# Inference of population structure from ancient DNA

Tyler Joseph[1] and Itsik Pe'er[1,2,3]

[1] Department of Computer Science,
[2] Department of Systems Biology,
[3] Data Science Institute,
Columbia University, New York, NY 10027, USA
{tjoseph, itsik}@cs.columbia.edu

**Abstract.** Methods for inferring population structure from genetic information traditionally assume samples are contemporary. Yet, the increasing availability of ancient DNA sequences begs revision of this paradigm.We present Dystruct (Dynamic Structure), a framework and toolbox for inference of shared ancestry from data that includes ancient DNA, by explicitly modeling population history and genetic drift as a time-series, thereby by assigning most likely ancestry proportions under a realistic model. Formally, we use a normal approximation of drift, which allows a novel, efficient algorithm for optimizing model parameters using stochastic variational inference. We show that Dystruct outperforms the state of the art when individuals are sampled over time, as is common in ancient DNA datasets. We further demonstrate the utility of our method on a dataset of 92 ancient samples alongside 1941 modern ones genotyped at 222755 loci, and show it tends to present modern samples as the mixtures of ancestral populations they really are, rather than the artifactual converse of presenting ancestral samples as mixtures of contemporary groups.
**Availability:** Dystruct is implemented in C++, open-source, and available at
`https://github.com/tyjo/dystruct`.

# 1 Introduction

The sequencing of the first ancient human genome [27], first Denisovan genome [28], and first Neanthertal genome [27] — all in 2010 — has opened the floodgates for population genetic studies that include ancient DNA [18]. Ancient DNA grants a unique opportunity to investigate human evolutionary history, because it can provide direct evidence of historical relationships between populations around the world. Indeed, through combining ancient and modern samples, ancient DNA has driven many notable discoveries in human population genetics over the past ten years including the detection of introgression between anatomically modern humans and Neanderthals [23], evidence for the genetic origin of Native Americans [25], and evidence pushing the date of human divergence in Africa to over 250,000 years ago [29], among many others [2, 9, 18, 29, 32].

Nonetheless, incorporating new types of DNA into conventional analysis pipelines requires careful examination of existing models and tools. Ancient DNA is a particularly challenging example: individuals are sampled from multiple time points, from underlying populations that likely have experienced substantial genetic drift. Hence, allele frequencies are correlated over time. The current state of the art for historical inference uses pairwise summary statistics calculated from genome-wide data, called drift indices or F-statistics [20, 21], not to be confused with Wright's F-statistics, that measure the amount of shared genetic drift between pairs of populations. Drift indices have several desirable theoretical properties, such as unbiased estimators, and can be used to conduct hypothesis tests of historical relationships and admixture between sampled populations [20]. Combined with tree-building approaches from phylogenetics, drift indices can reconstruct complex population phylogenies [17] including admixture events that are robust to difference in sample times. However, computing drift indices requires identifying populations *a priori*, a challenging task given that multiple regions around the world experienced substantial population turnover.

One of the most ubiquitous approaches to genetic clustering is through the Pritchard-Stephens-Donnelly (PSD) model [22], implemented in the popular software programs *structure* and ADMIXTURE [1]. Under the PSD model, sampled individuals are modeled as mixtures of latent populations, where the genotype at each locus depends on the population of origin of that locus, and allele frequencies in the latent populations. Individuals can be clustered based on their mixture proportions, the proportion of sampled loci inherited from each population, which are interpreted as estimates of global ancestry [1]. ADMIXTURE computes maximum likelihood estimates of allele frequencies and ancestry proportions under the PSD model, while *structure* uses MCMC to compute posterior expectations of ancestry assignment. A key assumption of the PSD model is that populations are in Hardy-Weinberg equilibrium: the allele frequencies in each population are fixed. For ancient DNA, this assumption is clearly violated. The robustness of the PSD model to this violation remains under-explored.

In this paper, we develop a model-based method for inferring population structure and admixture of ancient and modern DNA – Dystruct (Dynamic Structure) – by extending the PSD model to time-series data. To efficiently infer model parameters, we leverage the close connection between the PSD model and another model from natural language processing: latent Dirichlet allocation [6]. The connection between the PSD model and LDA has long been known [3, 5], and applications of the statistical methodology surrounding LDA are beginning to enter the population genetics literature [11, 26]. Similar to the PSD model, LDA models documents as mixtures of latent topics, where each topic specifies a probability distribution over words. LDA has been successfully extended to a time-series model [5], where the word frequencies in topic distributions change over time in a process analogous to genetic drift. Thus, these dynamic topic models provide a natural starting point for models of population structure that incorporate genetic drift.

Our contributions are three-fold. First, we developed an efficient inference algorithm capable of parameter estimation under our time-series model. We combined the stochastic variational inference algorithm for the PSD model developed by [11] with variational Kalman filtering technique developed by [5]. We released software implementing our inference algorithm for general use. Second, we show that our model can lead to new insights on ancient DNA datasets: using simulations we demonstrate that Dystruct obtains more accurate ancestry estimates than ADMIXTURE on ancient DNA datasets; we then apply our model to a dataset of 92 ancient and 1941 modern samples genotyped at 222755 loci. Third, and more generally, our model opens the possibility for future model based approaches incorporating more complex demographic histories, complementing existing approaches for analyzing ancient DNA.

## 2 Methods

### 2.1 Preliminaries

Suppose we have genotypes of $D$ individuals across $L$ independent loci. Each individual $d$ is a vector of $L$ of the sum of two Bernoulli variables, $x_{dl} \in \{0, 1, 2\}$ counting the number of non-reference alleles at respective loci of $l = 1, \ldots, L$. Each individuals is assumed to have been alive during one of a finite set of time points $g[1], g[2], \ldots, g[T]$ measured as natural numbers of generations since the earliest time of interest. We further define $\varDelta g[t] = g[t] - g[t-1]$, the time in generations between time point $t$ and time point $t-1$. Each individual $d$ is assumed to have lived at specific time point $t_d \in \{1, 2, \ldots, T\}$.

Under the PSD model, each individual is a mixture from $K$ latent populations. Let $\boldsymbol{\theta}_d = (\theta_{d1}, ..., \theta_{dK})$ be the vector of probabilities that a locus in individual $d$ originated in population $k$. Thus, $\sum_k \theta_{dk} = 1$. Each population further has allele frequencies for each locus and each time point, $\beta_{kl}[t]$. Hence, $\boldsymbol{\beta_{kl}}$ is the vector of allele frequencies in population $k$ at locus $l$ over all time points. The model for generating genotypes for each individual is given by

$$\boldsymbol{\theta}_d \sim \text{Dirichlet}(\alpha_1, \alpha_2, ...\alpha_k) \tag{1}$$

$$x_{dl} \sim \text{Binomial}\left(2, \sum_k \theta_{dk}\beta_{lk}[t_d]\right) \tag{2}$$

This follows the recharacterization of the original PSD model by [1] and [11]. To generate an individual, we first draw mixture proportions $\boldsymbol{\theta}_d$ from a Dirichlet distribution with prior parameter $\boldsymbol{\alpha}$. Then we draw genotypes from a binomial distribution where the probability is a linear combination of the allele frequencies in the latent populations at time point $t_d$ (see Fig. 1).

