nature genetics

Article

https://doi.org/10.1038/s41588-025-02129-x

A likelihood-based framework for demographic inference from genealogical trees

Received: 23 October 2023

Accepted: 14 February 2025

Published online: 20 March 2025

Check for updates

Caoqi Fan[1,2] ✉, Jordan L. Cahoon[2,3], Bryan L. Dinh[1,2], Diego Ortega-Del Vecchyo[4], Christian D. Huber[5], Michael D. Edge[2], Nicholas Mancuso[1,2] & Charleston W. K. Chiang[1,2] ✉

The demographic history of a population underlies patterns of genetic variation and is encoded in the gene-genealogical trees of the sampled haplotypes. Here we propose a demographic inference framework called the genealogical likelihood (gLike). Our method uses a graph-based structure to summarize the relationships among all lineages in a gene-genealogical tree with all possible trajectories of population memberships through time and derives the full likelihood across trees under a parameterized demographic model. We show through simulations and empirical applications that for populations that have experienced multiple admixtures, gLike can accurately estimate dozens of demographic parameters, including ancestral population sizes, admixture timing and admixture proportions, and it outperforms conventional demographic inference methods using the site frequency spectrum. Taken together, our proposed gLike framework harnesses underused genealogical information to offer high sensitivity and accuracy in inferring complex demographies for humans and other species.

Accurately inferring the demographic history of humans not only has archeological and historical significance[1–6] but also lessens confounding effects in association studies by better accounting for genetic ancestry and serves as the null expectation of genetic variation when inferring about natural selection[7–11]. Given the complicated interplay of random processes related to the underlying demography and observed genotypes—including migration, coalescence, recombination, mutation and genotyping error—demographic inference is a challenging problem, often requiring simplifying model assumptions or relatively coarse data summaries. One approach to estimate population size histories, first popularized by the Pairwise Sequentially Markovian Coalescent (PSMC) model[12], uses a hidden Markov model (HMM) to describe the variation of haplotypes along the genome, in which the hidden states correspond to the underlying genealogical trees[12–15]. As the number of potential trees grows exponentially with sample size, these methods are only computationally tractable using a reduced representation of the underlying genealogy. As a result, these methods are typically constrained by small sample sizes (usually <100) and the assumption of a single, homogeneous population, although they are flexible with respect to population size trajectories over time. To accommodate larger samples informing recent histories and more complex demographic events, alternative approaches rely on a further reduced representation of the genealogy, such the pattern of haplotype sharing by descent[16,17], or more commonly, the site frequency spectrum[18–23] (SFS).

HMM-based and SFS-based methods are informed by observed genotypes or haplotypes. However, because neutral variation is related

[1]Center for Genetic Epidemiology, Department of Population and Public Health Sciences, Keck School of Medicine, University of Southern California, Los Angeles, CA, USA. [2]Department of Quantitative and Computational Biology, University of Southern California, Los Angeles, CA, USA. [3]Department of Computer Science, University of Southern California, Los Angeles, CA, USA. [4]Laboratorio Internacional de Investigación sobre el Genoma Humano, Universidad Nacional Autónoma de México, Querétaro, México. [5]Department of Biology, Penn State University, University Park, PA, USA. ✉e-mail: fcq1116@gmail.com; charleston.chiang@med.usc.edu

https://doi.org/10.1038/s41588-025-02129-x

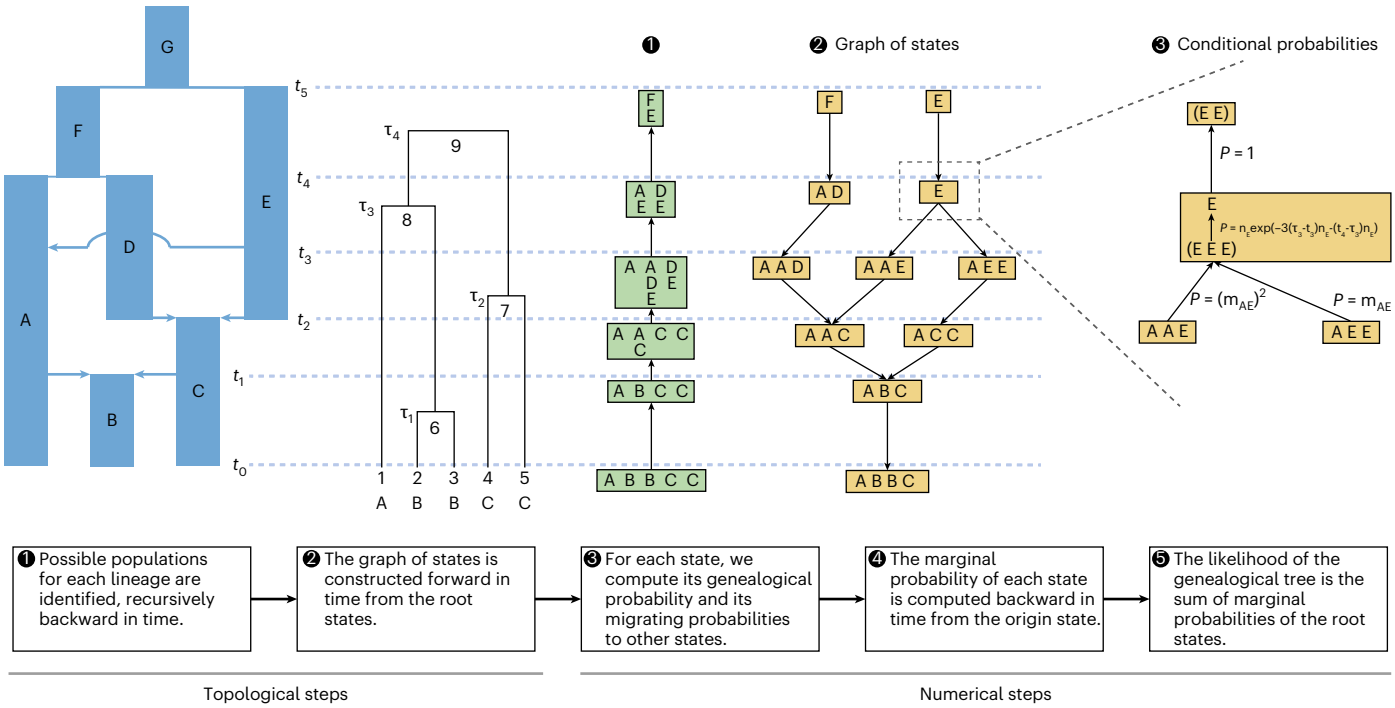


Fig. 1 | A schematic of the major steps of the gLike algorithm with examples. We develop a full methodology for the GOS around three key problems: (1) constructing a minimal GOS that contains all necessary states; (2) computing the conditional probabilities between connected states with considerations of migrations, coalescences and non-coalescences; and (3) propagating the marginal probabilities through the GOS to compute the total likelihood of the tree. Starting from a parameterized demography and an observed genealogical tree with known sample populations, gLike is operationally broken into two topological steps and three numerical steps. The topological steps construct the fundamental data structure in gLike: the GOS (constructed in step 2), which summarizes all possible scenarios for all lineages to move through the populations across history. GOS is guaranteed by a preparatory step 1 such that no redundant states will be generated, minimizing computational burden. The three numerical steps then follow to compute the conditional and marginal probabilities. Further operational details can be found in the Supplementary Notes, using this exact example.

to demographic history entirely through the genealogical processes, (unknown) genealogy arguably has a more direct relationship with the underlying demography than the downstream genotypes[24–26]. Moreover, the complete genealogy of a collection of samples, as represented by an ancestral recombination graph (ARG)[27–29], has richer information than the SFS because it includes the correlated coalescent histories between segments of a chromosome. Therefore, to the extent that ARG can be inferred accurately and consistently[30,31], a genealogy-based demographic inference method has the potential[32] to distinguish complex demographic histories.

Here, we introduce a genealogical likelihood framework named gLike to compute the likelihood of an observed genealogical tree under a parameterized demographic history. A genealogy does not imply the assortment history of any of its lineages (that is, the set of discrete population memberships that a particular lineage has traversed over time), requiring gLike to consider all possible combinations. Notably, gLike bears similarity to the independently proposed 'local ancestry path' problem[33], but instead of inferring the population membership distribution of each individual node, gLike aims to compute the total likelihood of all combinations. We demonstrate the advantage of genealogy-based demography inference by applying gLike to simulated and empirical scenarios of complicated admixture histories, such as three-way or four-way admixtures, and comparing gLike to SFS-based methods. For admixture scenarios across continental ancestries, our inference required no reference sample from the ancestral populations or explicit inference of local ancestries, information that is often not available or is imprecisely estimated for understudied populations with a complex recent history. As a first step towards a general-purpose statistical framework and towards using the information from the entire ARG, gLike is applicable to a variety of demographic events: migrations, splits, admixtures and population size variations, providing tools for model selection and parameter estimation.

Results

Genealogical likelihood under multi-population demography

A genealogical tree, despite being a complete record of the coalescent events of the sampled haplotypes within a chromosomal interval, does not specify the migration history of lineages. In a typical genetic study, samples (leaf nodes) are collected from known populations, which serves as the initial condition. The internal lineages could migrate, subject to the restriction that coalescences must happen within a population. Therefore, the probability of a given genealogical tree corresponds to the cumulative total of all migration scenarios that are compatible with that tree. Our proposed method, gLike, computes the likelihood of any given genealogical tree under a hypothesized demographic history (Methods). Operationally, it is broken into two topological steps to search for possible population memberships of lineages, followed by three numerical steps to compute the conditional and marginal probabilities (Fig. 1).

We define a 'state' as a specification of the population memberships of all lineages existing at a specific time. All possible states before each historical event (for example, occurring at $t_1, t_2, \ldots, t_5$ in Fig. 1) form a directed acyclic graph (Fig. 1, step 2), which we call the 'graph of states' (GOS), a complete representation of all possible migration scenarios. When a state specifies a lineage in an impossible population, it becomes a dead-end state that does not connect to the origin. For example, in step 2, if we imagine a state 'AA' at t_4 as a child of 'F', it will not connect to the origin state 'ABBCC' because the fourth and fifth samples cannot migrate from C to A per the hypothesized demographic model (Fig. 1). To reduce computation time, we avoid generating any

https://doi.org/10.1038/s41588-025-02129-x

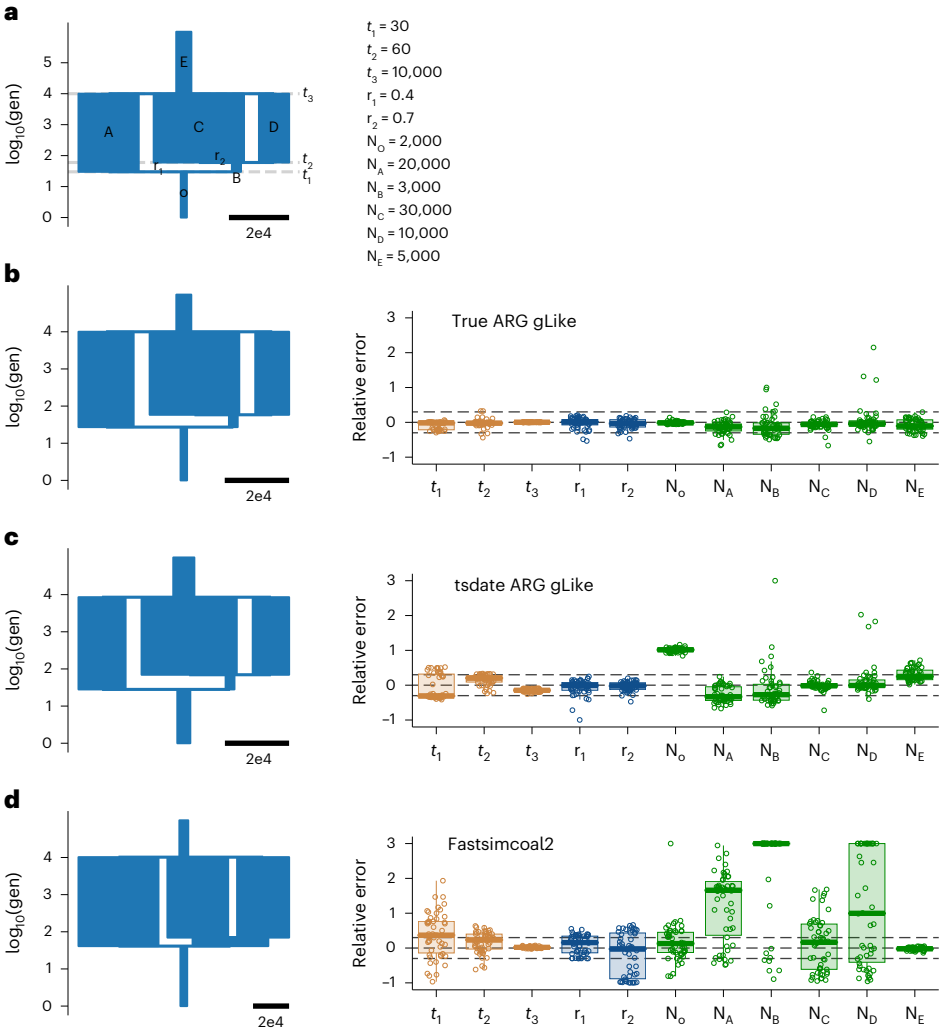


Fig. 2 | gLike accurately reconstructs three-way admixture without ancestral population samples. a, The true demography under which the genealogical trees and genotypes were simulated, with six populations involved: population O is admixed from A and B; B is the intermediate population admixed from C and D, where C is defined to be the major ancestor (proportion ≥ 0.5) without loss of generalizability; E is the ancestor of A, C and D. All population sizes are to scale. There are 11 parameters involved, including six population sizes as well as t_1, time of admixture of population O; t_2, time of admixture of population B; t_3, time of split from population E; r_1, admixture proportion of A in O; and r_2, admixture proportion of C in B. The true value of each parameter is provided on the right. **b**–**d**, The reconstructed demography using parameter estimates averaged over 50 independent simulations (left) and boxplots of relative errors ((estimated − true) / true) in each simulation (right). Boxplots are capped at 300% relative error for ease of visualization. Trees and genotypes of 1,000 haplotypes drawn from population O were simulated on a 30 Mb chromosome. The demographic parameters were estimated by gLike on the true trees (**b**), by gLike on the tsinfer + tsdate-reconstructed trees from the true genotypes (**c**), and by Fastsimcoal2 on the allele frequency spectra derived from true genotypes (**d**). A reference for the width of the population sizes equivalent to 20,000 is given in each panel.

dead-end states by a preliminary step (Fig. 1, step 1) that summarizes possible population memberships for each lineage. For example, in step 1 at t_4, lineage 8 may be in 'A' or 'E', and lineage 7 may be in 'D' or 'E'; thus, 'AA' is not a legal state in step 2 (Fig. 1). The GOS is then constructed from the root states ('F' or 'E' in this example) forward in time, by searching for child states according to both the specified migration events in the demography and the results in step 1.

