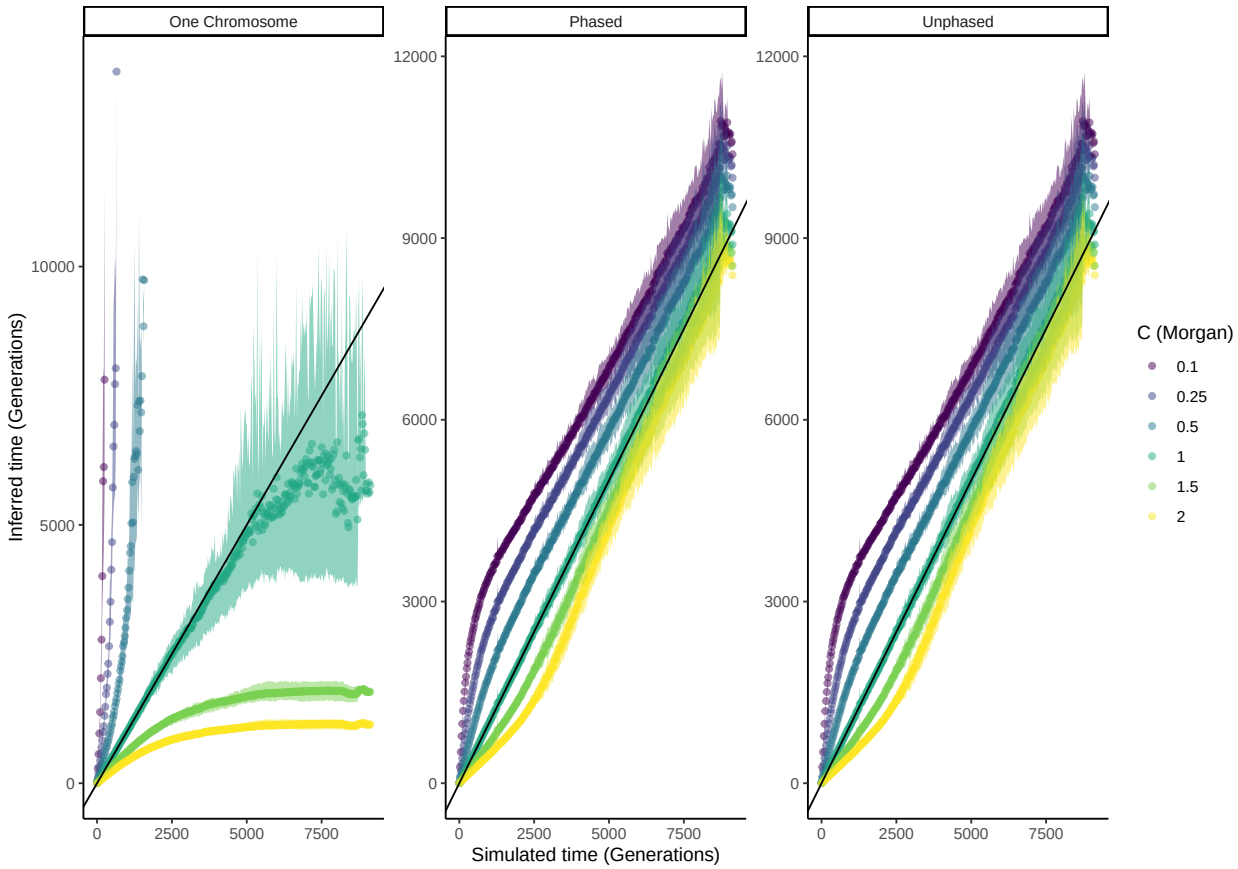

Fig 9. Sensitivity to C Estimated time since admixture for data simulated with $N = 1000$, $p = 0.5$ and $C = 1$. The solid black line indicates the simulated = estimated time, colors indicate the inferred time since admixture assuming a different rate of recombination than used to simulate. Dots indicate the mean time estimate across 10 replicates, where for each replicate, the time since admixture was inferred by averaging over 10 individuals, per time point. Shaded areas indicate the 95% range across these replicates. The three panels represent the three different methods available: using only information from a single chromosome, using phased information from two chromosomes, or using unphased information from two chromosomes.

