Computationally Efficient Demographic History Inference from Allele Frequencies with Supervised Machine Learning

Linh N. Tran,[1,2] Connie K. Sun,[2] Travis J. Struck ,[2] Mathews Sajan,[2] and Ryan N. Gutenkunst [2,*]

[1]Genetics Graduate Interdisciplinary Program, University of Arizona, Tucson, AZ 85721, USA

[2]Department of Molecular & Cellular Biology, University of Arizona, Tucson, AZ 85721, USA

***Corresponding author**: E-mail: rgutenk@arizona.edu

Associate editor: Kelley Harris

Abstract

Inferring past demographic history of natural populations from genomic data is of central concern in many studies across research fields. Previously, our group had developed dadi, a widely used demographic history inference method based on the allele frequency spectrum (AFS) and maximum composite-likelihood optimization. However, dadi's optimization procedure can be computationally expensive. Here, we present donni (demography optimization via neural network inference), a new inference method based on dadi that is more efficient while maintaining comparable inference accuracy. For each dadi-supported demographic model, donni simulates the expected AFS for a range of model parameters then trains a set of Mean Variance Estimation neural networks using the simulated AFS. Trained networks can then be used to instantaneously infer the model parameters from future genomic data summarized by an AFS. We demonstrate that for many demographic models, donni can infer some parameters, such as population size changes, very well and other parameters, such as migration rates and times of demographic events, fairly well. Importantly, donni provides both parameter and confidence interval estimates from input AFS with accuracy comparable to parameters inferred by dadi's likelihood optimization while bypassing its long and computationally intensive evaluation process. donni's performance demonstrates that supervised machine learning algorithms may be a promising avenue for developing more sustainable and computationally efficient demographic history inference methods.

Key words: **population genomics, demographic history inference, machine learning.**

Introduction

Inferring demographic history from genomic data has become routine in many research fields, from elucidating the anthropological origins and migration patterns of modern and archaic human populations (Gutenkunst et al. 2009; Bergström et al. 2020; Gopalan et al. 2022; Marchi et al. 2022), to inferring the population genetic trajectories of endangered animals (Mays Jr et al. 2018; Miller-Butterworth et al. 2021; Chavez et al. 2022). Accounting for demographic history is also essential in setting the appropriate background for detecting signals of natural selection (Nielsen et al. 2005; Boyko et al. 2008; Kim et al. 2017), disease associations (Mathieson and McVean 2012), and recombination hotspots (Johnston and Cutler 2012). Due to the wide range of possible demographic models and high dimensionality of genome sequence data, such analysis often involves computationally expensive modeling. As the size of genomic datasets rapidly grows to thousands of full genomes, there is a great need for more efficient and scalable methods for extracting information from such datasets.

One class of widely used methods infers demographic history from sequence data summarized as an allele frequency spectrum (AFS). An AFS is a multidimensional array with dimension equal to the number of populations being considered in a given demographic model. Each array entry is the number of observed single-nucleotide polymorphisms (SNP) with given frequencies in the sampled populations. For example, the [1,2] entry would count SNPs that were singletons in the first population and doubletons in the second. A major advantage of using the AFS as a summary statistic is the ease of scaling to whole-genome data (Marchi et al. 2021), as it efficiently reduces the high dimensionality of population genomic data. AFS-based inference methods are, therefore, often fast and suitable for exploring many demographic models (Spence et al. 2018). Given its wide use in countless empirical studies, much progress has been made towards understanding the theoretical guarantees and limitations of the AFS and AFS-based inference (Myers et al. 2008; Achaz 2009; Bhaskar and Song 2014; Terhorst and Song 2015; Baharian and Gravel 2018).

Received: May 24, 2023. **Revised:** April 8, 2024. **Accepted:** April 12, 2024

Demographic inference methods based on the AFS often work by maximizing the composite likelihood of the observed AFS under a user-specified demographic history model with parameters such as population sizes, migration rates, and divergence times (Coffman et al. 2016). The expected AFS can be computed via a wide range of approaches (Gutenkunst et al. 2009; Naduvilezhath et al. 2011; Lukić and Hey 2012; Excoffier et al. 2013; Jouganous et al. 2017; Kamm et al. 2017; Kern and Hey 2017) with varying degrees of computational expense, model flexibility, and scalability. Because this is the most computationally intensive step in the procedure, new methods developed thus far have focused on devising algorithms to speed up AFS calculation (Jouganous et al. 2017; Kamm et al. 2017, 2020). However, not much attention has been given to optimizing how the computed AFS is stored and used for inference. In a typical likelihood optimization procedure, hundreds to thousands of expected AFS are computed and compared to the data to obtain the best-fit parameter set. These generated AFS and their corresponding demographic parameters contain information regarding the mapping between the AFS and demographic parameters but are discarded after each optimization run. As there are often a few common demographic models regularly used across studies, if these simulated data could be captured, stored, and distributed for future use, individual groups as well as the research community as a whole could save a lot of time and computational effort by avoiding unnecessary repetition.

The mapping between the AFS and its associated demographic history model parameters can be efficiently captured by supervised machine learning (ML) algorithms. Given a training dataset with feature vectors (AFS—input) and labels (demographic history parameters—output), these algorithms can learn the function mapping from the input to the output. While training ML algorithms can be computationally intensive up front, subsequent inference from trained models will have minimal cost (Schrider and Kern 2018). ML algorithms have been widely adopted in population genetics over the past decade, thanks to their efficiency and flexibility. Several studies have used supervised ML algorithms such as random forest (RF) and multilayer perceptron (MLP) with AFS as training data for demographic model selection and demographic parameter inference (Sheehan and Song 2016; Smith et al. 2017; Lorente-Galdos et al. 2019; Mondal et al. 2019; Villanea and Schraiber 2019; Sanchez et al. 2021). In Smith et al. (2017) specifically, the RF algorithm was used to replace the rejection step in the approximate Bayesian computation (ABC) framework, significantly improving overall efficiency (Pudlo et al. 2016). This improvement in efficiency was in part due to more efficient use of simulated data. Whereas in a typical ABC procedure, any simulations beyond a threshold of difference to the data will be discarded, there all simulations were used as input for training the RF classification algorithm. The same principle can be applied in the maximum-likelihood optimization and regression framework, where an ML algorithm can be trained by simulated AFS to provide estimates of demographic parameter values, bypassing likelihood optimization.

New Approaches

Here, we introduce donni (Demography Optimization via Neural Network Inference), a supervised ML extension to dadi, a widely used AFS-based method for inferring models of demographic history (Gutenkunst et al. 2009) and natural selection (Kim et al. 2017). dadi computes the expected AFS by numerically solving a diffusion approximation to the Wright–Fisher model and uses composite-likelihood maximization to fit the model to the data. While the initial implementation of the software could only handle up to three populations, a recent update supports up to five populations (Gutenkunst 2021). donni uses dadi to generate AFS and demographic parameter labels for training Mean Variance Estimation (MVE) networks (Nix and Weigend 1994) (Fig. 1). Researchers can then use donni's trained MVE networks to instantaneously infer the parameter values and their associated uncertainty from future AFS input data, obviating the need for likelihood optimization. donni supports a wide range of common demographic parameters that dadi supports, including population sizes, divergence times, continuous migration rates, inbreeding coefficients, and ancestral state misidentification ratios. We show that donni has inference accuracy comparable to dadi but requires fewer computational resources, even after accounting for the cost of training the MVE networks. Our library of trained networks currently includes all demographic models in the dadi API as well as the models from Portik et al. (2017) pipeline. The supported sample sizes are 10, 20, 40, 80, and 160 haplotypes per population (up to 20 haplotypes only for three-population models). For users who only need to use the trained networks for available demographic models, almost no computation is required. For users who require custom models, we also provide our command-line interface pipeline for generating trained models that can save time compared to running likelihood optimization with dadi. Furthermore, the custom models produced can be contributed to our growing library and shared with the community.