After building the GOS, the relevant conditional probabilities are computed. As lineages are restricted to their respective population until a historical event, a state immediately before a historical event t_s is sufficient to specify the population memberships of all lineages between t_{s-1} and t_s. For example, the state 'EE' implies that not only the two lineages but also the subtrees under both lineages are all in population 'E' between t_3 and t_4. Given memberships of all lineages within the context of a state, we can compute the 'genealogical probability' of the state based on standard coalescent theory to describe the coalescence (or non-coalescence) events during the relevant interval on the tree. We also compute the 'migration probability' between a state and its child, which is the product of the migration probability of each lineage, according to the migration matrix of the historical event (Fig. 1, step 3). The 'marginal probability' of a state is then the probability conditional on the origin state and can be computed recursively (Fig. 1, step 4). Finally, we compute the likelihood of the genealogical tree as the sum of the marginal probabilities of the root states (Fig. 1, step 5).

Further operational details for each step with illustrative examples can be found in the Supplementary Notes and Supplementary Fig. 1. In practice, we apply gLike to a subsample of trees that are presumed independent, ideally from evolutionarily neutral sites distantly spaced across the genome (usually 10–100, depending on the computational resources), and the total likelihood is computed as the product over each individual tree and optimized. The final estimation of parameters is averaged over a number of subsamples with replacement. The variance across subsamples serves as an indicator of the uncertainty of the estimate.

https://doi.org/10.1038/s41588-025-02129-x

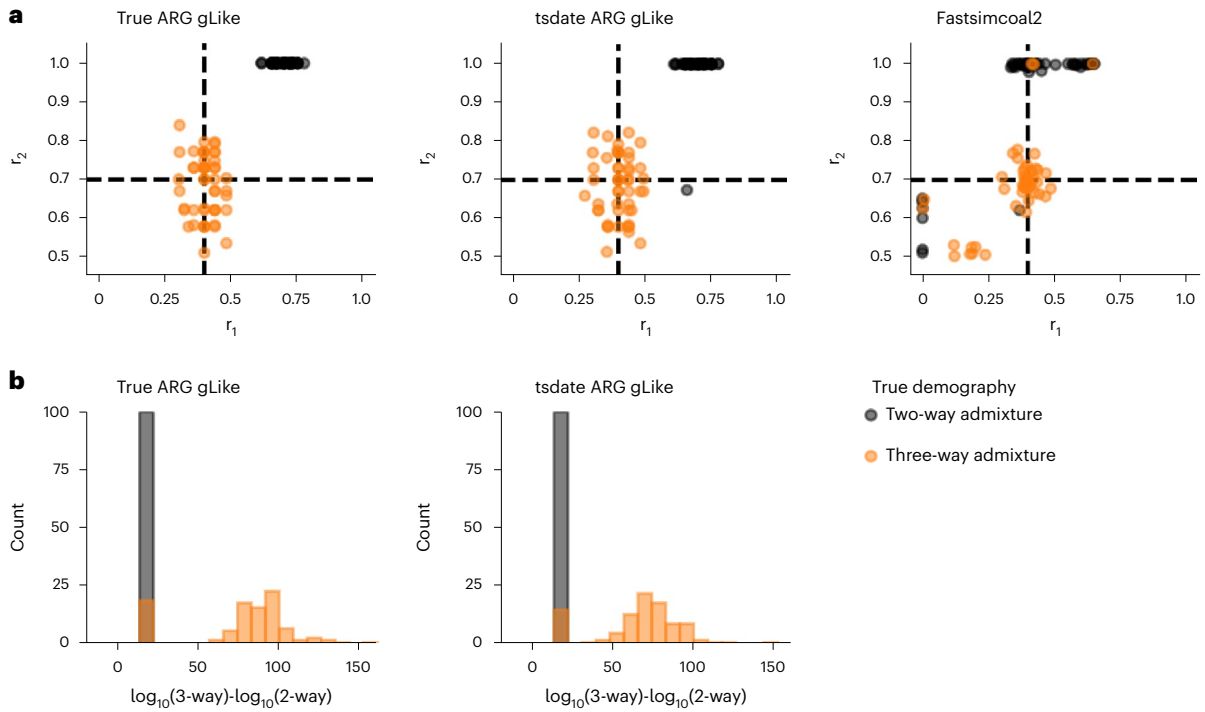


Fig. 3 | gLike distinguishes three-way admixture from two-way admixture. True (left) and tsifner + tsdate-reconstructed (middle) trees were obtained from simulated three-way (orange, same model as Fig. 2) and two-way (gray, r_2 was set to 1, removing contribution from population D) admixed populations. **a**, gLike was applied assuming a three-way admixture model. The estimated r_1 and r_2 values in each of 50 independent simulations are shown; dashed lines denote true values of r_1 and r_2 in three-way admixture simulations. **b**, gLike was first applied under a two-way admixture model, then the model is expanded into a three-way admixture and gLike likelihood is optimized while fixing shared parameters between two models (see Methods for technical details). The distributions of log-likelihood improvement after model expansion are shown as histograms. Model selection through AIC resulted in a classification accuracy of 92%.

gLike accurately infers three-way admixture demography

To showcase the performance of gLike to analyze complex admixture, we simulated 1,000 haplotypes on a 30 Mb chromosome from a population formed by two consecutive recent admixture events from three ancestral populations. Such a demography is parameterized by three event times, two admixture proportions and six population sizes, totaling up to 11 parameters (Fig. 2a). When true genealogical trees were available, the maximum likelihood estimates from gLike, averaged over 50 independent simulations, for all 11 parameters achieved an overall 3.8% relative error (Fig. 2b), with highly concordant distribution of the coalescence events (Extended Data Fig. 1). On the other hand, gLike on the tsdate-reconstructed trees achieved an overall 23.3% relative error (Fig. 2c). In this case, we found that t_1 and N_0 are the most overestimated parameters (by 35.6% and 97.3%, respectively) when using tsdate-reconstructed trees, possibly because of the tendency of tsdate to overestimate times of recent coalescences, prolonging the recent branches (Extended Data Fig. 2). Apart from t_1 and N_0, the other nine parameters are estimated with 13.7% relative error. We found that using as few as 30 haplotypes only marginally decreased accuracy when using true ARG (an overall relative error of 17.1%; Supplementary Fig. 2), although at least 30–100 samples are generally recommended for inference when using trees inferred by tsdate (Supplementary Fig. 3) and may vary by data quality and model complexity. We also found that accuracy is lowered when admixture events are older (Supplementary Figs. 4 and 5), probably because few lineages are left going farther back in time. However, this could be ameliorated with the presence of reference populations or ancient samples (see below, also see Discussion).

We benchmarked our method against Fastsimcoal2 (ref. 21), which is capable of flexibly inferring complex demography using site frequency spectra. Based on true genotypes and the same three-way admixture model, Fastsimcoal2 estimates had a relative error of 54.1%, which led to a visually distorted demography (Fig. 2d). This is in sharp contrast to Fastsimcoal2 showing comparable accuracy to gLike on a three-population split demography (Supplementary Fig. 6) and is probably partly driven by the lack of reference samples for admixture. We also found that gLike outperformed pg-gan[34], a generative-adversarial-network-based deep-learning approach (for example, Supplementary Fig. 7), although our experiments were not conducted with any specialized neural network hardware and thus we do not dismiss the potential for generative adversarial network as an emergent approach.

gLike detects components of admixture with high confidence

We examined the ability of gLike to distinguish two-way from three-way admixtures. We first applied gLike under a hypothesized three-way admixture model to estimate admixture proportions, r_1 and r_2. We found that when the true demography was a three-way admixture, the estimated admixture proportion for the third ancestry component, r_2, centered around the true value (0.7) and was always far from the boundaries (0.5 and 1.0). When the true demography was a two-way admixture, the estimated r_2 was almost always 1.0, with only one exception (Fig. 3a, left and middle). This indicates that gLike correctly reduced a three-way admixture model into a two-way model when it was indeed two-way admixed. By contrast, both r_1 and r_2 were estimated to be the boundary values around half of the time by Fastsimcoal2, regardless of the true demography (Fig. 3a, right panel).

We next evaluated the maximum likelihood achieved under a two-way admixture model and a three-way admixture model (Methods). Akaike's information criterion (AIC) model selection was applied on the log-likelihood differences between two models to select the more plausible model between the two-way and three-way admixtures. Across 100 independent simulations, the three-way admixture model was never

https://doi.org/10.1038/s41588-025-02129-x

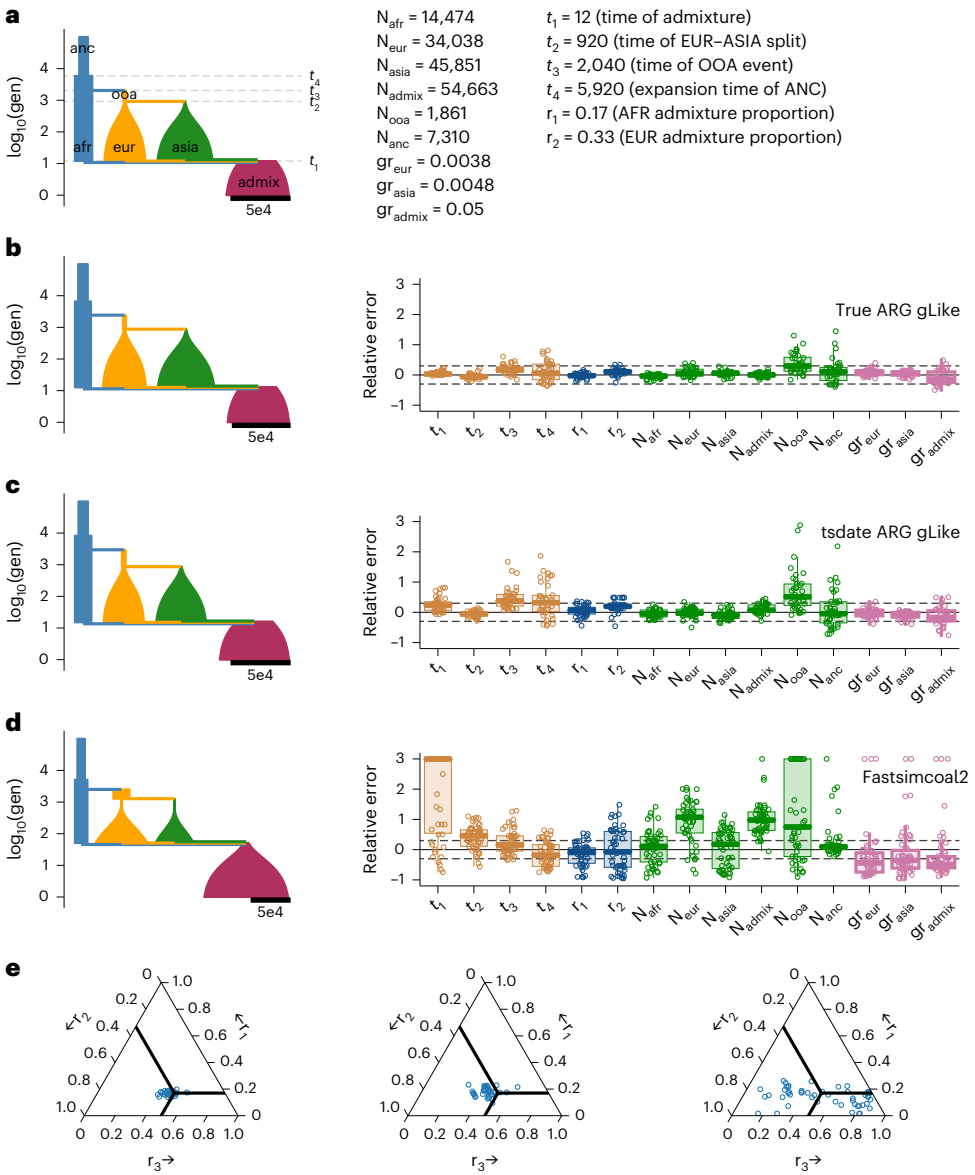


Fig. 4 | gLike reconstructs the American admixture demography. **a**, American admixture demography with parameters from stdpopsim model 4B11. All population sizes are drawn to scale. The true value for the sizes of each population (N), the growth rate (gr) and time of demographic events are given on the right. **b**–**d**, The reconstructed demography using estimations averaged over 50 replicate simulations (left) and boxplots of relative errors in each simulation (right). Trees and genotypes of 1,000 haplotypes from the admixed population were simulated on a 30 Mb chromosome, the demographic parameters were estimated by gLike on the true trees (**b**) or the tsinfer + tsdate-reconstructed trees (**c**), and by Fastsimcoal2 on the allele frequency spectra derived from true genotypes (**d**). Boxplots are capped at 300% relative error for ease of visualization. A reference for the width of the population sizes equivalent to 50,000 is given in each panel. **e**, Ternary plots showing admixture proportions estimated by gLike on the true trees (left), by gLike on the tsinfer + tsdate-reconstructed trees (middle) or by Fastsimcoal2 on the allele frequency spectra of the true genotypes (right), with slide lines indicating true parameters.

preferred when the true admixture was two-way, and the three-way admixture model was preferred over the two-way when it was the true model ~85% of the time with both true ARGs and tsdate-reconstructed ARGs (Fig. 3b), resulting in ~92% accuracy of classification.

gLike infers complex demographic histories from stdpopsim

We further evaluated the ability of gLike to reconstruct two additional demographic models with increasing complexity, as published in stdpopsim[35]: the American Admixture (stdpopsim model 4B11; Fig. 4) and the Ancient Europe (stdpopsim model 4A21; Fig. 5) demographies.

