A General Framework for Branch Length Estimation in Ancestral Recombination Graphs

Yun Deng[1], Yun S. Song*[1,2,3], and Rasmus Nielsen*[1,2,4,5]

[1]Center for Computational Biology, University of California, Berkeley, USA
[2]Department of Statistics, University of California, Berkeley, USA
[3]Computer Science Division, University of California, Berkeley, USA
[4]Department of Integrative Biology, University of California, Berkeley, USA
[5]Center for GeoGenetics, University of Copenhagen, Denmark

Abstract

Inference of Ancestral Recombination Graphs (ARGs) is of central interest in the analysis of genomic variation. ARGs can be specified in terms of topologies and coalescence times. The coalescence times are usually estimated using an informative prior derived from coalescent theory, but this may generate biased estimates and can also complicate downstream inferences based on ARGs. Here we introduce, POLEGON, a novel approach for estimating branch lengths for ARGs which uses an uninformative prior. Using extensive simulations, we show that this method provides improved estimates of coalescence times and lead to more accurate inferences of effective population sizes under a wide range of demographic assumptions. It also improves other downstream inferences including estimates of mutation rates. We apply the method to data from the 1000 Genomes Project to investigate population size histories and differential mutation signatures across populations. We also estimate coalescence times in the HLA region, and show that they exceed 30 million years in multiple segments.

1 Introduction

The genetic relationships between the DNA sequences in a sample can be described by a tree. However, in the presence of recombination, the trees might differ from position to position in the genome. The set of all trees in the genome, and corresponding change points (Figure 1A), is represented by the Ancestral Recombination Graph (ARG) [11, 12, 16, 28, 40]. The generative ancestral process creating ARGs is the coalescent with recombination. Simulations of ARGs under this model is relatively straightforward [17, 21], but inferring ARGs from DNA sequence data is known as a notoriously difficult problem [24, 31]. The difficulty of ARG inference is mainly caused by the high dimensionality of the ARG; there are a lot of possible topologies for each local tree and the genome might contain a large number of such trees. Every recombination event will generate a new tree that might have a new topology. Even when topologies do not change after recombination, branch lengths might change [8] and ARG inference contains a dual problem of inferring both topology and branch lengths.

One important feature of ARGs is the"node persistence" property, i.e. that a node can be shared by multiple marginal trees. For example, in Figure 1A, all three trees share node 7. Each

*To whom correspondence should be addressed: yss@berkeley.edu, rasmus_nielsen@berkeley.edu

node in the ARG represents a genomic segment of ancestral genetic material, which might span over multiple local trees. Node persistence reduces the dimensionality of the ARG inference problem and explains how a tree in one position of the genome is informed by trees in adjacent segments of the genome. This property also facilitates the efficient storage and simulation of ARGs implemented in the package tskit, which leverages this property in its "tree sequence" format [21, 22].

There are many methods for estimating ARGs including ARGweaver [31], Relate [37], tsinfer [23], KwARG [18], ARG-Needle [42], and SINGER [7]. Several of these methods require a branch length estimation step after the topology has been inferred. Relate [37] can estimate branch lengths for each locally inferred topology in the genome under a specified demographic model, tsdate [39] dates the node ages in the inferred ARG topology from tsinfer [23], and ARG-Needle [42] normalizes node ages to ensure that they match a distribution obtained using simulations under a given demography model. To our knowledge, only Relate can estimate branch length and the population size history jointly. The "EstimatePopulationSize.sh" module in Relate achieves this using a Markov Chain Monte Carlo (MCMC) algorithm to re-estimate branch length given a demographic prior and subsequently estimates a new demographic model from the new branch length. Relate alternates between updating branch length and the demography model similarly to an EM-algorithm until convergence is achieved (the default is 10 iterations). This alternation between branch length and demography estimation is computationally demanding and will, in most applications, be the computational bottleneck in the inferences.

The use of a coalescent prior in previous work is a natural approach for estimating coalescence time (or equivalently branch lengths) in ARGs. However, it may bias the estimates when the assumed coalescent prior is mis-specified, which typically will be the case. The true distribution of coalescence times in natural populations is highly complicated and depends on the history of population sizes and population structure. Here we introduce a novel approach, POLEGON (**P**rior-**O**blivious **L**ength **E**stimation in **G**enealogies with **O**riented **N**etworks) for estimating branch lengths for ARGs using an uninformative uniform prior for coalescence times, which currently works for ARG topology inferred by SINGER [7]. POLEGON can also, with small adjustments, be applied to other methods for ARG topology inference. Using simulations, we show that it provides improved estimates of coalescence times under a wide range of demographic assumptions.

2 Results

2.1 Methodology overview

An ARG can be converted to a directed acyclic graph (DAG) of only the coalescence nodes (without recombination nodes), by identifying the nodes with the same index across local trees, and the branches with the same child and parent node (Figure 1B). This representation was introduced as the "genome-ARG" in Wong et al. [40]. We note that branch length estimation is equivalent to node age estimation here, because the length of a branch is the age difference of the parent and child nodes associated with the branch. In the following, we will describe our new method in terms of node age estimation in the DAG.

To estimate the node ages, we will use an MCMC algorithm algorithm that iteratively updates the marginal node ages node by node, while the ages of all other nodes are fixed. Let nodes i and n be adjacent nodes (i.e., there is an edge between them) in the DAG, then we define the span, $s_{(i,n)}$, as the total length of the genomic segments where i and n are adjacent (Figure 1C). We assume that mutations can be mapped on edges in the ARG under the assumption of an infinite sites model, using a parsimony criterion, or using some other algorithm (see e.g. [7, 23, 37]). We denote the number of mutations mapped to the branch connecting node i and n by $k_{(i,n)}$, where i

and n are unordered. The age of node n is denoted by t_n. For each edge, i adjacent to node n, we then assume that $k_{(i,n)}$ follows a Poisson distribution:

$$k_{(i,n)} \sim \text{Poisson}(\Theta_{(i,n)}), \quad \text{with } \Theta_{(i,n)} = \mu s_{(i,n)} |t_i - t_n|, \tag{1}$$

where μ is the mutation rate, i.e., the expected number of mutations per site per time unit. Note that t_n is not a completely free parameter; it is constrained by the ages of its children nodes and its parent nodes (Figure 1D). Let l_n be the age of node n's oldest child node and u_n be the age of node n's youngest parent node. Then,

$$l_n < t_n < u_n. \tag{2}$$

Therefore, we define the marginal likelihood of the age of node n, conditionally on it being in the interval (l_n, u_n), as:

$$\mathcal{L}(t_n; D) \propto \mathbb{I}(l_n < t_n < u_n) \prod_{i \in \mathcal{N}(n)} \Theta_{(i,n)}^{k_{(i,n)}} \exp(-\Theta_{(i,n)}), \tag{3}$$

where $\mathbb{I}(l_n < t_n < u_n)$ is the indicator function of the condition in (2), $\mathcal{N}(n)$ is the set of all neighbors of node n, and D denotes the mutation mapping data. We use an improper uniform prior (from 0 to ∞) on the node ages, so the marginal posterior is proportional to (3). We note this is equivalent to assuming the coalescent times follow an improper uniform distribution (regardless of the number of lineages), in a single coalescent tree and the full ARG (see Supplementary Section A.2 for detailed discussion). So this avoids the usage of informative coalescent priors as in previous efforts such as Relate and tsdate [37, 39].