Results

Choice of MVE Network for Demographic History Model Parameter Estimation with Uncertainty

We wanted to develop a supervised ML method that can infer not only the demographic history parameters but also their associated uncertainties. Uncertainty estimation has not been the focus of previous supervised neural-network-based approaches in demographic history inference (Sheehan and Song 2016; Flagel et al. 2019). There are several techniques for constructing a prediction interval from neural-network-based point estimation as reviewed by Khosravi et al. (2011). Among them, the MVE method is one of the most conceptually straightforward and least

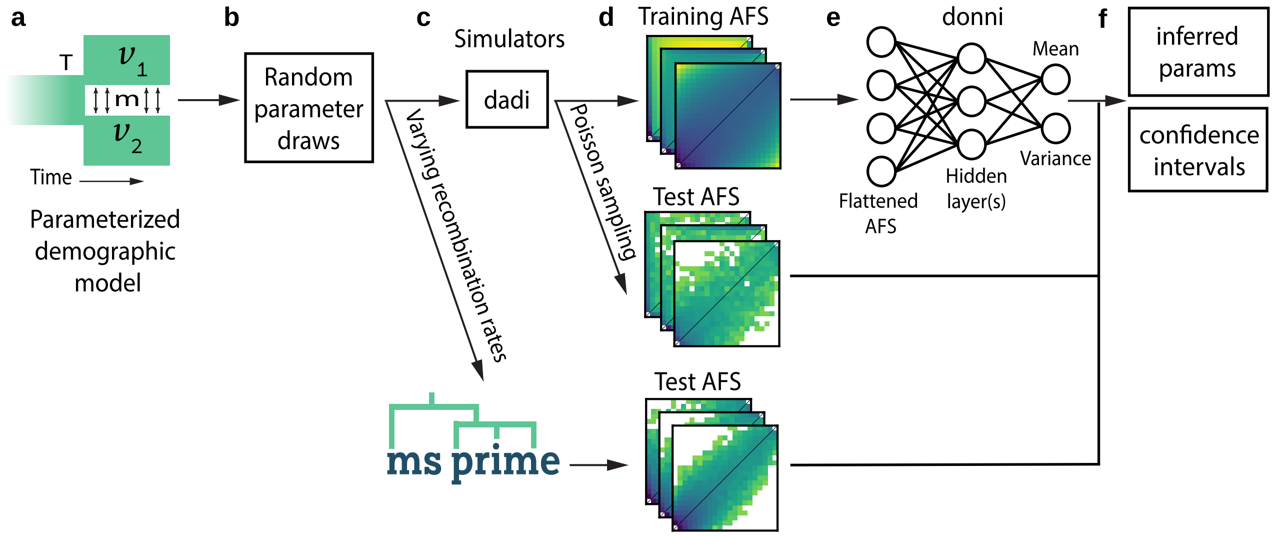

Fig. 1. Schematic of the workflow for training and testing donni. For a given demographic model a), we drew sets of model parameters b) from a biologically relevant range (supplementary table S1, Supplementary Material online). Each parameter set represents a demographic history and corresponds to an expected AFS. These parameters were input into simulator programs c) to generate training and test AFS d). We use the expected AFS simulated with dadi and their corresponding parameters as training data for donni's MVE networks e). We generated test data either by Poisson sampling from dadi-simulated AFS or by varying recombination rates with msprime, resulting in a change in test data variance compared to training AFS. The output of donni's trained networks includes both inferred parameters and their confidence intervals f).

computationally demanding, which are important factors for our goal of improving efficiency.

An MVE network is a feedforward neural network with two output nodes, one for the prediction mean and one for the prediction variance (Fig. 1e). This approach provides an uncertainty estimate in a regression setting by assuming that the errors are normally distributed around the mean estimation. For demographic history inference, the mean is the value of the demographic history model parameter we want to infer. We can construct confidence intervals using the normal distribution defined by the output mean and variance estimates. There are different implementations of the feedforward network architecture for MVE network (Sluijterman et al. 2023). Our implementation is a fully connected network, similar to the MLP, in which all hidden layer weights are shared by the mean and variance output nodes.

Variance in Allele Frequencies Affects donni Training and Performance

Since the AFS is the key input data in our method, we first considered how different levels of variance in the training data AFS might affect training and performance of the MVE networks underlying donni. While the expected AFS computed by dadi under a given set of demographic model parameter gives the mean value of each AFS entry, AFS summarized from observed data will have some variance. We asked whether training the network on AFS with some level of variance or AFS with no variance would lead to better overall performance. When generating AFS simulations, we modeled such variance in the AFS by Poisson-sampling from the expected AFS (examples in supplementary figs. S1 and S2B-D, Supplementary Material online.) We implemented four levels of AFS data variance: none, low, moderate, and high in AFS used for training and testing. We then surveyed the inference accuracy for all pairwise combinations for each type of variance in training sets versus test sets.

Overall, we found that networks trained on AFS with no to moderate level of variance perform similarly across all variance levels in test AFS (supplementary figs. S3–S6, Supplementary Material online for the split-migration model). High variance in training AFS led to substantially poorer inference accuracy in parameters that are more difficult to infer, such as time and migration rate. The population size change and ancestral state misidentification parameters were the least affected by AFS variance, and inference accuracy remained similarly high across all variance scenarios. For the time parameter, training on AFS with moderate variance produced the best-performing accuracy across all test cases (supplementary fig. S4, Supplementary Material online). However, for the migration rate parameter, training on AFS with no variance produced the overall best-performing accuracy (supplementary fig. S5, Supplementary Material online). We concluded that for subsequent analyses and model library production for donni, we would train using AFS with no variance, since there was no clear benefit from adding an extra variance simulation step in training. For test AFS, we would use AFS with moderate level of data variance to better match real data.

donni is Efficient and has Comparable Inference Accuracy to dadi

Since we built donni to be an alternative to dadi's likelihood optimization, we compared with dadi in our performance analysis. We validated the inference accuracy of donni for three models: a two-population model with an ancestral population split and symmetric migration