The American Admixture model consists of four populations. Following stdpopsim, three ancestral populations were labeled AFR, EUR and ASIA to represent ancestries from the African, European and Asian continents, respectively. ADMIX is the population formed by a very recent admixture from the three ancestral populations. This model has 15 parameters, including four event times, two admixture proportions, six population sizes and three exponential growth rates (Fig. 4). We simulated 1,000 haplotypes from population ADMIX on a 30 Mb chromosome. gLike on the true trees inferred all 15 parameters with overall 11.3% relative error (Fig. 4b). In particular, N_{ooa}, the size of the out-of-Africa predecessor of the EUR population, was overestimated by 38.5%. This is a result of approximations that gLike undertook because the number of connections between states exceeded the predefined threshold (see Methods), and the bias can be mitigated by increasing this threshold (Extended Data Fig. 3). gLike on the tsdate-reconstructed trees inferred parameters with overall 23.5% relative error (Fig. 4c).

https://doi.org/10.1038/s41588-025-02129-x

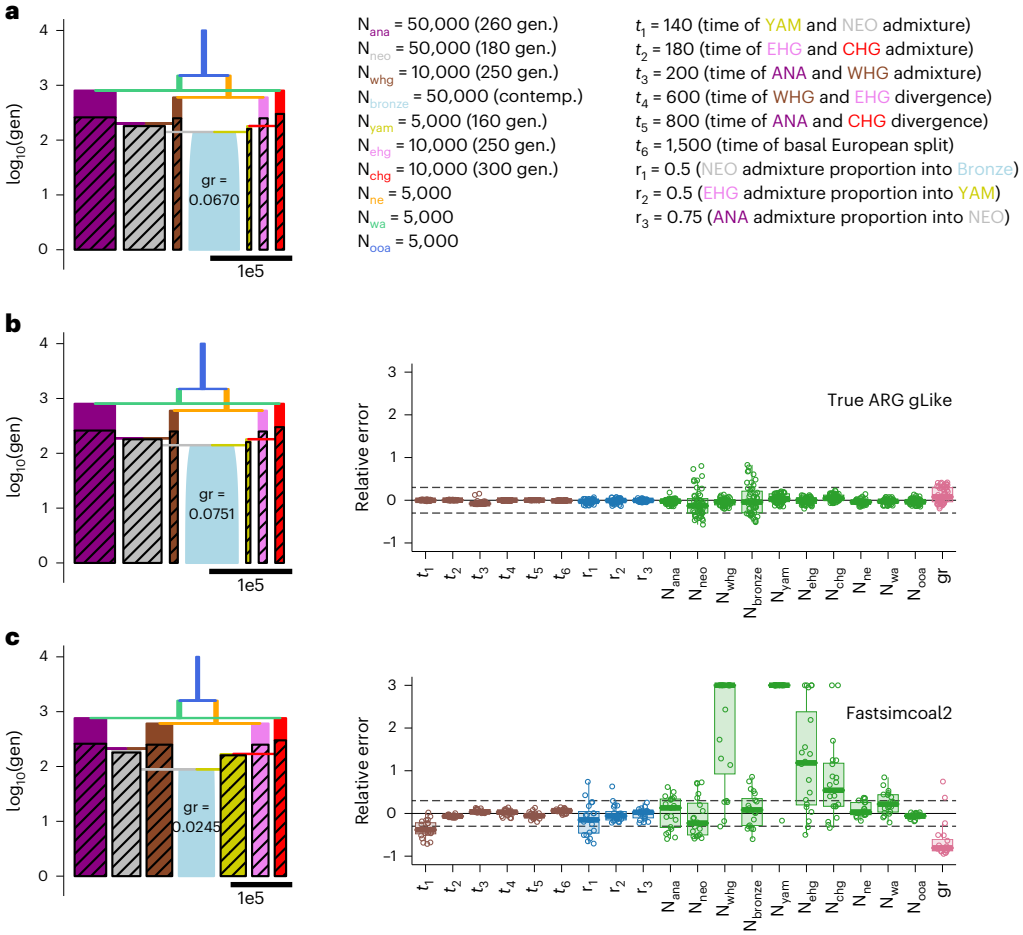


Fig. 5 | gLike reconstructs the ancient Europe demography. a, Ancient Europe demography with parameters from stdpopsim model 4A21. Populations are labeled per stdpopsim model: OOA/ooa, out-of-Africa; NE/ne, Northern European; WA/wa, West Asian; CHG/chg, Caucasus hunter-gatherer; ANA/ana, Anatolian; WHG/whg, western hunter-gatherer; EHG/ehg, eastern hunter-gatherer; YAM/yam, Yamnaya; NEO/neo, Neolithic; Bronze/bronze, Bronze Age. The Bronze Age population is plotted with initial size true to scale, but the growth rate (gr) is shown as text to avoid a disproportionate figure. All other population sizes are constant size and drawn to scale. True parameters for simulation are shown on the right. For sampled populations, the sampling times are shown inside brackets following the corresponding population sizes. **b**,**c**, The reconstructed demography using estimates averaged over 50 replicate simulations (left) and boxplots of percentage errors in each simulation (right). Trees and genotypes were simulated on a 30 Mb chromosome. A total of 220 haplotype samples (100 contemporary samples descended directly from the Bronze Age population and 20 ancient samples each from the six ancient populations) were drawn at collection times as described by stdpopsim. The demographic parameters were estimated by gLike on the true trees (**b**) or by Fastsimcoal2 on the site frequency spectra of the true genotypes (**c**). Boxplots are capped at 300% relative error for ease of visualization. A reference for the width of the population sizes equivalent to 100,000 is given in each panel. For each ancestral population, the period after reference sample collection is marked by hash lines, to indicate that reference samples do not provide information on this part of history.

Except from the overestimation of N_{ooa} by 77.8%, the error concentrated on the AFR branch. Fastsimcoal2, by comparison, estimated the same set of parameters with 46.0% relative error (Fig. 4d). Fastsimcoal2 estimated the AFR proportion fairly accurately, but appears to be unable to distinguish between the EUR and ASIA proportions (Fig. 4e). Providing Fastsimcoal2 with 500 additional haplotypes from each ancestral population and multidimensional site frequency spectra improved the accuracy and consistency of Fastsimcoal2's estimation of almost all parameters (an average of 17.2% relative error), which is comparable to the performance of gLike based on the inferred trees (16.7% relative errors), although gLike on true trees (5.8% relative errors) was still more accurate in capturing the histories of these populations (Supplementary Fig. 8).

To test the performance of gLike on intra-continental admixtures, we also evaluated the Ancient Europe model from stdpopsim (2A21). This model is a four-way admixture model in which the two intermediate ancestors of the Bronze Age ('bronze' in Fig. 5a) population are each in turn admixed from two ancestors. For such a complex demography that involves six ancestral populations and relatively old admixtures, inference without any reference sample appeared challenging (Supplementary Fig. 9). Therefore, we simulated 100 haplotypes from the present-day population that descended from the Bronze Age and 20 from each of the ancient populations, sampled according to the times specified by stdpopsim. These sample sizes were chosen to roughly match the sample sizes currently available for an ancient DNA sample and for computational consideration (see 'Runtime considerations' section below). Applying gLike to the true trees resulted in estimates of the 20 parameters with overall 3.1% relative error (Fig. 5b). Fastsimcoal2 estimated all parameters with an average relative error of 44.5% (Fig. 5c). Notably, Fastsimcoal2 significantly overestimated N_{yam} because the reference samples were only informative for about 20 generations in the model (sampled 160 generations ago and admixed 180 generations ago). We did not test gLike using inferred trees, as ancient samples are not yet fully incorporated during ARG inference by tsdate (see Supplementary Fig. 10 for an illustrative example). However, we anticipate that gLike will substantially improve over Fastsimcoal2 in

https://doi.org/10.1038/s41588-025-02129-x

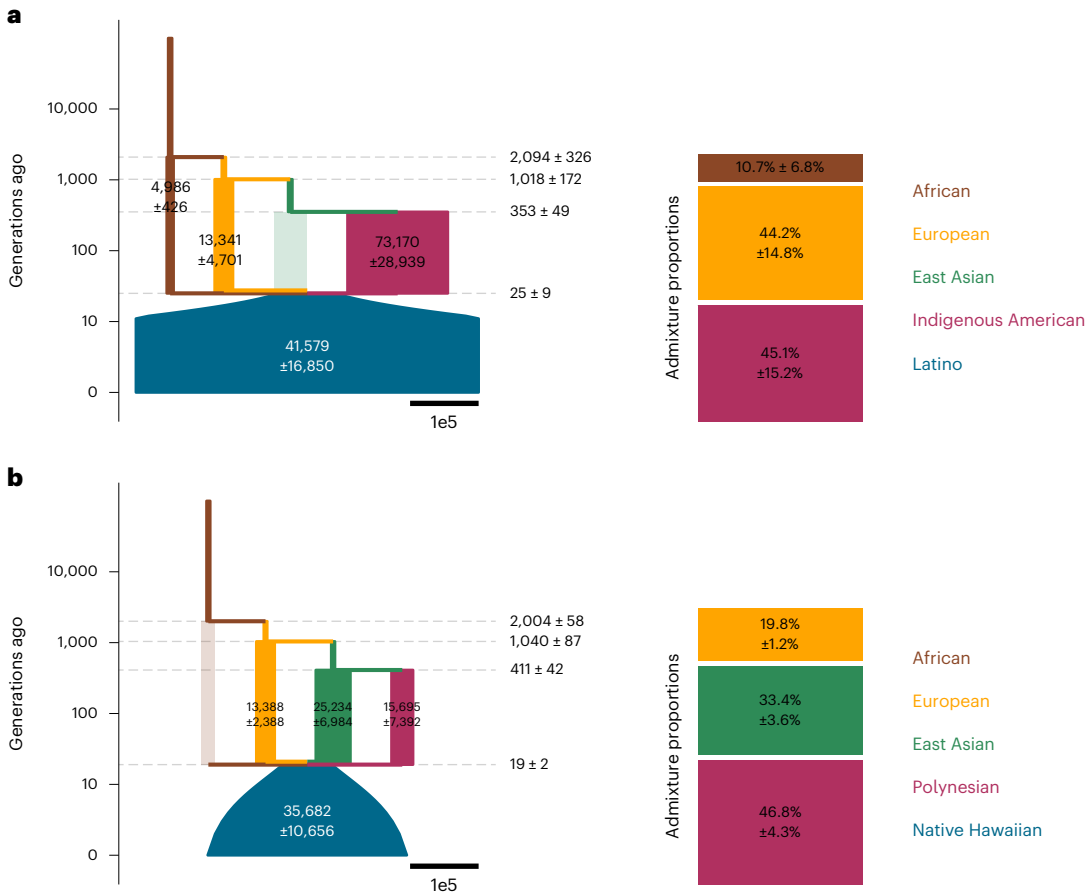


Fig. 6 | Parameter estimations for the demographic histories of Latinos and Native Hawaiians. a,**b**, gLike was applied under a potential four-way admixture model reminiscent of stdpopsim model 4B11 for both the Latino (**a**) and Native Hawaiian (**b**) data, each with 500 individuals. The four potential ancestral populations are African, European, East Asian and Indigenous American (for Latinos) or Polynesian (for Native Hawaiians). gLike was run in 20 independent parallel threads, each making an inference based on ten randomly sampled trees. The reconstructed demographic diagrams are to scale, marked with relevant parameters and estimation uncertainties based on the mean and standard deviation of the 20 threads, respectively. Ancestral populations estimated to have 0% admixture proportion are shown as translucent because their sizes cannot be estimated. Stacked barplots show the estimated admixture proportions of ancestral populations.

accuracy for some parameter estimates if inferred ARGs can accurately incorporate ancient samples, and expect that gLike can generally handle intra-continental admixtures when ancestral populations may be relatively closely related.

Inferring admixture history of Latinos and Native Hawaiians

We applied gLike to investigate populations with complex demographic histories using genome-wide array genotyping data from Latinos and Native Hawaiians. We estimated 16 parameters of a four-way admixture model consisting of Africans, Europeans, East Asians and a fourth ancestral population representing either the Indigenous Americans (for Latinos) or the Polynesians (for Native Hawaiians) (Fig. 6 and Supplementary Table 1). In both cases, the inferred demography reduced down to a three-way admixed model (Fig. 6), with estimates of admixture proportions broadly consistent with those from previous studies[36,37] and an alternative supervised approach (Supplementary Table 2). We estimated the Native Hawaiians to be more recently admixed than the Latinos (19 compared to 25 generations ago), with a slightly smaller initial population size (35,682 ± 10,656 compared to 41,579 ± 16,851, but both are probably overestimated; see Discussion) and smaller growth rate (0.078 ± 0.009 compared to 0.132 ± 0.012) since the admixture. We found that European ancestries participated in both admixture events, with similar population sizes (13,388 ± 2,388 and 13,341 ± 4,702) and timing of its divergence with the East Asians (1,018 ± 172 and 1,041 ± 87 generations ago), suggesting a similar underlying population that colonized the Americas and Polynesia. Note that this ancestry should be more appropriately interpreted as the colonizing population, which is less genetically diverse than the entire European continent currently or at the time. On the other hand, the Indigenous American ancestry was estimated to have larger sizes (73,170 ± 28,939) than the Polynesian ancestry (15,695 ± 7,393), which may reflect greater population sizes or more extensive structure in the indigenous ancestors of the Latinos than their counterpart of the Native Hawaiians.

We estimated the uncertainties of these parameter estimates through parametric bootstrapping. We found the resampled error intervals captured the true parameters approximately 83% of the time overall, although empirical coverage differed across parameters (Supplementary Table 3). Considering the potential errors during the ARG-reconstruction process (as have been seen in Figs. 2 and 4), biases resulting from approximations for computational efficiency (Extended Data Fig. 3) and the lack of high-quality sequencing data (Supplementary Table 4), these point estimates of the demographic parameters for both populations should be taken with caution. Nevertheless, our results suggest that gLike is able to qualitatively capture known features of the demographic history of Latinos and Native Hawaiians without reference data from their ancestral populations, and the results stand to improve as ARG-reconstruction approaches advance.

