Evaluating ARG-estimation methods in the context of estimating population-mean polygenic score histories

Dandan Peng[1], Obadiah J. Mulder[1], and Michael D. Edge*[1]

[1]Department of Quantitative and Computational Biology, University of Southern California

May 24, 2024

Abstract

Scalable methods for estimating marginal coalescent trees across the genome present new opportunities for studying evolution and have generated considerable excitement, with new methods extending scalability to thousands of samples. Benchmarking of the available methods has revealed general tradeoffs between accuracy and scalability, but performance in downstream applications has not always been easily predictable from general performance measures, suggesting that specific features of the ARG may be important for specific downstream applications of estimated ARGs. To exemplify this point, we benchmark ARG estimation methods with respect to a specific set of methods for estimating the historical time course of a population-mean polygenic score (PGS) using the marginal coalescent trees encoded by the ancestral recombination graph (ARG). Here we examine the performance in simulation of six ARG estimation methods: `ARGweaver`, `RENT+`, `Relate`, `tsinfer+tsdate`, `ARG-Needle/ASMC-clust`, and `SINGER`, using their estimated coalescent trees and examining bias, mean squared error (MSE), confidence interval coverage, and Type I and II error rates of the downstream methods. Although it does not scale to the sample sizes attainable by other new methods, `SINGER` produced the most accurate estimated PGS histories in many instances, even when `Relate`, `tsinfer+tsdate`, and `ARG-Needle/ASMC-clust` used samples ten times as large as those used by `SINGER`. In general, the best choice of method depends on the number of samples available and the historical time period of interest. In particular, the unprecedented sample sizes allowed by `Relate`, `tsinfer+tsdate`, and `ARG-Needle/ASMC-clust` are of greatest importance when the recent past is of interest—further back in time, most of the tree has coalesced, and differences in contemporary sample size are less salient.

*Correspondence: edgem@usc.edu

1 Introduction

The ancestral recombination graph, or ARG (Griffiths & Marjoram, 1996), is a rich representation of the history of a sample of haplotypes, including all the mutation, recombination, and common-ancestry events that affect contemporary variation. Thus, the ARG encodes all historical information that can be extracted from a sample of contemporary genomes, much of it in gene genealogies or coalescent trees (R. Hudson, 1990; Wakeley, 2016) for every location in the genome, termed "local" or "marginal" trees. The true ARG is generally unknown, and estimation of the ARG is a very challenging problem. Nonetheless, the last ten years have witnessed major advances in ARG estimation, with new methods that provide estimated ARGs with unprecedented accuracy, scalability, or both (Deng et al., 2024; Kelleher et al., 2019; Mirzaei & Wu, 2017; Rasmussen et al., 2014; Speidel et al., 2019; Zhang et al., 2023). These advances have produced a great deal of excitement about the potential of estimated ARGs in evolutionary biology and beyond (Brandt et al., 2024; Harris, 2019, 2023; Lewanski et al., 2024).

The promise of estimated ARGs depends on their performance in downstream applications. Many of the available ARG estimation methods have been benchmarked in general terms (Deng et al., 2021; Y C Brandt et al., 2022), revealing tradeoffs between accuracy and scalability. (We refer to all the methods we consider here as "ARG estimation" methods, regardless of whether they estimate the full ARG, including recombinations between marginal trees, or just a set of marginal trees.) However, early indications are that the actual performance of ARG estimators in downstream applications can vary in ways that are not necessarily predicted by a generic accuracy vs. scalability tradeoff (Fan et al., 2022, 2023). Thus, it seems as though the performance of estimated ARGs in downstream procedures may depend on specific features of the ARG and how well they are estimated. Because the ARG is complicated, it is possible in principle to extract some features of it accurately (e.g. recent coalescence times) while other features are not estimated accurately.

To explore this point in depth, we conducted thorough benchmarking of ARG estimators with respect to a set of methods for studying polygenic traits. These methods take estimated marginal trees as input. Polygenic traits—traits influenced by genetic variants from across the genome—are a promising area for applications of estimated ARGs (Edge & Coop, 2019; Link et al., 2023; Stern et al., 2021; Zhang et al., 2023). Understanding the evolution of polygenic traits is challenging in part because signals of selection are spread across many loci. Researchers can study the history of polygenic traits by examining fossil records (Kappelman, 1996) or ancient DNA (Mathieson et al., 2015) where available. However, many traits do not leave a fossil record, and ancient DNA is not always available. An alternative approach is to examine the genomes of contemporary individuals, perhaps in combination with the estimated effects of contemporary variants on target traits (Berg & Coop, 2014; Field et al., 2016; Racimo et al., 2018; Robinson et al., 2015; Uricchio et al., 2019), examining either allele-frequency differences among groups or traces of selection in patterns of within-population genetic diversity. If allele-frequency changes can be estimated

from contemporary genetic data, then those allele-frequency changes, in combination with information about associations between alleles and traits of interest, can be used to study selection on traits.

Edge & Coop (2019) proposed a set of methods to estimate the historical time course of a predicted population-mean trait level using local coalescent trees embedded in an ARG. The trait prediction is known as a polygenic score (PGS) or polygenic index. Although some of their methods are applicable to any trait prediction formed from genetic data, Edge & Coop focused on a population-mean PGS expressed as a weighted sum of population allele frequencies at unlinked loci, $Z(t) = 2\sum_{i=1}^{k} \beta_i p_i(t)$, where the weight β_i is the additive effect size on a trait of interest of an allele at locus i, and $p_i(t)$ is the frequency of the effect allele at locus i at time t. The estimators operate by estimating allele-frequency changes at the loci contributing to the PGS. Under neutrality, the proportion of lineages at time t in a coalescent tree subtending derived alleles in the contemporary sample is an unbiased estimator of the derived allele frequency at time t. Under selection, the ancestors of the sample are a biased sample from the ancestral population, but ideas from phylodynamics can be borrowed to form noisy estimates of the number of carriers of each allelic type in a specified time period.

To gain a deeper understanding of the performance of different ARG estimation approaches in estimating population-mean PGS histories, we applied the Edge & Coop framework to estimate local trees from six pieces of software. For comparison with the work of Edge & Coop (2019), we evaluated `RENT+` (Mirzaei & Wu, 2017), which they used in their original paper. We also evaluated the performance of `ARGweaver` (Rasmussen et al., 2014), which predates `RENT+` and provides more accurate estimates on smaller samples, as well as `Relate` (Speidel et al., 2019), `tsinfer+tsdate` (Kelleher et al., 2019; Wohns et al., 2022), `ARG-Needle/ASMC-clust` (Zhang et al., 2023), and `SINGER` (Deng et al., 2024), all of which scale to larger samples. We also consider the effect of sample size on the resulting estimates.

2 Methods

We simulated derived-allele frequency trajectories at all unlinked loci that affect a trait. The true population-mean PGS trajectory was computed as the weighted sum $2\sum_{i=1}^{k} \beta_i p_i(t)$ of these allele-frequency trajectories, where the weights β_i are additive effect sizes and $p_i(t)$ is the frequency of the effect allele at locus i at time t. Taking these simulated allele-frequency trajectories as input, we generated corresponding coalescent trees and haplotypes using `mssel` (Berg & Coop, 2015). The haplotypes were fed to all of the tree estimation software packages considered here, and the estimated trees were saved. Upon obtaining both the true and estimated trees, we applied the allele-frequency estimators from Edge & Coop (2019) to them, and the estimated allele-frequency trajectories were used to compute estimated population-mean PGS trajectories using the additive effect sizes. The estimated

population-mean PGS trajectory was then compared with the true trajectory using various metrics (Fig. 1).