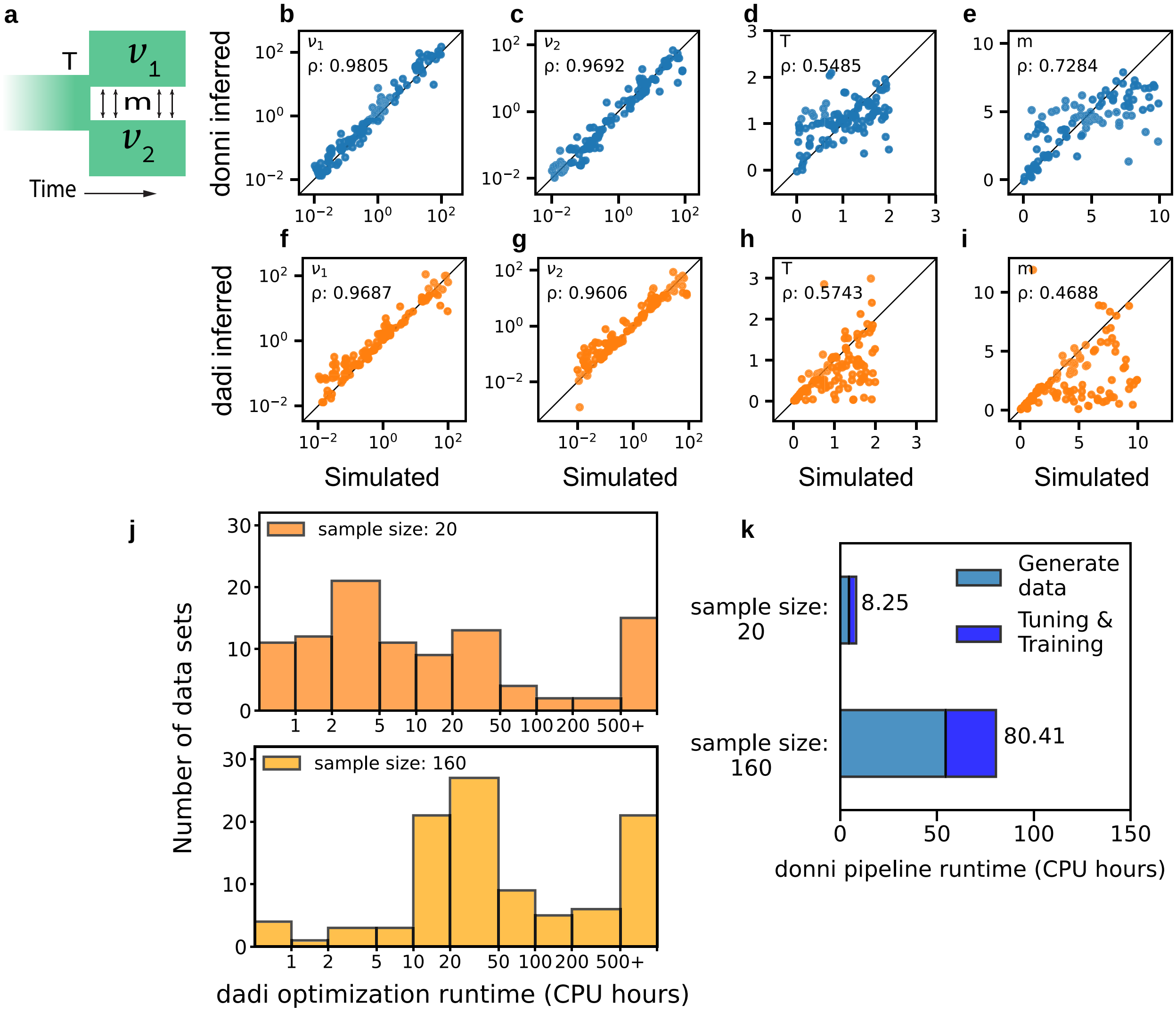

Fig. 2. Inference accuracy and computing time of donni and dadi for a two-population model. a) The two-population split-migration model with four parameters: ν_1 and ν_2 are relative sizes of each population to the ancestral, T is time of split, and m is the migration rate. b-i) Inference accuracy by donni b-e) and dadi f-i) for the four parameters on 100 test AFS (sample size of 20 haplotypes). j) Distribution of optimization times among test datasets for dadi. k) Computing time required for generating donni's trained networks for two sample sizes. Generate data include computing time for generating 5,000 dadi-simulated AFS as training data. Tuning & training is the total computing time for hyperparameter tuning and training the MVE network using the simulated data.

between the populations (split-migration model, Fig. 2a), a one-population model with one size change event (two epoch model, Fig. 3a), and a three-population model for human migration out-of-Africa (the OOA model, Fig. 5a) from Gutenkunst et al. (2009). We also compared the computational efficiency of donni and dadi for two different sample sizes of the split-migration model.

For the split-migration model, donni was able to infer all demographic history parameters with accuracy comparable to dadi (Fig. 2b-i). The population size change parameters ν_1 and ν_2 were inferred very well by both donni (Fig. 2b,c) and dadi (Fig. 2f,g). The time parameter T (Fig. 2d,h) and migration rate m (Fig. 2e,i) were more difficult to accurately infer for both methods, with dadi having trouble optimizing parameter values close to the specified parameter boundary (Fig. 2e). We used Spearman's correlation coefficient ρ to quantify the monotonic relationship between the true and the inferred parameter values, similar to Flagel et al. (2019). For a more direct measurement of inference accuracy, we also provide the RMSE scores for all models in supplementary table S1, Supplementary Material online.

To compare the efficiency of donni and dadi, we benchmarked the computational resources required by each method to infer demographic parameters from the same 100 test AFS (Fig. 2j-k). Since inferring parameters with donni's trained networks is computationally trivial, we instead measured the resources required by donni to generate trained networks. For both methods, computation was substantially more expensive as the sample size increased from 20 haplotypes to 160. For dadi (Fig. 2j), there was a spread of optimization runtime among the 100 test AFS, with several difficult spectra requiring more than 500 CPU hours to

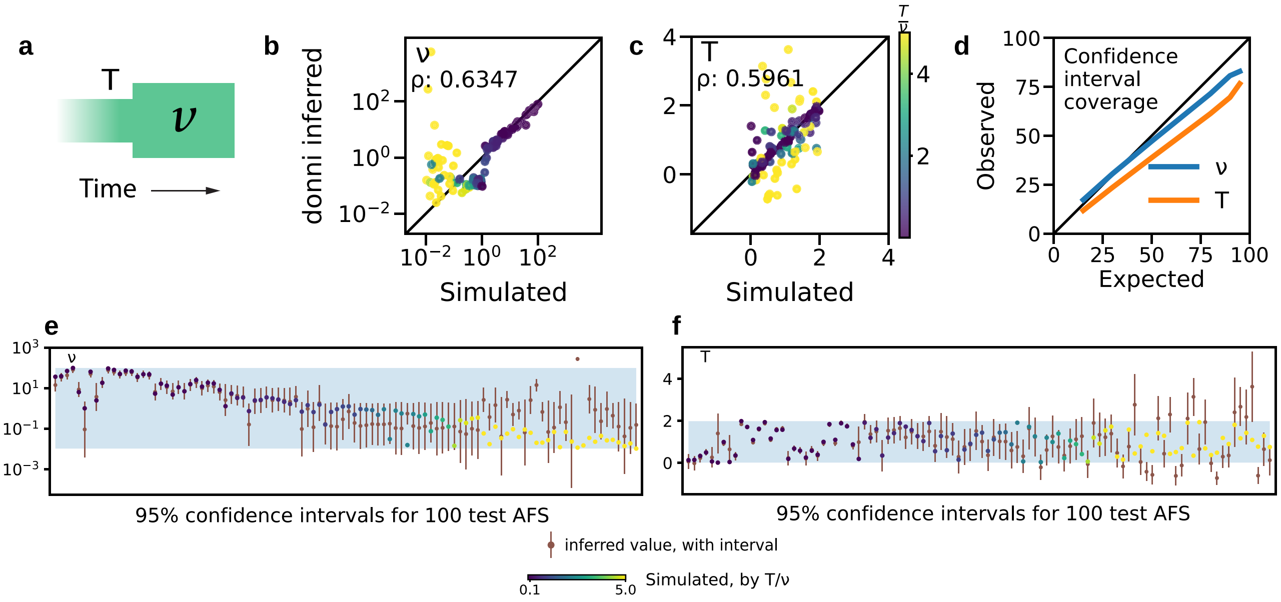

Fig. 3. Uninformative AFS affecting inference accuracy and uncertainty quantification method validation. a) The one-population two-epoch model with two parameters, ν for size change and T for time of size change. b-c) Inference accuracy for ν and T by donni on 100 test AFS, colored by simulated $\frac{T}{\nu}$ values. d) Confidence interval coverage for ν and T by donni. The observed coverage is the percentage of test AFS that have the simulated parameter values captured within the corresponding expected interval. e-f) As an example, we show details of the 95% confidence interval data points from panel d for 100 test AFS. The simulated values for ν e) and T f) of these AFS are colored by their $\frac{T}{\nu}$ values, similar to panels b-c. donni's inferred parameter values and 95% confidence interval outputs are in brown. The percentage of simulated color dots lying within donni's inferred brown interval gives the observed coverage at 95%. The light shades are the simulated parameter range (supplementary table S2, Supplementary Material online) used in simulating training and test AFS. The 100 test AFS are sorted along the x axis using true $\frac{T}{\nu}$ values.

reach convergence for both sample sizes. By comparison, the computation required for donni (Fig. 2k), including generating training data with dadi, hyperparameter tuning and training, was less than the average time required for running dadi optimization on a single AFS. This result suggests that donni may benefit many cases where dadi optimization can take a long time to reach convergence.

