Molecular Biology and Evolution, 2025, **42**, 1–9
https://doi.org/10.1093/molbev/msaf257
Advance access publication 5 November 2025
Methods

MBE OXFORD

GHIST 2024: The First Genomic History Inference Strategies Tournament

Travis J. Struck [1,†], **Andrew H. Vaughn** [2,3,†], **Austin Daigle** [4,5], **Dylan D. Ray** [5], **Ekaterina Noskova** [6,7], **Jaison J. Sequeira** [8], **Svetlana Antonets** [9], **Elizaveta Alekseevskaya** [10], **Elizaveta Grigoreva** [11], **Evgenii Raines** [12,13], **Eilish S. McMaster** [14], **Toby G.L. Kovacs** [14], **Aaron P. Ragsdale** [15], **Andrés Moreno-Estrada** [16], **Katie E. Lotterhos** [17], **Adam Siepel** [18], **Ryan N. Gutenkunst** [1,*]

[1]Department of Molecular and Cellular Biology, University of Arizona, Tucson, AZ, USA
[2]Center for Computational Biology, University of California, Berkeley, CA, USA
[3]Bakar Computational Health Sciences Institute, University of California, San Francisco, CA, USA
[4]Curriculum in Bioinformatics and Computational Biology, University of North Carolina, Chapel Hill, NC, USA
[5]Department of Genetics, University of North Carolina, Chapel Hill, NC, USA
[6]Department of Biology, University of Fribourg, Fribourg, Switzerland
[7]Institute of Integrative Biology, ETH Zürich, Zürich, Switzerland
[8]Department of Applied Zoology, Mangalore University, Mangaluru, Karnataka, India
[9]Department of Bioinformatic Data Processing, Genotek Ltd., Moscow, Russia
[10]Mechanisms of Cellular Senescence Laboratory, Institute of Cytology of the Russian Academy of Sciences, St. Petersburg, Russia
[11]Gregor Mendel Institute of Molecular Plant Biology, Austrian Academy of Sciences, Vienna, Austria
[12]Institute of Applied Computer Science, ITMO University, St. Petersburg, Russia
[13]Department of Systems Immunology, Weizmann Institute of Science, Rehovot, Israel
[14]School of Life and Environmental Sciences, University of Sydney, Camperdown, Australia
[15]Department of Integrative Biology, University of Wisconsin-Madison, Madison, WI, USA
[16]Aging Research Center, Cinvestav Sede Sur, Center for Research and Advanced Studies of the National Polytechnic Institute, Mexico City, Mexico
[17]Department of Marine and Environmental Sciences, Northeastern University Marine Science Center, Nahant, Massachusetts, USA
[18]Simons Center for Quantitative Biology, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, USA

[†]These authors contributed equally to this work.
***Corresponding author**:E-mail:rgutenk@arizona.edu.
Associate editor: Kelley Harris

Abstract

Evaluating population genetic inference methods is challenging due to the complexity of evolutionary histories, potential model misspecification, and unconscious biases in self-assessment. The Genomic History Inference Strategies Tournament (GHIST) is a community-driven competition designed to evaluate methods for inferring evolutionary history from population genomic data. The inaugural Genomic History Inference Strategies Tournament competition ran from July to November 2024 and featured four demographic history inference challenges of varying complexity: a bottleneck model, a split with isolation model, a secondary contact model with demographic complexity, and an archaic admixture model. Data were provided as error-free VCF files, and participants submitted numerical parameter estimates that were scored by relative root-mean-squared error. Approximately 60 participants competed, using diverse approaches. Results revealed the current dominance of methods based on site frequency spectra, while highlighting the advantages of flexible model-building approaches for complex demographic histories. We discuss insights regarding the competition and outline the next iteration, which is ongoing with expanded challenge diversity. By providing standardized benchmarks and highlighting areas for improvement, Genomic History Inference Strategies Tournament represents a substantial step toward more reliable inference of evolutionary history from genomic data.

Keywords: population genomics, demographic history, competition

Population genetic inference aims to reconstruct the recent evolutionary history of populations from genomic variation data. This field has seen explosive growth, driven by the increasing availability of whole-genome sequencing data from diverse groups of humans and other species (Pool et al. 2010). But population genetic inference is inherently challenging. First, the stochasticity of the evolutionary process means that the same history can produce different genetic patterns. Second, different histories can produce similar patterns of genetic variation, creating an identifiability problem (Myers et al. 2008; Lapierre et al. 2017; Lawson et al. 2018; Rosen et al. 2018). Third, real populations rarely conform to the simplified models typically used for inference, leading to potential biases when models are misspecified (Loog 2021; Momigliano et al. 2021). Finally, computational constraints often necessitate approximations that may impact accuracy.

Received: August 11, 2025. **Revised:** October 1, 2025. **Accepted:** October 2, 2025

Many methods for population genetic inference exist. For example, site frequency spectrum (SFS) methods examine the distribution of allele frequencies within and among populations (Marth et al. 2004; Gutenkunst et al. 2009; Excoffier et al. 2021). Linkage-based approaches analyze patterns of linkage disequilibrium or identity-by-descent (IBD) segments (Harris and Nielsen 2013; Browning and Browning 2015). Markovian coalescent methods reconstruct recent genealogical relationships among samples (Li and Durbin 2011; Schiffels and Durbin 2014), while ancestral recombination graph (ARG) methods explicitly reconstruct the genealogical history including recombination events (Rasmussen et al. 2014; Kelleher et al. 2019; Speidel et al. 2019). More recently, machine learning approaches apply supervised learning to haplotype matrices or summary statistics (Schrider and Kern 2018; Flagel et al. 2019; Sanchez et al. 2021; Tran et al. 2024). Each approach captures only a portion of the information contained in genomic data, and different methods excel in different scenarios.

Papers describing new inference methods typically benchmark against existing approaches, but these self-assessments are often biased (Norel et al. 2011; Boulesteix 2015), if unconciously. First, method developers naturally focus on scenarios where their approaches excel, potentially masking weaknesses. Second, developers have intimate knowledge of optimal parameter settings for their own methods but may use default parameters for competing methods, leading to unfair comparisons. Finally, developers benchmarking their own tools know the ground truth they simulated, enabling unconscious bias toward that truth. Best-practice guidelines for benchmarking studies (Boulesteix 2015; Lotterhos et al. 2022) can reduce, but not eliminate, these biases.