To extend the model to time series data, we allow the allele frequencies to change at each time point using a normal approximation to genetic drift [7] (Fig. 1):

$$\beta_{kl}[t] \sim \text{Normal}\left(\beta_{kl}[t-1], \frac{\varDelta g[t]}{12N_k}\right) \tag{3}$$

where $N_k$ is the effective population size in population $k$. Initial allele frequencies $\beta_{kl}[0]$ and $N_k$ are parameters of the model. Initial allele frequencies $\beta_{kl}[0]$ are estimated from data, while $N_k$ are treated as known and fixed. Our simulations demonstrate that this is not a serious limitation.

The state space model here is slightly different than the traditional normal approximation to the binomial for genetic drift. Under the Wright-Fisher model of genetic drift, changes in allele frequencies can be viewed as a Markov Chain where alleles in future generations are chosen by

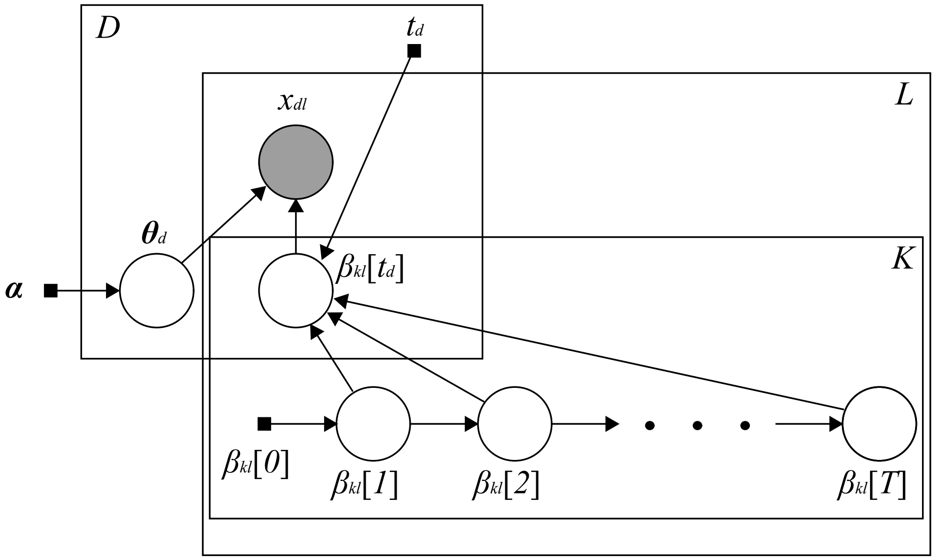


Fig. 1: Graphical model depicting Dystruct's generative model. $D$ individuals are genotyped at $L$ loci from $K$ populations. Each individual is is time stamped with a sample time point $t_d$. Each genotype in each individual, $x_{dl}$, is a binomial observation that depends on: i) ancestry proportions, $\boldsymbol{\theta}_d$, and allele frequencies $\beta_{kl}[t_d]$ at time point $t_d$. Allele frequencies, $\beta_{kl}[t]$, depend on the allele frequency at the previous time point $\beta_{kl}[t-1]$. Plates denote how many times each random variable is generated.

sampling from the previous generation with replacement. The total number of individuals with a particular allele, say $A$, one generation in the future is a Binomial$(2N_k, p)$ random variable, where $N_k$ is the population size, and $p$ is the fraction of individuals in the previous generation with allele $A$. Using the normal approximation to the binomial, and modeling allele frequencies rather than counts of individuals, the variance in allele frequencies $\Delta g[t]$ generations in the future given the current allele frequency $\beta_{kl}[t-1]$ is $\sigma^2 = \frac{\Delta g_t \beta_{kl}[t-1](1-\beta_{kl}[t-1])}{2N_k}$ We approximate $\sigma^2$ by averaging the variance over the interval $(0, 1)$. That is

$$\int_0^1 \frac{\Delta g[t]\beta_{kl}[t-1](1-\beta_{kl}[t-1])}{2N_k}\, d\beta_{kl}[t-1] = \frac{\Delta g[t]}{12N_k} \tag{4}$$

In practice, through simulations, we found that we were able to obtain accurate estimates despite this approximation. In this context, effective population size is not the parameter of interest, and we instead focus on ancestry estimates.

### 2.2 Posterior Inference

We take a Bayesian approach by inferring ancestry proportions through the posterior distribution $p(\boldsymbol{\theta}_{1:D}, \boldsymbol{\beta}_{1:K,1:L} | \boldsymbol{x}_{1:D,1:L})$. Direct posterior inference is intractable because the normal distribution is not a conjugate prior for the binomial. Following [5], we derive a variational inference algorithm that approximates the true posterior. We hereby summarize the variational inference approach for completion.

Variational inference methods [4,14,33] approximate the true posterior by specifying a computationally tractable family of approximate posterior distributions indexed by variational parameters,

$\boldsymbol{\phi}$. These parameters are then optimized to minimize the Kullback-Leibler (KL) divergence between the true posterior and its variational approximation. The key to variational inference algorithms relies on the observation that given some distribution of latent parameters $q(\boldsymbol{z})$, the log likelihood of the observations $\boldsymbol{x}$ can be decomposed into two terms:

$$\log p_{\boldsymbol{\eta}}(\boldsymbol{x}) = \int \log p_{\eta}(\boldsymbol{x}) q_{\boldsymbol{\phi}}(\boldsymbol{z})\, d\boldsymbol{z} = \int \log \left( \frac{p_{\boldsymbol{\eta}}(\boldsymbol{x},\boldsymbol{z})}{p_{\boldsymbol{\eta}}(\boldsymbol{z}|\boldsymbol{x})} \frac{q_{\boldsymbol{\phi}}(\boldsymbol{z})}{q_{\boldsymbol{\phi}}(\boldsymbol{z})} \right) q_{\boldsymbol{\phi}}(\boldsymbol{z})\, d\boldsymbol{z} \tag{5}$$