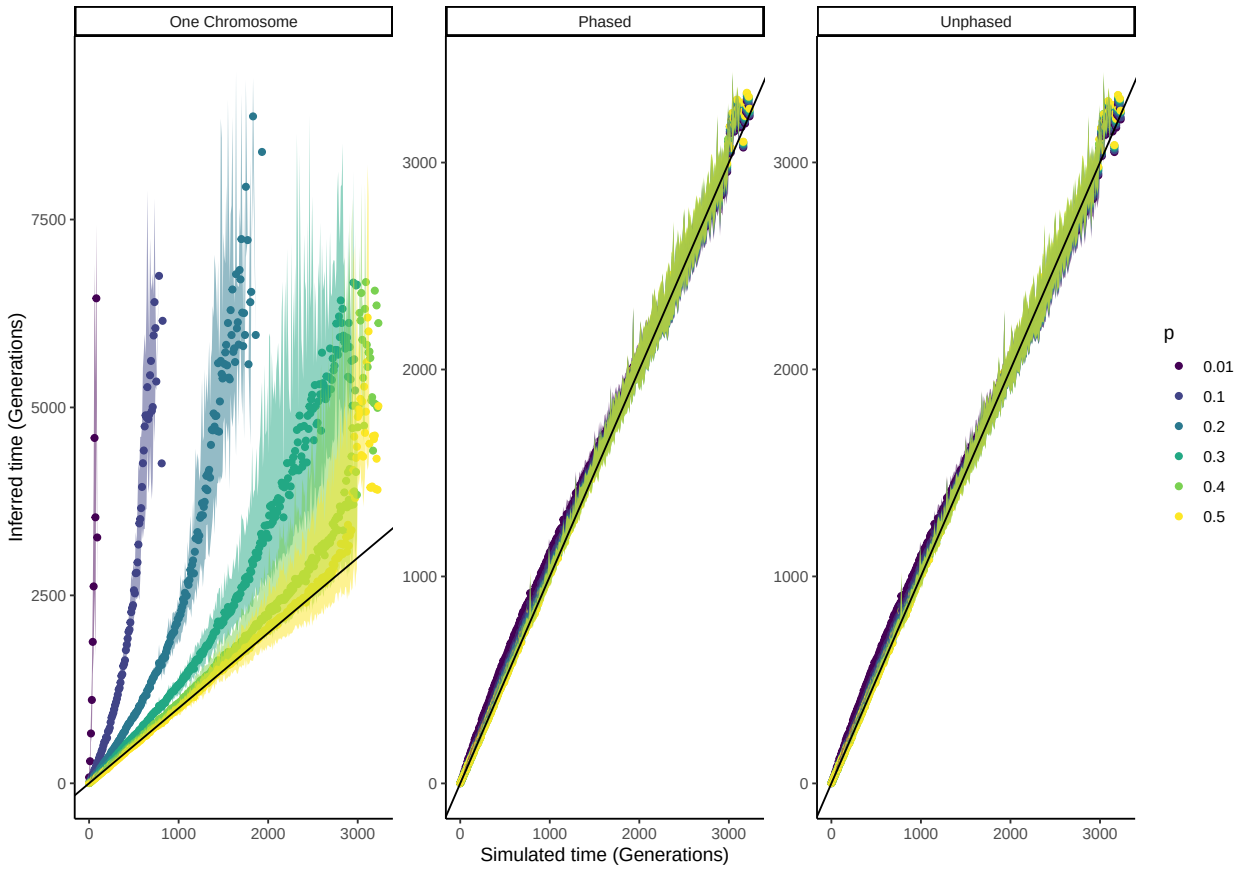

Fig 10. Sensitivity to p. Estimated time since admixture for data simulated with $N = 1000$, $p = 0.5$ and $C = 1$. The solid black line indicates the simulated = estimated time, colors indicate the inferred time since admixture assuming a different value of p than used to simulate. Dots indicate the mean time estimate across 10 replicates, where for each replicate, the time since admixture was inferred by averaging over 10 individuals, per time point. Shaded areas indicate the 95% range across these replicates. The three panels represent the three different methods available: using only information from a single chromosome, using phased information from two chromosomes, or using unphased information from two chromosomes.