Independent benchmarking studies can provide more reliable conclusions than developer-driven benchmarking (Boulesteix et al. 2013), and they have been conducted in population genomics, but limitations remain. While developing a data simulation framework for the community, the stdpopsim project compared methods for inferring demographic history, distributions of fitness effects, and selective sweeps, although not systematically (Adrion et al. 2020; Gower et al. 2025). For demographic history inference, parametric SFS-based methods have been compared with nonparametric SFS-based (Lapierre et al. 2017) and Markovian coalescent methods (Beichman et al. 2017). The confounding effects of background selection on such inference have been studied for SFS-based and Markovian coalescent methods (Johri et al. 2021) and ARG-based methods (Marsh and Johri 2024). Brandt et al. (2022) evaluated the accuracy of ARG inference methods in estimating coalescence times, Peng et al. (2025) evaluated ARG-based methods for predicting historical polygenic scores, and Patton et al. (2019) evaluated nonparametric methods for demographic history inference under varying genome assembly quality. Although these studies have investigated many different tools, each has been carried out by a small group of authors, and their expertise in the tools tested can strongly influence benchmark results (Lotterhos et al. 2016; Weber et al. 2019). And because each of these studies is singular, it is difficult to assess progress in the field from them.

Community-based competitions have proven effective at driving innovation across multiple domains of computational biology (Meyer et al. 2011). The Critical Assessment of Protein Structure Prediction (CASP), running since 1994, is perhaps the most successful (Moult et al. 1995). By providing semi-annual blind tests of protein structure prediction methods, CASP has catalyzed remarkable improvements, culminating in the 14th competition with AlphaFold 2's breakthrough performance that revolutionized structural biology (Jumper et al. 2021; Kryshtafovych et al. 2021). Similarly, challenges from DREAM (Dialogue for Reverse Engineering Assessment and Methods) have addressed diverse problems in systems biology and genomics, from gene regulatory network inference to disease prediction (Stolovitzky et al. 2007; Marbach et al. 2012; Saez-Rodriguez et al. 2016). More recently, the Critical Assessment of Genome Interpretation focuses on predicting phenotypic consequences of genetic variants, driving improvements in variant effect prediction (Critical Assessment of Genome Interpretation Consortium 2024). PrecisionFDA challenges evaluate methods for variant calling, genome assembly, and other genomics tasks, setting standards for precision medicine applications (Olson et al. 2022). These examples illustrate the power of competition-based assessments of computational biology methods. In evolutionary inference, real-world data for which the truth is known is typically lacking (see Randall et al. 2016 for an exception), but modern simulators capture enough features of real data to provide valuable insights (Baumdicker et al. 2022; Haller and Messer 2023).

The Genomic History Inference Strategies Tournament (GHIST[a]) adapts the successful competition model to address the specific challenges of population genetic inference. Here, we report results from the first competition, which consisted of four challenges focused on inferring demographic history. The competition attracted many participants, demonstrated the feasibility of the model, and revealed current community practices.

Methods

The organizing team (authors Struck and Gutenkunst) and the design committee (authors Lotterhos, Moreno-Estrada, Ralph, and Siepel) collaborated closely to develop the structure of the first GHIST competition. Although creating highly complex challenges was tempting, we prioritized accessibility to ensure early community engagement and success. We chose demographic history inference as the competition's focus, because it is foundational to many other population genetic analyses, it allows comparison across a variety of established methods, and aligns with the organizers' expertise. To further encourage participation, we outlined a proactive communication strategy and offered authorship on the resulting paper as recognition for the top-performing competitors.

The structure of the first GHIST competition was developed in collaboration between the organizing team (authors Struck and Gutenkunst) and design committee (authors Lotterhos, Moreno-Estrada, Ralph, and Siepel). While it was tempting to develop extremely complex challenges, accessibility was deemed important for early success of the competition. It was decided to focus on demographic history inference, because it is foundational for other population genetics inference tasks, there are many methods to compare, and because the organizers have specific expertise. To engage community participation, a preliminary plan for communications was also developed. Finally, to incentivize participation, it was decided that top competitors would be offered authorship on the resulting paper.

The inaugural GHIST competition consisted of four demographic history inference challenges. These were a simple

bottleneck, a simple split with migration, a complex split with secondary contact, and a complex archaic admixture scenario. Competitors could submit to any challenge(s) they chose, in any order. The scenarios were parameterized such that existing methods were expected to have good statistical power and sample sizes were set to be similar to contemporary non-human data sets. For all four challenges, the data were simulated using the Wright-Fisher coalescent method msprime (Baumdicker et al. 2022) and distributed as error-free Variant Call Format (VCF) files (Danecek et al. 2011), with only biallelic sites and correct ancestral states provided. To minimize complexity, mutation and recombination rates were uniform across the simulated regions, and selection was absent.

For each challenge, competitors reported estimates for a small number of key population genetic parameters, such as population sizes, divergence times, or admixture proportions. They were told the total size of the simulated region and the true simulated mutation and recombination rates. Entries were scored based on the relative root-mean-squared error between submitted $\hat{\theta}$ and true parameter values θ:

$$\mathrm{RRMSE} = \sqrt{\sum_i \left(\frac{\hat{\theta}_i - \theta_i}{\theta_i}\right)^2}. \quad (1)$$

This interpretable metric allowed comparison across parameters of different scales and penalized both over- and underestimation equally. For each challenge, the leaderboard was ranked based on RRMSE scores, with lower scores indicating better performance. To allow methodological refinement, competitors were allowed five submissions for each challenge. In addition to their inferences, competitors were asked to submit a brief write-up of their approach, including software tools used and the logical flow of their analyses. The scripts for generating the data and scoring submissions are available at https://github.com/tjstruck/GHIST-2024-paper.

The competition was hosted on the Synapse platform developed by Sage Bionetworks, a not-for-profit organization that promotes open science and collaborative research. Synapse provided automated handling of competitor submissions, including timestamps, versioning, validation, and real-time leaderboards. The integrated wiki functionality was used for competition documentation and tutorials, and discussion boards enabled competitors to ask questions of the organizers. The Synapse site for the first GHIST competition is available at https://synapse.org/Synapse:syn51614781, and the main GHIST website is at https://ghist.bio.

The inaugural GHIST competition ran from July to November 2024, to span the summer conference season and the beginning of the academic term (Fig. 1). It began with a kick-off workshop at the Society for Molecular Biology and Evolution (SMBE) conference in Puerto Vallarta, Mexico, where participants were introduced to the competition, analyzed data from the Bottleneck challenge using dadi-cli (Huang et al. 2023), and submitted their inferences. The competition extended into the academic term to enable new students to participate as a training opportunity. The competition was promoted in-person at the SMBE and Evolution conferences, through posts to the Evoldir, dadi user, and fastsimcoal user mailing lists, and through targeted emails to specific investigators known to the organizers. It was also promoted through posts on X and Bluesky by the organizers and SMBE.