$$= \mathbb{E}_q \left[ \log \left( \frac{p_{\boldsymbol{\eta}}(\boldsymbol{x},\boldsymbol{z})}{q_{\boldsymbol{\phi}}(\boldsymbol{z})} \right) \right] + \mathbb{E}_q \left[ \log \left( \frac{q_{\boldsymbol{\phi}}(\boldsymbol{z})}{p_{\boldsymbol{\eta}}(\boldsymbol{z}|\boldsymbol{x})} \right) \right] = L(\boldsymbol{\eta},\boldsymbol{\phi};\boldsymbol{x}) + \mathrm{KL}\left( q_{\boldsymbol{\phi}}(\boldsymbol{z}) \big|\big| p_{\boldsymbol{\eta}}(\boldsymbol{z}|\boldsymbol{x}) \right) \tag{6}$$

where $\boldsymbol{\eta}$ are the model parameters and $\boldsymbol{\phi}$ are the variational parameters. The term on the right is the KL divergence between the true posterior and the variational approximation. The KL divergence is non-negative, hence the $L$ term in (6) is a lower bound on the log likelihood, called the evidence lower bound (ELBO). In practice, $L$ is maximized with respect to the variational parameters $\boldsymbol{\phi}$, minimizing the KL divergence between the true and approximate posterior, and potentially with respect to model parameters $\boldsymbol{\eta}$, giving maximum likelihood estimates of $\boldsymbol{\eta}$.

The posterior distribution in our model is given by

$$p(\boldsymbol{\beta}_{1:K,1:L}, \boldsymbol{\theta}_{1:D} | \boldsymbol{x}_{1:D,1:L}) \propto \prod_{d=1}^{D} p(\boldsymbol{\theta}_d) p(x_{dl} | \beta_{kl}[t_d], \boldsymbol{\theta}_d) \prod_{t=1}^{T} \prod_{k=1}^{K} \prod_{l=1}^{L} p(\beta_{kl}[t] | \beta_{kl}[t-1]) \tag{7}$$

which we approximate by the variational posterior

$$q(\boldsymbol{\beta}_{1:K,1:L}, \boldsymbol{\theta}_{1:D}) = \prod_{d=1}^{D} q(\boldsymbol{\theta}_d | \hat{\boldsymbol{\theta}}_d) \prod_{k=1}^{K} \prod_{l=1}^{L} q(\beta_{kl}[1], ..., \beta_{kl}[T] \big| \hat{\beta}_{kl}[1], ..., \hat{\beta}_{kl}[T]) \tag{8}$$

where $q(\boldsymbol{\theta}_d; \hat{\boldsymbol{\theta}}_d)$ specifies a Dirichlet($\hat{\boldsymbol{\theta}}_d$) distribution. In the next section we elaborate on the form of the $q(\beta_{kl}[1], ..., \beta_{kl}[T]; \hat{\beta}_{kl}[1], ..., \hat{\beta}_{kl}[T])$ distribution.

### 2.3 Variational Kalman Smoothing

Successful variational inference algorithms depend on formulating an approximate posterior close in form to the true posterior, and such that the expectations that make up the ELBO are tractable. Variational approximations commonly rely on the mean-field assumption that posits the variational posterior as the product of independent distributions for each latent variable. However, this assumption is invalid for our model, as it ignores the dependence of allele frequencies over time. [5] introduce variational Kalman filtering as a technique to construct variational approximations to nonconjugate normal state space models. Variational Kalman filtering uses variational parameters $\hat{\beta}_{kl}[t]$ that are pseudo-observations from the state space model:

$$\hat{\beta}_{kl}[t] \sim \mathrm{Normal}(\beta_{kl}[t], \nu^2) \tag{9}$$

with additional variational parameter $\nu$. Given the pseudo-observations, standard Kalman filtering and smoothing equations can be used to calculate marginal means $\widetilde{m}_{kl}[t]$ and marginal variances $\widetilde{v}_{kl}[t]$ of the latent variables $\beta_{kl}[1:T]$ given the pseudo-observations $\hat{\beta}_{kl}[1:T]$. The variational approximation takes the form

$$\beta_{kl}[t] \sim \mathrm{Normal}(\widetilde{m}_{kl}[t], \widetilde{v}_{kl}[t]) \tag{10}$$

The ELBO is then maximized with respect to the pseudo-observations.

### 2.4 Parameter Inference

Variational inference algorithms in the setting above often rely on optimizing parameters through coordinate ascent: each parameter is updated iteratively while the others remain fixed. Coordinate ascent can be computationally expensive, especially as the size of the data becomes large. For this reason, stochastic variational inference [4,13] is a popular alternative. Briefly, stochastic variational inference distinguishes global variational parameters, such as $\hat{\boldsymbol{\theta}}_d$, whose coordinate ascent update requires iterating through the entire dataset, with local parameters, $\hat{\boldsymbol{\beta}}_{kl}$, whose update only depends on a subset of the data.

We optimize variational parameters using stochastic variational inference (Algorithm 1 - see appendix), along with a conjugate gradient algorithm for the local parameters $\hat{\boldsymbol{\beta}}_{kl}$ which do not have a closed form update, using a surrogate lower bound on the ELBO [11]. We first subsample a particular locus $l$, update the pseudo-outputs for that locus, then update the variational parameters $\hat{\boldsymbol{\theta}}_d$. This process continues until the $\hat{\boldsymbol{\theta}}_d$ converge. See the appendix for full details.

Estimates for ancestry proportions are computed by taking the posterior expectation of $\boldsymbol{\theta}_d$:

$$\mathbb{E}_q[\theta_{dk}] = \frac{\hat{\theta}_{dk}}{\sum_s \hat{\theta}_{ds}} \tag{11}$$

We further optimized our implementation, obtaining an order of magnitude speed up over a naive implementation.

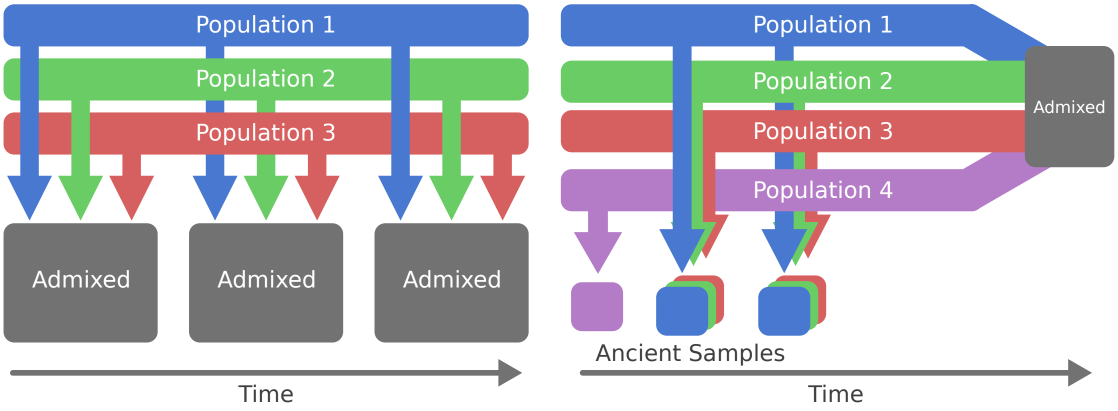
### 2.5 Simulated Data

We designed simulations to test the ability of our method to assign ancient samples into populations under two historical scenarios (Fig. 2a, Fig. 2b). In each scenario, we simulated $K$ populations according to the Wright-Fisher model for genetic drift: we drew initial allele frequencies from a Uniform$(0.2, 0.8)$ distribution and simulated discrete generations by drawing $2N_k$ individuals randomly with replacement from the previous generation. When then drew a subset of admixed individuals at specified time points with genotypes and ancestry proportions specified by the generative model. Note that we are not simulating data under the normal approximation.