We use a Metropolis-Hastings algorithm to iteratively sample from this posterior for all nodes in the graph marginally (Section 4.1). When processing the MCMC output, it will often be convenient to take an average over MCMC samples, and it is convenient in this context that the posterior sample averages of the ages of the nodes will satisfy the DAG constraints from Eq. (2). This is because each set of node ages in one posterior sample satisfies the constraints and their averages, as linear combinations, will consequently satisfy the constraints as well.

If the true coalescence process follows the coalescent with recombination [16], or any other common coalescent process, the improper uniform prior on node ages is a misspecified prior, in the sense that it does not match the known/assumed generative process. How to address the problem of prior misspecification is an active topic in Bayesian statistics, and there has been substantial recent research in this area, including solutions based on empirical Bayes methods [29], the Saerens-Latinne-Decaestecker algorithm [32], EM algorithms [5] and more [27, 36]. In the current context, the ARG rescaling technique introduced in SINGER [7] can help correct the posterior for misspecification of the prior distribution. Briefly, this rescaling technique relies on the assumption of a constant mutation rate through time, which can be used to define a monotonic transformation of the node ages such that the mutation rate is estimated to be constant from the ARG with adjusted node ages. As such, ARG rescaling attempts to bring coalescence times closer to the posterior distribution that would have been observed, if the true coalescent process had been used as a prior for the node ages (Figure 2A). A caveat of this approach is that it assumes that the aggregate, overall mutation rate remains constant over time. The full algorithm of POLEGON includes such an ARG rescaling step applied to the samples obtained from the MCMC chain (Supplementary Section A.1).

It is worth noting that ARG rescaling does not change the relative order of nodes in the DAG, even when multiple orderings are permitted by the topology. For example, node 5 in Figure 1B can

be older or younger than node 6 without violating the DAG constraints. The MCMC algorithm (Section 4.1) samples possible relative orderings, while the rescaling step assigns ages given the ordering of nodes.

Relate [37] also provides an MCMC method for obtaining a posterior for the node ages of an ARG. However, Relate only enforces constraints for a marginal coalescent tree given the nodes in that tree, i.e. it does not use the node-persistence property of the ARG. In contrast, POLEGON enforces constraints from the full DAG. Another major difference is in the choice of prior, with Relate enforcing a demographic prior, while POLEGON is using an improper uniform prior with subsequent posterior rescaling. This increases the speed of the program as joint estimation of both demographic parameters and node ages is not necessary, and arguably makes POLEGON more robust and flexible in terms of demographic models it can accommodate. Finally, POLEGON can accommodate spatial variation in mutation rate along the length of the genome.

In terms of demography inference, Relate infers ARG-based population size history by alternatively estimating branch length or coalescence rates while fixing the other (by default using 10 iterations). Here, we estimate the population size history based on pairwise TMRCA distribution, directly after estimating branch lengths with POLEGON, without having to re-estimate branch lengths given the demography model (details in Section 4.2). The full details of POLEGON's algorithm can be found at Supplementary Section A.1.

2.2 Simulation benchmarks

To benchmark the performance of POLEGON, we carry out simulations using msprime [21], with parameters that are realistic for humans, $\mu = 1.2 \times 10^{-8}, r = 1.2 \times 10^{-8}, L = 10$ Mb and with a sample size of 100 genomes, for 10 replicates. We simulate under three models: (1) constant-size ($N_e = 10^4$) and (2) CEU and (3) YRI models from SMC++ (details in Supplementary Section B.1).

To demonstrate how the combination of uninformative prior and ARG rescaling (Figure 2A) works, we first analytically solve the pairwise coalescent case with unknown constant population size (Supplementary Section A.3). In this case, we show analytically that the rescaled posterior distribution, in expectation, is exactly identical to the correct distribution (Supplementary Section A.3), even when the population size history is unknown.

For more complex generative processes of the coalescent with recombination, we use simulations to demonstrate that the ARG rescaling provides well-calibrated posterior distributions. We simulate under the aforementioned three different models, constant N_e, a CEU model, and a YRI model (details in Supplementary Section B.1). We provide POLEGON with the true topology from the simulations, and use a rank plot to diagnose the quality of the posterior samples [3, 41]. If the posterior sampling is perfect, then the relative rank of the true node age among the age samples, when simulating data from the prior, should be uniformly distributed between 0 and M, where M is the number of samples from the MCMC chain [3]. This is, to a close approximation, indeed the case for all three different generative models, as evidenced by a low Kullback-Leibler divergence between the empirical rank distribution and a uniform distribution (Figure 2B). Similarly, the estimates of the distribution of pairwise coalescence times from POLEGON match the true distributions well, but only after ARG rescaling (Figure 2C). We also examined the same demographic models but with $\mu/\rho = 0.25$ or 4, to show that this conclusion generalizes to other parameter choices (Supplementary Figure S2).

We compare the inference accuracy of SINGER+POLEGON to that of SINGER and Relate, to evaluate the accuracy gain from additionally using POLEGON. SINGER has previously been shown to provide more accurate estimates in some aspects than other ARG inference methods [7]. In the comparisons, we examine the inference accuracy of pairwise time to the most recent

common ancestor (pairwise TMRCA), the distribution of pairwise TMRCA, local diversity (the average pairwise TMRCA in genome windows), and local mutation density (the average tree branch length in genome windows). These are statistics similar to those used to evaluate ARG inference performance in [7]. In all simulations, we first run SINGER to infer the ARG topology and we then run POLEGON on the MCMC samples from SINGER. Posterior averages of these statistics among samples were calculated and compared to ground truth to evaluate accuracy. The details of the simulation benchmarks can be found at Supplementary Section B.5.

We compute the local diversity and local mutation density in 10 kb windows, following the branch-length-based definition of [30] (using "diversity" and "segregating_sites" API in tskit). We used the mean squared error (MSE) to characterize the inference accuracy for all statistics. For evaluating the distribution of pairwise TMRCAs, we categorize the coalescence times into 20 bins so that each bin occupies 5% of the probability mass in the ground truth distribution, and then compute the symmetrized Kullback-Leibler divergence (KLD), also known as the Jeffreys divergence, between the simulated and inferred distribution with the discretized distributions. For all four tasks, POLEGON+SINGER improves the accuracy over that of SINGER alone, which is itself better than Relate (Figure 3). Importantly, the pairwise TMRCA distribution is more accurately inferred by SINGER+POLEGON than using SINGER alone (Figure 3B), which will lead to improved inferences of demographic history.