Figure 2k also suggests that generating the expected AFS with dadi is computationally expensive, often equivalent to if not more so than tuning and training a network. Such expensive operations are indeed what we aimed to minimize with donni. During each dadi optimization, a large number of expected AFS are also calculated. As opposed to discarding all these expensive calculations after each dadi optimization, donni's trained network effectively captures the mapping between the expected AFS and demographic history model parameter values in its network weights, which can be reused instantaneously in the future.

donni Accurately Estimates Uncertainty of Inferred Parameter Values

Sometimes, demographic model parameters may be unidentifiable, because multiple parameter sets generate nearly identical AFS. As a simple example, we considered the one-population two epoch model (Fig. 3a), which is parameterized by the relative size ν of the contemporary population and the time at which the population size changed T. For this model, donni inferences are inaccurate when T/ν is large (Fig. 3b-c). In this parameter regime, over the time T after the size change, the AFS relaxes back to that of an constant-sized equilibrium population. Therefore, in this case, the true parameters are unrecoverable because the AFS itself does not have the appropriate signal to infer them. While this problem may be avoided if users follow the best practice for model selection of exploring simpler models before complex ones (Marchi et al. 2021), it also highlights the need for uncertainty quantification, where a wide confidence interval would appropriately indicate problematic inference.

Using the prediction variance output from the trained MVE networks, donni can calculate any range of confidence intervals specified by the user for each inferred parameter. We validated our uncertainty quantification approach by measuring the observed coverage for six confidence intervals: 15, 30, 50, 75, 80, and 95% intervals (details in Materials and Methods). For the two-epoch model, our approach provided well-calibrated confidence intervals (Fig. 3d). Considering individual test AFS, the uninformative AFS yielded appropriately wide confidence intervals (Fig. 3e-f, yellow points). We found that confidence intervals were similarly well calibrated for the split-migration model (supplementary fig. S7, Supplementary Material online).

donni is not Biased by Linkage between Alleles

The Poisson Random Field model underlying dadi (Sawyer and Hartl 1992) and thus donni assumes independence of all genomic loci in the data, which is equivalent to assuming infinite recombination between any pair of loci. But loci within close proximity on the same chromosome are

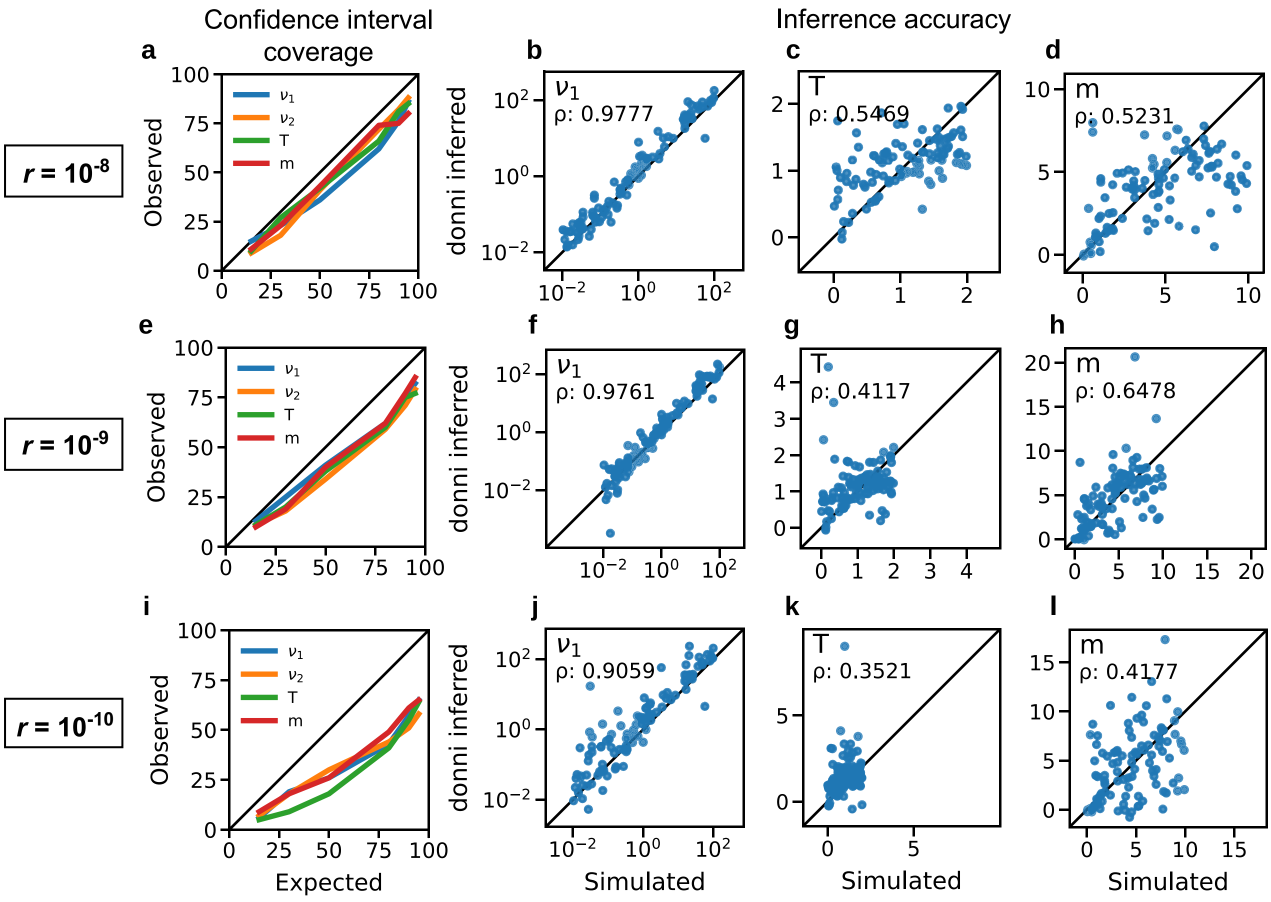

Fig. 4. donni's inference accuracy and uncertainty quantification coverage on msprime-simulated test AFS with linkage. Each row shows the confidence interval coverage and inference accuracy for select parameters of the split-migration demographic model (Fig. 2a) at varying recombination rate. Recombination rate decreases from top to bottom row, corresponding to increased linkage and data variance in the msprime-simulated test AFS. The same networks (trained on dadi-simulated AFS) were used in this analysis as in Fig. 2f-i.

likely sorted together during recombination and therefore linked. To assess how linkage affects donni inference, we tested donni's networks that were trained on dadi-simulated AFS without linkage on test AFS simulated with msprime, a coalescent simulator that includes linkage (Baumdicker et al. 2022). These msprime-simulated test AFS (examples in supplementary figs. S1B and S2E-G, Supplementary Material online) represent demographic scenarios similar to those in dadi but also include varying levels of linkage under a range of biologically realistic recombination rates. Since smaller recombination rates lead to more linkage and further departure from the training data assumption, we tested donni on AFS with decreasingly small recombination rates down to $r = 10^{-10}$ crossover per base pair per generation, which is two orders of magnitude smaller than the average recombination rate in humans.