One concern is that our model assumes allele frequencies are away from 0 or 1, while the allele frequencies in the Wright-Fisher model are guaranteed to fix given sufficient time. We allowed allele frequencies to fix in our simulations to test our model's robustness to violating this assumption, though most allele frequencies do not reach fixation in our simulations. We fixed effective population to $N_e = 2500$. To generalize our results across different effective population sizes, we measured time in coalescent units (1 coalescent unit = $2N_e$ generations). We denote the total simulation time in coalescent units by $\tau$. Each simulation was run across $\tau \in \{0.02, 0.04, 0.08, 0.16\}$. We simulated 10000 independent loci.

In the *baseline* simulation scenario (Fig. 2a), we sampled 40 individuals from $K = 3$ populations at 3 evenly spaced time points. We drew ancestry proportions from a Dirichlet$(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ distribution, ensuring that the majority of any one individual's genome originated in a single population, with smaller ancestry proportions from the remaining populations.

In the *historical* scenario (Fig. 2b) we performed simulations under more realistic conditions. We assumed a modern population resulted from the instantaneous admixture from $K = 4$ ancestral populations. Ancient individuals were sampled pre-admixture and modern individuals were sampled post-admixture. Current datasets comprise a small number of ancient samples when compared with modern samples. Thus, we simulated 508 samples where approximately 4.5 % (23) of the samples were ancient and the remaining samples were modern (485) reflecting the balance of samples in 2.6.



(a) Baseline simulations

(b) Historical simulations

Fig. 2: Simulation scenarios explored with Dystruct. (a) Baseline simulation scenario. Three populations were simulated that admixed at three time points. Individuals were sampled from the admixed populations. (b) Historical simulation scenario. Ancient individuals were sampled from four unadmixed populations that merged to form a modern admixed population. Modern samples were from the admixed population.

All ancient samples occurred before time $\frac{\tau}{2}$. One of the four ancient populations was observed in the oldest ancient sample only, but appeared in modern populations.

We repeated each simulation scenario 10 times for a total of 80 simulations, and compared the ability of our model to infer the parameter $\boldsymbol{\theta}_d$ with that of ADMIXTURE (v1.3.0). Since effective population size is a fixed parameter in Dystruct, we tested Dystruct on several effective population sizes. We ran Dystruct with $N_k = 1000, 2500, 5000, 10000$ for all simulation scenarios. For each simulation, we computed the root-mean-square error (RMSE) between the true ancestry proportions, and parameters inferred by Dystruct and ADMIXTURE:

$$\text{RMSE}(\boldsymbol{\theta}^{true}, \boldsymbol{\theta}^{inf}) = \sqrt{\frac{1}{DK}\sum_{d=1}^{D}\sum_{k=1}^{K}(\theta_{dk}^{true} - \theta_{dk}^{inf})^2}$$

## 2.6 Real Data

[12] analyze a hybrid dataset of modern humans from the Human Origins dataset [16,20], 69 newly sequenced ancient Europeans, along with 25 previously published ancient samples [10,31], to study population turnover in Europe. Ancient samples included several Holocene hunter gatherers (∼5000-6000 years old), Neolithic farmers (∼5000-8000 years old), Copper/Bronze age individuals (∼3100-5300 years old), and an Iron Age individual (∼2900 years old). In addition, the data include three Pleistocene hunter-gatherers — ∼45000 year old Ust-Ishim [8], ∼30000 year old Kostenki14 [30], and ∼240000 year old MA1 [25] — the Tyrolean Iceman [15], the hunter-gatherers LaBrana1 [19] and Loschbour [16], and the Neolithic farmer Stuttgart [16].

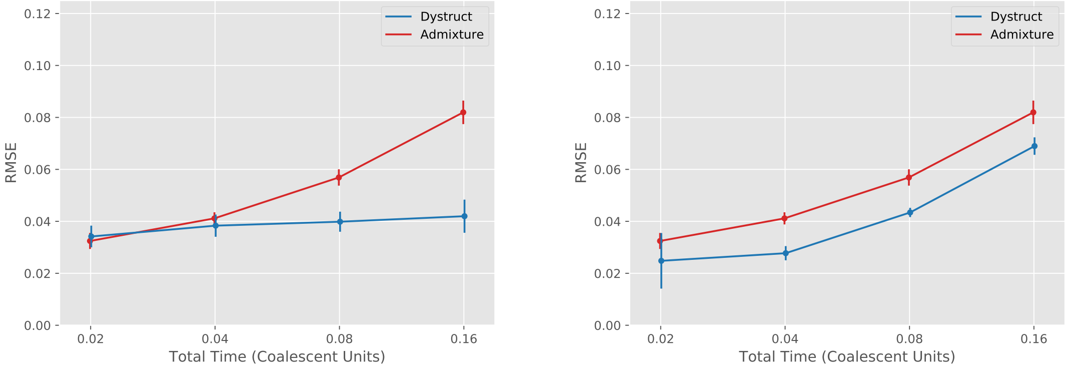
We analyzed the publicly available dataset from `https://reich.hms.harvard.edu/datasets`. After removing related individuals identified in [12], and removing samples from outside the scope of their paper, we were left with a dataset consisting of 92 ancient samples and 1941 modern samples genotyped at 354212 loci. Again following [12], we pruned this original dataset for linkage disequilibrium in in PLINK [24] (v1.07) using --indep-pairwise 200 25 0.5, leaving 222755 SNPs. Each reported ancient sample includes confidence intervals for radiocarbon date estimates. To convert radiocarbon dates to generation time required by Dystruct, we assumed a 25 year generation time, and took the midpoint of the radiocarbon dates as point estimates divided by 25 for ancient samples. We further grouped time points for samples together if they were within the 95 % confidence interval for radiocarbon date estimates, and were part of the same culture. We assigned the year 2015 to modern samples. The final dataset spanned 1800 generations.