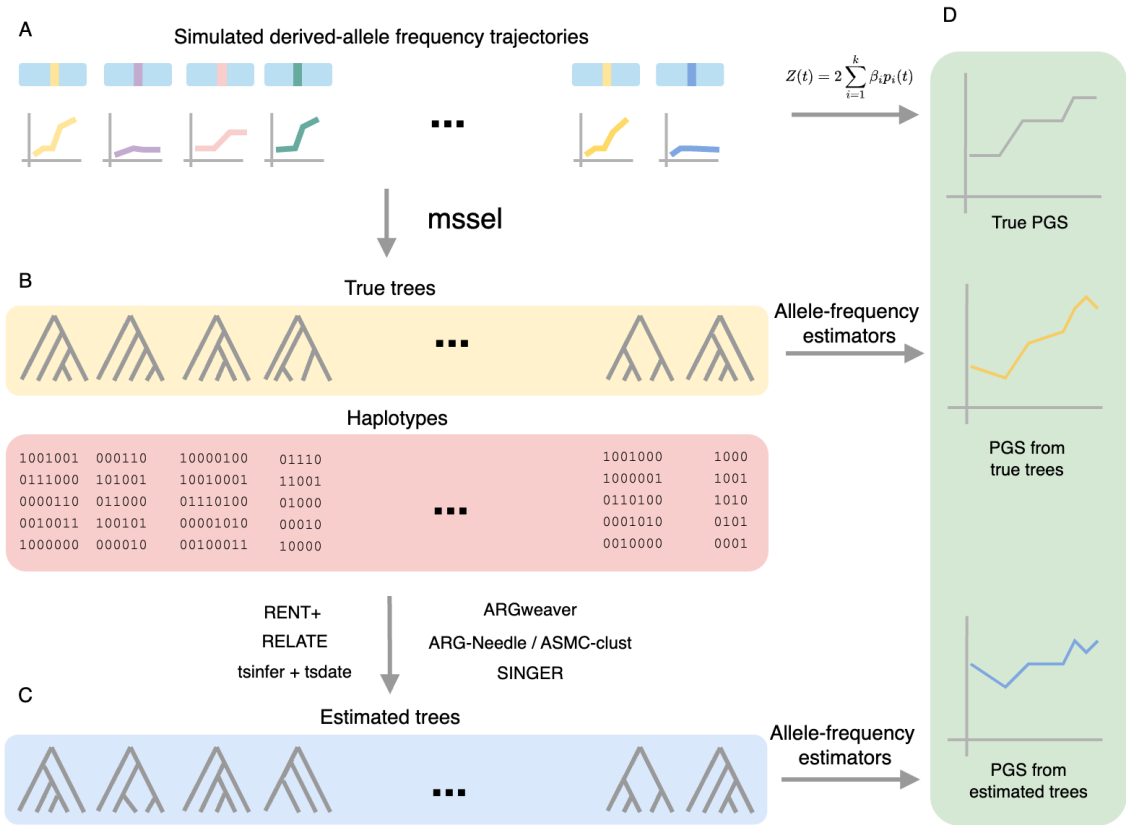

Figure 1: Method overview. A). Simulate derived allele frequency trajectories at unlinked loci. B). Generate true trees and haplotypes from those trajectories. C). Obtain estimated trees from tree estimation software. D). Apply allele-frequency estimators to true and estimated trees, then compare the true and estimated population-mean PGS trajectories.

2.1 Simulations

We simulated the population-mean PGS trajectories of traits additively determined by 100 unlinked loci under two scenarios: (i) neutral evolution, and (ii) trait-increasing directional selection occurring from 0.04 to 0.02 coalescent units ago, with neutrality at other time points. The PGS is a weighted sum of the allele frequencies, with weights equal to the additive effect sizes. There are no environmental effects or interactions—in our simulations, there is no distinction between the PGS and the trait. Further, we treat all the effect sizes as known.

For each trait, we simulated the derived allele-frequency histories of 100 trait-associated loci. For each locus, an effect size for the derived allele was drawn from a normal distribution $\mathcal{N}(0, h^2\sigma^2 w/n)$ in which the heritability h^2 and the contemporary variance σ^2 of the trait are set to 1, w is a modified version of Watterson's constant calculated as $\sum_{i=\lceil 2Nc \rceil}^{\lfloor (1-c)2N \rfloor} 1/i$ (N is the effective population size; c is the minimum minor allele frequency being drawn, we set $c = 0.01$), and n is the number of loci affecting the trait, set to 100.

Next, we simulated allele-frequency trajectories for each locus. For both neutral and selected traits, the majority of the history of each locus was simulated backward in time—for neutral traits, simulation was entirely backward, and for selected traits, simulation was backward up to the onset of selection. We set the probability of obtaining a derived-allele frequency $k/2N$ at the more recent end of the period simulated backward in time to be inversely proportional to k, where $k \in \{1, ..., 2N-1\}$. This sampled allele frequency serves as the present derived-allele frequency under neutrality and as the initial allele frequency at the onset of the selection in the recent-selection scenario.

Given a derived-allele frequency $p_i(t)$, the allele frequency at $1/(2N)$ coalescent time units in the past was drawn from $\mathcal{N}(p_i(t)(2 - 1/2N), p_i(t)(1 - p_i(t))/2N)$ (Berg & Coop, 2015; Lee & Coop, 2017; Przeworski et al., 2005). During the selection period, we simulated the derived-allele frequency forward in time in steps of $1/(2N)$ coalescent units. The frequency $p_i(t+1)$ was drawn from $\mathcal{N}(p_i(t)+sp_i(t)(1-p_i(t)), p_i(t)(1-p_i(t))/2N)$ conditional on the frequency $p_i(t)$, where s is the selection coefficient on the derived allele at time t. The value of s is $\alpha\beta$, where α is the selection gradient on the trait at time t and β is the effect size of the derived allele (we set the value of α according to the expected trait variance at the onset of selection, which was 1).

After the allele-frequency trajectory was simulated, the polygenic-score trajectory calculated based on the allele-frequency trajectory was retained if the difference in population-mean PGS between the onset and end of the selection was within 5% of the expected change $2N\delta tS$, where δt is the duration of selection in coalescent units and S is the selection differential on the PGS.

We used an unpublished modified version of `ms` (R. R. Hudson, 2002) called `mssel` (Berg & Coop, 2015) to produce simulated local coalescent trees and haplotypes across a region flanking the causal variant. In `mssel`, the sample size was set to 2,000 and the number of derived chromosomes was drawn as a binomial random variable with a size of 2,000 and success frequency equal to the contemporary allele frequency. We selected an effective population size N of 10,000, and a haplotype length of 200,000 (with the effect locus at position 100,000). The per-base-pair mutation rate was set as 2e-8, and the per-base-pair recombination rate was set as 2.5e-8. These values were transformed to population-scaled `mssel` inputs of `-r 199.5` and `-t 159.68`.

2.2 Software specifications

The coalescent trees produced by `mssel` for the selected sites are the true trees, and the haplotypes corresponding to the true trees were used as input to ARG estimation software to generate estimated trees. A brief description of each piece of estimation software can be found in the supplementary text. The ARG estimation programs vary with respect to the total sample size that can be used. We thus down-sampled the 2,000 simulated haplotypes generated in each simulation as necessary. We used 20 randomly drawn haplotypes as input to `ARGweaver`, and 20 and 200 haplotypes for `RENT+` and `SINGER`. For comparison with these methods, we also used the same subsamples of 20 and 200 haplotypes as input to `Relate` and `tsinfer+tsdate`. We did not use subsamples for `ARG-Needle`/`ASMC-clust`, which requires sample sizes of at least 300 haplotypes.

`ARGweaver`

`ARGweaver` site input files were generated by a custom `Python` function. We ran the `arg-sample` program with the same values in the simulation (`-N 10000`, `-r 2.5e-8`, `-m 2e-8`) with the SMC' model (`-smcprime`). The total number of sample iterations (`-n`) was set to 700 with the first 200 iterations as burn-in, and 50 estimated trees were extracted for one locus, one from every 10th iteration (`-sample-step 10`).

`RENT+`

Haplotypes generated by `mssel` were converted to `RENT+` format. We ran `RENT+` using `-t` to estimate branch lengths for local trees and `-l` to specify the proportional positions on the chromosomes.

`Relate`

We wrote a script to generate `Relate` haps, map, and sample files from simulated haplotypes. `Relate` was run with `-mode All`, `-m 2e-8` (mutation rate), and `-N 20000` (haploid effective population size). We used the add-on module `RelateExtract` to extract the tree corresponding to the selected site.

`tsinfer+tsdate`

The `tsinfer` input file was generated by calling the `add_site` function. Coalescent trees were estimated with `tsinfer` using default settings. The age of nodes in the tree was estimated using `tsdate` with the same parameter values as in the simulations (`Ne = 10000`, `mutation_rate = 2e-8`). We converted the tree at the selected site to Newick format with the `as_newick` method in tskit.

ARG-Needle/ASMC-clust

The haps, map, and sample input file of ARG-Needle were generated by custom R functions. We set the -normalize parameter as the default value and used the constant 20K-sized (haploid population size) demography for -normalize_demography. ARG-Needle requires a "decoding file" containing information about the demographic model, time discretization, and allele-frequency information. We created our own decoding file with the prepare_decoding function in Python package asmc – derived allele frequencies were calculated on the basis of the simulated haplotypes; the discretized time intervals were set as 14 intervals of 30 generations each starting in the present (0, 30, 60..., 420), followed by 16 more ancient intervals of 100 generations (520, 620, 720,..., 1920); the demography file was built with a constant population size ($N_e = 20,000$). The output tree was converted to Newick format with the arg_to_newick function from the arg_needle_lib package. To run ASMC-clust, we set the parameter -asmc_clust to 1. Other parameters were kept at the default values.