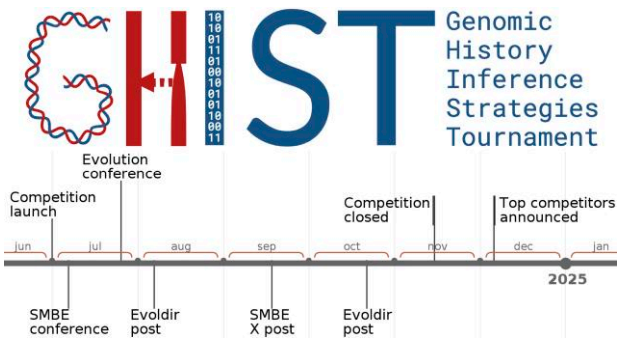

Fig. 1. Timeline of the first GHIST competition, including notable promotion events.

Results

The inaugural GHIST competition attracted approximately 60 participants spanning career stages from graduate students to senior faculty. Participation varied across challenges, with more entries for the simpler challenges. Competitors employed a variety of approaches, with top competitors mostly relying on the site frequency spectrum (SFS). A variety of software was employed, including custom pipelines.

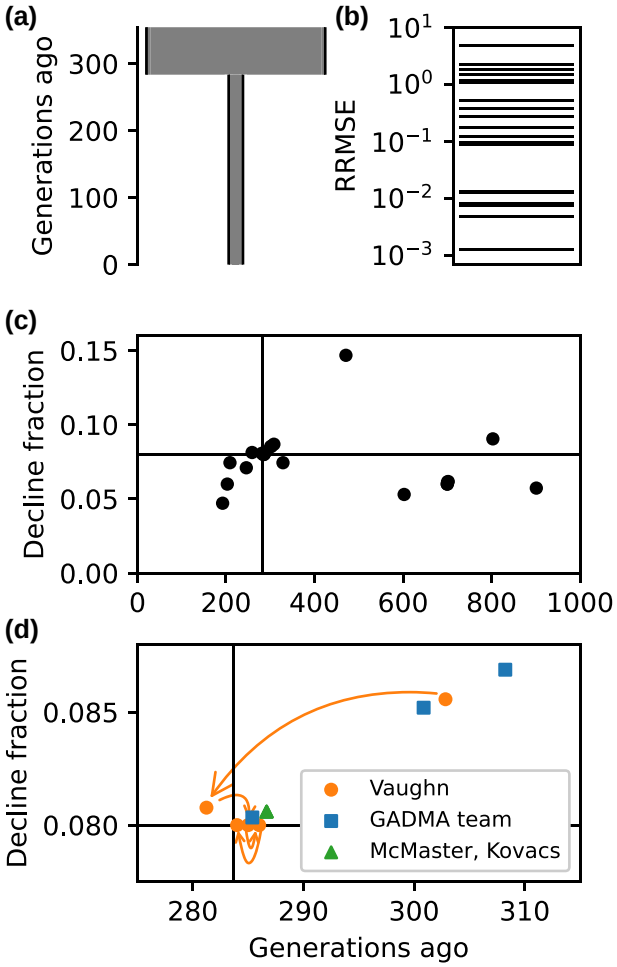

Fig. 2. GHIST 2024 Bottleneck challenge. a) True simulated demographic history. b) Relative root mean square error scores of submissions. c) Parameter inferences of majority of submissions. True values are indicated by solid lines. d) Parameter inferences zoomed close to true values to indicate top competitors. Arrows indicate competitor Vaughn's leaderboard optimization procedure.

Bottleneck Challenge

The first challenge involved a simple bottleneck (Fig. 2a), with competitors inferring the timing and magnitude of the population decline. Competitors were given 100 megabases (Mb) of data from 20 diploid individuals, yielding 219 thousand biallelic variants.

Submissions for the bottleneck challenge showed a range of strategies and accuracy. The RRMSE values of submissions spanned orders of magnitude (Fig. 2b). Almost all submissions successfully identified the presence of a bottleneck (Fig. 2c), but only a few were highly accurate.

The most accurate submissions for the Bottleneck challenge used site frequency spectrum (SFS) approaches (Fig. 2d). Competitor Vaughn developed a custom approach using mushi's code for analytically calculating the the expected SFS for piecewise constant demographic histories (DeWitt et al. 2021) and the Kullback-Leibler (KL) divergence to measure differences between model and data spectra. Competitors McMaster and Kovacs used the SFS-based methods dadi-cli (Huang et al. 2023) and fastsimcoal2 (Excoffier et al. 2021) and the Markovian coalescent tool SMC++ (Terhorst et al. 2017) for their submissions. Competitor Noskova led a team using her GADMA (Noskova et al. 2020, 2023) framework, using the dadi (Gutenkunst et al. 2009), moments (Jouganous et al. 2017), and momi2 (Kamm et al. 2020) engines for calculating model spectra.

A surprise was that top competitor Vaughn metagamed the challenge by using the leaderboard to optimize his submissions. He made an excellent first submission (Fig. 2d) based on the provided data, but this would not have been enough to win the challenge. To improve his result, he correctly deduced that the challenge simulation used round parameter values, and he used his remaining four submissions to search through the parameter space using the leaderboard RRMSE score to converge on nearly the exact values. Fundamentally, the public leaderboard leaked information by enabling competitors to know whether subsequent submissions were approaching the true simulated parameter values. Competitor Vaughn used the leaderboard and his allowed multiple guesses to iteratively optimize his submissions beyond what his initial data analysis enabled.

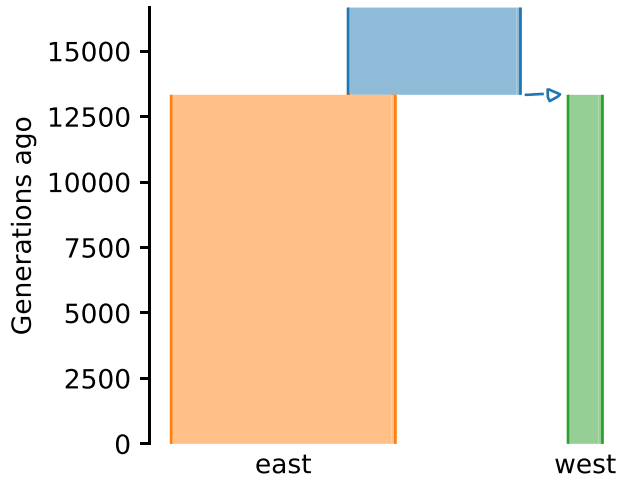

Fig. 3. True demographic history for the Split with Isolation challenge, represented using demesdraw (Gower et al. 2022).