Population size parameters ν were inferred well no matter the recombination rate, but the inference accuracy for T and m decreased as the recombination rate decreased (Fig. 4). Confidence intervals were well calibrated at the higher recombination rates (Fig. 4a,e), but too small at the lowest recombination rate (Fig. 4i). These patterns are similar to those we found when testing the effects of AFS variance by Poisson-sampling from expected AFS with dadi (supplementary figs. S3–S7, Supplementary Material online), where accuracy decreased with higher variance, and confidence intervals were underestimated at the highest variances. Note that at $r = 10^{-10}$, linkage disequilibrium often extends entirely across the simulated test regions, so in this regime methods assuming zero recombination, such as IMa3 (Hey et al. 2018), may be more appropriate. Importantly, even though more linkage did lead to higher prediction variance in the estimated parameter values, we did not observe bias in our inferences.

Comparison with dadi for the OOA Model

We tested donni on the three-population OOA model with six size change parameters, four migration rates, and three time parameters (Fig. 5a). In general, we observed a similar pattern to previous models; size change parameters were often easier to infer than times or migration rates (Fig. 5). For example, both donni and dadi showed near perfect inference accuracy for ν_{Af} (Fig. 5b,g). They both also performed well for the for ν_{Eu}, ν_{As}, and *misid* parameters (supplementary fig. S8, Supplementary Material online). But several parameters were challenging for both methods, including some size change parameters,

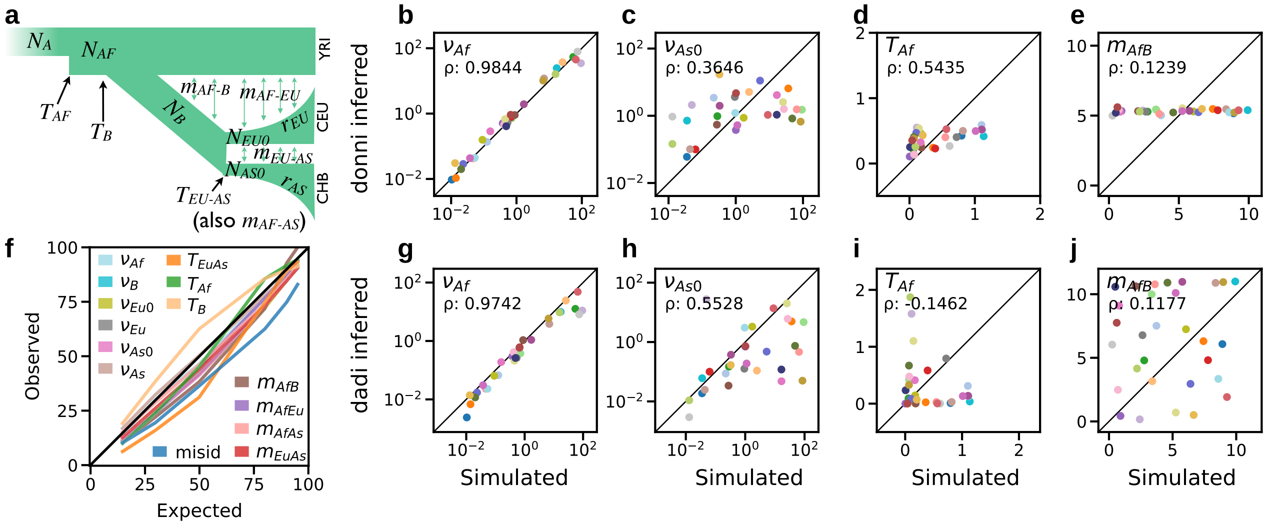

Fig. 5. Inference accuracy compared with dadi and confidence interval coverage by donni for the OOA demographic model. a) The three-population OOA model with 14 demographic history parameters. b-e) Inference accuracy for representative parameters on 30 simulated test AFS inferred by donni. g-j) Inference accuracy for the same parameters and 30 test AFS inferred by dadi. Each of the 30 test AFS is represented by a different color dot. For the accuracy of the rest of the parameters see supplementary fig. S8 and table S1, Supplementary Material online. f) donni confidence interval coverage for all model parameters.

such as ν_{As0} (Fig. 5c,h), ν_B, and ν_{Eu0} (supplementary fig. S8, Supplementary Material online). The time parameters proved to be the most challenging with relatively lower accuracy for both methods, with T_{Af} (Fig. 5d,i) and T_B (supplementary fig. S8, Supplementary Material online) being particularly difficult. Overall, both methods agree on the parameters that are easy versus difficult to infer.

However, when inference accuracy is poor on difficult parameters, dadi and donni tend to have different failure patterns. For instance with the m_{AfB} parameter, dadi tended to get stuck at the parameter boundaries for many AFS (Fig. 5j), while donni essentially inferred the average value for all test AFS (Fig. 5e). This indicates a failure by donni to learn any information from the training AFS for this particular parameter. For all other migration rate parameters in the model, donni performs well, matching dadi (supplementary fig. S8, Supplementary Material online).

While performance varied between the two methods among parameters, donni still had comparable accuracy to dadi in most cases. donni was also able to produce well calibrated confidence intervals for all parameters (Fig. 5f). Due to the computational expense of dadi optimization for this model, we only analyzed 30 test AFS for direct comparison between donni and dadi. Since donni is not as computationally constrained, we also tested donni on all 1,000 test AFS per our standard procedure, finding similar results (supplementary table S1, Supplementary Material online).

Finally, we investigated the empirical AFS data from Gutenkunst et al. (2009) using donni's trained MVE networks for the OOA demographic model (supplementary table S3, Supplementary Material online). We found that donni's estimates differ from dadi's to varying degrees across the parameters. The similarity in accuracy pattern between donni and dadi in Fig. 5 and supplementary fig. S8, Supplementary Material online does not translate to similar inference values between the two approaches on these data. For example, donni and dadi have similarly high accuracy patterns for ν_{As} but have very different estimates on the empirical AFS data ($\nu_{As} = 7.29$ for dadi and $\nu_{As} = 1.276$ for donni). For this model, donni also tends to infer a stronger migration rate than dadi does, with a higher estimate across all four migration rate parameters. Despite these differences in the estimated parameter values, dadi's estimates are within donni's 95% confidence intervals for all parameters.

donni's Trained Networks are Accessible

Given its speed, we expect that donni will be useful for quickly exploring many demographic scenarios given a user's dataset. To support this, we have produced trained networks for a large collection demographic history models. These include five one-population and eight two-population models from the current dadi API, plus the 34 two-population and 33 three-population models from Portik et al. (2017). For each of these models, we provide trained networks for unfolded and folded AFS for each of five sample sizes (only two sample sizes for three-population models). For large-scale production, we developed a comprehensive command-line interface pipeline for generating training data, tuning hyperparameters, and assessing the quality of the trained networks. donni's pipeline is open-source and available on GitHub (https://github.com/lntran26/donni) for users interested in training custom models. The trained network library is available on CyVerse (Center 2011; Merchant et al. 2016) and donni's command-line interface will automatically download appropriate networks. The library also includes all accuracy and confidence interval coverage plots for all supported demographic history models.

Discussion

We addressed dadi's computationally intensive optimization procedure by developing donni, a new inference method based on a supervised ML algorithm, the MVE network. We found that donni's trained MVE networks can instantaneously infer many demographic history parameters with accuracy comparable to dadi on simulated data. Even when including computing time required for training the network networks, for many cases donni is faster than dadi's maximum-likelihood optimization. Users are also provided a confidence interval for each inferred demographic history model parameter value from donni. Through examples of one-, two-, and three-population demographic models, we demonstrated that donni's uncertainty quantification method works well for a wide range of demographic parameters. We also showed that donni works well for AFS simulated by msprime, which includes linkage.