SINGER

The VCF file was generated from the simulated data by a custom R function. The mutation rate (-m) was set to 2e-8 and the ratio (-ratio) was set to 1.25. The population size (-Ne) was set to 10,000. SINGER was run for 700 iterations, with the first 200 iterations as burn-in. We took 50 samples from the rest, thinning every 10 iterations (-step 10). The estimated trees were stored as tree sequences. The marginal tree at the selected site was converted to Newick format with the as_newick method in tskit.

2.3 Branch length units

The ARG estimation tools we considered use different units for branch length. We standardized branch lengths to units of $2N$ generations for all trees. True trees from mssel are produced in units of $4N$ generations, so we multiplied the branch lengths by 2. In the manuscript describing RENT+, there is some ambiguity about the units, which are described as "standard coalescent units" (typically units of $2N$ generations), but in some calculations appear as if they are in units of $4N$ generations (Supplementary Table S1-S2). We applied the estimators on both the original branch lengths (assuming units of $2N$ generations) and a rescaled branch length (assuming the reported values are in units of $4N$ generations). We show the rescaled version in comparison with the other methods—although Edge & Coop (2019) treated the reported branch lengths as if they were in units of $2N$ generations, we believe that the makers of RENT+ intended the branch lengths to be interpreted as in units of $4N$ generations. However, assuming $2N$-generation units leads to slightly better performance on average (Supplementary Fig. S1-S2, Supplementary Table S1). The ARGweaver, tsinfer+tsdate, ARG-Needle/ASMC-clust, and SINGER trees are reported in units of generations, and we divided their original branch length by $2N$. Relate trees use

years as the unit and assume 28 years per generation, so we divided the original branch length by $28 * 2N$. The comparison of the tMRCA from the original branch length and the scaled branch length is listed in Supplementary Tables S1 and S2.

2.4 Applying population-mean PGS trajectory estimators to the estimated trees

We applied the three estimators proposed in Edge & Coop (2019): the "proportion-of-lineages" estimator, the "waiting-time" estimator, and the "lineages-remaining" estimator (further description of the estimators in Supplementary text). All three estimators were applied to the estimated trees to obtain the estimated allele-frequency trajectories for each locus. For `ARGweaver` and `SINGER` trees, since there is more than one estimated tree for each locus, we applied the estimators to each sampled tree and took the average result from 50 trees as the estimated frequency trajectory for that locus. As discussed in Edge & Coop (2019), the proportion-of-lineages estimator is expected to perform well under neutral evolution but to be biased during periods of directional selection. The other two estimators are expected to be approximately unbiased given the true trees but to be much more variable. The waiting-time estimator, as written in Edge & Coop (2019) is not applicable in cases in which there are polytomies in a coalescent tree. We devised a scheme to address this limitation (see Supplementary text).

As a general note about the estimators of Edge & Coop (2019), we emphasize that the relationship between changes in the population-mean PGS and trait changes is not straightforward. Even if PGS histories are estimated perfectly on an accurate PGS, phenotypes may be influenced by environmental changes and gene-by-gene or gene-by-environment interactions. Additionally, the linkage disequilibrium (LD) between markers and causal variants will change over time, causing the relationship between markers and the phenotype to change with it (Martin et al., 2017). Finally, some variants affecting the trait in the past will have been lost in the present (Carlson et al., 2022) or may no longer be detectable in GWAS. Nonetheless, Edge & Coop's methods can be used to identify and roughly date patterns of allele-frequency change that are not consistent with neutral evolution.

2.5 Benchmarking metrics

We evaluated the performance of the ARG-estimation software as inputs to various approaches to PGS-history estimation, comparing them in terms of bias, mean squared error (MSE), 95% confidence interval coverage, and type I error rates and power (at the 0.05 level) of the T_X statistic for testing directional selection. The T_X statistic is computed as a sum of squared, standardized changes in an estimated population-mean PGS between pre-specified time points. That is, $T_X = \sum_{j=0}^{w} X_j^2$, with

$$X_j = \frac{Z(t_j) - Z(t_{j-1})}{\sqrt{2V_A(t_j - t_{j-1})}},$$

where $Z(t_j)$ is the population-mean PGS at time t_j, and V_A is the additive genetic variance of the PGS ($2\sum_{i=1}^{k}\beta_i^2 p_i(1-p_i)$). Under neutral evolution, X_j approximately follows a standard normal distribution, and the allele-frequency changes are independent in distinct time intervals. Thus T_X, the sum of the X_j^2, should approximately follow the $\chi^2(w)$ distribution. However, under directional selection, the population-mean PGS changes more quickly than predicted under neutrality, leading to T_X values that are large compared with the expected $\chi^2(w)$ distribution.

The efficacy of Edge & Coop's estimators depends on the accuracy of branch lengths in the estimated trees. To elucidate the differences in software performance, we extracted the pairwise coalescent times from true trees and estimated trees. Then we compared the point estimate of coalescent times between the two, and the overall distribution of pairwise coalescent times from estimated trees against the expected exponential distribution, similar to the benchmarking recently performed by Brandt et al. (2022). To obtain the samples of pairwise coalescent times, we gathered trees from 10 randomly selected traits, resulting in a total of 1,000 trees. Specifically, for `Relate`, `tsinfer`, and `ARG-Needle/ASMC-clust`, we collected pairwise coalescence times from their 2,000-sample trees. For `RENT+` and `SINGER`, we used 200-sample subtrees, and for `ARGweaver`, we used 20-sample subtrees. In the case of `ARGweaver` and `SINGER`, we averaged pairwise coalescence time across 50 sampled trees per locus. The averaged times correspond to the same pairs of tips in each sampled tree. Due to memory constraints, we could not plot all pairwise coalescence times from one-thousand 200-sample and 2,000-sample trees. Instead, we randomly sampled 190,000 pairwise coalescence times, which is 1,000 times the total number of pairs of tips on a 20-tip tree (i.e. 1,000 times $\binom{20}{2}$).

We also examined the impact of a selection event on the distribution of pairwise coalescence times by setting the selection intensity to ∼0.004, causing the derived-allele frequency to increase from approximately 0.3 to 0.7 during the selection period. Then 100 estimated trees with 300 samples were generated on the basis of this allele-frequency trajectory. We also compared the MSE of the estimates from true trees with different sample sizes (20, 200, and 2,000 haplotypes) to assess the impact of the increased sample size.

3 Results

3.1 Runtime comparison

The tools vary substantially in their runtime. One important difference is between tools that incorporate MCMC sampling (`ARGweaver` and `SINGER`) and those that do not. Both `ARGweaver` and `SINGER` sample trees from the posterior distribution and the reported runtime represent 700 MCMC samples (200 burn-in). `SINGER` is, as expected, substantially faster than `ARGweaver` (Fig. 2A). Additionally, `tsinfer` has a clear advantage in speed with larger sample sizes. Further details of the runtime of other tools can be found in Supplementary Table S4.

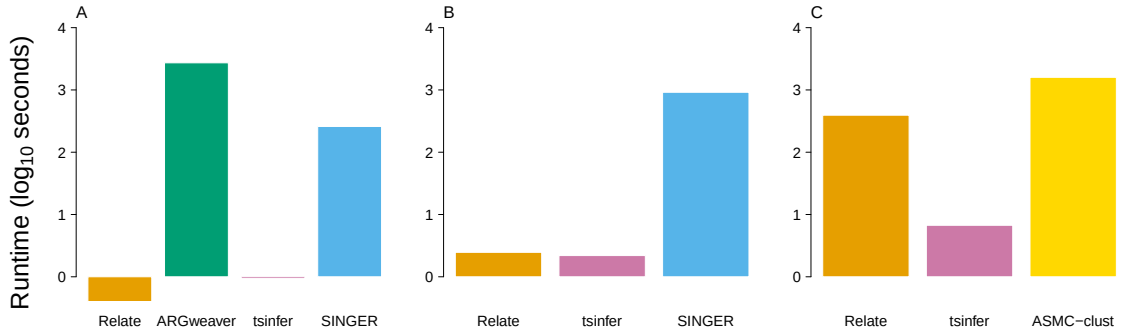

Figure 2: Average runtime for estimating ARGs for one hundred 200kb segments of 20, 200, and 2,000 samples (A, B, C) from the ARG-estimated tools. Runtimes are measured in seconds and displayed on a logarithmic (base 10) scale.

3.2 Bias, mean Squared Error, and confidence Interval coverage under neutral evolution

When traits are simulated under neutral evolution and the true trees are known, the proportion-of-lineages procedure can be viewed as a maximum-likelihood estimator for the allele-frequency trajectories (Edge & Coop, 2019). The waiting-time and lineages-remaining estimators have previously been observed to be more variable than the proportion-of-lineages estimator with either the true trees or `RENT+` trees (Edge & Coop, 2019).