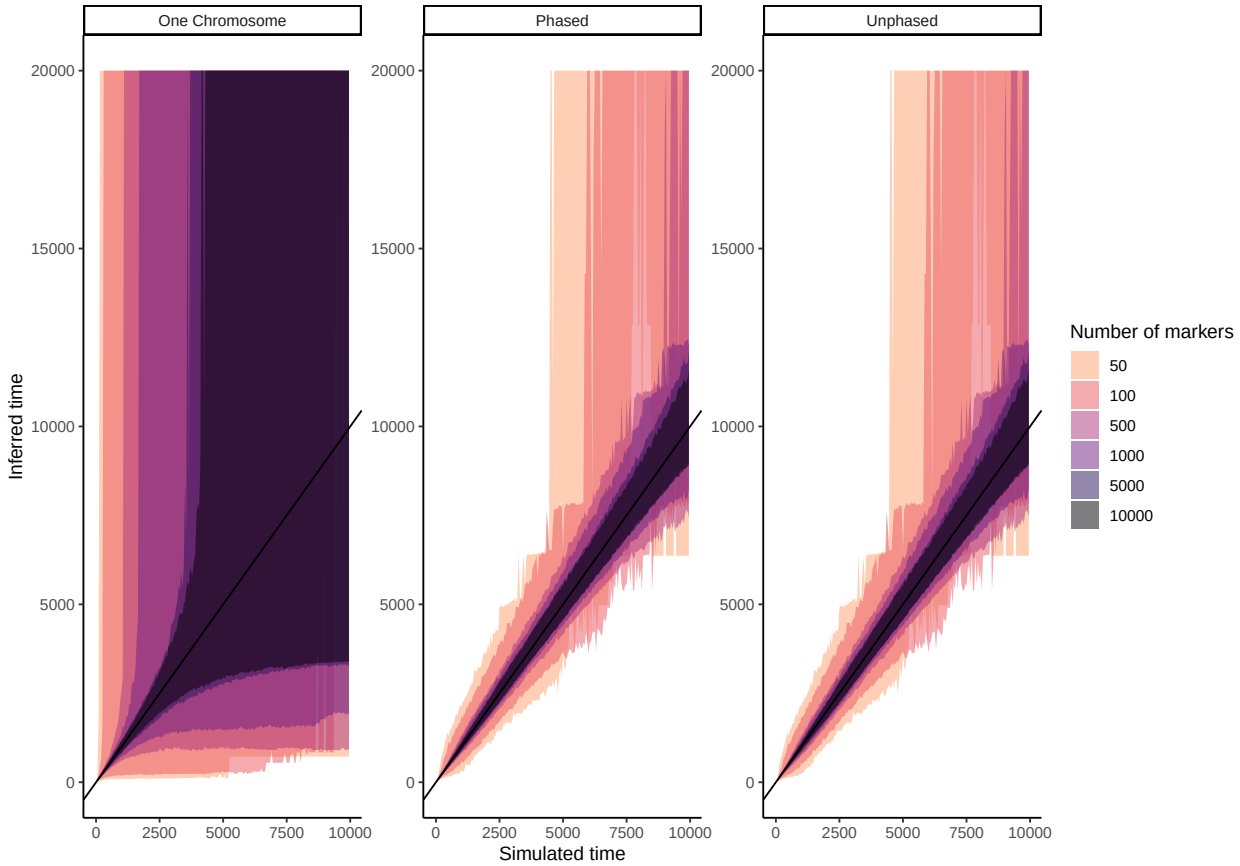

Fig 11. Sensitivity to n. Estimated time since admixture for data simulated with $N = 1,000$, $p = 0.5$ and $C = 1$. The solid black line indicates the simulated = estimated time, colors indicate the inferred time since admixture assuming different numbers of markers. Shown are the median estimates across 100 replicates, and the 95% envelope across . The time since admixture was inferred for 10 separate individuals per replicate, per time point. The three panels represent the three different methods available: using only information from a single chromosome, using phased information from two chromosomes, or using unphased information from two chromosomes.

S3 Appendix: Phasing error

We extend the analysis done in the main text by an analysis where we explore the error in inference of admixture time using only 500 or 1,000 markers. In the main text we only focused on 10,000 markers, which is much larger than the simulated time. In this scenario, incorrectly phased markers *always* introduce a fake junction, but never remove a junction. In contrast, if the number of junctions is lower or equal to the simulated time, an incorrectly phased marker might also remove an observed junction. Instead of focusing on the same percentage of phasing error, we have chosen to use the same number of incorrectly phased markers - in order to avoid having no incorrectly phased markers at all (0.25% of 500 markers has an expected value lower than 1). We simulate again with a population of 10,000 individuals, for $p = 0.5$ and $C = 1$. We find that including of incorrectly phased markers tends to increase the age estimate, even when the number of markers is relatively low. Please note that the error might seem incredibly large, but this might be due to the fact that the absolute number of incorrect markers is high compared to the total number of markers.