Split with Isolation Challenge

The second challenge involved two populations that diverged without subsequent gene flow, representing geographic isolation (Fig. 3), with competitors inferring the contemporary population sizes and the timing of the split. They were given 100 Mb of data from 22 and 18 individuals from the two populations, yielding 1.2 million biallelic variants.

Performance on this challenge was generally strong, with several competitors achieving high accuracy for all three parameters (Table 1). SFS-based methods performed well in this challenge, with competitor Vaughn using msprime and tskit (Kelleher et al. 2016; Wong et al. 2024) to calculate expected spectra by averaging over multiple simulations and KL divergence to fit the model and McMaster and Kovacs using dadi-cli for inference. The Team of Daigle and Ray used a machine learning approach. They first used dadi to identify the relevant ranges of parameter values, then simulated data over those ranges with msprime, and then used scikit-allel (Miles et al. 2024) and pylibseq (Thornton 2003) to calculate summary statistics, including statistics based on the SFS, haplotypes, and LD decay. These summary statistics were then passed to a multi-layer perceptron for inference. However, their best-scoring submission for this challenge simply employed dadi. As in the first Challenge, competitor Vaughn achieved the top score by strategically rounding his inferences.

Secondary Contact Challenge

The third challenge involved secondary contact between isolated populations, with complexity in population size histories that no parametric model was expected to capture (Fig. 4a). Competitors were tasked with inferring the contemporary population sizes, timing of the split and recontact, and the rate of migration after recontact. Competitors were again given 100 Mb of data, from 22 diploid mainland individuals and 8 island individuals, for a total of 842 thousand biallelic sites.

As expected, this challenge was more difficult than the previous two, with no submission accurately estimating all parameters (Table 2). The team of Daigle and Ray did well with their machine learning approach based on summary statistics. Competitor Vaughn's top submissions were all based on leaderboard optimization after his initial inference. All these submissions assumed simple constant population size histories, like the truth in the Split with Isolation challenge. The best performance came from the GADMA team, using the moments engine (Table 2). GADMA automatically builds and refines models of increasing complexity, and their best model allowed for growth in both populations (Fig. 4b), perhaps enabling their model to account for some of the effects of the true complex population size changes.

Table 1 Top submissions for the Split with Isolation challenge

RRMSE	N_{east}	N_{west}	T	Competitor	Approach
truth	130,000	20,000	13,333		
0.005	130,000	20,100	13,300	Vaughn	metagaming
0.027	126,686	19,973	13,233	McMaster, Kovacs	dadi-cli
0.029	126,941	19,867	13,119	Daigle, Ray	dadi
0.031	125,992	20,109	13,301	Vaughn	msprime SFS

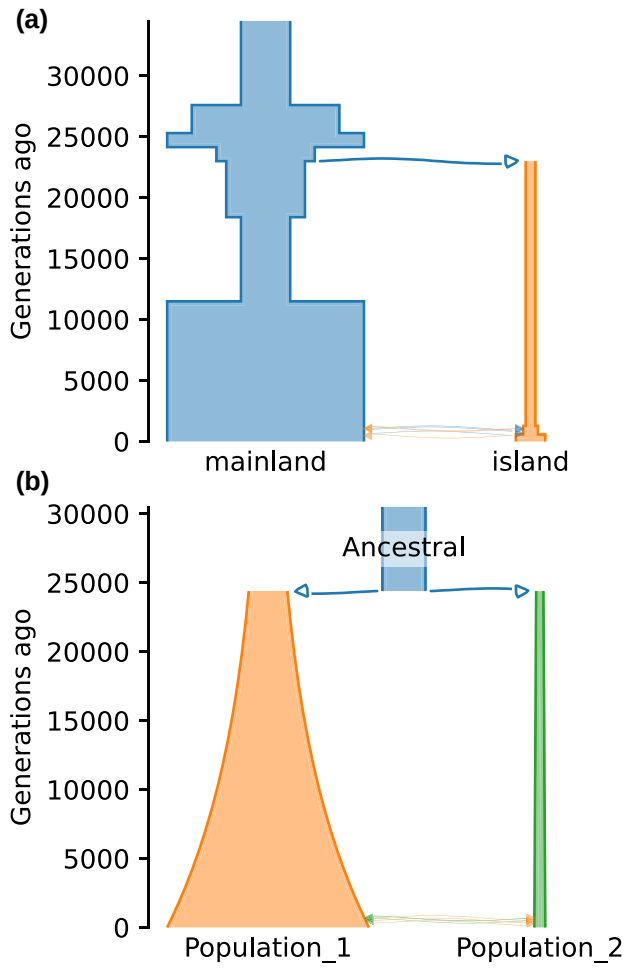

Fig. 4. GHIST 2024 Secondary Contact challenge. a) True demographic history, represented using demesdraw (Gower et al. 2022). b) Top-scoring model, from the GADMA team.

Archaic Admixture Challenge

To probe a distinct but related form of inference, the final challenge involved archaic admixture. Competitors were tasked with inferring the timing and magnitude of admixture into two modern populations (Fig. 5). They were given 250 Mb of data from 20 and 16 samples for the modern populations, along with 1 to 3 samples from each of the potential archaic contributors, sampled 17,500 to 100,000 simulated years ago, for a total of 1.7 million biallelic sites.

The top competitors accurately estimated admixture proportions but were less accurate when estimating timings (Table 3). For each modern population, competitor Vaughn used msprime simulations to simulate a two-population model with archaic admixture from a ghost population and fit that to the SFS from the modern population. He then optimized the leaderboard to refine his estimates. The GADMA software does not support ancient samples, so it could not be applied to this challenge. But the GADMA team used momi2 (Kamm et al. 2020) directly to fit models involving all five sample groups, achieving superior accuracy before metagaming.

Competitor Feedback

Competitors noted several lessons from the competition, and they overall found it valuable. For those new to population genetics inference, engaging with the challenges was difficult. Poor documentation was noted for some popular tools, such as dadi and dadi-cli, which made it difficult to get started without video tutorials or personal mentorship. Competitors also noted that different tools used different definitions of parameter values, especially migration rates, so care was needed in translating between them. They also noted that parameter optimization was often more difficult than they expected, so that close monitoring of tool runs was required to achieve best results. Finally, several student competitors highlighted how much they learned through the experience.

Discussion

The inaugural GHIST competition demonstrated the feasibility and value of a community-driven evaluation framework for population genetic inference methods. The Synapse platform proved robust and capable, and the range of challenges enabled accessibility while pushing the limits of existing inference methods. The conference-based launch and extended timeframe facilitated participation from diverse researchers, including students.