With respect to bias, we find that, in line with previous results, none of the estimators display much bias when estimated with the true trees or `RENT+` under neutrality. In addition, almost all newly tested tools/algorithms (`Relate`, `tsinfer+tsdate`, `ARGweaver`, `ASMC-clust`, and `SINGER`) follow similar patterns. Differences in bias resulting from trees estimated with these different pieces of software, especially during the recent past, and especially with the proportion-of-lineages estimator, are relatively minor (Fig. 3A-C, Supplementary Fig. S3-S5). We report the results calculated from `ASMC-clust` trees instead of `ARG-Needle` trees due to the latter showing a higher bias than other estimated trees (Supplementary Fig. S5). The likely reason for this is that `ARG-Needle`'s algorithm does not guarantee that mutations map to particular branches in the ARG. Instead, carriers of the derived allele may form polyphyletic groups on the local marginal tree. The frequency of non-monophyly among tips carrying the derived allele is much higher in `ARG-Needle` trees than in trees estimated by any other method (Supplementary Table S5). Although such polyphyly may be acceptable for many purposes, it can cause major problems for the estimators of Edge & Coop (2019), particularly in the ancient past. However, if `ARG-Needle` trees are manually modified to force monophyly among derived-allele-carrying tips, then the bias of `ARG-Needle` is similar to other methods (Supplementary Fig. S7).

The advantage of the proportion-of-lineages estimator over the waiting-time and lineages-remaining estimators under neutrality is more apparent when examining estimated MSE. In the three MSE plots (Fig. 3D-F), MSE tends to increase from the recent past into the distant past for all estimators and all software packages. This is unsurprising. Under neutrality, the variance of the proportion-of-lineages estimator is inversely related to the number of lineages ancestral to the sample (Edge & Coop, 2019). Thus, the fact that most lineages coalesce in the recent past implies that the estimator is more variable in the distant past. The proportion-of-lineages estimator (Fig. 3D) exhibits a much lower MSE than the other two estimators. The true trees, again unsurprisingly, provide the best MSE. However, among the estimated ARGs, the ones produced by software that scales to large samples do not necessarily show markedly better performance. Across much of the range examined, `SINGER`, which uses samples of only 200 haplotypes, has the lowest MSE, and in the more distant past, `ARGweaver`, which uses samples of only 20 haplotypes, is comparable with `ASMC-clust`, which uses samples of 2,000 haplotypes. In the MSE plots of the waiting-time and lineages-of-remaining estimators, the estimates derived from `ARGweaver` and `SINGER` trees consistently show relatively low MSE values (Fig. 3E, F).

We assessed credible interval coverage with `SINGER` and `ARGweaver` and confidence interval coverage with the true trees and all other methods (Fig 3G-I). The true trees produce acceptable coverage with all estimators. Credible intervals from `SINGER`-estimated trees also show consistently high coverage in most cases, though they drop below 95% coverage in the recent past with the proportion-of-lineages estimator. Other estimated trees, however, all produced somewhat lower coverage, with a tendency for coverage to decline into the more distant past.

3.3 Bias, mean squared error, and confidence interval coverage under recent directional selection

Under selection, the proportion-of-lineages estimator is known to be biased, as the ancestors of the sample are not representative of the ancestral population from which they are drawn (Edge & Coop, 2019). The waiting-time and lineages-remaining estimators were developed to avoid this bias.

When there is a burst of directional selection in the recent past, the performance differences among ARG-estimation software packages are more pronounced. In contrast to the neutral scenario, and as expected, the proportion-of-lineages estimator is more strongly biased, has a higher MSE, and has a lower confidence-interval coverage than the other estimators (Fig. 4A-I, Supplementary Fig. S7-S9), particularly during the period of selection and more anciently, as expected. Looking backward in time, the bias of the proportion-of-lineages estimator between the present and the end of the period in which selection occurred is low. This basic pattern also appears in the MSE and confidence-interval coverage plots, with good performance between the present and the end of directional selection, declining into the past during the period in which selection occurred, and then recovering

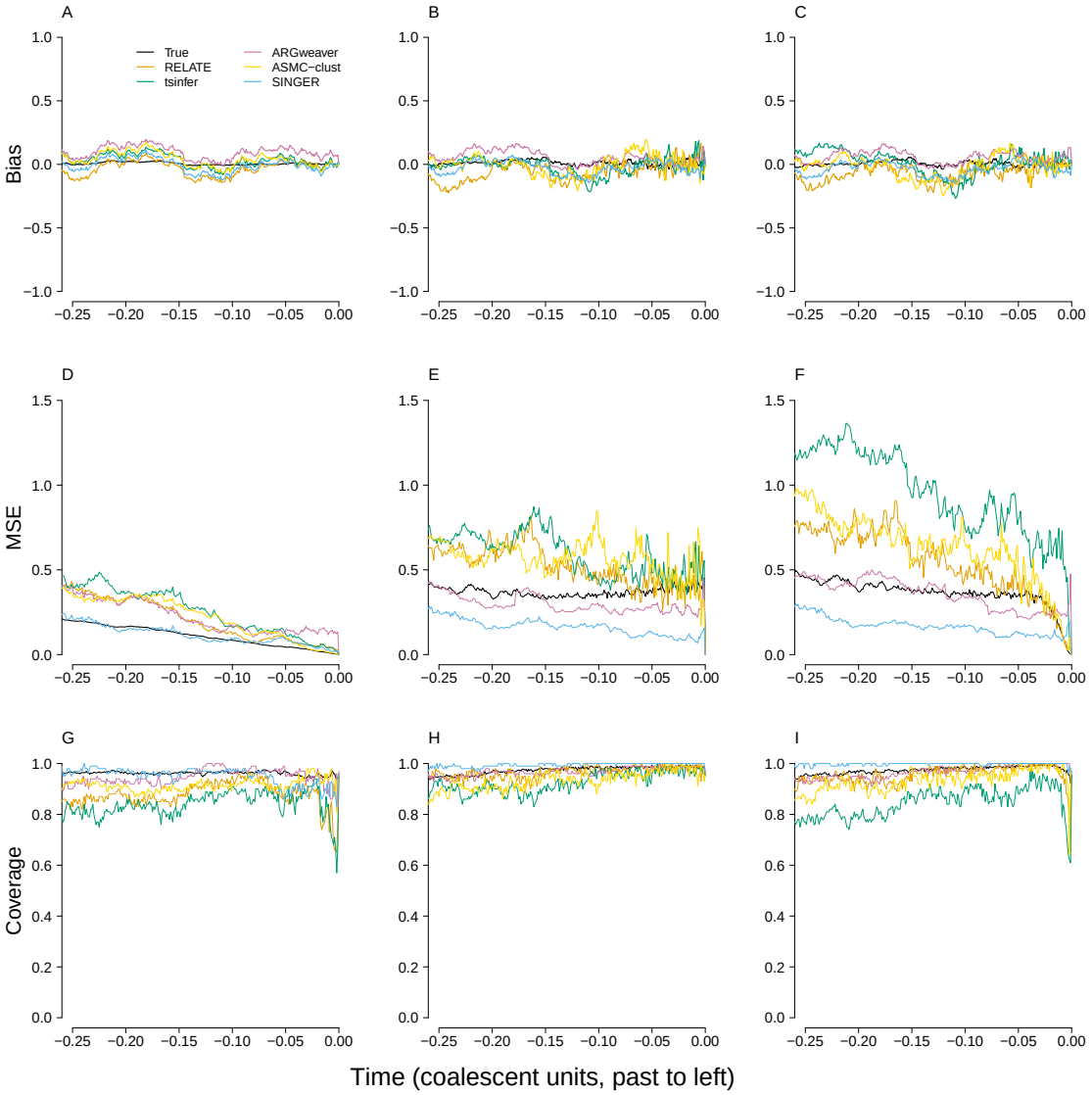

Figure 3: The bias (A-C), MSE (D-F), and confidence-interval coverage (G-I) of the proportion-of-lineages (left column), waiting-time (middle column), and lineages-remaining estimators (right column), with the true trees and estimated trees from each ARG-estimation method as input. For the estimates computed from true coalescent trees (black lines), 1000 simulations were performed with a sample of 2,000 chromosomes. In each simulation, the PGS was formed from 100 loci and evolved neutrally. `Relate`, `tsinfer`, and `ASMC-clust` were used to reconstruct trees with 2,000 chromosomes. `SINGER` was run with 200 chromosomes, and `ARGweaver` was run with 20 chromosomes. Lines from estimated trees are based on 100 simulations.

during the neutral period that preceded selection.