We then ran ADMIXTURE and Dystruct on the full data with effective population size of 7500 from $K = 2$ to $K = 16$, compared the results. We reported the results for $K = 11$ because it has the clearest historical interpretation.

## 3 Results

### 3.1 Simulated Data

When simulating data according to the *baseline* scenario, Dystruct consistently matches up with ADMIXTURE or significantly outperforms it (Fig. 3). ADMIXTURE performs much worse as the simulated coalescent time increases, from RMSE of 0.032 (averaged across the simulated population sizes) for $\tau = 0.02$ to RMSE of 0.082 at $\tau = 0.16$. Dystruct is less susceptible to this increase in error. Intuitively, the more coalescent time is considered, the more the drift, and hence, the more important it is to model its dynamics.



(a) Dystruct $N_k = 1000$

(b) Dystruct $N_k = 10000$

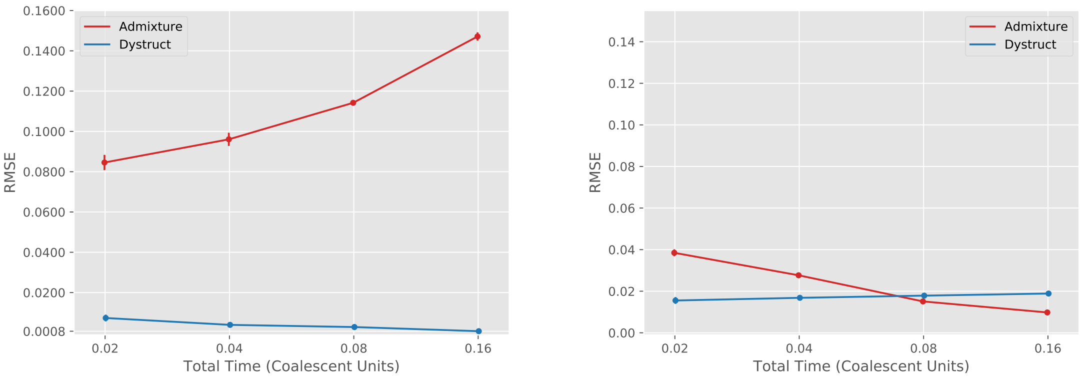
Fig. 3: RMSE for the *baseline* simulation scenarios across several effective population sizes provided to Dystruct.

The *historical* simulation scenario demonstrates a substantial advantage to Dystruct on ancient samples across population parameters. This suggests that a promising use case for our model is investigating hypotheses of historical admixture. Nonetheless a near zero RMSE for Dystruct is potentially misleading because ancient samples are not admixed.

Dystruct also performs well on the modern samples. Dystruct outperforms ADMIXTURE for $\tau = 0.02, 0.04$, while the converse is true for $\tau = 0.08, 0.16$.

### 3.2 Real Data

Dystruct shows good concordance with ADMIXTURE on modern data with known global populations (Fig. 5). In particular, African populations (Dark Blue; eg. Bantu, Mbuti, Yoruba), Asian



(a) RMSE for ancient samples

(b) RMSE for modern samples

Fig. 4: RMSE for ancestry estimates obtained by Dystruct ($N_k = 2500$) and ADMIXTURE for the *historical* simulation scenario. Results are similar across multiple $N_k$ (see Appendix).

populations (Red; eg. Han, Japanese, Korean), Native American populations (Dark Pink; eg. Mixe, Mayan, Zapotec), and Oceanian populations (Yellow; eg. Paupan) all form similar genetic clusters, among many other examples.

Dystruct and ADMIXTURE differ on the ancient samples. In Dystruct, most ancient samples are "pure," containing ancestry components from a single population, and modern day populations appear as mixtures of ancient populations. This is evident in the entropy across samples (Fig. 6). On ancient samples, Dystruct has a lower entropy than ADMIXTURE, while the opposite is true for modern samples. This is most apparent in the different ancestry assignments for the oldest samples: the Pliestocene hunter gatherers. MA1, Kostenki14, and Ust-Ishim differ substantially between Dystruct and ADMIXTURE. These are the samples where genetic drift is most prominent. In ADMIXTURE, MA1, Kostenki14, and Ust-Ishim all appears as mixtures of several modern day populations. In Dystruct, modern populations appear as mixtures of components derived from MA1, Kostenki14, and Ust-Ishim.

Most interestingly, the later ancient samples appear as mixtures of earlier samples in Dystruct, but not in ADMIXTURE. Late Neolithic, Bronze Age, and Iron Age samples appear as admixed between Yamnaya steppe herders (Orange), hunter-gatherers (Brown), and early Neotlithic (Green). Additionally, we see substantial shared ancestry between these groups and modern European populations. Both findings are consistent with [12], who obtained these results using alternative methods. Notably, Kostenki14 shares ancestry with the Yamnaya group, suggesting a possible source for steppe ancestry.

### 3.3 Running Time

Despite the added complexity, additional model parameters, and large dataset, Dystruct ran on the real data in approximately 6 days using 2 cores of a 2.9 GHz Intel Core i5 processor. ADMIXTURE ran in approximately 2 days using 2 cores. Using 1 core, Dystruct ran in approximately 30 minutes per replicate on the *baseline* scenario, and approximately 120 minutes per replicate on the *historical* scenario. ADMIXTURE ran in less than a minute on these scenarios. The advantages of stochastic variational inference are more apparent for larger datasets.

Dystruct's main computational consideration is the number of time points. During each iteration the parameters of a single locus are updated, then used to update ancestry estimates across all