Runtime considerations

The mean runtime of gLike inference on each experiment is summarized in Supplementary Table 5, ranging from 0.55 h for the three-way admixture model to 86.12 h for the Latino empirical data analysis. Factors influencing the computational burden of gLike include the number of states and their interconnections as well as the exact structure of the genealogical trees (see Supplementary Table 5 legend). Sample size is also an important factor affecting gLike runtime. In the three-way admixture model (Fig. 2), we found that increasing the sample size up to 30,000 led to a logarithmic increase in computational burden (Extended Data Fig. 4). Therefore, a sample size between 100 (see Supplementary Fig. 3) and 30,000 would be generally recommended for balancing performance and runtime. We also introduce a customizable parameter, κ, to directly control the scale of the GOS and restrict the runtime when enumerating all states may not be feasible for complex demographies (see Methods for details). In such cases, κ connections are randomly sampled to approximate the entire GOS. The default setting of 10,000 connections is usually sufficient for accurately estimating most parameters, but some parameters, such as the out-of-Africa population size, N_{ooa}, in the American Admixture model (Fig. 4) may require a higher κ to be estimated within 10% error (Extended Data Fig. 3).

Discussion

With recent advances in scalable ARG inference, a population-genetic approach that explicitly uses the ARG or its marginal trees is an exciting area of active research. In this study, we introduced a framework that explains the stochastic formation of the genealogical trees in a multi-population context and computes the full likelihood of each demographic scenario. Our results revealed that the history of cross-continental admixture can be clearly decoded from the genealogical trees of a single sample of admixed haplotypes. For many understudied diverse populations across the world, it is often unclear whether they are admixed and, if so, what the genetic properties of the ancestral populations may be. Even if the ancestral populations can be hypothesized, they may no longer exist or could be difficult to sample. For these populations, demographic inference using allele frequencies can be challenging[38]. gLike has the potential to provide new demographic insights for these understudied or ancient populations as well as for other species.

gLike has some commonality with approaches to species-tree inference based on gene trees, where gene trees can be used to estimate the topology and branch lengths of a phylogenetic tree[39]. Such methods estimate the whole topology, whereas we pre-specify the demographic history and estimate parameters related to it, including processes like admixture that do not feature as prominently in species-tree inference. Having to assume a parametric demographic model is a common approach shared by many existing demographic inference methods but it also underscores the importance of selecting an appropriate model for meaningful inference. The gLike package includes visualization utilities, such as coalescence distributions (Extended Data Fig. 1) and the most probable population label for each node of a tree, which could provide some intuition of the fit of data to the hypothesized demographic model. In comparing models with different waves of admixture (Fig. 3), we found that using AIC for model selection worked well, although alternative approaches, such as leaving alternate chromosomes out during model fitting, could also be sensible. Nevertheless, for complicated models, some prior knowledge will be useful for model construction or candidate model proposals. Therefore, developing comprehensive methods incorporating model selection, perhaps starting with approaches akin to the species-tree or admixturegraph inference[4,40–42] to obtain a skeleton topology, will be an important focus of future research.

We note that currently, gLike is not using the full information encoded in an ARG but rather relies on sets of presumed independent trees. HMM-based demographic inferences[12–15] are computationally intensive and have limited scalability because of their intricate handling of recombination events. We reasoned that although recombination events are essential for ARG inference, they are less informative for genealogy-based demographic inference. Given an accurately inferred ARG, recombination events can be modeled as a random breakpoint in the genealogical tree re-coalesced onto the rest of the tree. The random break is independent of demography, and the re-coalescence holds minimal information compared to the numerous coalescences already on the tree. In light of the limited gain in information from recombination events, gLike currently focuses on rigorously modeling lineage assortments and coalescent events within independent trees rather than the variability between neighboring trees to accommodate thousands of samples and multiple populations in the model.

One current limitation of gLike is that continuous migration is not supported because it drastically increases the number of states. In the American Admixture simulations (Fig. 4), we omitted the weak migrations (10^{-5}–10^{-4} per generation) between continental populations as originally specified by the stdpopsim model. Omitting the continuous migrations has no visible impact on estimating the remaining parameters unless they are ~100 times more intense than typically presumed rates between continental human populations (Extended Data Fig. 5). However, such frequent migrations (10^{-3}–10^{-2} per generation) may exist between intra-continental populations, where geographical separations are minimal. Estimating the migration rate itself is also of interest in ecological studies of other species, and a future focus will be extending gLike to incorporate continuous migration, perhaps through discretizing the continuous migration coupled with a more efficient random sampling technique on the states.

Future improvement on ARG inference methods (for example, ref. 43) may further expand the applications of gLike; here, we discuss three possible directions. First, the point estimates of coalescent times may be incomplete summaries of the data[30]. Incorporating the uncertainties or posterior distributions of the ARG might lead to more accurate and robust demographic inferences. Second, inferring genealogies with ancient DNA (aDNA) samples is of great interest in many applications. In simulations modeling the intra-continental demography of Ancient Europe, we found that as few as ten diploid aDNA samples significantly enhanced inference accuracy (compare Fig. 5 and Supplementary Fig. 9) and reduced computation time (Supplementary Table 4). This is because ancient samples are closer to the admixture events and thus experienced fewer coalescences. If aDNA samples can be accurately sequenced and phased, they offer more information about the admixtures and histories of ancestral populations than contemporary samples of the same size. However, this potential application is currently limited by the quality of ancient DNA data and the lack of methods to appropriately incorporate ancient DNA into the ARGs (see Supplementary Fig. 10). Finally, gLike on ARGs inferred by tsinfer + tsdate on simulated array data showed noticeable biases (Supplementary Tables S3 and S4). Implementing specialized correction procedures to account for ascertainment bias, either during ARG inference or demographic inference, could broaden the application of this method to other populations and species.

Lastly, we acknowledge that human migrations and admixtures exist on a continuum. In the current framework, we opted to model discrete populations and components of ancestries, as is customary when modeling the histories of multi-ancestry, recently admixed populations such as the Latinos. However, one of the advantages of an ARG-based view of human history may be to remove the notion of discrete populations. Enabling continuous rather than pulse-like migrations between populations to enhance gLike may be another step forward, and future developments of ARG-based demographic inference may emphasize the paradigm shift to represent human histories and structure on a continuum.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41588-025-02129-x.

References

1. Schlebusch, C. M. & Jakobsson, M. Tales of human migration, admixture, and selection in Africa. *Annu. Rev. Genom. Hum. Genet.* **19**, 405–428 (2018).
2. Chiang, C. W. K. et al. Genomic history of the Sardinian population. *Nat. Genet.* **50**, 1426–1434 (2018).
3. Micheletti, S. J. et al. Genetic consequences of the transatlantic slave trade in the Americas. *Am. J. Hum. Genet.* **107**, 265–277 (2020).
4. Patterson, N. et al. Ancient admixture in human history. *Genetics* **192**, 1065–1093 (2012).
5. Peter, B. M. Admixture, population structure, and *F*-statistics. *Genetics* **202**, 1485–1501 (2016).
6. Lipson, M. Applying f_4-statistics and admixture graphs: theory and examples. *Mol. Ecol. Resour.* **20**, 1658–1667 (2020).
7. Lohmueller, K. E. The distribution of deleterious genetic variation in human populations. *Curr. Opin. Genet. Dev.* **29**, 139–146 (2014).
8. Keinan, A. & Clark, A. G. Recent explosive human population growth has resulted in an excess of rare genetic variants. *Science* **336**, 740–743 (2012).
9. Wang, S. R. et al. Simulation of Finnish population history, guided by empirical genetic data, to assess power of rare-variant tests in Finland. *Am. J. Hum. Genet.* **94**, 710–720 (2014).
10. Medina-Muñoz, S. G. et al. Demographic modeling of admixed Latin American populations from whole genomes. *Am. J. Hum. Genet.* **110**, 1804–1816 (2023).
11. Uricchio, L. H., Zaitlen, N. A., Ye, C. J., Witte, J. S. & Hernandez, R. D. Selection and explosive growth alter genetic architecture and hamper the detection of causal rare variants. *Genome Res.* **26**, 863–873 (2016).
12. Li, H. & Durbin, R. Inference of human population history from individual whole-genome sequences. *Nature* **475**, 493–496 (2011).
13. Sheehan, S., Harris, K. & Song, Y. S. Estimating variable effective population sizes from multiple genomes: a sequentially Markov conditional sampling distribution approach. *Genetics* **194**, 647–662 (2013).
14. Schiffels, S. & Durbin, R. Inferring human population size and separation history from multiple genome sequences. *Nat. Genet.* **46**, 919–925 (2014).
15. Terhorst, J., Kamm, J. A. & Song, Y. S. Robust and scalable inference of population history from hundreds of unphased whole genomes. *Nat. Genet.* **49**, 303–309 (2017).
16. Palamara, P. F., Lencz, T., Darvasi, A. & Pe'er, I. Length distributions of identity by descent reveal fine-scale demographic history. *Am. J. Hum. Genet.* **91**, 809–822 (2012).
17. Browning, S. R. et al. Ancestry-specific recent effective population size in the Americas. *PLoS Genet.* **14**, e1007385 (2018).
18. Gutenkunst, R. N., Hernandez, R. D., Williamson, S. H. & Bustamante, C. D. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genet.* **5**, e1000695 (2009).
19. Bhaskar, A., Wang, Y. X. R. & Song, Y. S. Efficient inference of population size histories and locus-specific mutation rates from large-sample genomic variation data. *Genome Res.* **25**, 268–279 (2015).
20. Kamm, J., Terhorst, J., Durbin, R. & Song, Y. S. Efficiently inferring the demographic history of many populations with allele count data. *J. Am. Stat. Assoc.* **115**, 1472–1487 (2020).
21. Excoffier, L. et al. fastsimcoal2: demographic inference under complex evolutionary scenarios. *Bioinformatics* **37**, 4882–4885 (2021).
22. Excoffier, L., Dupanloup, I., Huerta-Sánchez, E., Sousa, V. C. & Foll, M. Robust demographic Inference from genomic and SNP data. *PLoS Genet.* **9**, e1003905 (2013).
23. Liu, X. & Fu, Y.-X. Stairway Plot 2: demographic history inference with folded SNP frequency spectra. *Genome Biol* **21**, 280 (2020).
24. McVean, G. A genealogical interpretation of principal components analysis. *PLoS Genet.* **5**, e1000686 (2009).
25. Opgen-Rhein, R., Fahrmeir, L. & Strimmer, K. Inference of demographic history from genealogical trees using reversible jump Markov chain Monte Carlo. *BMC Evol. Biol.* **5**, 6 (2005).
26. Fan, C., Mancuso, N. & Chiang, C. W. K. A genealogical estimate of genetic relationships. *Am. J. Hum. Genet.* **109**, 812–824 (2022).
27. Hudson, R. R. Gene genealogies and the coalescent process. In *Oxford Surveys in Evolutionary Biology* Vol. 7 (eds Futuyma, D. & Antonovics, J.) 1–44 (Oxford Univ. Press, 1990).
28. Griffiths, R. C. & Marjoram, P. Ancestral inference from samples of DNA sequences with recombination. *J. Comput. Biol.* **3**, 479–502 (1996).
29. Lewanski, A. L., Grundler, M. C. & Bradburd, G. S. The era of the ARG: an introduction to ancestral recombination graphs and their significance in empirical evolutionary genomics. *PLoS Genet.* **20**, e1011110 (2024).
30. Brandt, Y. C. et al. Evaluation of methods for estimating coalescence times using ancestral recombination graphs. *Genetics* **221**, iyac044 (2022).
31. Peng, D., Mulder, O. J. & Edge, M. D. Evaluating ARG-estimation methods in the context of estimating population-mean polygenic score histories. *Genetics* https://doi.org/10.1093/genetics/iyaf033 (2025).
32. Brandt, D. Y. C., Huber, C. D., Chiang, C. W. K. & Ortega-Del Vecchyo, D. The promise of inferring the past using the ancestral recombination graph. *Genome Biol. Evol.* **16**, evae005 (2024).
33. Pearson, A. & Durbin, R. Local ancestry inference for complex population histories. Preprint at *bioRxiv* https://doi.org/10.1101/2023.03.06.529121 (2023).
34. Wang, Z. et al. Automatic inference of demographic parameters using generative adversarial networks. *Mol. Ecol. Resour.* **21**, 2689–2705 (2021).
35. Adrion, J. R. et al. A community-maintained standard library of population genetic models. *eLife* **9**, e54967 (2020).
36. Sun, H. et al. The impact of global and local Polynesian genetic ancestry on complex traits in Native Hawaiians. *PLoS Genet.* **17**, e1009273 (2021).
37. Jeon, S. et al. Genome-wide trans-ethnic meta-analysis identifies novel susceptibility loci for childhood acute lymphoblastic leukemia. *Leukemia* **36**, 865–868 (2022).
38. Myers, S., Fefferman, C. & Patterson, N. Can one learn history from the allelic spectrum? *Theor. Popul. Biol.* **73**, 342–348 (2008).
39. Kubatko, L. S., Carstens, B. C. & Knowles, L. L. STEM: species tree estimation using maximum likelihood for gene trees under coalescence. *Bioinformatics* **25**, 971–973 (2009).
40. Pickrell, J. K. & Pritchard, J. K. Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genet.* **8**, e1002967 (2012).
41. Molloy, E. K., Durvasula, A. & Sankararaman, S. Advancing admixture graph estimation via maximum likelihood network orientation. *Bioinformatics* **37**, i142–i150 (2021).
42. Nielsen, S. V. et al. Bayesian inference of admixture graphs on Native American and Arctic populations. *PLoS Genet.* **19**, e1010410 (2023).
43. Deng, Y., Nielsen, R. & Song, Y. S. Robust and accurate Bayesian inference of genome-wide genealogies for large samples. Preprint at *bioRxiv* https://doi.org/10.1101/2024.03.16.585351 (2024).