In real data analyses, the underlying demographic model is likely different from a constant-size panmictic population model, and the demographic model typically needs to be inferred from the data. There exist many demographic inference methods based on different frameworks, for example using the Site Frequency Spectrum (SFS) [19, 20, 26], the Sequentially Markovian Coalescent (SMC) for pairs of genomes [25, 33, 34], etc. Each of these methods have drawbacks. Reducing population genomic data to an SFS leads to a substantial loss of information, as all linkage/haplotype information is lost. Classical SMC methods typically only handle 2 to 8 genomes [25, 33, 34]. ARGs provides a powerful alternative for estimating population size changes [37], if the node ages can be estimated accurately. Here we compare the demography inference results from SINGER+POLEGON with other competing methods including MSMC2 [33, 34] (an SMC based method), Relate [37] (based on ARGs), and FitCoal [15] (an SFS based method). We find that, the accuracy of POLEGON is at least comparable to MSMC2 and Relate, and better than that of FitCoal, especially on the resolution of the change history (Figure 4A, B, Supplementary Figure S6). A likely reason is that FitCoal only accommodates a few changes in N_e (it assumes a piece-wise exponential growth/decline), while other methods typically allow for more change points. We also note that using the full information in the ARG as opposed to only a few pairs of sequences can lead to more stable estimates, as indicated by Supplementary Figure S3.

In order to evaluate the performance under a structured demography model, we additionally simulate a model with two populations that split $t = 10,000$ generations ago (Supplementary Section B.3), and compare the performance of SINGER+POLEGON to that of Relate in terms of accuracy of inferred cross-population coalescence rates. We find that SINGER+POLEGON infers the cross-coalescence rate dropping towards zero slightly faster after the true divergence time than Relate (Supplementary Figure S4).

2.3 Application to real data

Population size history in YRI and CEU. Non-African populations share an bottleneck roughly 60kya [33] due to the "Out-Of-Africa" migration, which is not shared by African populations. However, it has recently been proposed that African populations experienced a rather

extreme bottleneck around 900kya [15]. This signal has only been identified by FitCoal but not other methods, such as SMC++, PSMC, MSMC, Relate and so on. Efforts to reproduce their results with other methods also failed [4, 38]. More recently, it was argued that the severe bottleneck is a statistical artifact [6].

We applied SINGER to the CEU and YRI samples in the 1000 Genomes Project high coverage dataset [2] to obtain estimates of ARG topology. We then used POLEGON to estimate branch lengths and population size history (details in Supplementary Section C.1). We compared our results to those inferred with MSMC2 [33, 34] and Relate [37], and neither shows evidence of the sudden, severe bottleneck. All methods are able to infer the Out-of-Africa bottleneck in CEU and its absence in YRI. They also all observed a mild ancient bottleneck around 1 million years ago (Figure 4C, D). Importantly, this mild bottleneck is also detected in both CEU and YRI, which makes sense because it predates the divergence time of CEU and YRI. [6] also reported that the severe bottleneck is likely a statistical artifact and that FitCoal tends to infer a sharp bottleneck in the presence of only mild reductions of population size.

Differential mutation signatures in YRI and CEU. Several different approaches have shown that human populations differ in certain mutation-type specific mutation rates [9, 10, 13, 14, 37]. Most notably, a signature of increased mutation rate of TCC to TTC in European populations has been identified using many different approaches [9, 13, 14, 37].

Some methods for estimating mutation rate trajectories are based on allele age estimation from inferred genealogies, and as such, the quality of inferred genealogies and their branch lengths has substantial impact on allele age estimates [37]. Additionally, as mutations can only be dated to a specific branch in the ARG, the associated allele age is often assumed to be uniformly distributed along the length of that branch. However, this is only valid under a constant mutation rate model, and it does not apply to the case of temporally varying mutation rates. Therefore, rate estimation based on a uniform distribution on branches tends to be inaccurate if there truly is rate variation over time.

To illustrate this point, we performed a simulation with an elevated mutation rate from 300 to 3,000 generations ago (details in Supplementary Section B.1). We note here that we fix the node ages from the simulation and only infer the mutation rate trajectory. Even when using the ground truth topology and mutation mappings from simulations, the inferred mutation rate trajectory underestimates the true temporal rate heterogeneity (Figure 5A), when assuming a uniform placement of mutations along the length of a branch. We therefore introduce a simple iterative algorithm to account for time-heterogeneous mutation rates (details in Section 4.3), which can be shown to be much more accurate than when assuming uniform mutation placement (Figure 5A).

We have incorporated this new mutation rate trajectory estimation to the inferred genome-wide genealogies with SINGER+POLEGON in CEU and YRI and recover the signal of elevated TCC to TTC mutations specific to CEU (Figure 5C) and absent in YRI. As expected, using the uniform placement of mutations on branches leads to underestimation of mutation rate variation (Figure 5B). Our estimates of the timing of elevated mutation rates agrees well with those previously reported in Speidel et al. [37]. However, we note that Speidel et al. [37] inferred a somewhat larger difference in the mutation rate of TCC to TTC mutations between modern Europeans and Africans. Previous literature [14] argued that the rate difference is not clearly observed in rare variants and suggested a potential recent loss of the mutation rate modifier. The difference between our estimates and those of Speidel et al. [37] is likely due to the artifact induced by uniform placement of mutation on branches in Relate, similarly to that observed in Figure 5A. As noted earlier, we here assume a constant aggregate mutation rate (across all sub-types), which the ARG rescaling relies on, allowing us only to infer temporal changes in the relative mutation rates of different sub-types.

Coalescence times in the HLA region. The HLA locus has been shown to harbor trans-species polymorphisms hypothesized to be under strong balancing selection [1]. The presence of trans-species polymorphism would imply coalescence times in this region that are older than the divergence time between humans and chimpanzees, and Deng et al. [7] indeed shows this to be the case.

In [7] the mutation rate was assumed to be constant, which ignores the heterogeneity of mutation rate along the genome. Here we reanalyze the HLA data with POLEGON on the SINGER topologies [7] with a mutation map specific to the HLA region estimated with Roulette [35]. The methods used to extract and process the mutation map can be found in Supplementary Section C.4. The updated estimates are older than the the previous estimates from SINGER (Figure 5B), with some alleles having average pairwise coalescence time older than 30Mya. To validate that the estimates from POLEGON+SINGER indeed make more sense than using SINGER alone, we compared the observed mutation density versus the predicted mutation density from the inferred ARG (Figure 5D, E). With inferred genealogies, the predicted mutation density in a window is defined as the integral of the product of local tree branch length and local mutation rate over the window, divided by the window length. With 10kb genome windows, the mutation density predictions from using POLEGON+SINGER clearly correlate better than using SINGER alone, which suggests better inference quality (Figure 5D, E). Notably, there is a slight underestimation bias in SINGER with the higher mutation density regions, but it is largely corrected with SINGER+POLEGON. The bias in SINGER is expected as it uses a standard exponential coalescence prior that will tend to shrink the estimates towards zero.

3 Discussion

In this article, we introduced POLEGON, a novel approach for estimating branch lengths in ancestral recombination graphs. POLEGON infers branch lengths and population size history given a fixed ARG topology using an uninformative uniform prior. The use of an uninformative prior provides increased accuracy and flexibility when the populations studied are not panmictic populations of constant size.

We applied POLEGON to SINGER-inferred genealogies for CEU and YRI populations from the 1000 Genomes Projects and estimated the demographic history of these groups to re-evaluate the evidence for a severe bottleneck in the YRI [15]. Our results largely agreed with previous findings, and we did not find evidence of a strong bottleneck. Instead, consistent with other methods, we inferred a much milder bottleneck shared by CEU and YRI.