The waiting-time and the lineages-remaining estimators show substantially lower bias (Fig. 4B-C). The true trees generally show acceptable performance throughout the time period examined, with only slight bias during the period of selection and fairly uniform performance across time on other desiderata. However, all estimated trees produce noticeably worse performance on all criteria (Fig. 4B-I). Overall, among the estimated trees, the `SINGER` trees produce the lowest bias and MSE and interval coverage closest to the nominal level. `ARGweaver` ranks second across most of the investigated time. `ASMC-clust` and `Relate` are competitive in the recent and distant past respectively.

3.4 Power of T_X

Next, we examined the type I error rate and power of tests of neutrality using T_X, a test statistic sensitive to changes in a population-mean PGS that are larger than expected under neutrality. T_X can be understood as a version of the Q_X statistic (Berg & Coop, 2014) applied to a population-mean PGS from one population through time, rather than multiple populations sampled at the present (Edge & Coop, 2019). To use T_X, one picks a set of time points to calculate the statistic $X_j = \frac{Z(t_j)-Z(t_{j-1})}{\sqrt{2V_A(t_j-t_{j-1})}}$,where $Z(t_j)$ is the PGS at time t_j and V_A is the additive genetic variance of the PGS. With the true allele-frequency trajectories, under neutrality, the sum across time points $Tx = \sum_{j=1}^{w} X_j^2$ computed from w distinct intervals is approximately $\chi^2(w)$ distributed (Edge & Coop, 2019).

In line with previous results, T_X has an acceptable type I error rate when compared against the the χ^2 distribution only when the proportion-of-lineages estimator is used to form the statistic (Table 1). With the proportion-of-lineages estimator, the observed type I error rates for the true trees, `tsinfer+tsdate` trees, `ASMC-clust` trees, and `SINGER` trees do not differ significantly from the nominal rate (`RENT+` and `ARG-Needle` trees in Supplementary Table S6). For `Relate`, the rate is slightly higher. However, with all estimators and tree estimation software, calibrated type I error rates can be recovered by using a permutation distribution instead of the theoretical χ^2 distribution, as expected.

We also tested T_X's power in simulations that included a period of directional selection between 0.02 and 0.04 coalescent units in the past. In line with the previous results, T_X calculated from the proportion-of-lineages-estimated allele frequencies is much more powerful than when the other two estimators of allele frequency are used. Considering proportion-of-lineages T_X, power was approximately 98% with the true trees. Except for `SINGER`, most estimated trees lead to a substantial loss in power. Comparing against the permutation distribution, `SINGER` trees led to an observed power of 90%, whereas `ASMC-clust` produced T_X statistics with a power of 69%, and `Relate` and `tsinfer+tsdate` trees are only 14% and 19% relatively. Power obtained with `RENT+` trees with different branch length units can be found in Supplementary Table S3.

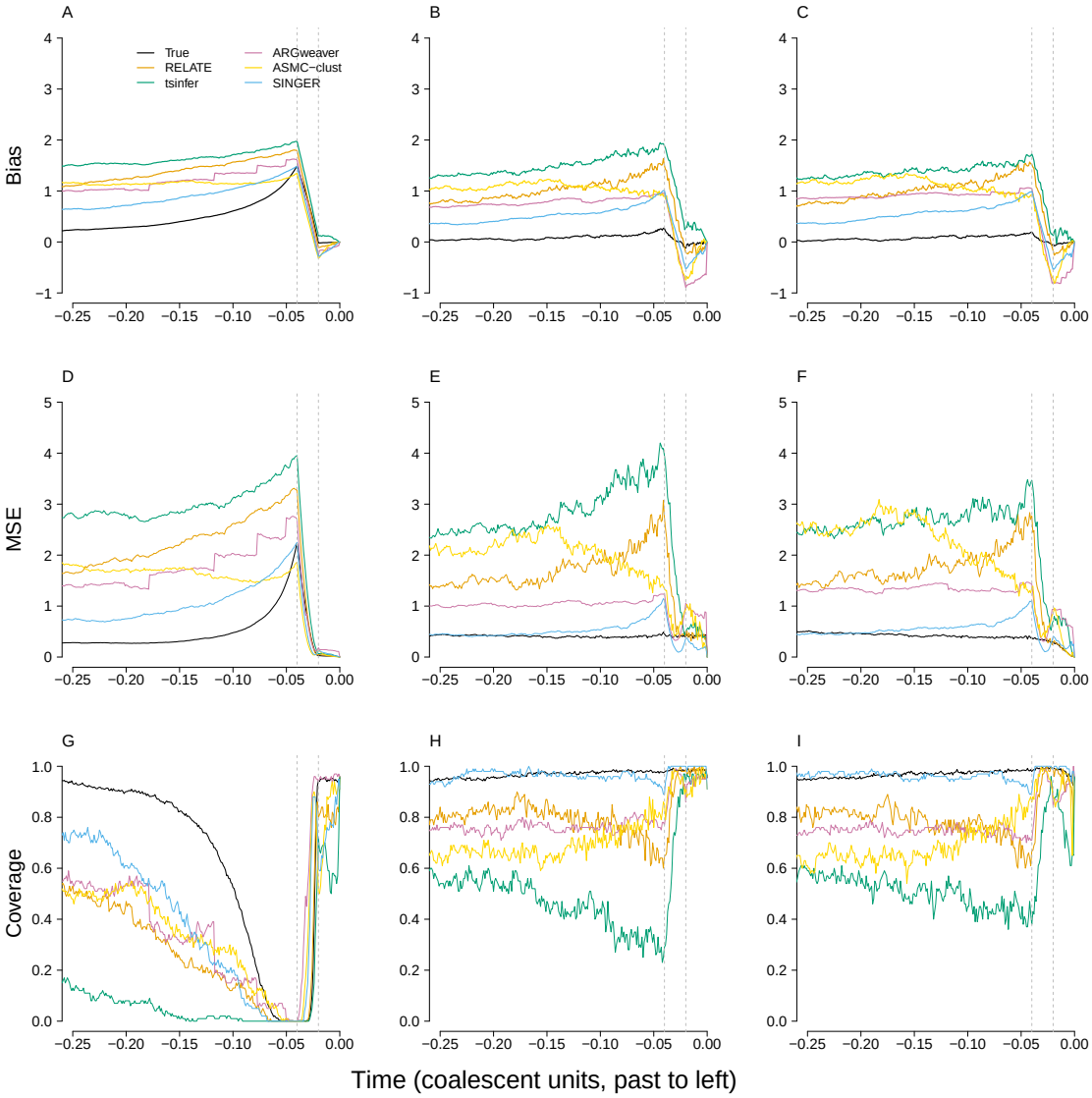

Figure 4: The bias (A-C), MSE (D-F), and confidence-interval coverage (G-I) of the proportion-of-lineages (left column), waiting-time (middle column), and lineages-remaining estimators (right column), with the true trees and estimated trees from each ARG-estimation method as input. The PGS is influenced by a period of directional selection that occurred from 0.04 to 0.02 coalescent units ago. The parameters used to run all software are otherwise identical to those applied under neutrality (Fig. 3).

Table 1: Type-I error/Power of various implementations of the T_X statistic

Input		Proportion-of-lineages		Waiting-time		Lineages-remaining	
		χ^2 distribution	Permutation distribution	χ^2 distribution	Permutation distribution	χ^2 distribution	Permutation distribution
Neutral	True trees	0.069	0.054	0.078***	0.061	0.045	0.045
	`Relate`	0.1*	0.07	0.23***	0.05	0.04	0.05
	`tsinfer`	0.07	0.04	0.12**	0.08	0.01	0.07
	`ASMC-clust`	0.04	0.02	0.02	0.03	0.03	0.08
	`SINGER`	0.09	0.05	0.35***	0.04	0.13	0.09
Selection	True trees	0.982	0.981	0.09	0.077	0.062	0.084
	`Relate`	0.14	0.21	0.28	0.05	0.08	0.1
	`tsinfer`	0.19	0.07	0.29	0.06	0	0.01
	`ASMC-clust`	0.69	0.57	0.01	0.05	0.04	0.06
	`SINGER`	0.9	0.85	0.56	0.14	0.3	0.15

For the type I error simulations, asterisks indicate whether the observed type I error rate differs significantly from the nominal rate of 0.05: $*p < 0.05; **p < 0.01; ***p < 0.001$.

3.5 Comparison between true and simulated pairwise coalescence time

A major contributor to differences in performance among the tree-estimation procedures is the accuracy of the estimated coalescence times. We compared the pairwise coalescence times from estimated trees with their true values. As measured by the Spearman correlation between true and estimated pairwise coalescence times, `SINGER` outperforms other software packages, with `ARGweaver` and `ASMC-clust` close behind. It is also possible to see finer-grained patterns in figure 5. `ARGweaver` and `ASMC-clust` show a tendency to underestimate small coalescence times (Fig. 5A,D). Both `Relate` and `tsinfer+tsdate` estimates appear biased for moderate-length coalescence times (between 0.1 and 1); `Relate` tends to underestimate these values, whereas `tsinfer+tsdate` tends to overestimate them (Fig. 5B-C). Additionally, `Relate` suffers more in short time estimation than `tsinfer+tsdate`, as previously observed by Fan et al. (Fan et al., 2023). Finally, the time discretizations used by `ARGweaver` and `ASMC-clust` are visible as horizontal bands. These patterns are qualitatively similar when examined in simulations performed under selection (Supplementary Fig. S11) and in general distributions of pairwise coalescence times (Supplementary Fig. S12-S13.