Methods

Probability of a genealogical tree under a demography

The demographic history of K populations can be represented by the interplay between two stochastic processes affecting the lineages: coalescence and movement among populations. The coalescence rate, $n_a(t)$, of each population a as a function of time t is:

$$n_a(t) = \frac{1}{kN_a(t)}, a \in \{1, \ldots, K\}, t \in (0, \infty),$$

where N_a is the effective population size and k is ploidy. The migration probability matrix, m, at each of the S historical events is:

$$m_{ab}(t_s), a, b \in \{1, \ldots, K\}, s \in \{1, \ldots, S\},$$

where t_s is the time of the s^{th} historical event and $m_{ab}(t_s)$ is the instantaneous probability for a lineage to move (backward in time) from population a to b.

The demography is thus defined as:

$$\mathcal{D} = (n, m) = (\{n_a\}, \{m_{ab}\}),$$

a size-K vector of coalescence rates defined on continuous time and a $K \times K$ matrix of migration probabilities defined on a discrete set of times. Although gLike currently does not explicitly incorporate continuous migration, it can potentially be represented as a series of historical events through discretization.

A genealogical tree with N nodes can be defined by the time and children of each node as:

$$\mathcal{G} = \{(\tau_i, \pi_i) | i \in \{1, \ldots, N\}\},$$

where τ_i is the time of the node i (or, equivalently, the emergence of lineage i) and π_i is the set of its child nodes (which is empty if i is a leaf node). The end time ω_i of lineage i can be calculated as the time of its parent node (that is, $\omega_i = \tau_j$ if $i \in \pi_j$) or ∞ if it has no parent. Our goal is to compute $\mathbb{P}(\mathcal{G}\mathcal{D})$ for arbitrary $\mathcal{G}$ and $\mathcal{D}$, and we will omit thereafter the 'conditional on $\mathcal{D}$' notation, which is always implied.

It is helpful to define the set of lineages existing at time t as:

$$L(t) = \{i | \tau_i \le t < \omega_i\}$$

and the lineages emerging between t and t' as:

$$L(t, t') = \{i | t < \tau_i, \omega_i < t'\}$$

Migration trajectory and states

The population identity of a lineage i during its existence,

$$x_i(t), t \in [\tau_i, \infty)$$

is a time-dependent variable taking values from $\{1, \ldots, K\}$ that describes how this lineage, or its ancestor lineage when $t > \omega_i$, migrates in history. For convenience, the value of $x_i(t)$ at exactly the time of a historical event is defined as the left limit $x_i(t_s) = \lim_{t \to t_s^-} x(t)$, so that $x(t)$ is left-continuous.

The population identity of all lineages existing at any time throughout history is:

$$x(t) = \{x_i(t), |, i \in L(t)\}, t \in [0, \infty),$$

which gives a complete migration trajectory of the genealogical tree. The genealogical tree itself does not dictate x, and the probability of it should be computed as the sum over all possible trajectories,

$$\mathbb{P}(\mathcal{G}) = \sum_x \mathbb{P}(\mathcal{G} \cap x).$$

To compute $\mathbb{P}(\mathcal{G})$ recursively over time, we define $\mathcal{G}(0, t)$ as the genealogical history in $\mathcal{G}$ until time t and define a 'state' as:

$$\mathcal{G}(0, t) \cap x(t).$$

For example, the state 'ABCC' in Fig. 1 at t_1 contains $\mathcal{G}(0, t_1)$, which indicates that lineages 2 and 3 coalesced at τ_1 but all other possible coalesces has not happened at t_1, and $x(t_1) =$ ABCC, which indicates that the remaining four lineages (1, 6, 4 and 5) are in populations A, B, C and C, respectively, at t_1.

Now $\mathbb{P}(\mathcal{G})$ can be expressed as the sum of the probability of root states:

$$\mathbb{P}(\mathcal{G}) = \mathbb{P}(\mathcal{G}(0, \infty)) = \sum_{x(\infty)} \mathbb{P}(\mathcal{G}(0, \infty) \cap x(\infty)).$$

Conditional probability between states

The conditional probability between states,

$$\mathbb{P}(\mathcal{G}(0, t_{s+1}) \cap x(t_{s+1}) | \mathcal{G}(0, t_s) \cap x(t_s))$$
$$= \mathbb{P}(\mathcal{G}(0, t_s) \cap x(t_{s+1}) | \mathcal{G}(0, t_s) \cap x(t_s)) \mathbb{P}(\mathcal{G}(0, t_{s+1}) \cap x(t_{s+1}) | \mathcal{G}(0, t_s) \cap x(t_{s+1})),$$

consists of a migration probability and a genealogical probability.

The migration probability,

$$\mathbb{P}(\mathcal{G}(0, t_s) \cap x(t_{s+1}) | \mathcal{G}(0, t_s) \cap x(t_s)) = \prod_{i \in L(t_s)} m_{x_i(t_s) x_i(t_{s+1})}(t_s),$$

describes the migration of each lineage i from $x_i(t_s)$ to $x_i(t_{s+1})$ at time t_s.

The genealogical probability, $\mathbb{P}(\mathcal{G}(0, t_{s+1}) \cap x(t_{s+1}) \mathcal{G}(0, t_s) \cap x(t_{s+1}))$, describes how likely the genealogical tree grows, according to $\mathcal{G}$, backward in time from t_s to t_{s+1}, given population identities $x(t_{s+1})$. This requires that every coalescence in $\mathcal{G}$ happened exactly at its time in $\mathcal{G}$ (which we call the coalescence probability) and that any other possible coalescence did not happen (which we call the non-coalescence probability).

The coalescence probability is:

$$\prod_{i \in L(t_s, t_{s+1})} [n_{x_i(t_{s+1})}(\tau_i)]^{\max(0, |\pi_i| - 1)},$$

where $n_{x_i(t_{s+1})}(\tau_i)$ is the coalescence rate of lineage i's population when it emerges. Note that the lack of migration between τ_i and t_{s+1} guarantees $x_i(\tau_i) = x_i(t_{s+1})$; $\max(0, |\pi_i| - 1)$ is the number of coalescences at the emergence of i (for example, a binary node is formed with one coalescence, a ternary node can be viewed as two coalescences at the same moment and a leaf node or unary node does not have coalescence).

The non-coalescence probability is:

$$\prod_{a \in \{1, \ldots, K\}} \exp\left(-\int_{t_s}^{t_{s+1}} \binom{l_a(t)}{2} \cdot n(t)\, dt\right),$$

where

$$l_a(t) = |\{i | i \in L(t), x_i(t_{s+1}) = a\}|$$

is the number of lineages in population a at time t (if population identities are specified by $x_i(t_{s+1})$), which is a step function that jumps when lineages emerge or coalesce; $\binom{l_a(t)}{2} = \frac{l_a(t)(l_a(t)-1)}{2}$ is the number of lineage pairs in a that are possible to coalesce; and the exponential term is the probability that none of them actually coalesced during (t_s, t_{s+1}), which is derived from a nonhomogeneous Poisson process with rate

$\lambda(t) = \binom{l_a(t)}{2} \cdot n(t)$. Note that $n(t)$ can be any integrable function, enabling flexibility to the population size variation in the demographic model.

We conclude that the conditional probability between states is:

$$\begin{aligned}
&\mathbb{P}\left(\mathcal{G}(0,t_{s+1}) \cap x(t_{s+1}) \,|\, \mathcal{G}(0,t_s) \cap x(t_s)\right) \\
&= \left(\prod_{i \in L(t_s)} m_{x_i(t_s)x_i(t_{s+1})}(t_s)\right) \cdot \left(\prod_{i \in L(t_s,t_{s+1})} \left[n_{x_i(t_{s+1})}(\tau_i)\right]^{\max(0,|\pi_i|-1)}\right) \\
&\cdot \left(\prod_{a \in \{1,\ldots,K\}} \exp\left(-\int_{t_s}^{t_{s+1}} \binom{l_a(t)}{2} \cdot n(t)\, dt\right)\right) \\
&= (\text{migration probability}) \cdot (\text{coalescence probability}) \\
&\cdot (\text{noncoalescence probability}) \\
&= (\text{migration probability}) \cdot (\text{genealogical probability})
\end{aligned}$$

Practically, the migration probability has to be computed between any parent–child state pair, but the genealogical probability is independent from the child state and needs to be calculated only once for every state. As a boundary condition, the origin state at the bottom (that is, leaves) of the tree has a probability of 1:

$$\mathbb{P}\left(\mathcal{G}(0,0) \cap x(0)\right) = \mathbb{P}\left(x(0)\right) = 1,$$

where $x(0)$ specifies the population identities of each individual in the study samples.

The minimal GOS

All possible states at all times of all historical events $t_1, t_2, \ldots, t_S$ form a directed acyclic graph, named the GOS, whereby states in adjacent layers (one at t_s and the other at t_{s+1}) are connected with their conditional probability as introduced above. A state with zero marginal probability will not contribute to the marginal probability of its parent state and is redundant in the graph. A GOS without redundant states is called a minimal GOS.

The coalescence probability and non-coalescence probability are always >0, because population sizes cannot be zero or infinity. This means that to judge whether a state is possible or not, we only have to check the migration probabilities, which are decomposable into migrations of each individual lineage. In other words, a state is possible if every lineage is in a possible population. To put it mathematically, we have:

$$\mathbb{P}\left(\mathcal{G}(0,t_s) \cap x(t_s)\right) > 0 \Longleftrightarrow \left[I(x_i(0)) \prod_{1 \le r \le s} m(t_r)\right]_{x_i(t_s)} > 0, \forall i \in L(0),$$

where $I(x_i(0))$ is a size-K indicator vector with value 1 at the population $x_i(0)$ from which sample was collected, and all other elements are zero; $\prod_{1 \le r \le s} m(t_r)$ is the transition matrix summarizing the first s historical events; and $\left[I(x_i(0)) \prod_{1 \le r \le s} m(t_r)\right]_{x_i(t_s)}$ is the probability that lineage migrated from $x_i(0)$ to $x_i(t_s)$. Step 1 in Fig. 1 can be understood as the non-zero elements in $I(x_i(0)) \prod_{1 \le r \le s} m(t_r)$ for every s.

Sampling connections between states

The computational time and memory cost for a gLike evaluation of a tree depends on the number of states and the number of connections between states or, in the language of graph theory, the number of vertices and edges of the GOS. We introduce a customizable parameter, κ, that controls the maximum number of connections between the current layer of states to the next (moving forward in time). If the number of connections is prohibitive for enumeration even in the minimal GOS, gLike creates a sampled GOS (sGOS) by randomly sampling a number of connections to approximate the complete GOS. Specifically, gLike controls the total number of connections between adjacent layers:

$$\sum_{x(t_{s+1})} \sum_{x(t_s)} \nu\left(x(t_{s+1}), x(t_s)\right) = \kappa,$$

where $\nu = 1$ if the states $x(t_{s+1})$ and $x(t_s)$ are connected in the sGOS, and $\nu = 0$ otherwise. The hyperparameter κ is intuitively the 'throughput' of the sGOS and controls the trade-off between time and performance. By default, it is set to 10,000. All connections between two adjacent layers in the original GOS are equally likely to be sampled into the sGOS. That is,

$$\begin{aligned}
&\mathbb{P}_\kappa\left(\nu\left(x(t_{s+1}), x(t_s)\right) = 1\right) = \min\left(1, \frac{\kappa}{K_s}\right), \\
&\text{if } \mathbb{P}\left(\mathcal{G}(0,t_s) \cap x(t_{s+1}) \,|\, \mathcal{G}(0,t_s) \cap x(t_s)\right) > 0,
\end{aligned}$$

where

$$K_s = \sum_{x(t_{s+1})} \sum_{x(t_s)} 1_{\mathbb{P}\left(\mathcal{G}(0,t_s) \cap x(t_{s+1}) \,|\, \mathcal{G}(0,t_s) \cap x(t_s)\right) > 0}$$

is the number of connections between t_{s+1} and t_s in the original GOS.

A set of ν values gives an instance of the sGOS, which represents a probability measure $\mathbb{Q}_\nu$ different from $\mathbb{P}$. The conditional probability in $\mathbb{Q}_\nu$ is

$$\begin{aligned}
&\mathbb{Q}_\nu\left(\mathcal{G}(0,t_{s+1})x(t_{s+1}) \,|\, \mathcal{G}(0,t_s)x(t_s)\right) \\
&= \mathbb{P}\left(\mathcal{G}(0,t_{s+1})x(t_{s+1}) \,|\, \mathcal{G}(0,t_s)x(t_s)\right) \cdot \nu\left(x(t_{s+1}), x(t_s)\right) \cdot \max\left(1, \frac{K_s}{\kappa}\right).
\end{aligned}$$

It is straightforward by induction that the equality,

$$\begin{aligned}
&\mathbb{E}_\kappa\left(\mathbb{Q}_\nu\left(\mathcal{G}(0,\infty) \cap x(\infty) \,|\, \mathcal{G}(0,t_s) \cap x(t_s)\right)\right) \\
&= \mathbb{P}\left(\left(\mathcal{G}(0,\infty) \cap x(\infty) \,|\, \mathcal{G}(0,t_s) \cap x(t_s)\right)\right),
\end{aligned}$$

holds for any state. Applying this to the origin state yields

$$\begin{aligned}
&\mathbb{E}_\kappa\left(\mathbb{Q}_\nu\left(\mathcal{G}(0,\infty)\right)\right) = \mathbb{E}_\kappa\left(\mathbb{Q}_\nu\left(\mathcal{G}(0,\infty)x(\infty) | \mathcal{G}(0,0)x(0)\right)\right) \\
&= \mathbb{P}\left(\mathcal{G}(0,\infty)x(\infty) | \mathcal{G}(0,0)x(0)\right) = \mathbb{P}\left(\mathcal{G}(0,\infty)\right),
\end{aligned}$$

which means $\mathbb{Q}_\nu\left(\mathcal{G}(0,\infty)\right)$ is an unbiased estimator of $\mathcal{G}(0,\infty)$.

In practice, K_s can be quickly determined using the migration matrix $m(t_s)$. If $K_s < \kappa$, no approximation is conducted; if $K_s > \kappa$, all states between t_{s+1} and t_s are sampled without replacement with a probability κ/K_s. The migration probability is multiplied by K_s/κ to keep the unbiasedness of the sGOS.