The GHIST competition provided several insights into the relative performance of inference approaches. Approaches based on the site frequency spectrum were most common and successful, because they are both accessible from established software tools and powerful for demographic inference. For the Bottleneck, Split with Isolation, and Secondary Contact challenges, the GADMA team directly compared SFS-based engines with the moments.LD engine that uses multi-population linkage disequilibrium statistics (Ragsdale and Gravel 2019, 2020), achieving better scores with SFS-based engines. Approaches based on machine learning showed promise but were not widely used by competitors. As those approaches become more accessible, we expect their representation and success to increase. Almost all approaches applied assumed prespecified parametric models, which may not capture the complexity of real demographic histories (Loog 2021). The exception was GADMA, and its success in the Secondary Contact challenge (Fig. 4a), which was designed to violate typical pre-specification, highlights the importance of model flexibility when dealing with complex histories. A caution is that GHIST cannot distinguish between performance properties of methods in theory and how they are used by competitors in practice. User expertise affects the outcomes of genomic analysis (Lotterhos et al. 2016), and feedback from some GHIST competitors emphasized the challenges in adopting some methods due to limited or overly sophisticated documentation. Increased participation may enable distinctions to be drawn between the performance of typical versus expert users of different methods.

There were notable gaps in the methods employed by participants. Approaches based on ancestral recombination graphs show great promise for population genetics (Rasmussen et al. 2014; Kelleher et al. 2019; Speidel et al. 2019; Deng et al. 2025), but they were not applied to this competition, perhaps because of their high computational cost or complexity. The Archaic Admixture challenge (Fig. 5) was designed to encourage the use of specialized methods based on lengths of admixture tracts (Pool and Nielsen 2009; Gravel 2012), but no competitors used them, perhaps due to insufficient outreach to the relevant subset of researchers. Methods for demographic history inference based on the Markovian coalescent (Li and Durbin 2011; Schiffels and Durbin 2014) that don't depend on a user-specified parametric model were also underrepresented

Table 2 Top submissions for Secondary Contact challenge

RRMSE	N_{main}	N_{island}	T_{split}	T_{mig}	m	Competitor	Approach
truth	240000	36000	23000	1277	5.0		
0.92	284284	15567	24397	950	8.2	GADMA team	GADMA w/ moments
0.97	150000	35000	30000	200	5.0	Vaughn	metagaming
1.05	200000	12000	18000	310	5.0	Vaughn	metagaming
1.05	220000	60000	17000	300	5.0	Vaughn	metagaming
1.30	211999	10945	15818	215	1.8	Daigle, Ray	summary stats perceptron

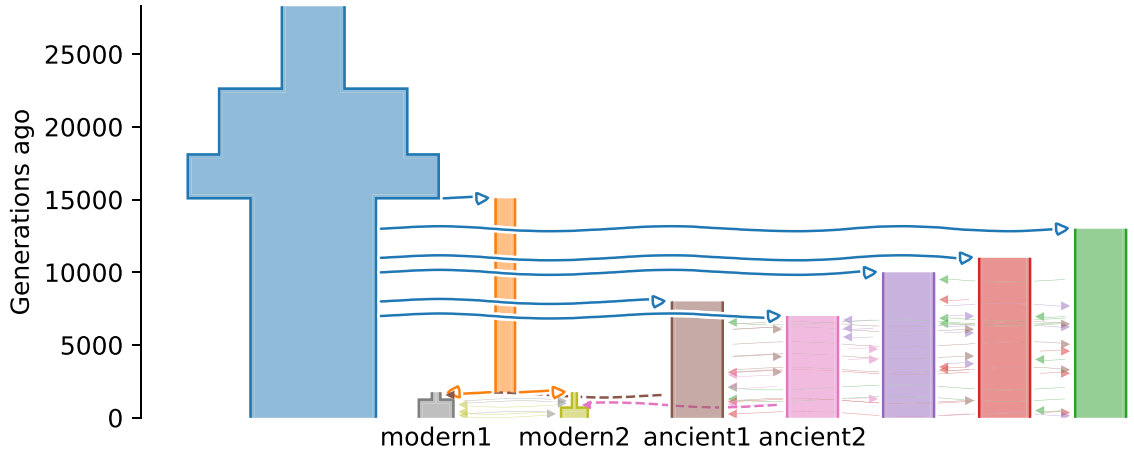

Fig. 5. True demographic history for Archaic Admixture challenge, represented using demesdraw (Gower et al. 2022).

Table 3 Top submissions for Archaic Admixture challenge

RRMSE	$\%admix_1$	$Tadmix_1$	$\%admix_2$	$Tadmix_2$	Competitor	Approach
truth	1.10	1566	0.20	883		
0.10	1.00	1500	0.20	900	Vaughn	metagaming
0.78	1.01	1096	0.19	252	GADMA team	momi2
0.81	1.02	1066	0.16	250	GADMA team	momi2
1.05	0.59	1000	0.37	1000	GADMA team	momi2
1.93	1.06	2500	0.22	2500	Vaughn	msprime SFS

relative to their popularity in the literature. They differ fundamentally from methods like dadi (Gutenkunst et al. 2009) and momi2 (Kamm et al. 2020) that output explicit demographic history model inferences, because they typically output coalescence or cross-coalescence rates. Changes in these rates are often interpreted in terms of demographic history, but that interpretation is an additional subjective step toward submitting inferences from these tools to the present competition.

The second GHIST competition launched at the Evolution conference in June 2025 in Athens, Georgia and runs through November 2025, with expanded challenge types and refinements based on lessons from the inaugural tournament. To discourage metagaming while preserving the benefits of iterative submission, for each challenge there are now two data sets. Unlimited submissions are allowed on a testing data set, to enable exploration of different approaches. But only a single submission is allowed on the final data set, to avoid leaderboard metagaming. To increase the realism and difficulty of demographic history inference, two of the challenges include background selection, leveraging the stdpopsim framework for simulation (Gower et al. 2025). To expand the range of tasks, four challenges involve inferring single or multiple hard selective sweeps (Stephan 2019), under simple and complex demographic scenarios and with and without background selection. Finally, to increase accessibility, simple web applications were developed to enable users to manually fit bottleneck models to site frequency spectra and to detect selective sweeps using summary statistics. The Synapse site for this second competition is at https://synapse.org/Synapse:syn65877330.

The first and second GHIST competitions use simple metrics for evaluating submissions based on parameter values or sweep locations, but future competitions could use more complex metrics. Within Synapse, submissions are scored using custom code executed on a cloud instance, so in principle anything that can be calculated can be scored. For example, agreement with the complex true population size history in the Secondary Contact challenge (Fig. 4) could be assessed more completely by a integrated deviation between submitted and true population sizes over time. For some applications, distributions of coalescent times might be more relevant, which could be simulated from submitted demographic history models and compared with those from the true simulated model. Either more complex evaluation would require submitters to provide complete models, either in a standardized format like Demes (Gower et al. 2022) or as runnable Docker images, which would substantially increase complexity of submission.