Our approach of using supervised ML to reduce the computational expense of the maximum-likelihood optimization step is similar in spirit to Smith et al. (2017) using RFs to improve the efficiency of the computationally intensive ABC procedure. While Smith et al. (2017) developed a classification approach for demographic model selection, our method is a regression approach, where we provide a suite of pretrained regressors for many commonly used demographic history models. Users can quickly explore many possible scenarios and get an estimate for several demographic parameters based on their input AFS data. However, we caution users to always start with simpler models first before trying more complex ones, to avoid exacerbating the uninformative parameter space problem. While we have implemented an accompanying uncertainty quantification tool to aid in identifying such problematic scenarios, best practices in model-based inference should still be followed.

Our choice of AFS as input data for training the network algorithm has several limitations. First, because the size of the AFS depends on the sample size but the network requires a fixed input size, we have to train a different set of networks for different sample sizes within the same demographic history model. Different sets of trained networks are also required for unfolded versus folded AFS. We have limited our trained network library to sample sizes of 10, 20, 40, 80, and 160 haplotypes per population. User data that does not match exactly these sample sizes will be automatically down-projected (Marth et al. 2004) by donni to the closest available option, leading to some data loss. It is, however, possible to use donni's pipeline to train custom models that can support a different sample size. We also verified that donni still provides accurate inference and well-calibrated confidence intervals on down-projected data (supplementary fig. S9, Supplementary Material online).

Second, for optimal network performance, we need to normalize the AFS data for training, leading to the loss of information about the parameter $\theta = 4N_a\mu L$, where N_a is the ancestral effective population size, μ is the mutation rate, and L is the sequence length. Estimating θ is required for converting all demographic parameters in genetic units to absolute population sizes and divergence times. While donni can provide a point estimate for θ, it cannot provide the uncertainty, which is necessary for estimating the uncertainty of absolute parameter values. This limitation can be overcome with a hybrid approach between donni and dadi, where donni's inferred parameter outputs become the starting point for dadi's optimization procedure and uncertainty estimation (Coffman et al. 2016). While this approach requires running likelihood optimization, a good starting value provided by donni should reduce overall computing time. donni trains a separate MVE network for each parameter in a given demographic history model, even though the model parameters are correlated. This is a limitation of our implementation, because the canonical MVE network architecture includes only one node for the prediction mean and one node for the prediction variance. It may be possible to add additional nodes to output prediction means, variances, and covariances from a single network, but we found that this often affects the overall inference quality of the trained MVE network. Additionally, we tested an alternative multioutput regression approach (the scikit-learn Multilayer-Perceptron Regressor) and found that our single-output approach provided similarly accurate estimates (supplementary fig. S10, Supplementary Material online). To our knowledge, existing methods for estimating uncertainties of multioutput neural network regressions are limited.

At its heart, the neural network approach of donni corresponds to a nonlinear regression of model parameters on AFS entries, in contrast to existing approaches which typically maximize a composite likelihood through optimization. Neural networks can be used to estimate likelihoods (e.g. Tejero-Cantero et al. 2020), which could then be optimized or sampled over, but here we prefer the more direct regression approach. Although dadi and donni display comparable overall accuracy (Figs. 2 and 5), they may differ when applied to any given dataset (supplementary table S3, Supplementary Material online), reflecting differences between regression and composite-likelihood optimization.

In conclusion, our results indicate that using supervised ML algorithms trained with AFS data is a computationally efficient approach for inferring demographic history from genomic data. Despite implementation limitations discussed above, the AFS is fast to simulate compared with other types of simulated data such as genomic sequence images (Flagel et al. 2019; Sanchez et al. 2021) or coalescent trees (Kelleher et al. 2016; Baumdicker et al. 2022). Furthermore, while ignoring linkage may be a weakness of AFS-based methods, it can also be a strength in that it is more species-agnostic and therefore trained models are transferable among species. A major challenge for AFS-based methods such as ours is the poor scaling to large sample sizes and number of populations, where the AFS matrix becomes high dimensional and sparse, and simulation becomes prohibitively expensive. While we limited this study to three-population models, there have been major improvements in AFS-based methods that can handle more (Jouganous et al. 2017; Kamm et al. 2017, 2020; Gutenkunst 2021). Given our results, a supervised ML approach might be a promising next step to extend to such AFS-based methods to further improve their computational efficiency.

Materials and Methods

Simulations with dadi

We used dadi v.2.3.0 (Gutenkunst et al. 2009) to simulate AFS for training and testing the networks. For each demographic model, we uniformly drew parameter sets from a biologically relevant range of parameters (supplementary table S2, Supplementary Material online). We then generated each expected AFS by specifying the demographic model and parameters in dadi. We calculated the extrapolation grid points used for dadi integration based on the number of haplotypes per population according to Gutenkunst (2021) for one-population models. For models with more than one population, we used the same formula but also increased the grid points by a factor of 1.5 for each additional population. The demographic model parameter values are used as labels for the generated AFS data. To simulate AFS with different levels of data variance, we started with the original expected AFS set (no variance). We then Poisson-sampled from the expected AFS to generate a new AFS with variance. We controlled the level of variance by the parameter θ, by which we multiplied the expected AFS before sampling. We used $\theta = 10{,}000$, 1,000, and 100 corresponding to low, moderate, and high levels of variance, respectively (supplementary figs. S3–S7, Supplementary Material online). Intuitively, modifying $\theta = 4N_a\mu L$ is equivalent to altering the effective number of sites surveyed L. Assuming $\mu \sim 10^{-8}$ and $N_a \sim 10^4$, $\theta = 1{,}000$ is equivalent to $L \sim 2.5 \times 10^6$ sites. Smaller θ is equivalent to fewer sites surveyed, hence noisier AFS. Finally, we normalized both expected and Poisson-sampled AFS for training and testing. The results shown in Figs. 2, 3, 5 and supplementary fig. S8, Supplementary Material online are based on unfolded AFS with sample size of 20 haplotypes per population.

Simulations with msprime

We used msprime v1.2.0 (Baumdicker et al. 2022) to simulate AFS from demographic history models while including linkage. We first specified dadi-equivalent demography in msprime for the two epoch and split-migration models. This included the population size change ratio ν and time of change T parameters for the two epoch model, and population size change ratios ν_1 and ν_2, time T of split, and migration rate m for the split-migration model. We then specified additional parameters required for msprime to yield $\theta = 4N_A L\mu = 40{,}000$, with ancestral population size $N_A = 10{,}000$, sequence length $L = 10^8$ base pairs, and mutation rate $\mu = 10^{-8}$ per base pair per generation. We used three recombination rates 10^{-8}, 10^{-9}, and 10^{-10} per base pair per generation to simulate different levels of linkage and variance in the AFS. We then generated tree-sequence data with msprime before converting to the corresponding unfolded AFS of sample size 20 haplotypes per population and normalizing for testing with trained networks.

Network Architecture and Hyperparameter Optimization

We used TensorFlow v2.12.1 and Keras v2.12.0 to generate all-trained MVE networks for donni. These networks have two fully connected hidden layers containing between 4 and 64 neurons. The exact number of neurons in each hidden layer are hyperparameters that were automatically selected via our tuning procedure described below. The input layer is a flattened AFS with varying sizes depending on the sample size and whether it is a folded or unfolded AFS. The output layer has two nodes for the prediction mean and variance of one demographic history parameter. For tuning and training the network, we implemented a custom loss function based on the negative log-likelihood of a normal distribution:

$$L(\theta) = \sum_{i=1}^{N} \frac{1}{2} \log\left(\sigma_\theta^2(x_i)\right) + \frac{1}{2} \frac{(y_i - \mu_\theta(x_i))^2}{\sigma_\theta^2(x_i)}.$$

Here θ denotes the set of network parameters (edge weights and node biases), and the sum is over training data AFS x_i corresponding to true demographic parameter values y_i. For each training AFS, the MVE network outputs a prediction mean $\mu_\theta(x_i)$ and variance $\sigma_\theta^2(x_i)$. During training, the network parameters θ are optimized to minimize the loss function, thus both improving parameter prediction accuracy (through μ_θ) and uncertainty estimation (through σ_θ^2).