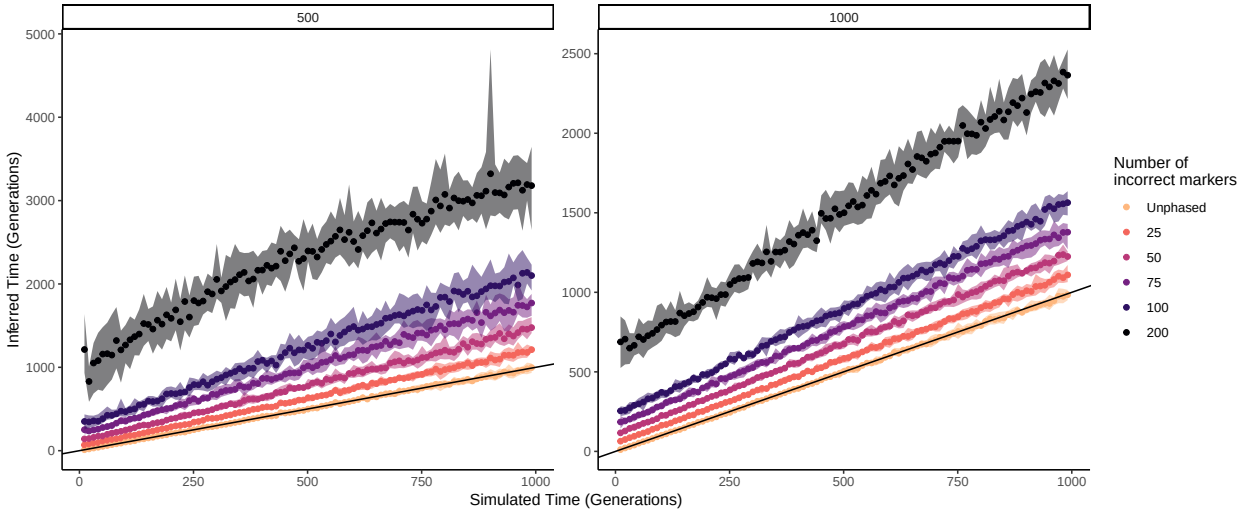

Fig 12. Effect of phasing error on the estimated time since admixture. Estimated time since admixture for data simulated with $N = 10,000$, $p = 0.5$ and $C = 1$. The solid black line indicates the simulated = estimated time, colors indicate the mean across 10 replicates, with 10 individuals measured per replicate. Shown is the inferred time since admixture including a varying number of incorrectly phased markers. Shaded area indicates the 95% envelope across the replicates. The left panel shows results for $n = 500$, the right panel shows results for $n = 1000$.

S4 Appendix: Population Size in empirical datasets

3.7 Yeast

We have repeated the analysis in the main text, but varied the population size in [10, 100, 1000, 10000, 10000]. The results show (Fig 13) that with the age estimate is irrespective of population size, as long as the population size is above 100 individuals. Only for very small population sizes (10 individuals), does the age estimate increase, which brings the age estimates obtained with the two older recombination rate estimates (Cherry et al., 1997; Mancera et al., 2008) closer to the number of generations used in the experiment.