To examine the impact of selection on coalescence time estimation, we checked the distribution of pairwise coalescence time from 100 trees with 300 tips, each subject to a strong selection event that increases the allele frequency from 0.3 to 0.7. This selection event results in a high density of coalescences in the recent past within the derived-allele subtree, leading to a higher frequency of short coalescent times compared with the neutral simulations. This creates a distinct dip in the histogram of pairwise coalescence times (Fig. 6A). This pattern can also be observed, to varying degrees, in the distributions of pairwise coalescence times from estimated trees (Fig. 6B-E). Nevertheless, the dip in the histogram is wider and deeper when examining the true trees than when examining times from estimated trees. The density of recent coalescence times is also too low in estimated trees.

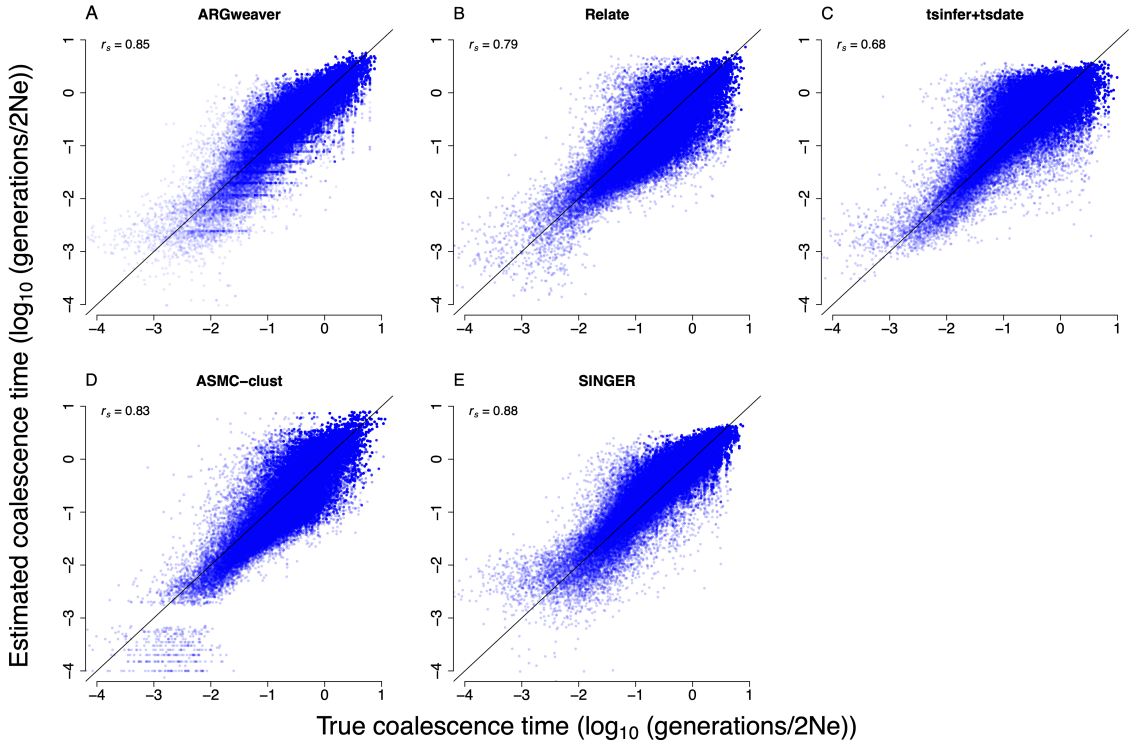

Figure 5: The comparison of pairwise coalescence times from true trees and estimated trees under neutrality. Each dot represents the true and estimated (log) coalescence time between a pair of samples. The diagonal line shows $x = y$, i.e. true times equal to estimated times. The values in the top-left corner show Spearman correlation coefficients.

3.6 Impact of sample size on estimation results

Another potential contributor to differences in performance among estimates derived from distinct ARG estimation procedures is sample size—ARG estimators differ in the sample sizes they can accommodate, and this variation is reflected in our simulations. Because large samples of haplotypes coalesce quickly in the recent past, we expect increased sample size to benefit allele-frequency trajectory estimation primarily in the recent past. We compared estimated PGS histories formed from the true trees obtained from samples of different sizes. The MSE of the proportion-of-lineages estimates decreases substantially when the sample size increases from 20 to 200. Notably, the MSE is not much larger for samples of size 200 than for samples of 2,000 lineages (Fig. 7A and D). This pattern extends to the other two estimators when applied under selection (Fig. 7E-F). But under neutrality, the waiting-time and lineages-remaining estimators perform approximately equally regardless of sample size (Fig. 7B-C). The pattern observed for true trees also applies to estimated

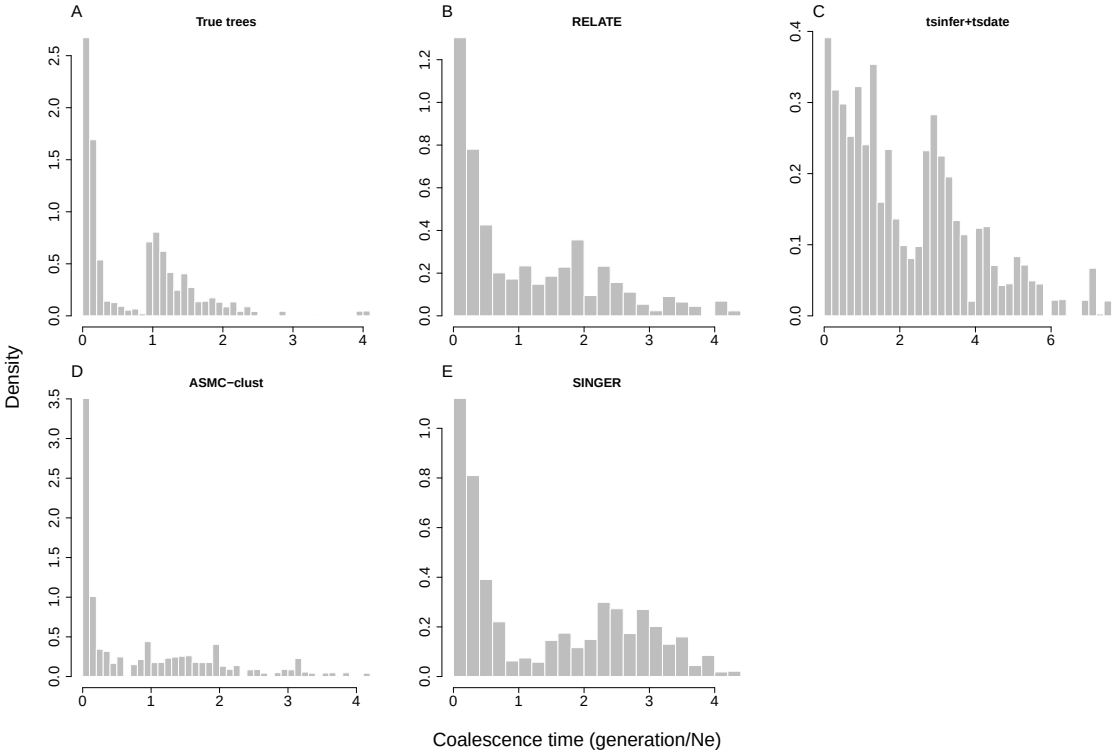

Figure 6: Distribution of the pairwise coalescence times from true and estimated trees at a locus undergoing strong selection.

trees (Supplementary Fig. S3-S5, S7-S9).