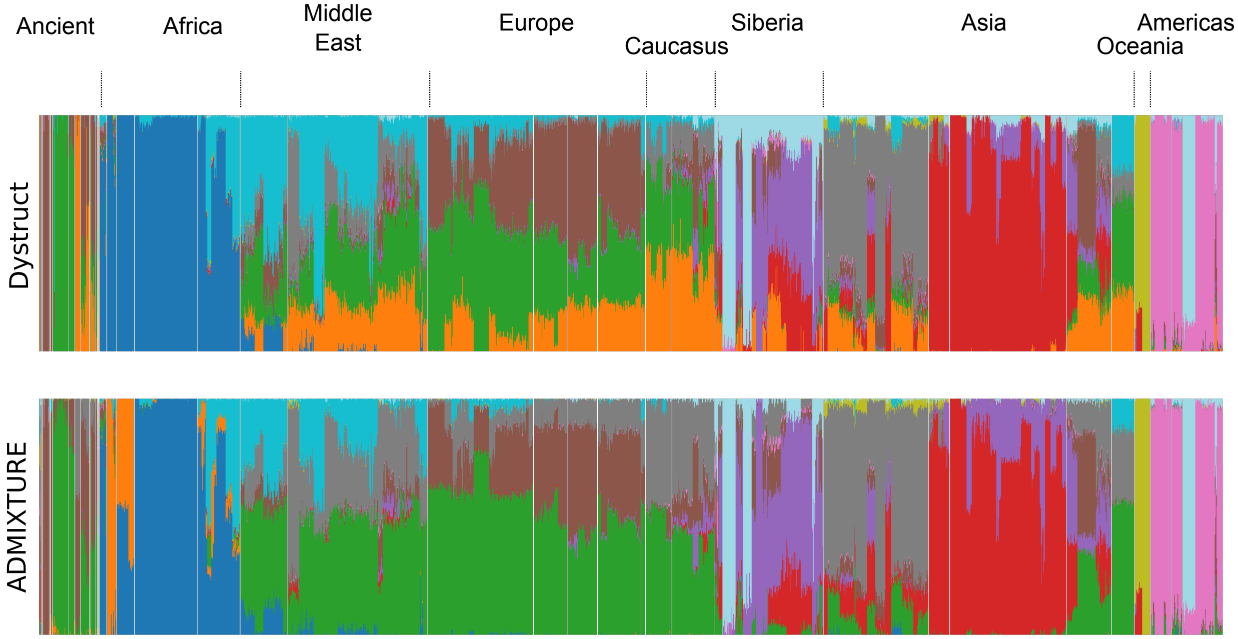


Fig. 5: Ancestry proportions inferred by Dystruct and ADMIXTURE across all samples. Top: Dystruct. Bottom: ADMIXTURE. Ancient samples and population colors correspond to Figure 7. ADMIXTURE and Dystruct agree on several major population clusters, but differ on modern day ancestry estimates from ancient samples.

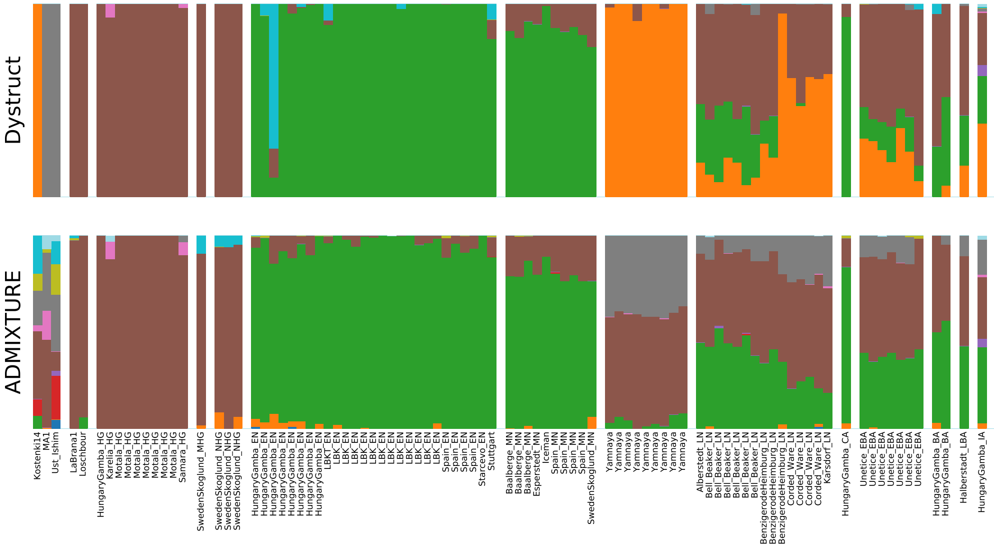


Fig. 7: Ancestry estimates for 92 ancient samples from [12]. PHG: Pleistocene Hunter-Gatherers. Dystruct estimates ancestry for modern populations as combinations of ancient samples. ADMIXTURE estimates ancestry for ancient samples as combinations of modern populations. Sample key: *EN: Early Neolithic; MN: Middle Neolithic; LN: Late Neolithic; HG: Hunter-Gatherer; BA: Bronze Age; EBA: Early Bronze Age; CA: Copper Age; IA: Iron Age.*

individuals. Estimates for ancestry parameters all, $\hat{\boldsymbol{\theta}}_d$, can be computed in closed form in $O(DK)$; however, the update for the parameters $\hat{\boldsymbol{\beta}}_{kl}$ is approximated numerically. Computing the gradient of the $\hat{\boldsymbol{\beta}}_{kl}$ at a locus takes $O(T^2+D)$ time because the marginal means $\widetilde{m}_{kl}[t]$ must be differentiated with respect to each pseudo-output. The gradient must be re-evaluated until the estimates for $\hat{\boldsymbol{\beta}}_{kl}$ converge.

## 4 Discussion

We have presented Dystruct, a model and inference procedure to understand population structure from ancient DNA. The novelty of the model is its explicit temporal semantics. This formalization of allele frequency dynamics facilitates perception of modern and more recent populations as evolved from more ancient ones or combinations thereof. We derived an efficient inference algorithm for the model parameters using stochastic variational inference, and released software for use by the broader community. We established the performance of our model on several simulation scenarios, and further demonstrated its utility for gaining insight from the analysis of real data.

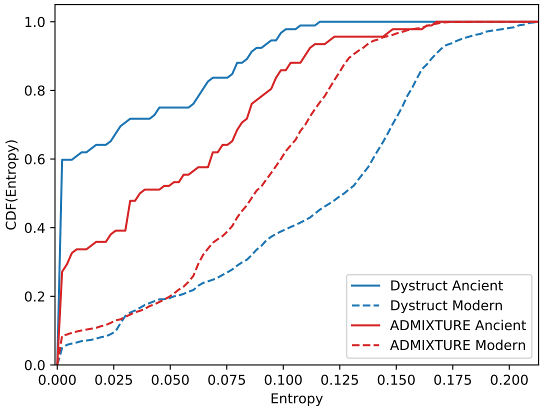


Fig. 6: Cumulative density function for entropy across ancient and modern samples. Dystruct has a lower entropy for ancient samples, while ADMIXTURE has a lower entropy for modern samples.

Our model outperforms the current standard modeling across a variety simulation scenarios. Encouragingly, our simulations show that Dystruct does a better job recovering population structure in the presence of genetic drift, an effect that hinders existing tool. Our method offers accuracy gains across many simulation scenarios, reaching as much as a 2-fold improvement, while never significantly reduces accuracy. In addition, simulations investigating increasing the sample size suggest that this gap will grow with additional samples analyzed across increasingly large datasets (Fig. 10).