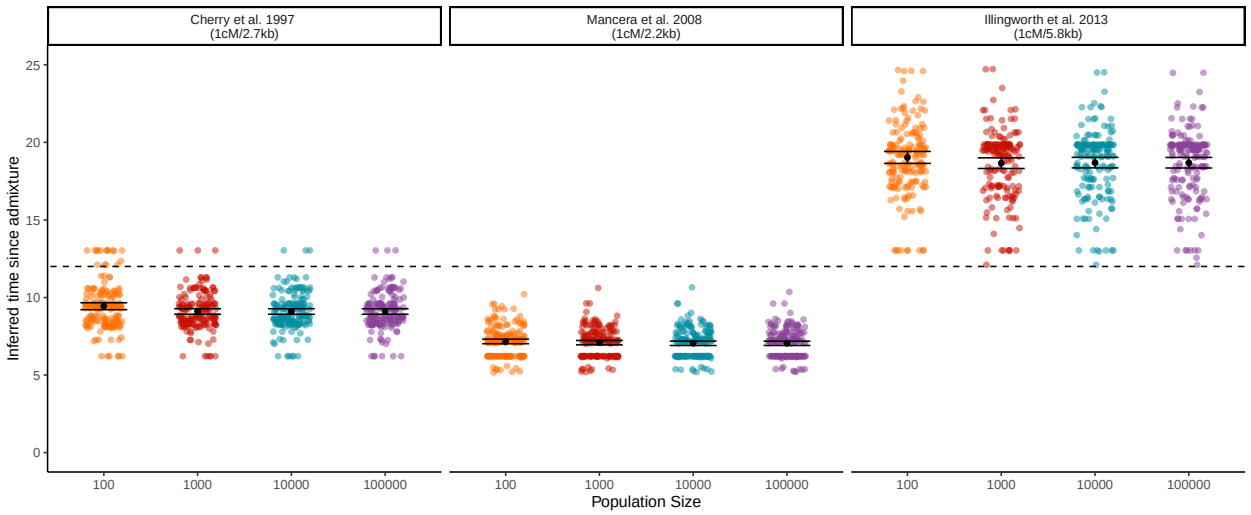

Fig 13. Inferred age for F12 Hybrid Yeast (*Saccharomyces cerevisiae*) individuals. (A) Inferred age for three different recombination rates: $1cM/2.7kb$ (Cherry et al., 1997), $1cM/2.2lb$ (Mancera et al., 2008) and $1cM/5.8kb$ (Illingworth et al., 2013). Shown is the distribution of inferred ages across 171 individuals, inferred using different population sizes. Solid black dot indicates the bootstrapped average across all individuals, black error bars indicate the 95% CI of these bootstraps.

3.8 Swordtail fish

We repeat the analysis shown in the main text, but with varying population size. Whereas the main text used the population size estimate of 1830 individuals as obtained from Schumer et al. (2014), here we explore the sensitivity of the age estimate to varying population size, by varying the value used for N in [5000, 10000, 100000, 1000000]. We find that using a (much) larger population size only reduces the age estimate to a small extent.

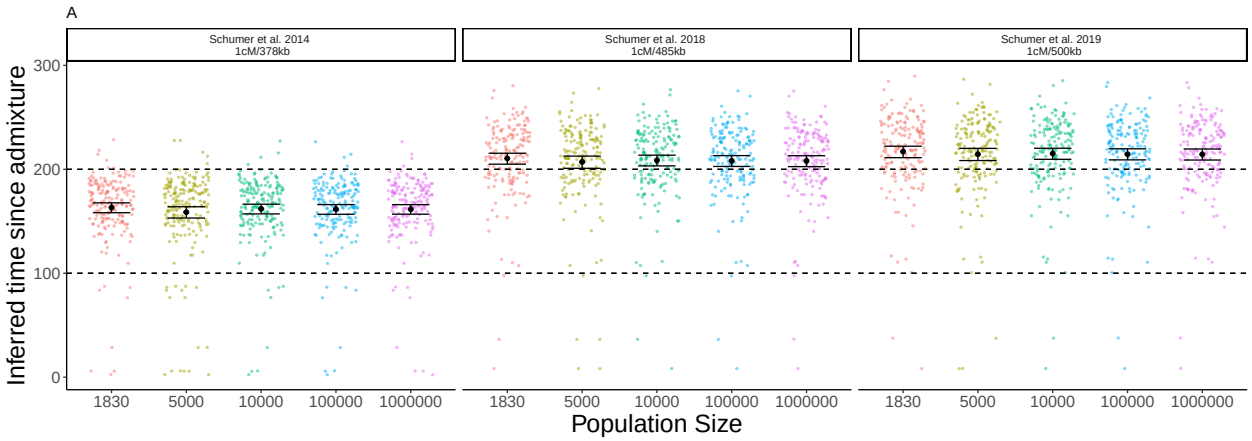

Fig 14. Inferred age for hybrid *Xiphophorus* fish from Tlatemaco (Mexico). Shown are the distribution of age inferences across 187 individuals based on three different recombination maps: $1cM/370kb$ (Schumer et al., 2014), $1cM/485kb$ (Schumer et al., 2018) and $1cM/500kb$ (Powell et al., 2020). Dotted lines indicate the hypothesized age limits of the population. Colors indicate different population sizes used during the age inference. Solid black dot indicates the bootstrapped average across all individuals, black error bars indicate the 95% CI of these bootstraps.