Demographic history inference is not only about estimating parameters; it also frequently entails model selection (Smith

et al. 2017; Johri et al. 2021). For example, the presence or absence of significant gene flow between natural populations is often of interest (Edwards et al. 2016; Momigliano et al. 2021). Particularly in humans, models differ in the number of pulses of introgression from archaic hominins into modern humans (Browning et al. 2018; Jacobs et al. 2019), and in the role of archaic introgression versus structure in ancestral African populations (Lorente-Galdos et al. 2019; Ragsdale et al. 2023). To evaluate model selection within a competition framework, competitors must be asked to analyze a large number of data sets that, for example, do and do not include gene flow. This would raise the burden on competitors, but it represents an important future direction for GHIST.

Ultimately, the success of GHIST depends on community participation. The more methods developers, users, educators, and students engage with the competitions, the more the community will learn. The space of potential challenges is vast, including inferences such as distributions of fitness effects (Eyre-Walker and Keightley 2007), spatial models (Bradburd and Ralph 2019), and polygenic selection (Barghi et al. 2020), and including complications such as low-pass data (Crawford and Lazzaro 2012), polyploidy (Dufresne et al. 2014), and biased gene conversion (Pouyet et al. 2018).

This initial competition has demonstrated the feasibility and utility of competitions for this community. Future competitions will enable deep insight into best practices for population genetics inference.

Note

a. GHIST is pronounced with a hard "g" sound and a soft "i" sound, like a blend of "gift" and "list".

Acknowledgments

We thank Pablo Meyer Rojas for introducing us to the Synapse platform and the Society for Molecular Biology and Evolution for hosting the kickoff workshop and promoting the competition on their social media channels. We thank all competitors for their participation. This paper was written with the assistance of generative artificial intelligence (AI). MacWhisper was used to transcribe the audio from author Gutenkunst's talk about GHIST at the 2024 Probabilistic Modeling in Genomics conference. Anthropic's Claude 4 Sonnet model was then given that transcript and asked to generate a detailed outline for the paper, including leveraging its existing knowledge base, within a Project containing previous papers by author Gutenkunst. Author Gutenkunst also used Claude's research mode to generate a report on published papers that independently compared population genetics inference approaches, which yielded a few studies he was previously unaware of. Author Gutenkunst edited or generated all text in the final manuscript and verified all references.

Funding

This work was supported by the National Institute of General Medical Sciences of the National Institutes of Health (R35GM149235 to R.N.G.). A.D. received support from NIGMS predoctoral training grant 5T32 GM067553 and from NIGMS grant R35GM154969 to Parul Johri. D.D.R was supported by the National Human Genome Research Institute of the National Institutes of Health under award number R01HG010774 to Daniel R. Schrider. K.E.L. was supported by National Science Foundation grant NSF-2043905. A.M-E. was supported by Chan Zuckerberg Initiative grant CZI-2024-354605.

Conflict of interest

None declared.

Data Availability

All data used in contest were simulated. Those data and the results from the competition are available at https://www.synapse.org/Synapse:syn51614781.

References

Adrion JR *et al.* A community-maintained standard library of population genetic models. *Elife*. 2020:9:e54967. https://doi.org/10.7554/eLife.54967.

Barghi N, Hermisson J, Schlötterer C. Polygenic adaptation: a unifying framework to understand positive selection. *Nat Rev Genet*. 2020:21:769–781. https://doi.org/10.1038/s41576-020-0250-z.

Baumdicker F *et al.* Efficient ancestry and mutation simulation with msprime 1.0. *Genetics*. 2022:220:iyab229. https://doi.org/10.1093/genetics/iyab229.

Beichman AC, Phung TN, Lohmueller KE. Comparison of single genome and allele frequency data reveals discordant demographic histories. *G3*. 2017:7:3605–3620. https://doi.org/10.1534/g3.117.300259.

Boulesteix AL. Ten simple rules for reducing overoptimistic reporting in methodological computational research. *PLoS Comput Biol*. 2015:11:e1004191. https://doi.org/10.1371/journal.pcbi.1004191.

Boulesteix AL, Lauer S, Eugster MJA. A plea for neutral comparison studies in computational sciences. *PLoS One*. 2013:8:e61562. https://doi.org/10.1371/journal.pone.0061562.

Bradburd GS, Ralph PL. Spatial population genetics: it's about time. *Annu Rev Ecol Evol Syst*. 2019:50:427–449. https://doi.org/10.1146/ecolsys.2019.50.issue-1.

Brandt DYC, Wei X, Deng Y, Vaughn AH, Nielsen R. Evaluation of methods for estimating coalescence times using ancestral recombination graphs. *Genetics*. 2022:221:iyac044. https://doi.org/10.1093/genetics/iyac044.

Browning SR, Browning BL. Accurate non-parametric estimation of recent effective population size from segments of identity by descent. *Am J Hum Genet*. 2015:97:404–418. https://doi.org/10.1016/j.ajhg.2015.07.012.

Browning SR, Browning BL, Zhou Y, Tucci S, Akey JM. Analysis of human sequence data reveals two pulses of archaic Denisovan admixture. *Cell*. 2018:173:53. https://doi.org/10.1016/j.cell.2018.02.031.

Crawford JE, Lazzaro BP. Assessing the accuracy and power of population genetic inference from low-pass next-generation sequencing data. *Front Genet*. 2012:3:66. https://doi.org/10.3389/fgene.2012.00066.

Critical Assessment of Genome Interpretation Consortium. CAGI, the critical assessment of genome interpretation, establishes progress and prospects for computational genetic variant interpretation methods. *Genome Biol*. 2024:25:53. https://doi.org/10.1186/s13059-023-03113-6.

Danecek P *et al.* The variant call format and VCFtools. *Bioinformatics*. 2011:27:2156–2158. https://doi.org/10.1093/bioinformatics/btr330.

Deng Y, Nielsen R, Song YS. Robust and accurate Bayesian inference of genome-wide genealogies for hundreds of genomes. *Nat Genet*. 2025:1–2.

DeWitt WS, Harris KD, Ragsdale AP, Harris K. Nonparametric coalescent inference of mutation spectrum history and demography. *Proceedings of the National Academy of Sciences*. 2021:118:e2013798118. https://doi.org/10.1073/pnas.2013798118.