4 Discussion

We studied the performance of a set of methods for estimating the history of population-mean PGS using estimated ARGs, with a particular interest in the relative performance of different ARG estimation procedures. We used a broad range of methods appearing over the past decade—`ARGweaver` (Rasmussen et al., 2014), `RENT+` (Mirzaei & Wu, 2017), `Relate` (Speidel et al., 2019), `tsinfer+tsdate` (Kelleher et al., 2019; Wohns et al., 2022), `ARG-Needle/ASMC-clust` (Zhang et al., 2023), and `SINGER` (Deng et al., 2024)—that vary in their approaches, runtimes, and scalability. Previous work on these methods considered only estimated marginal trees from `RENT+` (Edge & Coop, 2019). We also benchmarked the estimated coalescence times emerging from these methods in general terms under neutrality and selection, providing an update on the work of Brandt and colleagues (Y C Brandt et al., 2022), including several ARG estimation procedures that are newly released since their

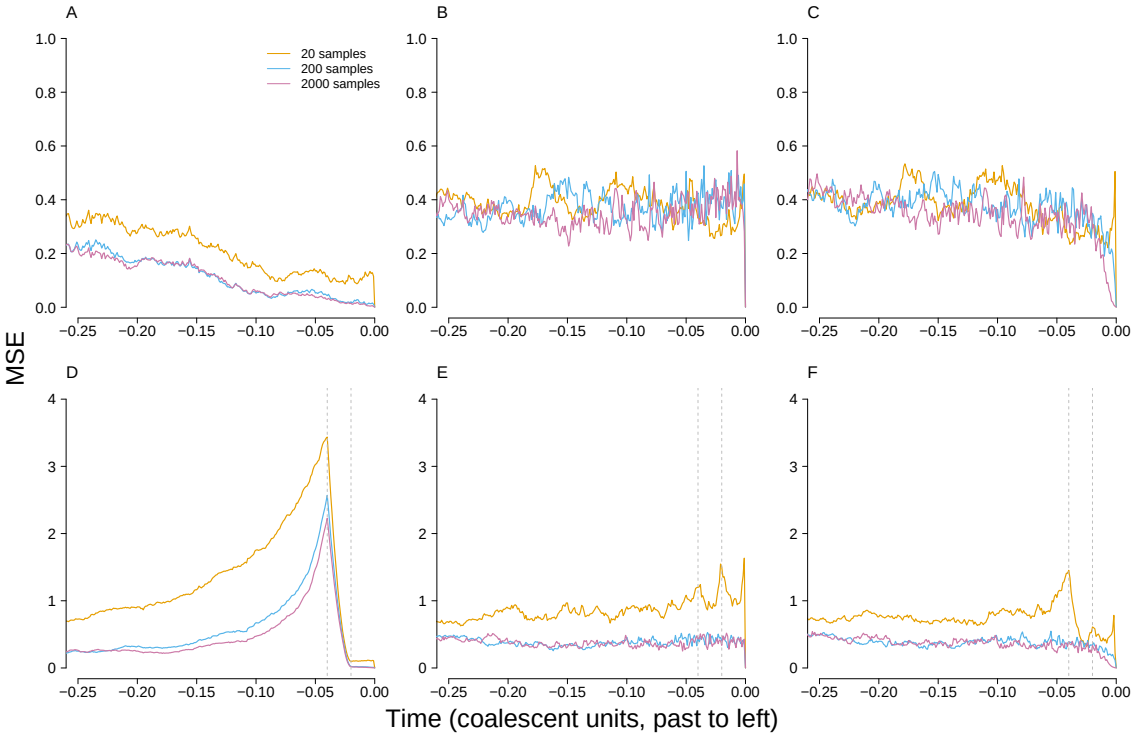

Figure 7: The MSE of the proportion-of-lineages (left column), waiting-time (middle column), and lineages-remaining estimators (right column) estimates from true trees with different sample sizes under neutrality (A-C) and directional selection (D-F).

work.

Many of the basic patterns we observed were consistent across all ARG or marginal-tree estimation procedures, as well as being consistent with previous work. For example, regardless of the tree estimation procedure used, the resulting PGS history estimates are more accurate under neutral evolution than under selection, and PGS estimates designed to work either under neutrality ("proportion-of-lineages") or under selection ("waiting-time" or "lineages-remaining") had the expected pattern of relative performance.

At the same time, there were considerable differences in the accuracy of estimated population-mean PGS histories depending on the method used for marginal tree estimation. This is expected given the rapid development of ARG and marginal-tree estimation methods in the past ten years, which have resulted in increased scalability of several orders of magnitude (Lewanski et al., 2024; Y C Brandt et al., 2022). However, bigger samples are not always better. In fact, the best performance overall from any set of estimated trees—whether measured in terms of mean squared error or confidence/credible interval coverage

of the estimated PGS histories, or in terms of power in tests of natural selection—came from `SINGER` trees estimated with 200 haplotypes, rather than any of the tree estimation procedures that were capable of scaling to 2,000 haplotypes. There were several other cases in which smaller samples fit by either `RENT+` or `ARGweaver` outperformed trees estimated from much larger samples.

The fact that trees estimated with much larger samples of haplotypes do not always outperform trees estimated with smaller samples may be counterintuitive. One part of the explanation is the trade-off between accuracy and scalability in ARG estimation. ARG estimation is a difficult problem, and scalability can be achieved with simplifications that can reduce accuracy (Brandt et al., 2024; Deng et al., 2021, 2024; Y C Brandt et al., 2022). In line with this, `ARGweaver` and `SINGER` trees tend to feature more accurate branch lengths than methods that scale to larger samples (Deng et al., 2024; Lewanski et al., 2024; Y C Brandt et al., 2022). The other important part of the explanation with respect to the estimators we explore here is that in large samples, coalescence initially happens very fast. Because the coalescence rate is proportional to $\binom{n}{2}$, with n the number of lineages that have not yet coalesced, large samples imply large amounts of coalescence in the very recent past. As a result, even very large samples will be represented by a small number of lineages in the recent past (H. Chen & Chen, 2013; Frost & Volz, 2010; Griffiths, 1984; Jewett & Rosenberg, 2014; Maruvka et al., 2011; Slatkin & Rannala, 1997; Tavaré, 1984; Volz et al., 2009), and therefore increasing the sample size beyond a few hundred haplotypes produces more precise allele-frequency estimation primarily in the very recent past. This is reflected in Fig. 7, which shows that, even with the true trees, increasing the sample size from 200 to 2,000 lowers the MSE of the Edge and Coop's (2019) estimators only in the very recent past.

The ARG estimation methods we have examined have all been benchmarked previously. However, most of this benchmarking has been done with respect to general indicators of performance, such as the overall accuracy of local tree topologies (Kelleher et al., 2019; Rasmussen et al., 2014); a generalized Robinson-Foulds distance for ARGs (Zhang et al., 2023), the distribution of distances between topologically distinct local trees (Deng et al., 2021), and the distribution of pairwise coalescent times (Deng et al., 2021; Y C Brandt et al., 2022). Although these general features are very important, it is not always straightforward to predict from them which methods will perform best when applied to a specific downstream task. For the estimators we explore here, for example, the accuracy of tree topology only matters to the extent it enables accurate estimation of the coalescence times on each allelic background. For other tasks, other specific features may assume outsize importance. For example, in estimating the expected genetic relatedness matrix (eGRM), the estimation of long coalescence times matters a great deal, since many pairs of lineages, and thus many entries of the eGRM, are related by these long times, leading `Relate` to outperform `tsinfer+tsdate` for this task (Fan et al., 2022). In contrast, for some demographic inference problems, more recent coalescence times are more important, leading `tsinfer+tsdate` to outperform `Relate` (Fan et al., 2023). In addition to the specific fea-

tures of the ARG that are important for a given downstream task, performance may vary according to the evolutionary scenario, e.g. the selection regime. In other words, there will not necessarily be an overall "best" method for ARG estimation—when choosing a method for ARG inference in empirical work, the specific strengths and weaknesses of particular methods may be more important than general considerations about overall accuracy or scalability.

This study adds to others suggesting that estimated ARGs and the marginal trees they encode are useful tools for the study of complex traits (M. Chen & Chiang, 2021; Edge & Coop, 2019; Link et al., 2023; Speidel et al., 2021; Stern et al., 2021; Zhang et al., 2023). As ARG estimation methods continue to improve in both accuracy and scalability, they will open new opportunities for mapping genetic variants contributing to phenotypes, revealing polygenic adaptation, and exploring the relationship between natural selection and population history.

5 Code availability

Code used in this paper can be found at https://github.com/dandanpeng/ARG_Benchmarking.git.

Acknowledgments

We thank members of the Edge, Mooney, and Pennell labs for helpful conversations. This work was supported by NIH grant R35GM137758 to MDE.

References

Berg, J. J., & Coop, G. (2014). A population genetic signal of polygenic adaptation. *PLoS Genet.*, *10*(8), e1004412.

Berg, J. J., & Coop, G. (2015). A coalescent model for a sweep of a unique standing variant. *Genetics*, *201*(2), 707–725.

Brandt, D. Y. C., Huber, C. D., Chiang, C. W. K., & Ortega-Del Vecchyo, D. (2024). The promise of inferring the past using the ancestral recombination graph. *Genome Biol. Evol.*, *16*(2), evae005.

Carlson, M. O., Rice, D. P., Berg, J. J., & Steinrücken, M. (2022). Polygenic score accuracy in ancient samples: Quantifying the effects of allelic turnover. *PLoS Genet.*, *18*(5), e1010170.