Implementation details and optimization

With the above-mentioned theory to calculate $\mathbb{P}\left(\mathcal{G} \,|\, \mathcal{D}_\theta\right)$ on a demographic model $\mathcal{D}_\theta$ parameterized by θ, the estimated parameter that best explains the observed $\mathcal{G}$ is

$$\theta^* = \underset{\theta}{\operatorname{argmax}}\, \mathbb{P}\left(\mathcal{G} | \mathcal{D}_\theta\right).$$

gLike encapsulates the likelihood computation and a simulated annealing-based optimization into an open-source Python package, alongside a C extension to accelerate Cartesian product operations when searching for child states (https://github.com/Ephraim-usc/glike). All probabilities are implemented in log scale, and sums of probabilities are calculated with the scipy logsumexp function, which are computationally relatively inexpensive (Extended Data Fig. 4). When multiple, presumed independent and neutrally evolving trees are provided, the final log likelihood is the sum of the log likelihoods of each

tree. We presume independence of trees, as the total likelihood would assume more complicated forms if trees were nearby and not independent. We also presume neutrality, as coalescence probabilities would deviate from the inverse of population sizes when there are variants under natural selection. We set a user-defined parameter to drop some proportion (default, 50%) of the lowest likelihood trees during optimization, as we found in practice that this filtering improves robustness against errors in tree reconstruction (such as erroneous coalescences) and migrations that are neglected in the demographic model.

A current limitation of gLike, which is a common problem in many demographic inference methods, is that certain parameters are not individually identifiable. These entangled parameters could only be optimized in combination if multiple combinations of the two parameters produce the same average coalescence rate. An example is that the effects of population size and growth rate are hard to separate if a population exists for only a short time (Extended Data Fig. 6). When applying gLike with simulated annealing-based optimization, the estimates of entangled parameters could be path-dependent. Therefore, a grid search on specific entangled parameters after a general optimization routine may be beneficial to an unbiased estimation of the demography.

Demographic inference in simulations

All simulations were performed on a 30 Mb chromosome with both recombination and mutation rates set to 10^{-8} per generation per base pair, with a sample size of 1,000 haplotypes from the admixed population. The demographic parameters are annotated in the corresponding figures or cloned from stdpopsim[35] models 4B11 (American Admixture) and 4A21 (Ancient Europe). In American Admixture simulations, we ignored the continuous migrations in our simulations and estimations. The extent to which hidden migrations potentially undermine gLike results was tested on additional simulations with 1×, 10× and 100× continuous migrations as reported by stdpopsim 4B11. In the Ancient Europe simulation, we additionally sampled 200 haplotypes from each ancestral population according to the collection times reported by stdpopsim, to mimic genetic studies with ancient DNA.

To evaluate gLike, ARGs and genotypes were simulated by msprime[44]. ARG reconstructions by tsinfer + tsdate (tsinfer v.0.3.0 and tsdate v.0.1.4)[45,46] or Relate (v.1.1.6)[47] were performed with all default parameters (including an effective population size (N_e) of 10,000) as suggested in the user manual. We observed that the inferred distribution of coalescences from tsinfer + tsdate is more robust to the choice of N_e than that from Relate, particularly for the relatively recent time period (~10–100 generations ago) when admixture events tend to occur (Extended Data Fig. 2). As such, we used $N_e = 10{,}000$ for all of our evaluations. In total, 100 evenly spaced trees across the chromosome were selected for gLike inference. The precision of the gLike parameter estimation (that is, the minimal step size during optimization by simulated annealing relative to the current estimate) was set to 2%. The parameters are initially set to uninformative values (for example, all 10,000 for population sizes and all 0.01 for growth rates) to avoid bias. The exact initial values and boundary conditions can be found in Supplementary Table 6. The absolute difference between the average estimate and the truth, divided by truth, is defined as the relative error. The average estimates across 50 (or 20 for analyses presented in supplementary figures) replicate simulations were used as the final pictorial representation of the reconstructed demography, with boxplots of the relative errors across 50 or 20 replicates also shown. All boxplots display the first, second (the median) and third quartiles of the data, with whiskers extending from the box to the farthest data point lying within 1.5× the inter-quartile range.

We find that in our application with gLike for the demographies we have studied, analyses using tsinfer + tsdate-estimated genealogical trees produced more accurately estimated demographies than using trees estimated by Relate. The difference in performance may trace to the fact that Relate does not accurately reproduce the coalescence distributions during the period between ~10–100 generations ago when admixture events happened (Extended Data Fig. 3b,c), thereby leading to mis-estimations in the gLike framework (even when using multiple trees sampled from the posterior of Relate; Supplementary Fig. 11). As a result, gLike on Relate-reconstructed trees was not tested further in this study. Notably, Relate may outperform tsdate in other applications using the genealogical trees, such as inferring the genome-wide expected relationship matrix[26], suggesting that current methods have respective strengths in capturing different aspects of the true ARGs.

To compare gLike to Fastsimcoal2 (v.2.8.0.0)[21], derived allele frequency spectra of 1,000 haplotypes (the same sample size as used when evaluating gLike) were computed on all simulated single nucleotide polymorphisms (SNPs) (including singletons), and parameter estimation was performed with 100,000 simulations and 40 ECM (expectation / conditional-maximization) loops, using the commands '-n 1 -s0 -d -k 1000000' for SFS simulation and '-n 100000 -d -M -L 40' for parameter estimation. Following the recommendation of the author for Fastsimcoal2, for a simulated dataset of a given demographic history, we performed 20 replicates of estimation (each initiated with an independent random seed). The single replicate with the highest likelihood was then taken as the inference result. This process is then repeated over 50 simulated datasets. Boxplots were made based on the best estimations for each of the 50 datasets. When multiple populations are present (for example, Fig. 5 and Supplementary Fig. 8), using the multidimensional SFS command (-n 100000 -d -M -L 40 -q –multiSFS -c12 -B12) showed an improvement over using pairwise two-dimensional SFS; however, because of computational limitation set by Fastsimcoal2 for multidimensional SFS, the sample sizes were proportionally reduced (in Fig. 5, 25 haplotypes from Bronze and five haplotypes from each ancestral populations; in Supplementary Fig. 8, 60 haplotypdes from ADMIX and 30 haplotypes from each ancestral population).

We also compared gLike performance to pg-gan[34] (v.9/27/22), a deep-learning demographic parameter inference method that uses generative adversarial networks to create realistic simulated training data. Genotypes from simulated ARGs of the same demographic model were used as training data and run for up to 300 training iterations with default training parameters. We used the same range for each demographic parameter to be consistent with the Fastsimcoal2 comparisons. As pg-gan gives multiple sets of parameter proposals at the end of training, the set of inferred demographic parameters with the lowest relative error compared to the true parameters was selected as the final estimate of this run. A total of 50 independent runs were conducted.

To characterize the impact of ARG reconstruction using array data instead of sequencing data, we performed an additional simulation experiment in which SNPs were retained with the probability

$$p(\mathrm{MAF}) = C_{\mathrm{ref}}(\mathrm{MAF})/C_{\mathrm{sim}}(\mathrm{MAF})$$

where MAF is the minor allele frequency of the simulated SNP, C_{ref} (MAF) is the number of occurrences of MAF in the Latinos array data and C_{sim} is the number of occurrences of MAF in a simulated genome (3,000 Mb). As expected, it was found that C_{sim} is greater than C_{ref} across all values of $\mathrm{MAF} \in [0, 0.5]$, which ensures that p is always less than one. We then inferred the ARG using tsinfer + tsdate using the simulated array data.

Model selection in simulations

To test for the existence of an additional ancestral component, gLike was applied under a two-way admixture model and a three-way admixture model, and the maximum likelihoods achieved under both models were compared. Specifically, the two-way admixture model structurally mimicked the three-way admixture as in Fig. 2a, but without population D, so that all lineages from population B entered population C. As such, the two-way admixture model had two fewer parameters: r_2

(admixture proportion from D) and N_D (population size of D). gLike was then applied in a two-step manner. First, the parameters were estimated under the two-way admixture model with the default hill-climbing optimization. Next, we applied gLike under the three-way admixture model and performed a grid search on r_2, N_C and N_D, while fixing other parameters at their two-way admixture estimates. Finally, the difference between the maximum log likelihoods achieved under two models was used for AIC model selection (with two degrees of freedom to account for the two extra parameters in the three-way admixture model), and the model with the higher AIC value was selected.

Latino and Native Hawaiian data processing

A total of 500 individuals were randomly chosen from each of the self-identified Native Hawaiians (from up to 5,382 individuals) and Latinos (from up to 3,659 individuals) subcohorts from the Multiethnic Cohort for empirical analysis using gLike. Written informed consent was obtained from all participants, and study protocols were reviewed and approved by the Institutional Review Boards at the University of Hawaii and the University of Southern California.

The two cohorts were genotyped on two separate genome-wide association study arrays: Illumina MEGA and Illumina Global Diversity Array. After taking the intersection of SNPs found on both arrays, the genotyping data were lifted to hg38 using triple-liftover[48] to ensure that alleles in inverted sequences between reference genome builds were properly lifted. We removed variants that were genotyped in fewer than 95% of individuals, variants out of Hardy–Weinberg equilibrium ($P < 10^{-6}$) and individuals with greater than 2% missing genotypes (although no one was removed with this threshold). After quality check, the Native Hawaiian and Latino datasets contained 990,549 and 1,093,693 SNPs, respectively. The data were phased without a reference using EAGLE (v.2.4.1)[49] and its default hg38 genetic map. We randomly subsampled 1,000 haploids and removed monomorphic SNPs, resulting in 879,040 and 927,254 SNPs in the Native Hawaiian and Latino datasets, respectively. The ancestral alleles were called by a comparison with the human ancestor GRCh38 e107 genome (ftp.ensembl.org/pub/release-86/fasta/ancestral_alleles). Tsinfer and tsdate were used with all default parameters as suggested in the user manual to reconstruct the ARG. The human neutralome[50] (that is, the regions of the human genome identified as probably selectively neutral) was converted into hg38 coordinates, and 319 neutral regions that are at least 5 Mb from each other were selected for gLike analysis. Ten trees were sampled in each gLike optimization thread, and 20 threads were run in parallel. The estimates of demographic parameters were averaged over 20 threads. The standard deviation across 20 threads serves as an indicator of the parameter uncertainties, as listed in Supplementary Tables 1 and 4. The precision of gLike parameter estimation was set to 5%, higher than 2% used in simulations. This choice is because of the broader span of the likelihood curve's plateau, which generally extends past 5%, wider than observed in simulations. Therefore, using smaller step sizes would increase computational costs with little gain in performance.

To compare the estimated ancestry proportion from gLike with that obtained from alternative approach, we also performed supervised ADMIXTURE (v1.3.0). AFR ($n = 678$), American (AMR; $n = 82$), East Asian (EAS; $n = 751$) and Non-Finnish European (NFE; $n = 648$) individuals as defined by gnomAD (v3.1.2) were used as references for African, Indigenous American, East Asian and European ancestries. AFR, EAS and NFE were used as references for both Latinos and Native Hawaiians. An additional 114 Native Hawaiian individuals from the Multiethnic Cohort previously estimated to have >90% Polynesian ancestry[36,51] were also added as Polynesian reference samples for Native Hawaiians, while the AMR individuals from gnomAD were used for Latinos. A total of 323,697 and 328,112 SNPs for Latinos and Native Hawaiians, respectively, were used for analysis after intersecting variants between the references and target cohorts and LD pruning (window size of 50 SNPs, shifting in ten SNP steps, and r2 threshold of 0.1) using PLINK (v.1.90b6.20). ADMIXTURE was run in supervised mode with $K = 4$ for both analyses.

Statistics and reproducibility

Choice for sample size and genomic region sizes in simulation were made to accommodate computational scale and feasibility while allowing rigorous insight to the evaluation of our method. For empirical analysis, we selected randomly 500 individuals from each of the Native Hawaiian and Latino subcohorts of the Multiethnic Cohort. No statistical method was used to predetermine sample size; this sample size was chosen to ensure the efficient demonstration of gLike's capabilities and because we found that sample sizes greater than 100 appear to be sufficient to provide accurate demographic inference using gLike (Supplementary Fig. 3). All samples passing quality controls were available for random selection. The researchers had no access to any information associated with the individual other than self-reported ethnicity for the purpose of forming analysis units and random subsets. Codes used for simulation and plotting can be found on Zenodo (https://doi.org/10.5281/zenodo.14708630)[52].

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The individual-level genetic data for Native Hawaiian and Latino datasets were derived from the Multiethnic Cohort and are available on dbGaP (accession numbers phs000220.v2.p2 and phs002183.v1.p1).

Code availability

The gLike package is available on its GitHub page (https://github.com/Ephraim-usc/glike). The version of gLike as well as codes used for simulation and plotting presented in this study can also be found on Zenodo (https://doi.org/10.5281/zenodo.14708630)[52].

References

44. Kelleher, J., Etheridge, A. M. & McVean, G. Efficient coalescent simulation and genealogical analysis for large sample sizes. *PLoS Comput. Biol.* **12**, e1004842 (2016).
45. Kelleher, J. et al. Inferring whole-genome histories in large population datasets. *Nat. Genet.* **51**, 1330–1338 (2019).
46. Wohns, A. W. et al. A unified genealogy of modern and ancient genomes. *Science* https://doi.org/10.1126/science.abi8264 (2022).
47. Speidel, L., Forest, M., Shi, S. & Myers, S. R. A method for genome-wide genealogy estimation for thousands of samples. *Nat. Genet.* **51**, 1321–1329 (2019).
48. Sheng, X. et al. Inverted genomic regions between reference genome builds in humans impact imputation accuracy and decrease the power of association testing. *HGG Adv.* **4**, 100159 (2023).
49. Loh, P.-R. et al. Reference-based phasing using the Haplotype Reference Consortium panel. *Nat. Genet.* **48**, 1443–1448 (2016).
50. Woerner, A. E., Veeramah, K. R., Watkins, J. C. & Hammer, M. F. The role of phylogenetically conserved elements in shaping patterns of human genomic diversity. *Mol. Biol. Evol.* **35**, 2284–2295 (2018).
51. Lin, M. et al. Population-specific reference panels are crucial for genetic analyses: an example of the CREBRF locus in Native Hawaiians. *Hum. Mol. Genet.* **29**, 2275–2284 (2020).
52. Fan, C. Ephraim-usc/glike: v1.0. *Zenodo* https://doi.org/10.5281/zenodo.14708630 (2025).