Dufresne F, Stift M, Vergilino R, Mable BK. Recent progress and challenges in population genetics of polyploid organisms: an overview of

current state-of-the-art molecular and statistical tools. *Mol Ecol*. 2014:23:40–69. https://doi.org/10.1111/mec.2013.23.issue-1.
Edwards T *et al*. Assessing models of speciation under different biogeographic scenarios; an empirical study using multi-locus and RNA-seq analyses. *Ecol Evol*. 2016:6:379–396. https://doi.org/10.1002/ece3.2016.6.issue-2.
Excoffier L *et al*. *fastsimcoal2*: demographic inference under complex evolutionary scenarios. *Bioinformatics*. 2021:37:4882–4885. https://doi.org/10.1093/bioinformatics/btab468.
Eyre-Walker A, Keightley PD. The distribution of fitness effects of new mutations. *Nat Rev Genet*. 2007:8:610–618. https://doi.org/10.1038/nrg2146.
Flagel L, Brandvain Y, Schrider DR. The unreasonable effectiveness of convolutional neural networks in population genetic inference. *Mol Biol Evol*. 2019:36:220–238. https://doi.org/10.1093/molbev/msy224.
Gower G *et al*. Demes: a standard format for demographic models. *Genetics*. 2022:222:iyac131. https://doi.org/10.1093/genetics/iyac131.
Gower G *et al*. Accessible realistic genome simulation with selection using stdpopsim. *Mol Biol Evol*. 2025:msaf236.
Gravel S. Population genetics models of local ancestry. *Genetics*. 2012:191:607–619. https://doi.org/10.1534/genetics.112.139808.
Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genet*. 2009:5:e1000695. https://doi.org/10.1371/journal.pgen.1000695.
Haller BC, Messer PW. SLiM 4: multispecies eco-evolutionary modeling. *Am Nat*. 2023:201:E127–E139. https://doi.org/10.1086/723601.
Harris K, Nielsen R. Inferring demographic history from a spectrum of shared haplotype lengths. *PLoS Genet*. 2013:9:e1003521. https://doi.org/10.1371/journal.pgen.1003521.
Huang X, Struck TJ, Davey SW, Gutenkunst RN. 2023. dadi-cli: Automated and distributed population genetic model inference from allele frequency spectra [preprint]. bioRxiv. https://doi.org/10.1101/2023.06.15.545182
Jacobs GS *et al*. Multiple deeply divergent Denisovan ancestries in papuans. *Cell*. 2019:177:1010. https://doi.org/10.1016/j.cell.2019.02.035.
Johri P *et al*. The impact of purifying and background selection on the inference of population history: problems and prospects. *Mol Biol Evol*. 2021:38:2986–3003. https://doi.org/10.1093/molbev/msab050.
Johri P *et al*. Recommendations for improving statistical inference in population genomics. *PLoS Biol*. 2021:20:e3001669. https://doi.org/10.1371/journal.pbio.3001669.
Jouganous J, Long W, Ragsdale AP, Gravel S. Inferring the joint demographic history of multiple populations: beyond the diffusion approximation. *Genetics*. 2017:206:1549–1567. https://doi.org/10.1534/genetics.117.200493.
Jumper J *et al*. Highly accurate protein structure prediction with AlphaFold. *Nature*. 2021:596:583–589. https://doi.org/10.1038/s41586-021-03819-2.
Kamm J, Terhorst J, Durbin R, Song YS. Efficiently inferring the demographic history of many populations with allele count data. *J Am Stat Assoc*. 2020:115:1472–1487. https://doi.org/10.1080/01621459.2019.1635482.
Kelleher J *et al*. Inferring whole-genome histories in large population datasets. *Nat Genet*. 2019:51:1330–1338. https://doi.org/10.1038/s41588-019-0483-y.
Kelleher J, Etheridge AM, McVean G. Efficient coalescent simulation and genealogical analysis for large sample sizes. *PLoS Comput Biol*. 2016:12:e1004842. https://doi.org/10.1371/journal.pcbi.1004842.
Kryshtafovych A, Schwede T, Topf M, Fidelis K, Moult J. Critical assessment of methods of protein structure prediction (CASP)-round XIV. *Proteins: Structure, Function, and Bioinformatics*. 2021:89:1607–1617. https://doi.org/10.1002/prot.v89.12.
Lapierre M, Lambert A, Achaz G. Accuracy of demographic inferences from the site frequency spectrum: the case of the yoruba population. *Genetics*. 2017:206:439–449. https://doi.org/10.1534/genetics.116.192708.
Lawson DJ, Van Dorp L, Falush D. A tutorial on how not to over-interpret STRUCTURE and ADMIXTURE bar plots. *Nat Commun*. 2018:9:3258. https://doi.org/10.1038/s41467-018-05257-7.
Li H, Durbin R. Inference of human population history from individual whole-genome sequences. *Nature*. 2011:475:493–496. https://doi.org/10.1038/nature10231.
Loog L. Sometimes hidden but always there: the assumptions underlying genetic inference of demographic histories. *Philosophical Transactions of the Royal Society B*. 2021:376:20190719. https://doi.org/10.1098/rstb.2019.0719.
Lorente-Galdos B *et al*. Whole-genome sequence analysis of a Pan African set of samples reveals archaic gene flow from an extinct basal population of modern humans into sub-saharan populations. *Genome Biol*. 2019:20:77. https://doi.org/10.1186/s13059-019-1684-5.
Lotterhos KE, Fitzpatrick MC, Blackmon H. Simulation tests of methods in evolution, ecology, and systematics: pitfalls, progress, and principles. *Annu Rev Ecol Evol Syst*. 2022:53:113–136. https://doi.org/10.1146/ecolsys.2022.53.issue-1.
Lotterhos KE, François O, Blum MG. 2016. Not just methods: user expertise explains the variability of outcomes of genome-wide studies [preprint]. bioRxiv. https://doi.org/10.1101/055046
Marbach D *et al*., DREAM5 Consortium. Wisdom of crowds for robust gene network inference. *Nat Methods*. 2012:9:796–804. https://doi.org/10.1038/nmeth.2016.
Marsh JI, Johri P. Biases in ARG-based inference of historical population size in populations experiencing selection. *Mol Biol Evol*. 2024:41:msae118. https://doi.org/10.1093/molbev/msae118.
Marth GT, Czabarka E, Murvai J, Sherry ST. The allele frequency spectrum in genome-wide human variation data reveals signals of differential demographic history in three large world populations. *Genetics*. 2004:166:351–372. https://doi.org/10.1534/genetics.166.1.351.
Meyer P *et al*. Verification of systems biology research in the age of collaborative competition. *Nat Biotechnol*. 2011:29:811–815. https://doi.org/10.1038/nbt.1968.
Miles A *et al*. scikit-allel: v1.3.13. 2024.
Momigliano P, Florin AB, Merilä J. Biases in demographic modelling affect our understanding of recent divergence. *Mol Biol Evol*. 2021:38:msab047. https://doi.org/10.1093/molbev/msab047.
Moult J, Pedersen JT, Judson R, Fidelis K. A large-scale experiment to assess protein structure prediction methods. *Proteins: Structure, Function, and Bioinformatics*. 1995:23:ii–v. https://doi.org/10.1002/prot.v23:3.
Myers S, Fefferman C, Patterson N. Can one learn history from the allelic spectrum? *Theor Popul Biol*. 2008:73:342–348. https://doi.org/10.1016/j.tpb.2008.01.001.
Norel R, Rice JJ, Stolovitzky G. The self-assessment trap: can we all be better than average? *Mol Syst Biol*. 2011:7:537. https://doi.org/10.1038/msb.2011.70.
Noskova E *et al*. GADMA2: more efficient and flexible demographic inference from genetic data. *Gigascience*. 2023:12:giad059. https://doi.org/10.1093/gigascience/giad059.
Noskova E, Ulyantsev V, Koepfli KP, O'Brien SJ, Dobrynin P. GADMA: genetic algorithm for inferring demographic history of multiple populations from allele frequency spectrum data. *Gigascience*. 2020:9:giaa005. https://doi.org/10.1093/gigascience/giaa005.
Olson ND *et al*. precisionFDA truth challenge V2: calling variants from short- and long-reads in difficult-to-map regions. *Cell Genom*. 2022:2:100129. https://doi.org/10.1016/j.xgen.2022.100129.
Patton AH *et al*. Contemporary demographic reconstruction methods are robust to genome assembly quality: a case study in Tasmanian devils. *Mol Biol Evol*. 2019:36:2906–2921. https://doi.org/10.1093/molbev/msz191.
Peng D, Mulder OJ, Edge MD. Evaluating ARG-estimation methods in the context of estimating population-mean polygenic score histories. *Genetics*. 2025:229:iyaf033. https://doi.org/10.1093/genetics/iyaf033.