Chen, H., & Chen, K. (2013). Asymptotic distributions of coalescence times and ancestral lineage numbers for populations with temporally varying size. *Genetics*, *194*(3), 721–736.

Chen, M., & Chiang, C. W. K. (2021). Allele frequency differentiation at height-associated SNPs among continental human populations. *Eur. J. Hum. Genet.*, *29*(10), 1542–1548.

Deng, Y., Nielsen, R., & Song, Y. S. (2024, March). *Robust and accurate bayesian inference of genome-wide genealogies for large samples.*

Deng, Y., Song, Y. S., & Nielsen, R. (2021). The distribution of waiting distances in ancestral recombination graphs. *Theor. Popul. Biol.*, *141*, 34–43.

Edge, M. D., & Coop, G. (2019). Reconstructing the history of polygenic scores using coalescent trees. *Genetics*, *211*(1), 235–262.

Fan, C., Cahoon, J. L., Dinh, B. L., Ortega-Del Vecchyo, D., Huber, C., Edge, M. D., Mancuso, N., & Chiang, C. W. K. (2023). A likelihood-based framework for demographic inference from genealogical trees. *bioRxiv*.

Fan, C., Mancuso, N., & Chiang, C. W. K. (2022). A genealogical estimate of genetic relationships. *Am. J. Hum. Genet.*, *109*(5), 812–824.

Field, Y., Boyle, E. A., Telis, N., Gao, Z., Gaulton, K. J., Golan, D., Yengo, L., Rocheleau, G., Froguel, P., McCarthy, M. I., & Pritchard, J. K. (2016). Detection of human adaptation during the past 2000 years. *Science*, *354*(6313), 760–764.

Frost, S. D. W., & Volz, E. M. (2010). Viral phylodynamics and the search for an 'effective number of infections'. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, *365*(1548), 1879–1890.

Griffiths, R. C. (1984). Asymptotic line-of-descent distributions. *J. Math. Biol.*, *21*(1), 67–75.

Griffiths, R. C., & Marjoram, P. (1996). Ancestral inference from samples of DNA sequences with recombination. *J. Comput. Biol.*, *3*(4), 479–502.

Harris, K. (2019). From a database of genomes to a forest of evolutionary trees. *Nature Genetics*, *51*(9), 1306–1307. https://doi.org/10.1038/s41588-019-0492-x

Harris, K. (2023). Using enormous genealogies to map causal variants in space and time. *Nature Genetics*, *55*(5), 730–731. https://doi.org/10.1038/s41588-023-01389-9

Hudson, R. (1990). Gene genealogies and the coalescent process. *Oxford surveys in evolutionary biology*, *7*, 1–44.

Hudson, R. R. (2002). Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics*, *18*(2), 337–338.

Jewett, E. M., & Rosenberg, N. A. (2014). Theory and applications of a deterministic approximation to the coalescent model. *Theor. Popul. Biol.*, *93*, 14–29.

Kappelman, J. (1996). The evolution of body mass and relative brain size in fossil hominids. *J. Hum. Evol.*, *30*(3), 243–276.

Kelleher, J., Wong, Y., Wohns, A. W., Fadil, C., Albers, P. K., & McVean, G. (2019). Inferring whole-genome histories in large population datasets. *Nat. Genet.*, *51*(9), 1330–1338.

Lee, K. M., & Coop, G. (2017). Distinguishing among modes of convergent adaptation using population genomic data. *Genetics*, *207*(4), 1591–1619.

Lewanski, A. L., Grundler, M. C., & Bradburd, G. S. (2024). The era of the ARG: An introduction to ancestral recombination graphs and their significance in empirical evolutionary genomics. *PLoS Genet.*, *20*(1), e1011110.

Link, V., Schraiber, J. G., Fan, C., Dinh, B., Mancuso, N., Chiang, C. W. K., & Edge, M. D. (2023). Tree-based QTL mapping with expected local genetic relatedness matrices. *Am. J. Hum. Genet.*, *110*(12), 2077–2091.

Martin, A. R., Gignoux, C. R., Walters, R. K., Wojcik, G. L., Neale, B. M., Gravel, S., Daly, M. J., Bustamante, C. D., & Kenny, E. E. (2017). Human demographic history impacts genetic risk prediction across diverse populations. *Am. J. Hum. Genet.*, *100*(4), 635–649.

Maruvka, Y. E., Kessler, D. A., & Shnerb, N. M. (2011). The birth-death-mutation process: A new paradigm for fat tailed distributions. *PLoS One*, *6*(11), e26480.

Mathieson, I., Lazaridis, I., Rohland, N., Mallick, S., Patterson, N., Roodenberg, S. A., Harney, E., Stewardson, K., Fernandes, D., Novak, M., Sirak, K., Gamba, C., Jones, E. R., Llamas, B., Dryomov, S., Pickrell, J., Arsuaga, J. L., de Castro, J. M. B., Carbonell, E., ... Reich, D. (2015). Genome-wide patterns of selection in 230 ancient eurasians. *Nature*, *528*(7583), 499–503.

Mirzaei, S., & Wu, Y. (2017). RENT+: An improved method for inferring local genealogical trees from haplotypes with recombination. *Bioinformatics*, *33*(7), 1021–1030.

Przeworski, M., Coop, G., & Wall, J. D. (2005). The signature of positive selection on standing genetic variation. *Evolution*, *59*(11), 2312–2323.

Racimo, F., Berg, J. J., & Pickrell, J. K. (2018). Detecting polygenic adaptation in admixture graphs. *Genetics*, *208*(4), 1565–1584.

Rasmussen, M. D., Hubisz, M. J., Gronau, I., & Siepel, A. (2014). Genome-wide inference of ancestral recombination graphs. *PLoS Genet.*, *10*(5), e1004342.

Robinson, M. R., Hemani, G., Medina-Gomez, C., Mezzavilla, M., Esko, T., Shakhbazov, K., Powell, J. E., Vinkhuyzen, A., Berndt, S. I., Gustafsson, S., Justice, A. E., Kahali, B., Locke, A. E., Pers, T. H., Vedantam, S., Wood, A. R., van Rheenen, W., Andreassen, O. A., Gasparini, P., ... Visscher, P. M. (2015). Population genetic differentiation of height and body mass index across europe. *Nat. Genet.*, *47*(11), 1357–1362.

Slatkin, M., & Rannala, B. (1997). Estimating the age of alleles by use of intraallelic variability. *Am. J. Hum. Genet.*, *60*(2), 447–458.

Speidel, L., Cassidy, L., Davies, R. W., Hellenthal, G., Skoglund, P., & Myers, S. R. (2021). Inferring population histories for ancient genomes using Genome-Wide genealogies. *Mol. Biol. Evol.*, *38*(9), 3497–3511.

Speidel, L., Forest, M., Shi, S., & Myers, S. R. (2019). A method for genome-wide genealogy estimation for thousands of samples. *Nat. Genet.*, *51*(9), 1321–1329.

Stern, A. J., Speidel, L., Zaitlen, N. A., & Nielsen, R. (2021). Disentangling selection on genetically correlated polygenic traits via whole-genome genealogies. *Am. J. Hum. Genet.*, *108*(2), 219–239.

Tavaré, S. (1984). Line-of-descent and genealogical processes, and their applications in population genetics models. *Theor. Popul. Biol.*, *26*(2), 119–164.

Uricchio, L. H., Kitano, H. C., Gusev, A., & Zaitlen, N. A. (2019). An evolutionary compass for detecting signals of polygenic selection and mutational bias. *Evol Lett*, *3*(1), 69–79.

Volz, E. M., Kosakovsky Pond, S. L., Ward, M. J., Leigh Brown, A. J., & Frost, S. D. W. (2009). Phylodynamics of infectious disease epidemics. *Genetics*, *183*(4), 1421–1430.

Wakeley, J. (2016, April). *Coalescent theory: An introduction*. Macmillan Learning.

Wohns, A. W., Wong, Y., Jeffery, B., Akbari, A., Mallick, S., Pinhasi, R., Patterson, N., Reich, D., Kelleher, J., & McVean, G. (2022). A unified genealogy of modern and ancient genomes. *Science*, *375*(6583), eabi8264.

Y C Brandt, D., Wei, X., Deng, Y., Vaughn, A. H., & Nielsen, R. (2022). Evaluation of methods for estimating coalescence times using ancestral recombination graphs. *Genetics*, *221*(1).

Zhang, B. C., Biddanda, A., Gunnarsson, Á. F., Cooper, F., & Palamara, P. F. (2023). Biobank-scale inference of ancestral recombination graphs enables genealogical analysis of complex traits. *Nat. Genet.*, *55*(5), 768–776.