We introduced a new way of inferring temporal mutation rate patterns that it is not biased by the assumption of uniform mutation placement on branches. Analyses of the CEU and YRI data confirmed the observation of a TCC to TTC pulse in the CEU but not YRI. While the timing of the pulse agrees well with the estimates of [37], the difference in rates between CEU and YRI before and after the pulse is smaller than those reported by [37], which assume uniform placement of mutation on branches. Third, we re-dated the coalescence times in the ARG of the HLA region from [7] with POLEGON, while accounting for mutation rate heterogeneity using a mutation map inferred by Roulette [35]. Our estimates suggest coalescence times as old as 30 million years in multiple different areas of the HLA region.

There are a few limitations to our work. Firstly, the demographic model we considered include only single population size changes in history and clean splits, and we have not investigated more complex models that include processes such as migration and admixture. Inferring node ages accurately in the presence of these factors may require an accurate reconstruction of ARG topology

incorporating these processes, in addition to an accurate dating methods. However, we note that the assumption of exchangeability among lineages in the standard coalescent model, and assumed in all common methods for estimating ARGs, corresponds to an uninformative prior on the tree topology, i.e. a prior that is not informed by population designation of leaf nodes.

Secondly, the current methodology is based on the assumption of constant aggregate mutation rate from all subtypes. Whether this is a reasonable assumption at all time scales remains to be investigated. If the pattern of temporal mutation rate variation has been inferred, then it is straightforward to incorporate into the MCMC algorithm and the subsequent ARG rescaling. It would perhaps make more sense to incorporate temporal mutation rate variation into the ARG estimation algorithm itself, rather than just at the branch length estimation step, but that is beyond the scope of this work focusing on branch length estimation for a given ARG topology.

Thirdly, even more accurate estimates of coalescence times could potentially be obtained by using an estimated demographic history as prior for the node age estimation, similarly to the approach applied by [37]. However, the use of an noninformative prior can provide unbiased (or at least less biased) estimates in regions with selection, as illustrated in our example for the HLA locus. Furthermore, it might also facilitate other downstream inferences by decoupling the estimates of coalescence times from the parametric assumptions regarding population history.

Finally, we note that the accuracy of coalescence time estimation is limited by the quality of the the topology inference. As such, future development of better methods for estimation topologies, can help improve coalescence time estimation.

4 Methodology

4.1 The MCMC algorithm

To sample the node ages in the ARG (or equivalently, the converted DAG), we perform MCMC with a Metropolis-Hastings algorithm for each node:

Algorithm 1 MCMC for node ages in Ancestral Recombination Graph

for $n = N$ to 1 **do**
  Propose a new node age t'_n from $f(\cdot)$ within interval (l_n, u_n) for node n
  Calculate the acceptance probability for node n:
  $p_a = \min\{1, \frac{f(t'_n)p(t_n)}{f(t_n)p(t'_n)}\}$
  if random() $\leq p_a$ **then**
    Accept t'_n for node n
  else
    Reject t'_n and keep t_n for node n
  end if
end for

Here, $f(\cdot)$ is the proposal function which determines how we propose new ages in the interval (l_n, u_n). When the upper bound, u_n, is not infinity, we simply propose uniformly at random in the interval (l_n, u_n):

$$f(t) = \frac{1}{u_n - l_n}, \quad \text{for } u_n < \infty.$$

This results in the following acceptance probability:

$$\begin{aligned} p_a &= \min\left\{1, \frac{f(t_n)p(t'_n)}{f(t'_n)p(t_n)}\right\} \\ &= \left\{\prod_i \Theta'^{k_{(i,n)}}_{(i,n)} \exp(-\Theta'_{(i,n)})\right\} \Big/ \left\{\prod_i \Theta^{k_{(i,n)}}_{(i,n)} \exp(-\Theta_{(i,n)})\right\} \end{aligned} \tag{4}$$

However, for the root node, $u_n = \infty$ and an alternative approach is needed. We then let $f(\cdot)$ equal an exponential distribution with a small rate parameter, λ (default at $\lambda = 0.1$ in POLEGON):

$$f(x) = \lambda \exp(-\lambda x), \quad \text{for } u_n = \infty. \tag{5}$$

The exponential distribution approaches an improper uniform on $[0, \infty)$ as $\lambda \to 0$. Now the acceptance probability is:

$$\begin{aligned} p_a &= \min\left\{1, \frac{\exp(-t_n)p(t'_n)}{\exp(-t'_n)p(t_n)}\right\} \\ &= \left\{\exp(-\lambda t_n)\prod_i \Theta'^{k_{(i,n)}}_{(i,n)} \exp(-\Theta'_{(i,n)})\right\} \Big/ \left\{\exp(-\lambda t'_n)\prod_i \Theta^{k_{(i,n)}}_{(i,n)} \exp(-\Theta_{(i,n)})\right\} \end{aligned} \tag{6}$$

We also note that we chose the order of node age updates to be the descending order of the node ages, so that when a node age is proposed, then the ages of its parent nodes will have already been updated. This leads to more efficient mixing compared to randomly choosing new nodes to update. The full details of the MCMC can be found in Supplementary Section A.1.

4.2 Population size history estimation

To infer demography, we first need to define change points in time when the population size (or equivalently, coalescence rate) can change. In this paper we chose 30 windows log-uniformly distributed from 100 to 200,000 generations ago (generation time by default at 28 years for human). We simply extract the empirical pairwise coalescence distribution from the ARG, and obtain an empirical survival function $S(t)$, i.e. the proportion of coalescence times larger than t.

Assume the time grid is $(0, t_1, t_2, \ldots, t_n, \infty)$, and the coalescence rates are $(\lambda_0, \lambda_1, \ldots, \lambda_n)$. The survival function should satisfy:

$$S(t_i) = \exp\left(-\sum_{j \le i} \lambda_{j-1}(t_j - t_{j-1})\right) \tag{7}$$

This can be turned into the following linear regression problem:

$$\log\Big(S(t_i)\Big) = \sum_{j \le i} \lambda_{j-1}(t_j - t_i) \tag{8}$$

We solve this problem to obtain estimates of coalescence rates for all time intervals and then subsequently convert them into an estimate of the population size history.

4.3 Estimation of temporal patterns of mutation rate variation

Here we introduce an iterative algorithm for estimating temporal mutation rate variation from genealogies. The core idea is to iteratively re-map mutations given the current estimates of mutation rates until the estimates have converged.

Assume the time grid is $(0, t_1, t_2, \ldots, t_K, \infty)$, the total ARG branch length in these windows are $(L_0, L_1, \ldots, L_K)$, and the current estimate of the mutation rates is $(\mu_0, \mu_1, \ldots, \mu_K)$.