We note the advantage of Dystruct increases with genetic drift and thus with coalescent time elapsed. This means that in practical situations, where samples are dated in years, Dystruct is most important when the effective population sizes are small and vary across populations. From statistical inference perspective, effective population size can be thought of as a regularizer that penalizes the difference between allele frequencies at each time point. Thus, as effective population size increases, alleles frequencies drift more slowly and become closer across time points. However, across all simulations and for real data we constrained the effective population size across all populations to be the same. Thus, the parameters converge to one of at least $K$ symmetric modes — population labels are exchangeable — and it is unclear how allowing different effective population sizes for different populations changes the log likelihood with respect to the parameter space. Future work should investigate this issue in more detail. Nonetheless, as we have demonstrated this is not a serious limitation to achieving reasonable estimates.

Notably, our results hold across a range of effective population sizes provided to Dystruct. Our results suggest that reasonable estimates can be obtained so long as the population size provided to Dystruct is within an order of magnitude of the true effective population size. We recommend taking conservative underestimate of the true effective population.

Our results on real data matched known population clusters on modern populations, and lead to new interpretations of the ancient dataset. Interestingly, the PSD model tends to describe the oldest ancient samples as mixtures of modern populations, while in Dystruct several modern populations appeared as mixtures of these ancient samples. This makes sense in light of the standard goal of maximizing overall variance explained, a quantity dominated by the majority of the samples, which are modern. In contrast, temporal semantics implicitly assume admixture occurs forward in time, putting the focus on labeled ancient populations. Dystruct can thus provide additional insight into such populations from ancient DNA.

There are several limitations to our approach. First, we model populations as independently evolving over time. This ignores historical relationships such as population splits and merges. Hence Dystruct will only capture one branch of a population phylogeny at a time. Nonetheless, Dystruct represents a first step toward incorporating more complex population histories, and future work should focus on this issue. Second, there is no clear procedure for choosing the correct number of populations $K$. We have deferred this issue to future work, but pose that this does not prevent a severe limitation: the current state of the art uses runs across multiple values of $K$, and interprets the results for each $K$.

More generally, we have presented a time-series model for population history with several promising extensions. Our method complements existing approaches, and can lead to new insights on ancient DNA datasets. Our method to performs best when modern populations are admixed between two or more ancestral populations. We predict a use case for our model is testing historical admixture between ancient and modern populations.

## A Stochastic Variational Inference

In this section we derive the inference algorithm for stochastic variational inference under the Dystruct model. We first derive the traditional coordinate ascent updates, then show how we can modify these updates for stochastic optimization. Finally, we extend the algorithm for missing data.

### A.1 Computing The ELBO

The ELBO is given by

$$\begin{aligned}
L &= \mathbb{E}_q[\log p(\boldsymbol{\beta}_{1:K,1:L}, \boldsymbol{\theta}_{1:D}, \boldsymbol{x}_{1:D,1:L})] - \mathbb{E}_q[\log q(\boldsymbol{\beta}_{1:K,1:L}, \boldsymbol{\theta}_{1:D})] && (12)\\
&= \sum_{t=1}^{T}\sum_{k=1}^{K}\sum_{l=1}^{L} \mathbb{E}_q[\log p(\beta_{kl}[t] \big| \beta_{kl}[t-1])] && (13)\\
&\quad + \sum_{d=1}^{D} \mathbb{E}_q[\log p(\boldsymbol{\theta}_d)] && (14)\\
&\quad + \sum_{d=1}^{D}\sum_{l=1}^{L} \mathbb{E}_q[\log p(x_{dl} | t_d, \boldsymbol{\theta}_d, \boldsymbol{\beta}_{1:K,l})] && (15)\\
&\quad - \sum_{k=1}^{K}\sum_{l=1}^{L} \mathbb{E}_q[\log q(\boldsymbol{\beta}_{kl} \big| \hat{\boldsymbol{\beta}}_{kl})] && (16)\\
&\quad - \sum_{d=1}^{D} \mathbb{E}_q[\log q(\boldsymbol{\theta}_d \big| \hat{\boldsymbol{\theta}}_d)] && (17)
\end{aligned}$$

The ELBO as written does not have a closed form due to the log sum terms that appear in $\mathbb{E}_q[\log p(x_{dl} \big| t_d, \boldsymbol{\theta}_d, \boldsymbol{\beta}_{1:K,l})] = \mathbb{E}_q[\log \text{Binomial}(2, \sum_k \beta_{kl}[t_d]\theta_{dk})]$:

$$x_{dl}\, \mathbb{E}_q \left[ \log \left( \sum_k \theta_{dk}\beta_{kl}[t_d] \right) \right] + (2 - x_{dl})\, \mathbb{E}_q \left[ \log \left( 1 - \sum_k \theta_{dk}\beta_{kl}[t_d] \right) \right] \tag{18}$$

Following [11], we optimize a surrogate lower bound by introducing auxiliary variational parameters $\boldsymbol{\phi}_{dl} = (\phi_{dl}[1], ..., \phi_{dl}[K])$ and $\boldsymbol{\zeta}_{dl} = (\zeta_{dl}[1], ..., \zeta_{dl}[K])$ whose vector components sums to 1. An application of Jensen's inequality shows

$$\log \left( \sum_k \theta_{dk}\beta_{kl}[t_d] \right) \geq \sum_k \phi_{dl}[k] \log \left( \frac{\theta_{dk}\beta_{kl}[t_d]}{\phi_{dl}[k]} \right) \tag{19}$$

$$\log \left( 1 - \sum_k \theta_{dk}\beta_{kl}[t_d] \right) \geq \sum_k \zeta_{dl}[t_d] \log \left( \frac{\theta_{dk}(1 - \beta_{kl}[t_d])}{\zeta_{dl}[k]} \right) \tag{20}$$

so we still maintain a lower bound on the log likelihood. The auxiliary parameters are optimized to provide a tight lower bound. Fixing all other parameters, the constrained optimization problem can be solved using an application of Lagrange multipliers

$$\phi_{dl}[k] \propto \exp\{\mathbb{E}_q[\log \theta_{dk}] + \mathbb{E}_q[\log \beta_{kl}[t_d]]\} \tag{21}$$

$$\zeta_{dl}[k] \propto \exp\{\mathbb{E}_q[\log \theta_{dk}] + \mathbb{E}_q[\log(1 - \beta_{kl}[t_d])]\} \tag{22}$$