Pool JE, Hellmann I, Jensen JD, Nielsen R. Population genetic inference from genomic sequence variation. *Genome Res*. 2010:20:291–300. https://doi.org/10.1101/gr.079509.108.

Pool JE, Nielsen R. Inference of historical changes in migration rate from the lengths of migrant tracts. *Genetics*. 2009:181:711–719. https://doi.org/10.1534/genetics.108.098095.

Pouyet F, Aeschbacher S, Thiéry A, Excoffier L. Background selection and biased gene conversion affect more than 95% of the human genome and bias demographic inferences. *Elife*. 2018:7:e36317. https://doi.org/10.7554/eLife.36317.

Ragsdale AP *et al.* A weakly structured stem for human origins in Africa. *Nature*. 2023:617:755. https://doi.org/10.1038/s41586-023-06055-y.

Ragsdale AP, Gravel S. Models of archaic admixture and recent history from two-locus statistics. *PLoS Genet*. 2019:15:e1008204. https://doi.org/10.1371/journal.pgen.1008204.

Ragsdale AP, Gravel S. Unbiased estimation of linkage disequilibrium from unphased data. *Mol Biol Evol*. 2020:37:923. https://doi.org/10.1093/molbev/msz265.

Randall RN, Radford CE, Roof KA, Natarajan DK, Gaucher EA. An experimental phylogeny to benchmark ancestral sequence reconstruction. *Nat Commun*. 2016:7:12847. https://doi.org/10.1038/ncomms12847.

Rasmussen MD, Hubisz MJ, Gronau I, Siepel A. Genome-wide inference of ancestral recombination graphs. *PLoS Genet*. 2014:10:e1004342. https://doi.org/10.1371/journal.pgen.1004342.

Rosen Z, Bhaskar A, Roch S, Song YS. Geometry of the sample frequency spectrum and the perils of demographic inference. *Genetics*. 2018:210:665. https://doi.org/10.1534/genetics.118.300733.

Saez-Rodriguez J *et al.* Crowdsourcing biomedical research: leveraging communities as innovation engines. *Nat Rev Genet*. 2016:17:470. https://doi.org/10.1038/nrg.2016.69.

Sanchez T, Cury J, Charpiat G, Jay F. Deep learning for population size history inference: design, comparison and combination with approximate Bayesian computation. *Mol Ecol Resour*. 2021:21:2645–2660. https://doi.org/10.1111/men.v21.8.

Schiffels S, Durbin R. Inferring human population size and separation history from multiple genome sequences. *Nat Genet*. 2014:46:919–925. https://doi.org/10.1038/ng.3015.

Schrider DR, Kern AD. Supervised machine learning for population genetics: a new paradigm. *Trends in Genetics*. 2018:34:301–312. https://doi.org/10.1016/j.tig.2017.12.005.

Smith ML *et al.* Demographic model selection using random forests and the site frequency spectrum. *Mol Ecol*. 2017:26:4562–4573. https://doi.org/10.1111/mec.2017.26.issue-17.

Speidel L, Forest M, Shi S, Myers SR. A method for genome-wide genealogy estimation for thousands of samples. *Nat Genet*. 2019:51:1321–1329. https://doi.org/10.1038/s41588-019-0484-x.

Stephan W. Selective sweeps. *Genetics*. 2019:211:5–13. https://doi.org/10.1534/genetics.118.301319.

Stolovitzky G, Monroe D, Califano A. Dialogue on reverse-engineering assessment and methods: the DREAM of high-throughput pathway inference. *Ann N Y Acad Sci*. 2007:1115:1–22. https://doi.org/10.1196/annals.1407.021.

Terhorst J, Kamm JA, Song YS. Robust and scalable inference of population history from hundreds of unphased whole genomes. *Nat Genet*. 2017:49:303–309. https://doi.org/10.1038/ng.3748.

Thornton K. libsequence: a C++ class library for evolutionary genetic analysis. *Bioinformatics*. 2003:19:2325. https://doi.org/10.1093/bioinformatics/btg316.

Tran LN, Sun CK, Struck TJ, Sajan M, Gutenkunst RN. Computationally efficient demographic history inference from allele frequencies with supervised machine learning. *Mol Biol Evol*. 2024:41:msae077. https://doi.org/10.1093/molbev/msae077.

Weber LM *et al.* Essential guidelines for computational method benchmarking. *Genome Biol*. 2019:20:125. https://doi.org/10.1186/s13059-019-1738-8.

Wong Y *et al.* A general and efficient representation of ancestral recombination graphs. *Genetics*. 2024:228:iyae100. https://doi.org/10.1093/genetics/iyae100.