If the branches with mutations mapped have lower node times of $(x_0, x_1, \ldots, x_n)$ and upper node times of $(y_0, y_1, \ldots, y_n)$, and they carry $(k_0, k_1, \ldots, k_n)$ mutations. When a mutation is mapped to a branch spanning over several time windows, it should contribute to each window proportional to its overlap with the window, similar to the treatment in the ARG re-scaling in SINGER [7]. So the mutation contribution from (x_j, y_j) in the i-th window will be:

$$m_{ji} = \frac{k_j \mu_i \mathbb{I}(x_j < t_{i+1}, y_j > t_i)[\min(y_j, t_{i+1}) - \max(x_j, t_i)]}{\sum_{i=1}^{K} \mu_i \mathbb{I}(x_j < t_{i+1}, y_j > t_i)[\min(y_j, t_{i+1}) - \max(x_j, t_i)]}$$

which means the average number of mutations mapped to the i-th interval is:

$$c_i = \sum_{j=1}^{n} m_{ji}$$

where $\mathbb{I}(x_j < t_{i+1}, y_j > t_i)$ is the indicator function of whether the time span (x_j, y_j) overlaps with the time window (t_i, t_{i+1}). This translates to the updated mutation rates in the i-th window:

$$\mu_i' = \frac{c_i}{L_i}$$

We initialize with a constant rate trajectory and carry out the above procedure iteratively for 5 iterations, which is empirically shown to be enough for convergence (Supplementary Figure S5).

Data availability

POLEGON is available at https://github.com/YunDeng98/POLEGON.git. The ARG samples from SINGER+POLEGON are available at: https://zenodo.org/records/14675005, https://zenodo.org/records/14674978, https://zenodo.org/records/14676049, https://zenodo.org/records/14676128, https://zenodo.org/records/14676158.

Acknowledgements

We thank William Dewitt for helpful discussion about uniform mutation placement on branches, Drew DeHaas and Kaiyuan Li for testing the software. This research is supported in part by NIH grants R56-HG013117 and R01-HG013117.

References

[1] Azevedo, L., Serrano, C., Amorim, A., Cooper, D.N., 2015. Trans-species polymorphism in humans and the great apes is generally maintained by balancing selection that modulates the host immune response. Human genomics 9, 1–6.

[2] Byrska-Bishop, M., Evani, U.S., Zhao, X., Basile, A.O., Abel, H.J., Regier, A.A., Corvelo, A., Clarke, W.E., Musunuri, R., Nagulapalli, K., et al., 2022. High-coverage whole-genome sequencing of the expanded 1000 genomes project cohort including 602 trios. Cell 185, 3426–3440.

[3] Cook, S.R., Gelman, A., Rubin, D.B., 2006. Validation of software for bayesian models using posterior quantiles. Journal of Computational and Graphical Statistics 15, 675–692.

[4] Cousins, T., Durvasula, A., 2025. Insufficient evidence for a severe bottleneck in humans during the early to middle pleistocene transition. Molecular Biology and Evolution , msaf041.

[5] Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the em algorithm. Journal of the royal statistical society: series B (methodological) 39, 1–22.

[6] Deng, Y., Nielsen, R., Song, Y.S., 2024a. A previously reported bottleneck in human ancestry 900 kya is likely a statistical artifact. bioRxiv URL: https://www.biorxiv.org/content/early/2024/10/02/2024.10.01.615851, doi:10.1101/2024.10.01.615851, arXiv:https://www.biorxiv.org/content/early/2024/10/02/2024.10.01.615851.full.pdf.

[7] Deng, Y., Nielsen, R., Song, Y.S., 2024b. Robust and accurate bayesian inference of genome-wide genealogies for large samples. bioRxiv , 2024–03.

[8] Deng, Y., Song, Y.S., Nielsen, R., 2021. The distribution of waiting distances in ancestral recombination graphs. Theoretical Population Biology 141, 34–43.

[9] DeWitt, W.S., Harris, K.D., Ragsdale, A.P., Harris, K., 2021. Nonparametric coalescent inference of mutation spectrum history and demography. Proceedings of the National Academy of Sciences 118, e2013798118.

[10] Gao, Z., Zhang, Y., Cramer, N., Przeworski, M., Moorjani, P., 2023. Limited role of generation time changes in driving the evolution of the mutation spectrum in humans. elife 12, e81188.

[11] Griffiths, R., 1981. Neutral two-locus multiple allele models with recombination. Theoretical Population Biology 19, 169–186.

[12] Griffiths, R.C., Marjoram, P., 1997. An Ancestral Recombination Graph, in: Donnelly, P., Tavare, S. (Eds.), Progress in Population Genetics and Human Evolution, IMA Volumes in Mathematics and its Applications, vol. 87. Springer, pp. 257–270. doi:10.1007/978-1-4757-2609-1{_}16.

[13] Harris, K., 2015. Evidence for recent, population-specific evolution of the human mutation rate. Proceedings of the National Academy of Sciences 112, 3439–3444.

[14] Harris, K., Pritchard, J.K., 2017. Rapid evolution of the human mutation spectrum. elife 6, e24284.

[15] Hu, W., Hao, Z., Du, P., Di Vincenzo, F., Manzi, G., Cui, J., Fu, Y.X., Pan, Y.H., Li, H., 2023. Genomic inference of a severe human bottleneck during the early to middle pleistocene transition. Science 381, 979–984.

[16] Hudson, R.R., 1983. Properties of a neutral allele model with intragenic recombination. Theoretical population biology 23, 183–201.

[17] Hudson, R.R., 2002. Generating samples under a wright–fisher neutral model of genetic variation. Bioinformatics 18, 337–338.

[18] Ignatieva, A., Lyngsø, R.B., Jenkins, P.A., Hein, J., 2021. Kwarg: parsimonious reconstruction of ancestral recombination graphs with recurrent mutation. Bioinformatics 37, 3277–3284.

[19] Kamm, J., Terhorst, J., Durbin, R., Song, Y.S., 2020. Efficiently inferring the demographic history of many populations with allele count data. Journal of the American Statistical Association 115, 1472–1487.

[20] Kamm, J.A., Terhorst, J., Song, Y.S., 2017. Efficient computation of the joint sample frequency spectra for multiple populations. Journal of Computational and Graphical Statistics 26, 182–194.

[21] Kelleher, J., Etheridge, A.M., McVean, G., 2016. Efficient Coalescent Simulation and Genealogical Analysis for Large Sample Sizes. PLoS Computational Biology 12, 1–22. doi:10.1371/journal.pcbi.1004842.

[22] Kelleher, J., Thornton, K.R., Ashander, J., Ralph, P.L., 2018. Efficient pedigree recording for fast population genetics simulation. PLOS Computational Biology 14, 1–21. URL: https://doi.org/10.1371/journal.pcbi.1006581, doi:10.1371/journal.pcbi.1006581.

[23] Kelleher, J., Wong, Y., Wohns, A.W., Fadil, C., Albers, P.K., McVean, G., 2019. Inferring whole-genome histories in large population datasets. Nature genetics 51, 1330–1338.

[24] Kuhner, M.K., Yamato, J., Felsenstein, J., 2000. Maximum likelihood estimation of recombination rates from population data. Genetics 156, 1393–1401.

[25] Li, H., Durbin, R., 2011. Inference of human population history from individual whole-genome sequences. Nature 475, 493–496.

[26] Liu, X., Fu, Y.X., 2015. Exploring population size changes using snp frequency spectra. Nature genetics 47, 555–559.

[27] Morningstar, W.R., Alemi, A., Dillon, J.V., 2022. Pacm-bayes: Narrowing the empirical risk gap in the misspecified bayesian regime, in: International Conference on Artificial Intelligence and Statistics, PMLR. pp. 8270–8298.

[28] Nielsen, R., Vaughn, A.H., Deng, Y., 2024. Inference and applications of ancestral recombination graphs. Nature Reviews Genetics , 1–12.