Acknowledgements

We would like to thank I. Mathieson, S. Mathieson and L. Speidel for discussions and advice. Research reported in this publication was supported by National Institute of Health under award number R35GM142783 and R01HG12605 to C.W.K.C., R35GM137758 to M.D.E., R01HG012133 and P01CA196569 to N.M. and by Programa de Apoyo

a Proyectos de Investigación e Innovación Tecnológica–Universidad Nacional Autónoma de México (PAPIIT–UNAM) under award number IN215524 to D.O.-D.V. The funders had no role in the study design, data collection and analysis, decision to publish or preparation of the manuscript. Computation for this work was supported by the University of Southern California's Center for Advanced Research Computing (https://carc.usc.edu).

Author contributions

C.W.K.C., D.O.-D.V. and C.D.H. conceived of the study. C.F. and C.W.K.C. designed the study. C.F., J.L.C. and B.L.D. performed the analysis. B.L.D. curated the data. C.F., M.D.E., N.A.M. and C.W.K.C. interpreted the data. C.F., J.L.C., M.D.E. and C.W.K.C. wrote the manuscript with input from all co-authors.

Competing interests

The authors declare no competing interests.

Additional information

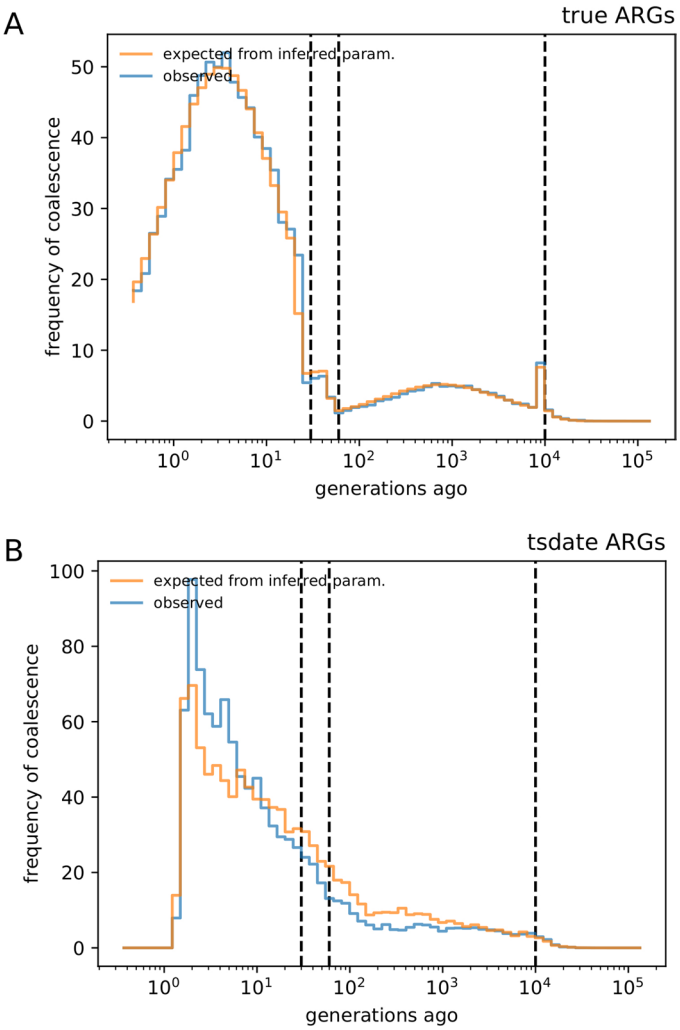
Extended data is available for this paper at https://doi.org/10.1038/s41588-025-02129-x.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41588-025-02129-x.

Correspondence and requests for materials should be addressed to Caoqi Fan or Charleston W. K. Chiang.

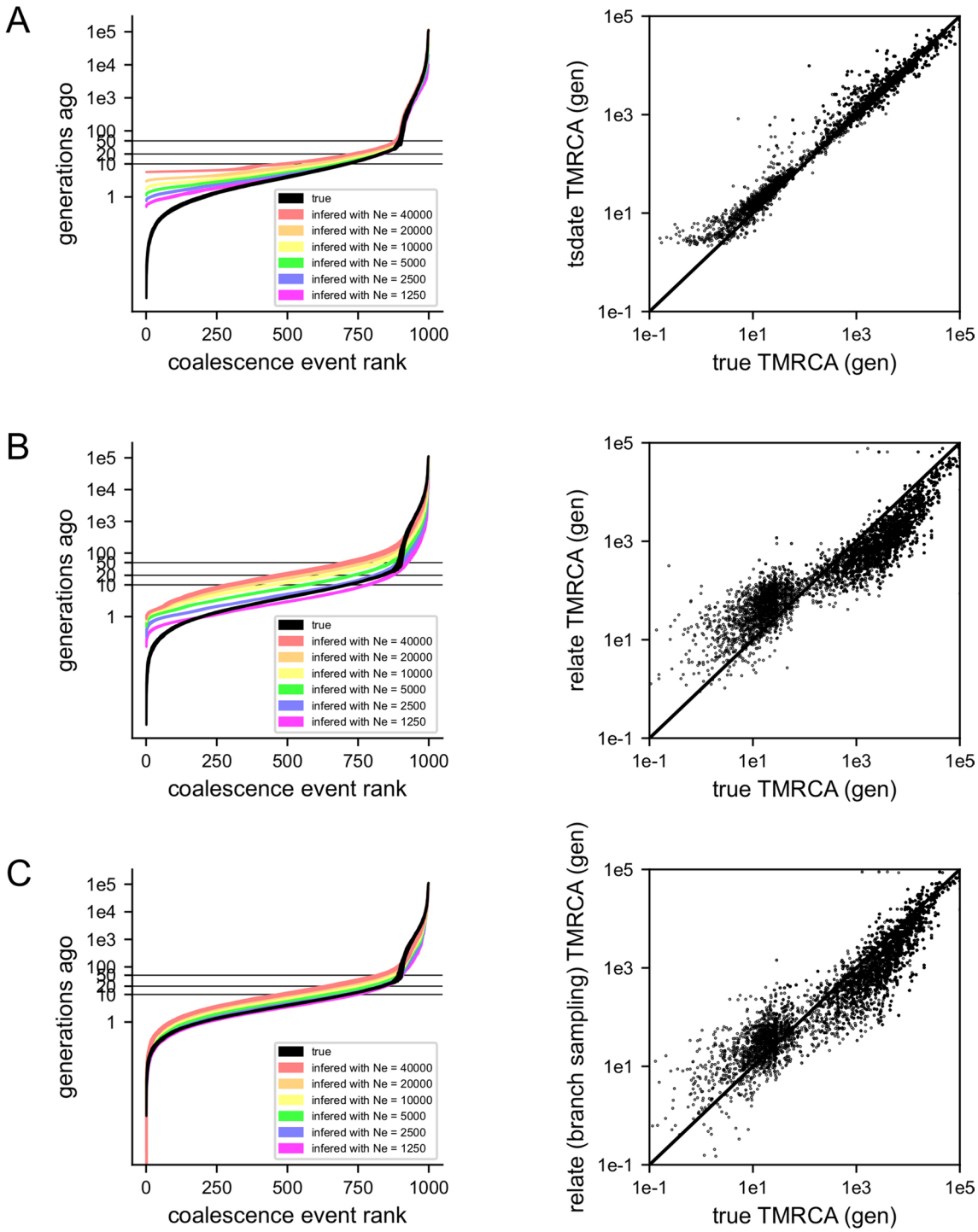
Peer review information *Nature Genetics* thanks Laurent Excoffier, Harald Ringbauer and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

https://doi.org/10.1038/s41588-025-02129-x



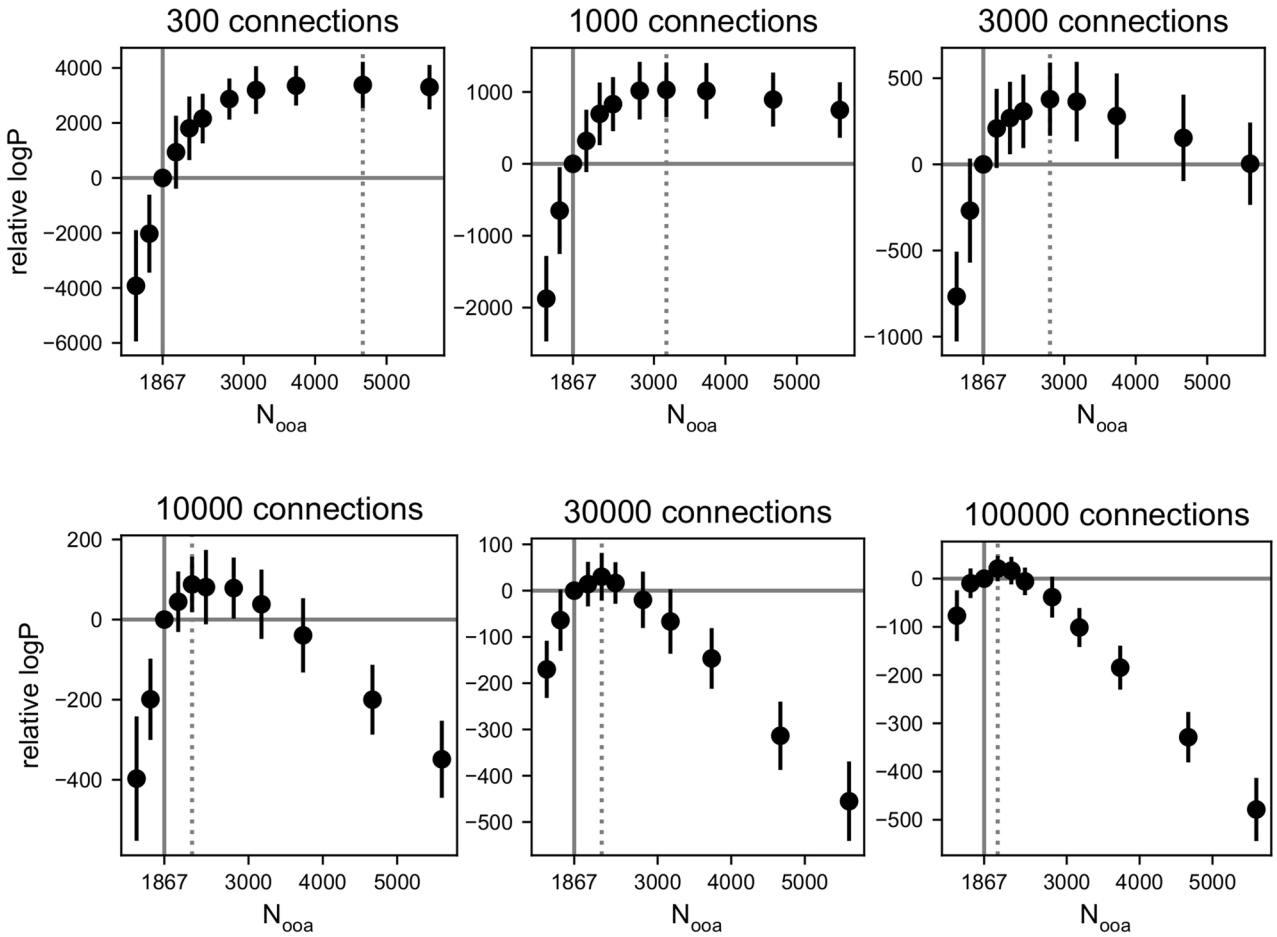
Extended Data Fig. 1 | The expected coalescence distribution based on the inferred demography matches the simulated input. (**A**) We simulated 100 equal-distant trees of 1000 haplotypes were simulated on a 30 Mb chromosome, under the same demography as in Fig. 2a. The demography is inferred by gLike on the true trees with default settings, and the expected coalescence distribution is computed by simulation of 10,000 trees under the inferred demography. The two distributions are highly consistent, except for small random fluctuations on the observed distribution. (**B**) The same experiment as in (**A**), but tsdate reconstruction is applied to the observed trees, the parameters are then inferred by gLike on the reconstructed trees, and tsdate reconstruction is again applied to the simulated trees under the inferred parameters. Vertical dash lines indicate t_1, t_2, and t_3 in the simulated demography, corresponding to the time of the more recent admixture event, the more distant admixture event, and the split of three ancestral populations, respectively.

https://doi.org/10.1038/s41588-025-02129-x



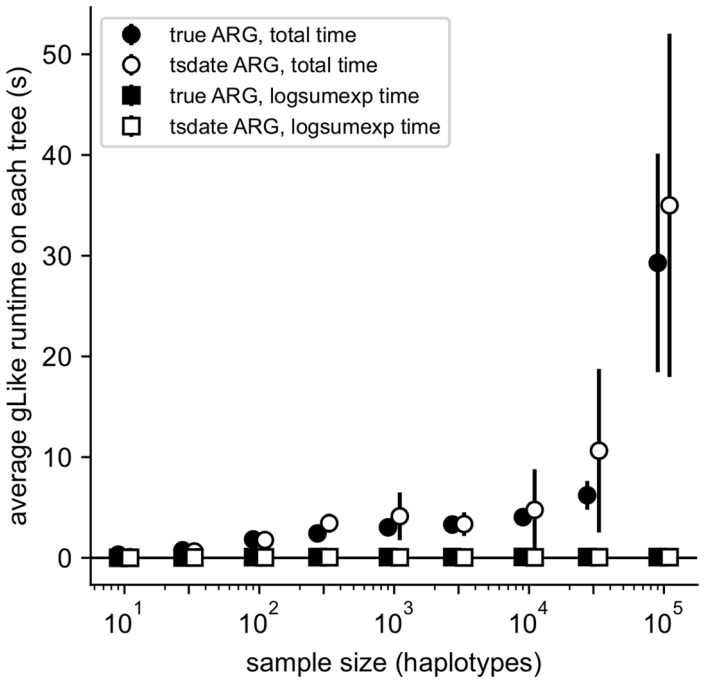
Extended Data Fig. 2 | The inferred distribution of coalescence under the three-way admixture demography as function of input parameter Ne during ARG inference. For ARG inference based on (**A**) tsdate, (**B**) Relate, and (**C**) Relate with branch sampling, the left panels show the times of coalescences (that is, inner nodes) in ascending order in a genealogical tree of 1000 haplotypes simulated under the three-way admixture demography as in Fig. 2a. Different color bands show 2 times standard deviation across 50 independent simulations. Right panels show TMRCA in the true tree versus the reconstructed tree, using Ne = 10,000, which we use as default for all ARG inference in this study. Results from 50 independent simulations are pooled for display.