S5 Appendix: Comparison with ANCESTRY HMM for larger population sizes

In the main text, figure 10 shows inference of the time since admixture for markers of varying quality, using $N = 1000$. Here, we repeat the analysis, but for $N = 10000$. Figure 15 shows that ANCESTRY HMM is more accurate in inferring the time since admixture when the population size is large, and that the differences between the inferred age by ANCESTRY HMM and our method converge towards the same estimate as the number of markers increases.

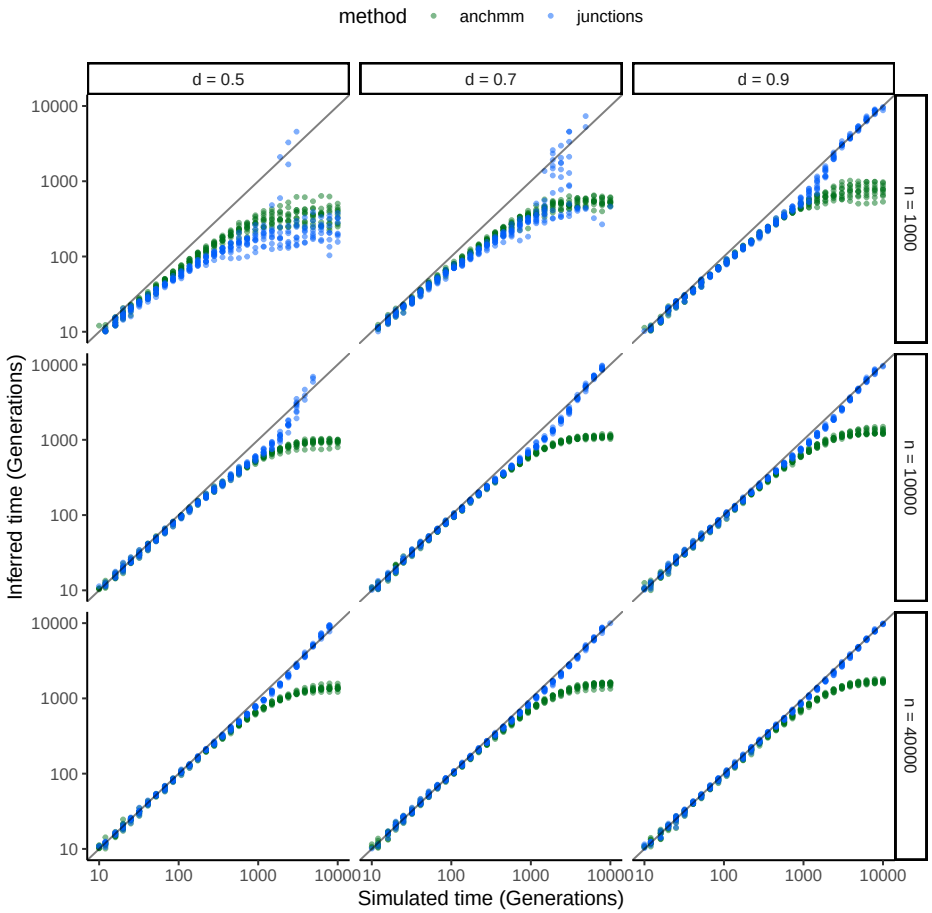

Fig 15. Comparison in estimating the time since admixture between ancestry hmm and the method proposed here, for a large population of $N = 10000$. The solid black line indicates the simulated = estimated time. Dots indicate the inferred ages, with the green dots representing ages inferred by ANCESTRY HMM and the blue dots indicate ages inferred by the junctions framework. Age estimates are based on simulated data with 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.

References

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.

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.

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, 2014.

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.

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. D. Shriver, M. W. Smith, L. Jin, A. Marcini, J. M. Akey, R. Deka, and R. E. Ferrell. Ethnic-affiliation estimation by use of population-specific dna markers. *American journal of human genetics*, 60(4):957, 1997.