[29] Petrone, S., Rizzelli, S., Rousseau, J., Scricciolo, C., 2014. Empirical bayes methods in classical and bayesian inference. Metron 72, 201–215.

[30] Ralph, P., Thornton, K., Kelleher, J., 2020. Efficiently summarizing relationships in large samples: a general duality between statistics of genealogies and genomes. Genetics 215, 779–797.

[31] Rasmussen, M.D., Hubisz, M.J., Gronau, I., Siepel, A., 2014. Genome-Wide Inference of Ancestral Recombination Graphs. PLoS Genetics 10. doi:10.1371/journal.pgen.1004342.

[32] Saerens, M., Latinne, P., Decaestecker, C., 2002. Adjusting the outputs of a classifier to new a priori probabilities: a simple procedure. Neural computation 14, 21–41.

[33] Schiffels, S., Durbin, R., 2014. Inferring human population size and separation history from multiple genome sequences. Nature genetics 46, 919–925.

[34] Schiffels, S., Wang, K., 2020. Msmc and msmc2: the multiple sequentially markovian coalescent, in: Statistical population genomics. Humana, pp. 147–165.

[35] Seplyarskiy, V., Koch, E.M., Lee, D.J., Lichtman, J.S., Luan, H.H., Sunyaev, S.R., 2023. A mutation rate model at the basepair resolution identifies the mutagenic effect of polymerase iii transcription. Nature Genetics 55, 2235–2242.

[36] Simchowitz, M., Tosh, C., Krishnamurthy, A., Hsu, D.J., Lykouris, T., Dudik, M., Schapire, R.E., 2021. Bayesian decision-making under misspecified priors with applications to meta-learning. Advances in Neural Information Processing Systems 34, 26382–26394.

[37] Speidel, L., Forest, M., Shi, S., Myers, S.R., 2019. A method for genome-wide genealogy estimation for thousands of samples. Nature Genetics 51, 1321–1329. URL: http://dx.doi.org/10.1038/s41588-019-0484-x, doi:10.1038/s41588-019-0484-x.

[38] Terhorst, J., 2024. Accelerated bayesian inference of population size history from recombining sequence data. bioRxiv .

[39] 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, eabi8264.

[40] Wong, Y., Ignatieva, A., Koskela, J., Gorjanc, G., Wohns, A.W., Kelleher, J., 2023. A general and efficient representation of ancestral recombination graphs. bioRxiv , 2023–11.

[41] YC 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, iyac044.

[42] 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. Nature Genetics , 1–9.

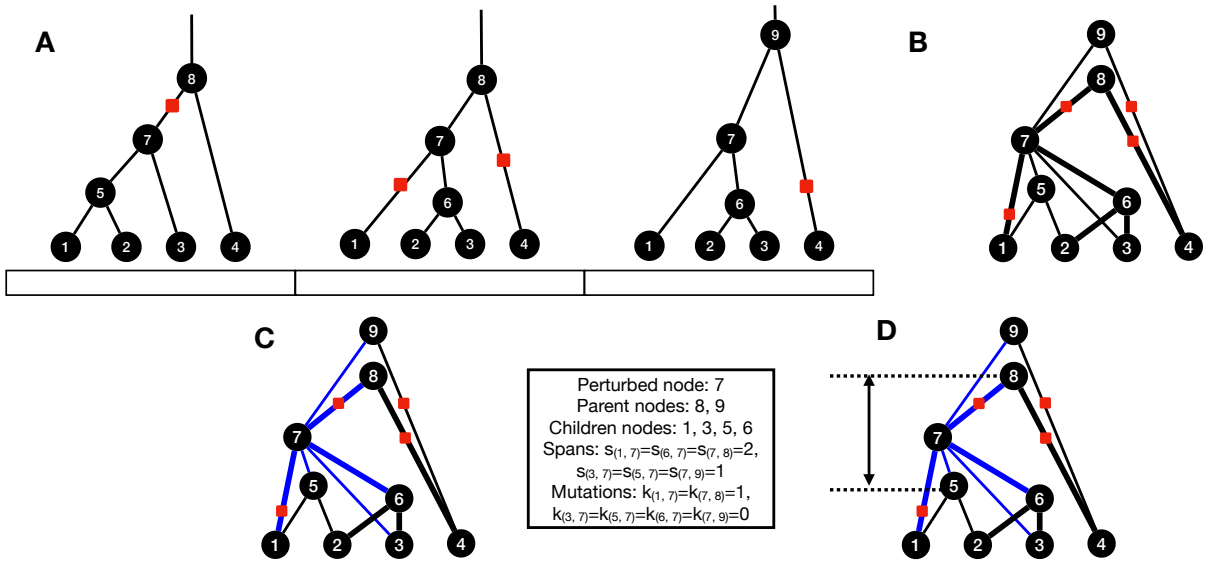

Figure 1: Methodology overview of POLEGON. The tree sequence format of local genealogies on each non-recombining block (A) can be converted to a DAG (B), by merging the nodes with the same index across trees. The width of the edges indicates their spatial spans in the tree sequence. When updating the age of a particular node (C), its age can only move within the interval (D) determined by Eq. (2), as the blue branches must have positive branch length. The age of the node is perturbed in this interval with an MCMC algorithm to sample from Eq. (3). We denote the number of mutations mapped to the branch connecting node u and v by $k_{(u,v)}$, and the total length of the genomic segments where node u and v are adjacent by $s_{(u,v)}$, where u and v are unordered.