https://doi.org/10.1038/s41588-025-02129-x



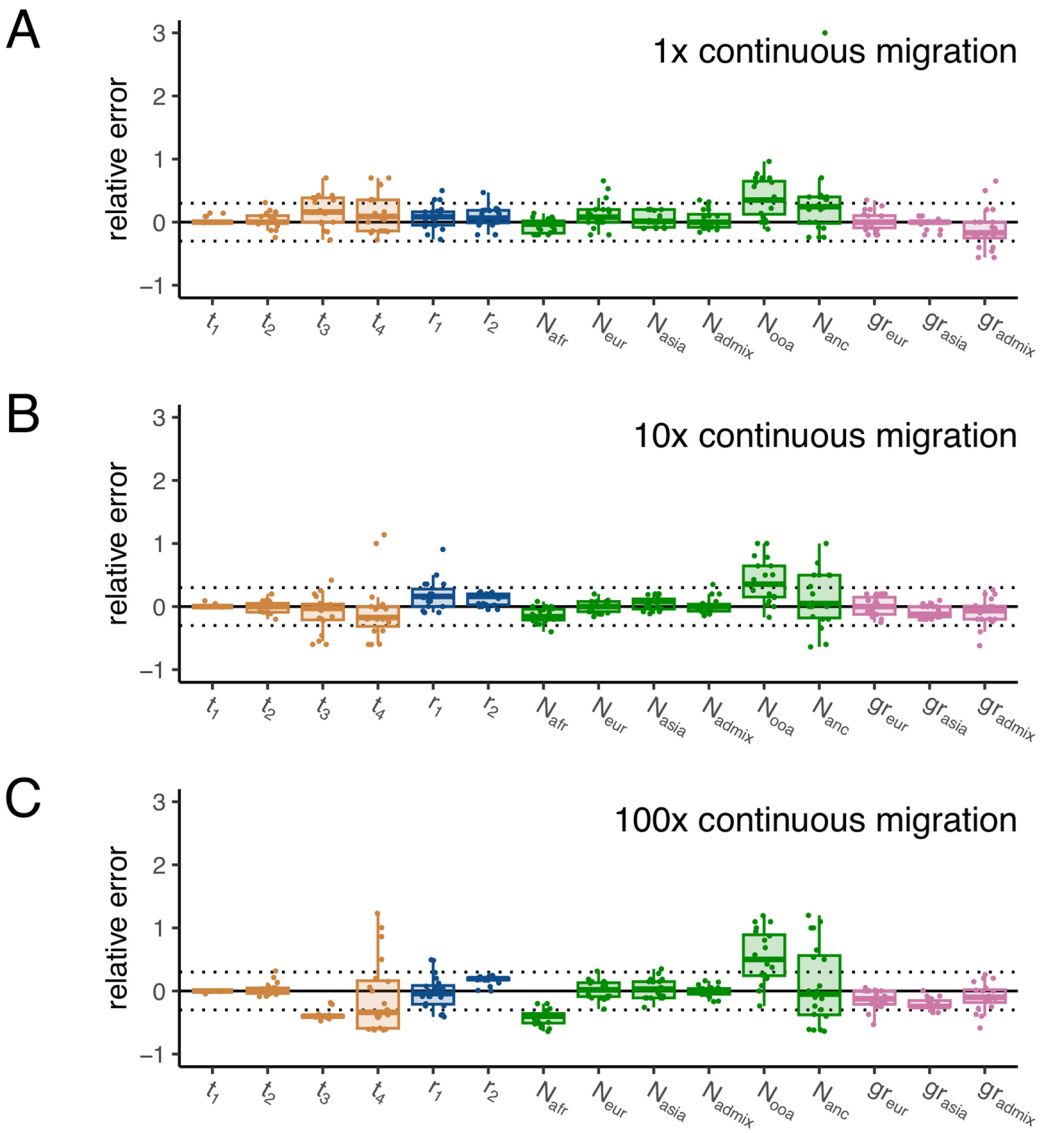
Extended Data Fig. 3 | Log likelihood distribution around N_{ooa} values for different thresholds of maximum number of edges connecting all states between two time points. For computational efficiency, if the total connections between two adjacent layers of the GOS exceeds a customizable hyperparameter, κ, gLike will approximate via sampleing (see **Methods** for details). Here we evaluate the impact of setting this threshold, κ, on the apparent biased estimate of N_{ooa} parameter in Fig. 4. A total of 50 replicate experiments were conducted in each panel. Solid circles and error bars indicate mean and standard deviation, respectively, across the replicates. In each replicate experiment, 100 equally distant trees of 1000 haplotypes were simulated on a 30 Mb chromosome from population ADMIX under the same demography as in Fig. 4. The log-likelihood (logP) of observing these 100 trees were calculated by gLike assuming different N_{ooa} values and all other demographic parameters fixed at true values. The logP calculated from the true N_{ooa} = 1867 were subtracted from all logP values, for comparability between replicates. As we increased the default threshold for connections before gLike begin approximating the likelihood, the maximum likelihood estimate (dashed line) also tended towards the true value (solid line), suggesting that the exact computation of likelihood is unbiased, though approximation for computational reasons could lead to bias.

https://doi.org/10.1038/s41588-025-02129-x



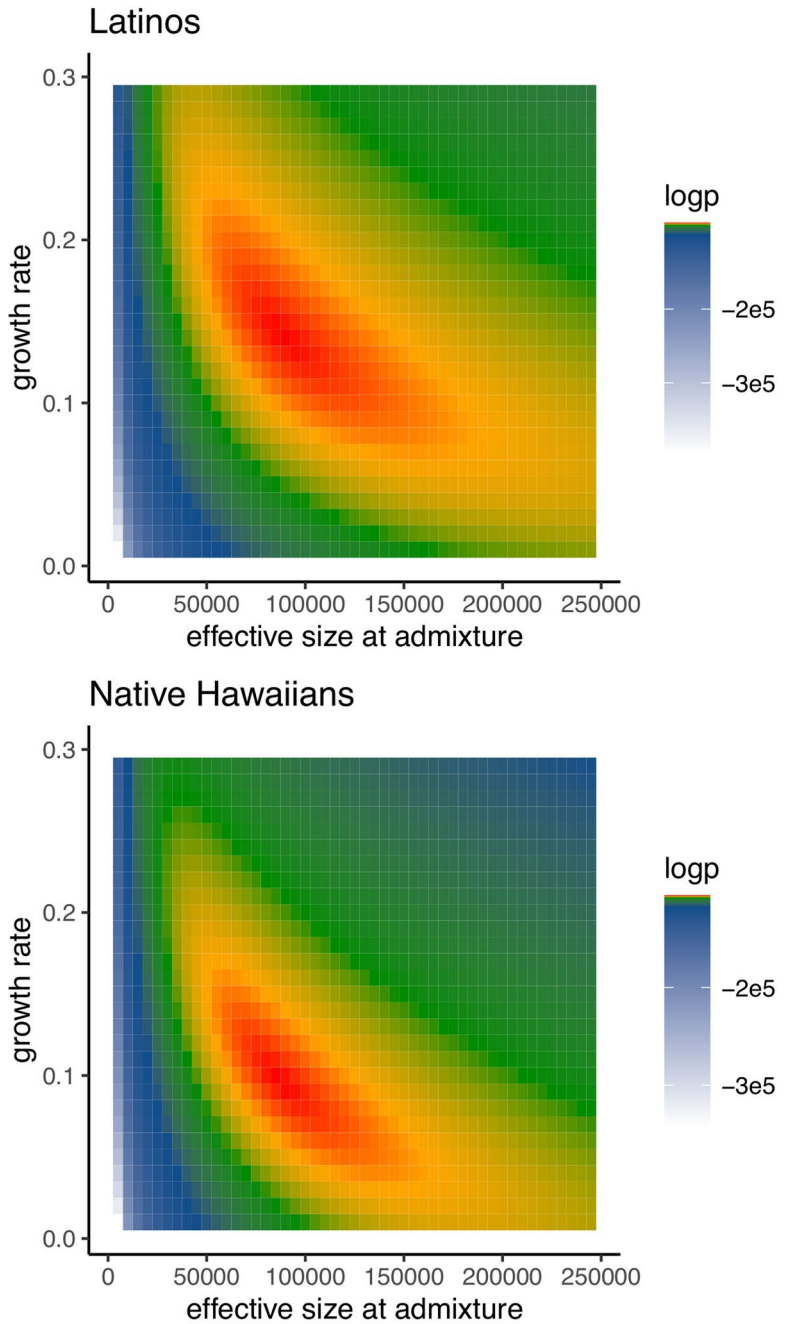
Extended Data Fig. 4 | Average gLike runtime on a single genealogical tree with varying sample sizes. 50 replicate experiments were conducted for each sample size. Solid circles indicate the average runtime on each tree, and squares indicate the average time spent on scipy logsumexp function for each tree. Error bars indicate the standard deviation across 50 replicates. In each replicate experiment, 100 equally distant trees of 1000 haplotypes were simulated on a 30 Mb chromosome under the same three-way admixture demography as in Fig. 2.

https://doi.org/10.1038/s41588-025-02129-x



Extended Data Fig. 5 | Robustness of gLike against misspecified continuous migrations. The same experiment in Fig. 4a except that the true demography contains AFR-EUR, AFR-ASIA, EUR-ASIA and AFR-OOA continuous migrations that are set to be 1x (**A**), 10x (**B**) and 100x (**C**) of their rates as in the stdpopsim 4B11 model. gLike was applied on the true trees in the same way as in Fig. 4a, assuming no continuous migrations. Note that the 1x continuous migrations have no visible impact on the results, while 100x continuous migrations lead to considerable underestimations of t_3, t_4 and N_{afr}, due to the accumulation of coalescences earlier than expected in a migration-free demography. Boxplots display the first, second (the median), and third quartiles of the data, with whiskers extending from the box to the farthest data point lying within 1.5x of the inter-quartile range.

https://doi.org/10.1038/s41588-025-02129-x



Extended Data Fig. 6 | Unidentifiability between population sizes and growth rates. The log-likelihood of the gLike model on the population sizes (at time of admixture) and growth rates of the Latinos and Native Hawaiians in a grid of possible parameters. All other parameters were fixed at their estimates shown in Fig. 6. This result indicates the potential bias when estimating entangled parameters, because the hill-climbing optimization could stop anywhere along the red curve, depending on the initial values.

nature portfolio

Corresponding author(s): FAN, CHIANG

Last updated by author(s): Jan 22, 2025

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐ A description of all covariates tested
☒	☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No data were collected
Data analysis	Analysis method and code are provided in a GitHub repository, https://github.com/Ephraim-usc/glike Minted version of gLike version 1.0 and simulation codes used in this study can be found at: https://www.doi.org/10.5281/zenodo.14708630 Other software versions: python: version 3.8.3 GCC: version 7.3.0 tsinfer: version 0.3.0 tsdate: version 0.1.4 relate: version 1.1.6 Fastsimcoal2: version 2.8.0.0, pg-gan: latest version as of Jun 28, 2023 ADMIXTURE: version 1.3.0 PLINK: version 1.90b6.20

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The individual level genetic data for Native Hawaiian and Latino datasets were derived from the Multiethnic Cohort (MEC), and are available on dbGaP (accession numbers: phs000220.v2.p2 and phs002183.v1.p1).

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	Not relevant
Reporting on race, ethnicity, or other socially relevant groupings	Following the convention established by the Multiethnic Cohort (MEC), from which we retrieved the data on Latinos and Native Hawaiians for analysis, labels and groupings of these individuals are based on self-report at baseline using terminology that was in practice in the early 1990s when the surveys for the Multiethnic Cohort were distributed to participants. In the survey, participants were asked to provide self-reported ethnic or racial background, marking all that applies, with the options of “Black or African-American”, “Chinese”, “Filipino”, “Hawaiian”, “Japanese (includes Okinawan)”, “Korean”, “Mexican or other Hispanic”, “White or Caucasian”, or “Other”. Most MEC analyses categorize participants into one of the five major racial/ethnic groups mentioned above that were targeted during recruitment, prioritizing group memberships in the following order: “African-American”, “Hawaiian”, “Latino”, “Japanese”, “White”. Thus, if an individual reports “Chinese” and “Hawaiian” in the survey, they would be classified as Native Hawaiian for these analyses.
Population characteristics	Not relevant
Recruitment	Not relevant
Ethics oversight	This study performed only secondary analysis on existing data; study is approved by the IRB at University of Southern California.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	500 randomly selected self-reported Latino and Native Hawaiians from the Multiethnic Cohort were included in the study to model the admixture history of these populations. No sample size calculation was done. 500 individuals were randomly chosen to efficiently demonstrate gLike's capability, as we have observed that sample sizes greater than 100 appears to be sufficient for accurate demographic inference.
Data exclusions	Individuals who were not randomly selected were excluded.
Replication	To validate the inferred demographic parameters empirically, we compared the genome-wide admixture proportion estimates from gLike with that based on a separate software, ADMIXTURE. This replication was conducted using the exact same set of 500 randomly selected individuals. gLike inference was without references, while ADMIXTURE was performed in supervised mode with references of ancestral populations. The results were broadly consistent.
Randomization	The selection of the subset of 500 individuals from each of Latinos and Native Hawaiian cohorts were made randomly.
Blinding	The researcher had no access to any outcome information associated with the individuals being studied, other than the self-reported ethnicity used for defining analysis population and random subsets.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Plants

Seed stocks

Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.

Novel plant genotypes

Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.

Authentication

Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosiacism, off-target gene editing) were examined.