For automatic hyperparameter tuning, we used the HyperBand and RandomSearch tuning algorithms available in keras-tuner v.1.4.6. The 5,000 AFS training dataset was split 80% for training and 20% for validation. For a given network, we first used HyperBand to optimize both the hidden layer size and learning rate. We then kept the MVE network from HyperBand with the best performance on the validation data, froze the hidden layer size, and then continued tuning only the learning rate using RandomSearch. The MVE network with the best performance on the validation data after RandomSearch is then selected for subsequent training on the full training data. All hyperparameter configurations and nondefault settings for the tuning algorithms are listed in supplementary table S4, Supplementary Material online.

Uncertainty Quantification Coverage

For uncertainty quantification, the trained MVE network outputs a prediction variance for each inferred demographic history parameter. The donni pipeline converts this variance to confidence intervals using the normal distribution. To validate our uncertainty quantification method, we first obtained the method's estimation for six confidence intervals, 15, 30, 50, 80, 90, and 95% on all test AFS. We then get the observed coverage by calculating the percentage of test AFS that have their corresponding simulated parameter value captured within the estimated interval. The expected versus observed percentages are plotted in our confidence interval coverage plots.

donni Training and Testing Pipeline

We used 5,000 AFS (no data variance) for training and tuning and 1,000 AFS (moderate data variance, $\theta = 1{,}000$) for accuracy and uncertainty coverage validation. For visualization, only 100 test AFS (30 AFS for the OOA model) are shown to compare with dadi. However, accuracy scores by donni on all 1,000 test AFS are provided in supplementary table S1, Supplementary Material online. Our pipeline tunes and trains one network for each demographic model parameter and sample size. For example, the two epoch model with two parameters v and T has 20 independently trained networks: 2 networks for v and T times 5 supported sample sizes times 2 polarization states.

Likelihood Optimization with dadi-cli

To infer demographic parameters for a large number of test AFS in parallel (100 AFS for the split-migration model and 30 AFS for the OOA model), we used dadi's command-line interface (Huang et al. 2023). We specified the upper and lower bound values for optimization based on the parameter range provided in supplementary table S2, Supplementary Material online. Optimization ran until convergence, as defined by $\delta log(L) = 0.0005$ for the OOA model and $\delta log(L) = 0.001$ for the split-migration model.

Benchmarking dadi Optimization and donni Pipeline

To benchmark the computational expense required for dadi optimization versus for training the networks, we used 10 CPUs on a single computing node for each task. For donni, the tasks are generating training AFS, hyperparameter tuning with HyperBand, and training using the tuned hyperparameters. Estimating demographic parameters for 100 test AFS with donni's trained networks is nearly instantaneous. For dadi, each test AFS is a task that was optimized until convergence, at which time was recorded, or until the specified cut-off time (50 h × 10 CPUs =500 CPU h).

Supplementary Material

Supplementary material is available at *Molecular Biology and Evolution* online.

Acknowledgments

This work was supported by the National Institute of General Medical Sciences of the National Institutes of Health (R01GM127348 and R35GM149235 to R.N.G.).

Data Availability

No new data were generated or analysed in support of this research.

References

Achaz G. Frequency spectrum neutrality tests: one for all and all for one. *Genetics*. 2009:**183**(1):249–258. https://doi.org/10.1534/genetics.109.104042.

Baharian S, Gravel S. On the decidability of population size histories from finite allele frequency spectra. *Theor Popul Biol*. 2018:**120**: 42–51. https://doi.org/10.1016/j.tpb.2017.12.008.

Baumdicker F, Bisschop G, Goldstein D, Gower G, Ragsdale AP, Tsambos G, Zhu S, Eldon B, Ellerman EC, Galloway JG, et al. Efficient ancestry and mutation simulation with msprime 1.0. *Genetics*. 2022:**220**(3): iyab229. https://doi.org/10.1093/genetics/iyab229.

Bergström A, McCarthy SA, Hui R, Almarri MA, Ayub Q, Danecek P, Chen Y, Felkel S, Hallast P, Kamm J, et al. Insights into human genetic variation and population history from 929 diverse genomes. *Science*. 2020:**367**(6484):eaay5012. https://doi.org/10.1126/science.aay5012.

Bhaskar A, Song YS. Descartes' rule of signs and the identifiability of population demographic models from genomic variation data. *Ann Stat*. 2014:**42**(6):2469–2493. https://doi.org/10.1214/14-AOS1264.

Boyko AR, Williamson SH, Indap AR, Degenhardt JD, Hernandez RD, Lohmueller KE, Adams MD, Schmidt S, Sninsky JJ, Sunyaev SR, et al. Assessing the evolutionary impact of amino acid mutations in the human genome. *PLoS Genet*. 2008:**4**(5):e1000083. https://doi.org/10.1371/journal.pgen.1000083.

Center D. The iPlant collaborative: cyberinfrastructure for plant biology. In: Chardon M, Vandewalle P (1991) Acoustico-Lateralis System Cyprinid Fishes; 2011.

Chavez DE, Gronau I, Hains T, Dikow RB, Frandsen PB, Figueiró HV, Garcez FS, Tchaicka L, de Paula RC, Rodrigues FH, et al. Comparative genomics uncovers the evolutionary history, demography, and molecular adaptations of South American Canids. *Proc Natl Acad Sci USA*. 2022:**119**(34):e2205986119. https://doi.org/10.1073/pnas.2205986119.

Coffman AJ, Hsieh PH, Gravel S, Gutenkunst RN. Computationally efficient composite likelihood statistics for demographic inference. *Mol Biol Evol*. 2016:**33**(2):591–593. https://doi.org/10.1093/molbev/msv255.

Excoffier L, Dupanloup I, Huerta-Sánchez E, Sousa VC, Foll M. Robust demographic inference from genomic and SNP data. *PLoS Genet*. 2013:**9**(10):e1003905. https://doi.org/10.1371/journal.pgen.1003905.

Flagel L, Brandvain Y, Schrider DR. The unreasonable effectiveness of convolutional neural networks in population genetic inference. *Mol Biol Evol*. 2019:**36**(2):220–238. https://doi.org/10.1093/molbev/msy224.

Gopalan S, Berl RE, Myrick JW, Garfield ZH, Reynolds AW, Bafens BK, Belbin G, Mastoras M, Williams C, Daya M, et al. Hunter-gatherer genomes reveal diverse demographic trajectories during the rise of farming in Eastern Africa. *Curr Biol*. 2022:**32**(8):1852–1860. https://doi.org/10.1016/j.cub.2022.02.050.

Gutenkunst RN. Dadi. cuda: accelerating population genetics inference with graphics processing units. *Mol Biol Evol*. 2021:**38**(5): 2177–2178. https://doi.org/10.1093/molbev/msaa305.

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**(10):e1000695. https://doi.org/10.1371/journal.pgen.1000695.

Hey J, Chung Y, Sethuraman A, Lachance J, Tishkoff S, Sousa VC, Wang Y. Phylogeny estimation by integration over isolation with migration models. *Mol Biol Evol*. 2018:**35**(11):2805–2818. https://doi.org/10.1093/molbev/msy162.

Huang X, Struck TJ, Davey SW, Gutenkunst RN. dadi-cli: automated and distributed population genetic model inference from allele frequency spectra. 2023.

Johnston HR, Cutler DJ. Population demographic history can cause the appearance of recombination hotspots. *Am J Hum Genet*. 2012:**90**(5):774–783. https://doi.org/10.1016/j.ajhg.2012.03.011.

Jouganous J, Long W, Ragsdale AP, Gravel S. Inferring the joint demographic history of multiple populations: beyond the diffusion approximation. *Genetics*. 2017:**206**(3):1549–1567. https://doi.org/10.1534/genetics.117.200493.