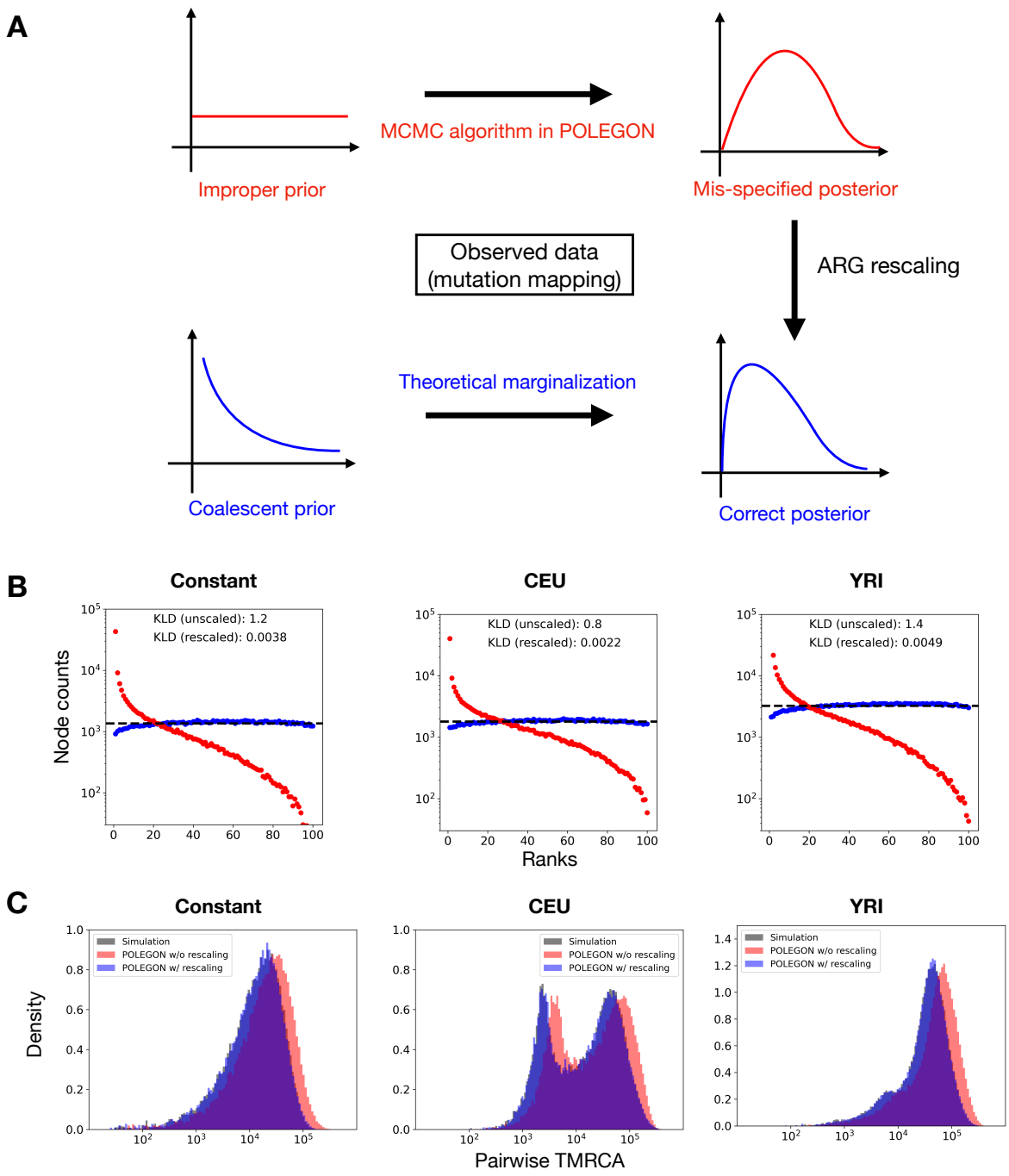

Figure 2: Calibration of distribution with ARG rescaling. (A) ARG rescaling transforms the original mis-specified posterior distribution closer to the correct distribution, even if the true coalescent prior is unknown; (B) The rank plot of the node ages against the node age samples from POLEGON, before (red) and after (blue) rescaling, under simulation under three different demography models: constant size, CEU model and YRI model; (C) The pairwise TMRCA distribution in simulation (black) compared to inferred with POLEGON, with (blue) and without (red) the ARG rescaling operation.

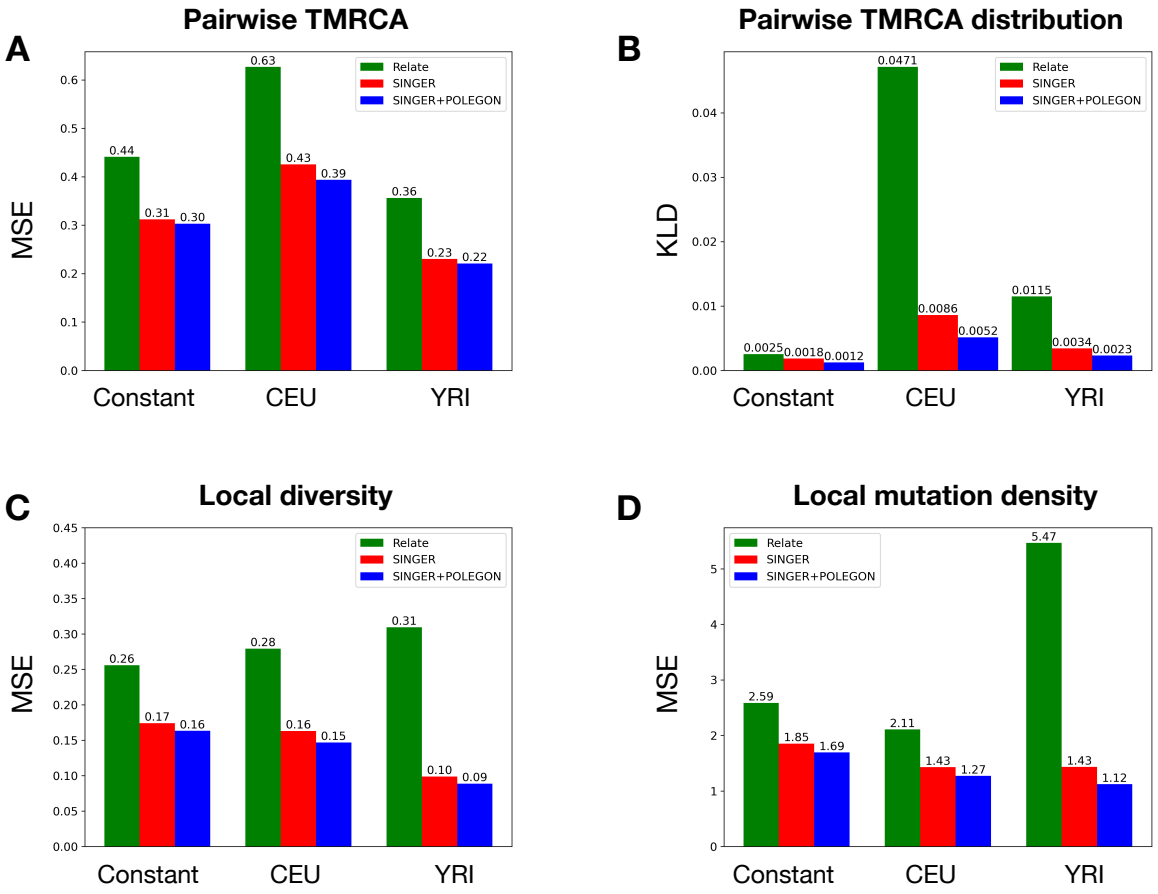

Figure 3: The inference accuracy comparison between Relate (green), SINGER (red) and SINGER+POLEGON (blue), in different aspects: pairwise TMRCA (A), pairwise TMRCA distribution (B), local diversity (C), and local mutation density (D).

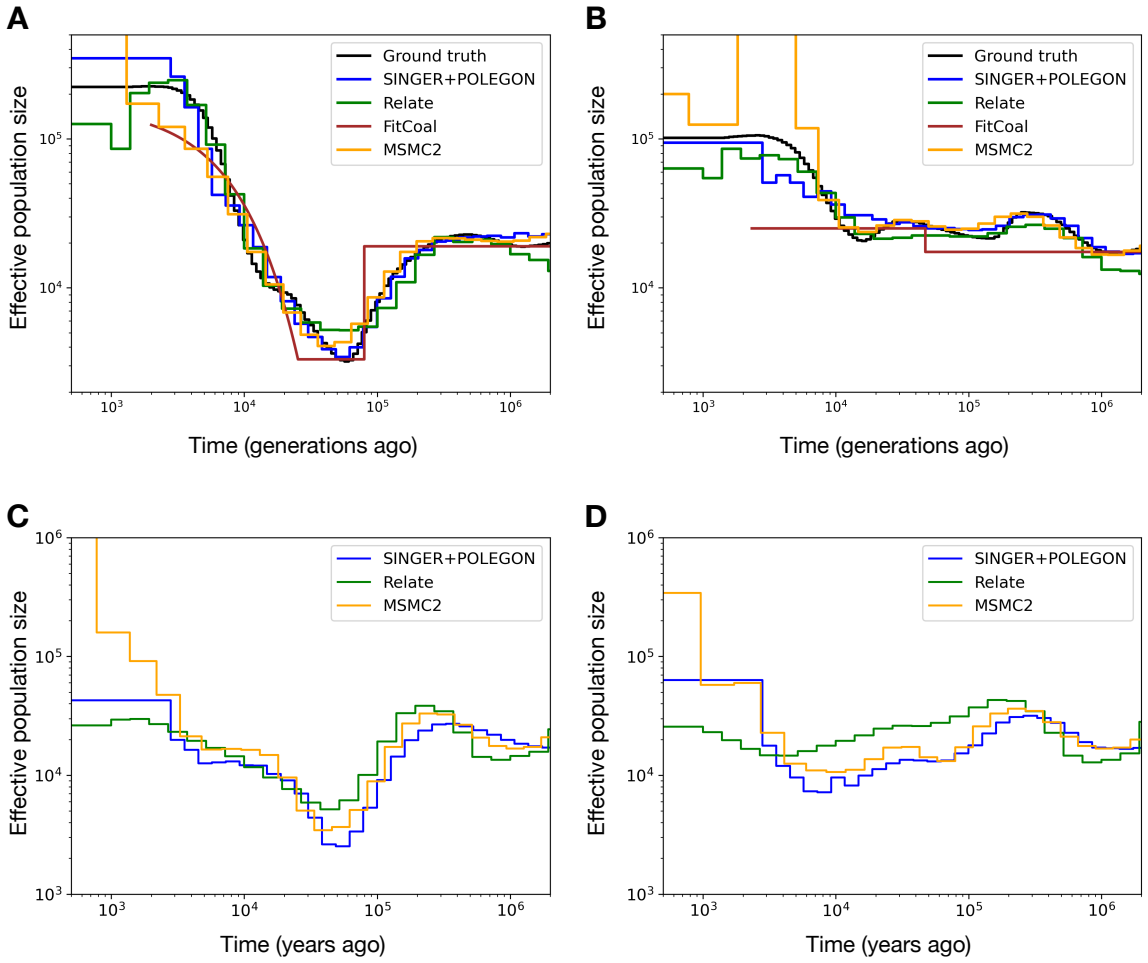

Figure 4: Demography inference results. Simulation benchmark of demography inference with the CEU model (A) and the synthetic model (B), with the inference error from each method provided in the brackets with coalescence rate divergence. We also inferred the population size history for CEU (C) and YRI (D) in the 1000 Genomes Project.

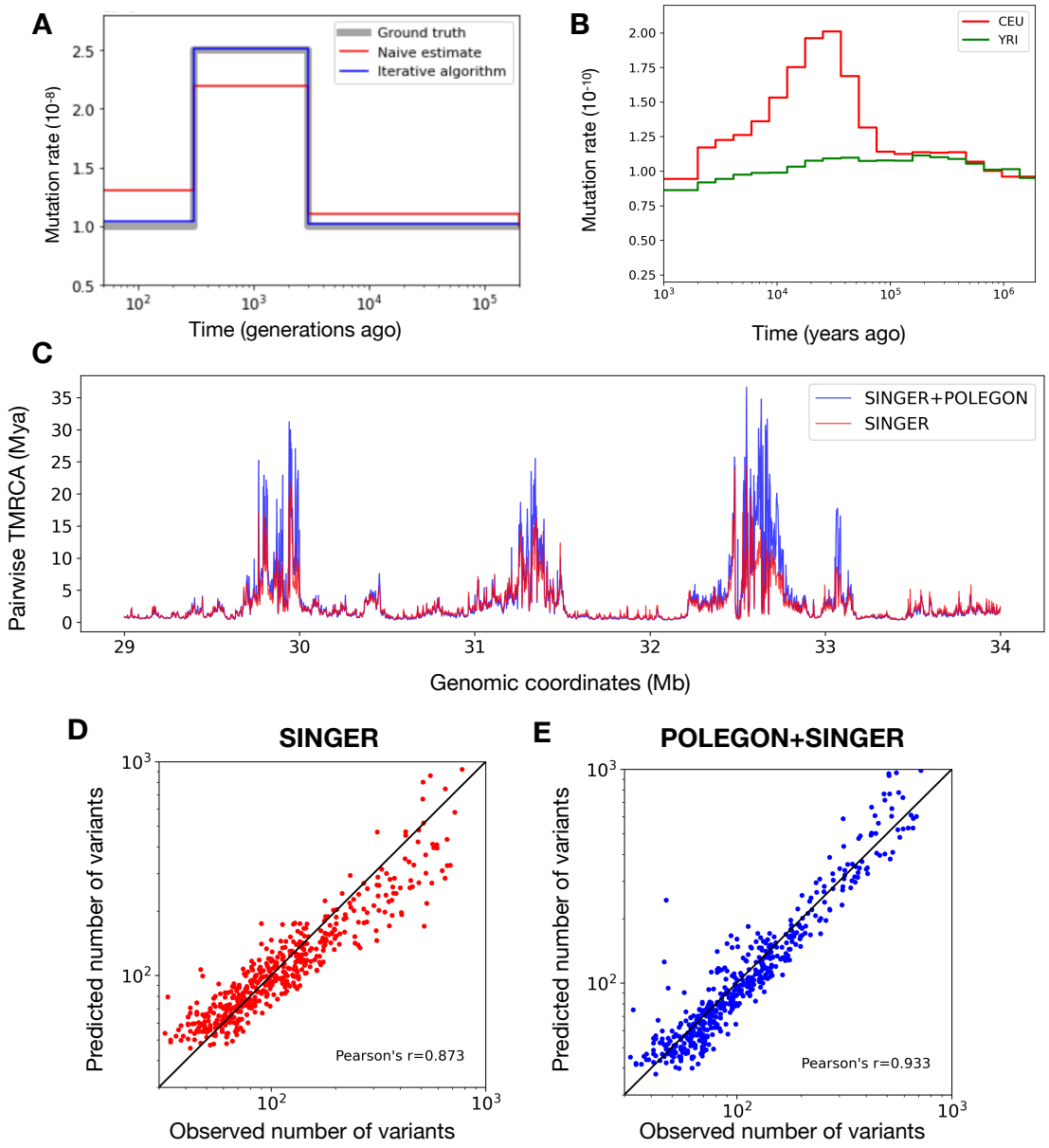

Figure 5: Mutation rate trajectory inference and ancient coalescence times at HLA inferred with SINGER+POLEGON. (A) Performance of mutation rate trajectory inference when assuming uniform placement (red) and using the proposed iterative algorithm (blue); (B) Inferred mutation rate trajectory for TCC → TTC in CEU and YRI, with the proposed iterative algorithm; (C) The estimated average pairwise TMRCA in HLA region using SINGER (red) compared to SINGER+POLEGON (blue); (D) The number of observed variants versus predicted from inferred ARGs with SINGER in 10kb windows; (E) The number of observed variants versus predicted from inferred ARGs with POLEGON+SINGER in 10kb windows.