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**(531):1472–1487. https://doi.org/10.1080/01621459.2019.1635482.

Kamm JA, Terhorst J, Song YS. Efficient computation of the joint sample frequency spectra for multiple populations. *J Comput Graph Stat*. 2017:**26**(1):182–194. https://doi.org/10.1080/10618600.2016.1159212.

Kelleher J, Etheridge AM, McVean G. Efficient coalescent simulation and genealogical analysis for large sample sizes. *PLoS Comput Biol*. 2016:**12**(5):e1004842. https://doi.org/10.1371/journal.pcbi.1004842.

Kern AD, Hey J. Exact calculation of the joint allele frequency spectrum for isolation with migration models. *Genetics*. 2017:**207**(1):241–253. https://doi.org/10.1534/genetics.116.194019.

Khosravi A, Nahavandi S, Creighton D, Atiya AF. Comprehensive review of neural network-based prediction intervals and new advances. *IEEE Trans Neural Netw*. 2011:**22**(9):1341–1356. https://doi.org/10.1109/TNN.2011.2162110.

Kim BY, Huber CD, Lohmueller KE. Inference of the distribution of selection coefficients for new nonsynonymous mutations using large samples. *Genetics*. 2017:**206**(1):345–361. https://doi.org/10.1534/genetics.116.197145.

Lorente-Galdos B, Lao O, Serra-Vidal G, Santpere G, Kuderna LF, Arauna LR, Fadhlaoui-Zid K, Pimenoff VN, Soodyall H, Zalloua P, 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**:1. https://doi.org/10.1186/s13059-019-1684-5.

Lukić S, Hey J. Demographic inference using spectral methods on SNP data, with an analysis of the human out-of-Africa expansion. *Genetics*. 2012:**192**(2):619–639. https://doi.org/10.1534/genetics.112.141846.

Marchi N, Schlichta F, Excoffier L. Demographic inference. *Curr Biol*. 2021:**31**(6):R276–R279. https://doi.org/10.1016/j.cub.2021.01.053.

Marchi N, Winkelbach L, Schulz I, Brami M, Hofmanová Z, Blöcher J, Reyna-Blanco CS, Diekmann Y, Thiéry A, Kapopoulou A, et al. The genomic origins of the worldś first farmers. *Cell*. 2022. https://doi.org/10.1016/j.cell.2022.04.008.

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**(1):351–372. https://doi.org/10.1534/genetics.166.1.351.

Mathieson I, McVean G. Differential confounding of rare and common variants in spatially structured populations. *Nat Genet*. 2012:**44**(3):243–246. https://doi.org/10.1038/ng.1074.

Mays Jr HL, Hung CM, Shaner PJ, Denvir J, Justice M, Yang SF, Roth TL, Oehler DA, Fan J, Rekulapally S, et al. Genomic analysis of demographic history and ecological niche modeling in the endangered sumatran rhinoceros dicerorhinus sumatrensis. *Curr Biol*. 2018:**28**(1):70–76.e4. https://doi.org/10.1016/j.cub.2017.11.021.

Merchant N, Lyons E, Goff S, Vaughn M, Ware D, Micklos D, Antin P. The iPlant collaborative: cyberinfrastructure for enabling data to discovery for the life sciences. *PLoS Biol*. 2016:**14**(1):e1002342. https://doi.org/10.1371/journal.pbio.1002342.

Miller-Butterworth CM, Diefenbach DR, Edson JE, Hansen LA, Jordan JD, Gingery TM, Russell AL. Demographic changes and loss of genetic diversity in two insular populations of bobcats (Lynx rufus). *Glob Ecol Conserv*. 2021:**26**:e01457. https://doi.org/10.1016/j.gecco.2021.e01457.

Mondal M, Bertranpetit J, Lao O. Approximate Bayesian computation with deep learning supports a third archaic introgression in Asia and Oceania. *Nat Commun*. 2019:**10**:1. https://doi.org/10.1038/s41467-018-08089-7.

Myers S, Fefferman C, Patterson N. Can one learn history from the allelic spectrum? *Theor Popul Biol*. 2008:**73**(3):342–348. https://doi.org/10.1016/j.tpb.2008.01.001.

Naduvilezhath L, Rose LE, Metzler D. Jaatha: a fast composite-likelihood approach to estimate demographic parameters. *Mol Ecol*. 2011:**20**(13):2709–2723. https://doi.org/10.1111/j.1365-294X.2011.05131.x.

Nielsen R, Williamson S, Kim Y, Hubisz MJ, Clark AG, Bustamante C. Genomic scans for selective sweeps using SNP data. *Genome Res*. 2005:**15**:1566–1575. https://doi.org/10.1101/gr.4252305.

Nix DA, Weigend AS. Estimating the mean and variance of the target probability distribution. In: *Proceedings of 1994 IEEE International Conference on Neural Networks (ICNN'94)*. Vol. 1. IEEE; 1994. p. 55–60.

Portik DM, Leaché AD, Rivera D, Barej MF, Burger M, Hirschfeld M, Rödel MO, Blackburn DC, Fujita MK. Evaluating mechanisms of diversification in a Guineo-Congolian tropical forest frog using demographic model selection. *Mol Ecol*. 2017:**26**(19):5245–5263. https://doi.org/10.1111/mec.2017.26.issue-19.

Pudlo P, Marin JM, Estoup A, Cornuet JM, Gautier M, Robert CP. Reliable abc model choice via random forests. *Bioinformatics*. 2016:**32**(6):859–866. https://doi.org/10.1093/bioinformatics/btv684.

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**(8):2645–2660. https://doi.org/10.1111/men.v21.8.

Sawyer SA, Hartl DL. Population genetics of polymorphism and divergence. *Genetics*. 1992:**132**(4):1161–1176. https://doi.org/10.1093/genetics/132.4.1161.

Schrider DR, Kern AD. Supervised machine learning for population genetics: a new paradigm. *Trends Genet*. 2018:**34**(4):301–312. https://doi.org/10.1016/j.tig.2017.12.005.

Sheehan S, Song YS. Deep learning for population genetic inference. *PLoS Comput Biol*. 2016:**12**(3):e1004845. https://doi.org/10.1371/journal.pcbi.1004845.

Sluijterman L, Cator E, Heskes T. Optimal training of mean variance estimation neural networks. arXiv 230208875. https://doi.org/10.48550/arXiv.2302.08875, 2023, preprint: not peer reviewed.

Smith ML, Ruffley M, Espíndola A, Tank DC, Sullivan J, Carstens BC. Demographic model selection using random forests and the site frequency spectrum. *Mol Ecol*. 2017:**26**(17):4562–4573. https://doi.org/10.1111/mec.2017.26.issue-17.

Spence JP, Steinrücken M, Terhorst J, Song YS. Inference of population history using coalescent HMMs: review and outlook. *Curr Opin Genet Dev*. 2018:**53**:70–76. https://doi.org/10.1016/j.gde.2018.07.002.

Tejero-Cantero A, Boelts J, Deistler M, Lueckmann JM, Durkan C, Gonçalves PJ, Greenberg DS, Macke JH. sbi: a toolkit for simulation-based inference. *J Open Source Softw*. 2020:**5**(52):2505. https://doi.org/10.21105/joss.

Terhorst J, Song YS. Fundamental limits on the accuracy of demographic inference based on the sample frequency spectrum. *Proc Natl Acad Sci USA*. 2015:**112**(25):7677–7682. https://doi.org/10.1073/pnas.1503717112.

Villanea FA, Schraiber JG. Multiple episodes of interbreeding between neanderthal and modern humans. *Nat Ecol Evol*. 2019:**3**:39–44. https://doi.org/10.1038/s41559-018-0735-8.