The first term in both equations is an expectation of a sufficient statistic, and therefore has a closed form: $\mathbb{E}_q[\log \theta_{dk}] = \Psi(\hat{\theta}_{dk}) - \Psi(\sum_k \hat{\theta}_{dk})$; $\Psi$ is the Digamma function. The two expectations in the second terms can be approximated by taking second order Taylor expansions around the marginal means of $\beta_{kl}[t_d]$ and $\widetilde{m}_{kl}[t]$ :

$$\mathbb{E}_q[\log \beta_{kl}[t_d]] \approx \log \widetilde{m}_{kl}[t_d] - \frac{\widetilde{v}_{kl}[t_d]}{2\widetilde{m}_{kl}[t_d]^2} \tag{23}$$

$$\mathbb{E}_q[\log(1 - \beta_{kl}[t_d])] \approx \log(1 - \widetilde{m}_{kl}[t_d]) \tag{24}$$

### A.2 Optimizing Ancestry Proportions

Note that the $q(\boldsymbol{\theta}_d \big| \hat{\boldsymbol{\theta}}_d)$ satisfy the mean field assumption - the $\boldsymbol{\theta}_d$ in the variational posterior are independent. Therefore they have optimal coordinate ascent updates of the form

$$q^*(\boldsymbol{\theta}_d) \propto \exp\{\mathbb{E}_q[\log p(\boldsymbol{\theta}_d \big| \boldsymbol{\beta}_{kl}[t_d], \boldsymbol{x}_{d,1:L})]\} \tag{25}$$

where we have used several conditional independencies to simplify the complete conditional of $\boldsymbol{\theta}_d$. Using the surrogate lower bound in the ELBO gives the optimal update

$$\hat{\theta}_{dk} = \alpha_k + \sum_{l=1}^{L} x_{dl}\phi_{dl}[k] + (2 - x_{dl})\zeta_{dl}[k] \tag{26}$$

matching the expression in [11].

### A.3 Optimizing allele frequencies

In variational Kalman filtering, the variational distribution for each $\beta_{kl}[t]$ is given by

$$\beta_{kl}[t] \sim \text{Normal}(\widetilde{m}_{kl}[t], \widetilde{v}_{kl}[t]) \tag{27}$$

where the mean and variance are the marginal means and posteriors given by the Kalman filtering and smoothing equations. Following the notation in [5], the forward (filtered) means and variances are given by

$$m_{kl}[t] = \frac{\nu^2}{v_{kl}[t-1] + \sigma_k^2[t] + \nu^2} m_{kl}[t-1] + \left(1 - \frac{\nu^2}{v_{kl}[t-1] + \sigma_k^2[t] + \nu^2}\right) \hat{\beta}_{kl}[t] \tag{28}$$

$$v_{kl}[t] = \left(\frac{\nu^2}{v_{kl}[t-1] + \sigma_k^2[t] + \nu^2}\right) (v_{kl}[t-1] + \sigma_k^2[t]) \tag{29}$$

where $\sigma_k^2[t] := \frac{\Delta g[t]}{12N_k}$. The initial conditions are $m_{kl}[0] = \beta_{kl}[0]$. The marginal (smoothed) means and variances are

$$\widetilde{m}_{kl}[t] = \left(\frac{\sigma_k^2[t]}{v_{kl}[t] + \sigma_k^2[t]}\right) m_{kl}[t] + \left(1 - \frac{\sigma_k^2[t]}{v_{kl}[t] + \sigma_k^2[t]}\right) \widetilde{m}_{kl}[t+1] \tag{30}$$

$$\widetilde{v}_{kl}[t] = v_{kl}[t] + \left(\frac{v_{kl}[t]}{v_{kl}[t] + \sigma_k^2[t]}\right)^2 (\widetilde{v}_{kl}[t+1] - v_{kl}[t] - \sigma_k^2[t+1]) \tag{31}$$

with initial conditions $\widetilde{m}_{kl}[T] = m_{kl}[T]$ and $\widetilde{v}_{kl}[T] = v_{kl}[T]$. The variational parameters $\hat{\beta}_{kl}[t]$ are optimized with respect to the ELBO, hence we need the partial derivatives of the marginal means $\widetilde{m}_{kl}[t]$ with respect to $\hat{\beta}_{kl}[t]$. These can be obtained using the forward-backward recurrence as in [5]. We will show the recurrence for initial frequencies $\beta_{kl}[0]$, which are not maximized in [5], and note that the other partial derivations can be obtained similarly. The recurrence is

$$\frac{\partial m_{kl}[t]}{\partial \beta_{kl}[0]} = \left(\frac{\nu^2}{v_{kl}[t-1] + \sigma_k^2[t] + \nu^2}\right) \frac{\partial m_{kl}[t-1]}{\beta_{kl}[0]} \tag{32}$$

$$\frac{\partial \widetilde{m}_{kl}[t]}{\beta_{kl}[0]} = \left(\frac{\sigma_k^2[t]}{v_{kl}[t] + \sigma_k^2[t]}\right) \frac{\partial m_{kl}[t]}{\partial \beta_{kl}[0]} + \left(1 - \frac{\sigma_k^2[t]}{v_{kl}[t] + \sigma_k^2[t]}\right) \frac{\partial \widetilde{m}_{kl}[t+1]}{\beta_{kl}[0]} \tag{33}$$

We optimize the $\hat{\boldsymbol{\beta}}_{kl}$ with respect to a single locus in a single population at time using a conjugate gradient algorithm, constraining the parameters to lie in the interval $(0, 1)$. The terms in the ELBO with respect to locus $l$ in population $k$ are

$$
\begin{aligned}
L_* &= \sum_{t=1}^{T} \mathbb{E}_q[\log p(\beta_{kl}[t] \big| \beta_{kl}[t-1]] - \mathbb{E}_q[\log q(\beta_{kl}[t] \big| \widetilde{m}_{kl}[t], \widetilde{v}_{kl}[t])] && (34) \\
&\quad + \sum_{d=1}^{D} \sum_{l=1}^{L} \mathbb{E}_q[\log p(x_{dl} \big| t_d, \beta_{kl}[t_d], \boldsymbol{\theta}_d)] \\
&= \frac{T}{2} \log 2\pi - \frac{1}{2} \sum_{t=1}^{T} \log \sigma_k^2[t] - \frac{1}{2} \sum_{t=1}^{T} \frac{1}{\sigma_k^2[t]} \mathbb{E}_q[(\beta_{kl}[t] - \beta_{kl}[t-1])^2] + \frac{T}{2} \log 2\pi && (35) \\
&\quad + \frac{T}{2} + \frac{1}{2} \sum_{t=1}^{T} \log \widetilde{v}_{kl}[t] \\
&\quad + \sum_{t=1}^{T} \sum_{l=1}^{L} \sum_{d:t_d=t} x_{dl} \phi_{dl}[k] \left( \log \widetilde{m}_{kl}[t] - \frac{\widetilde{v}_{kl}[t]}{2 \widetilde{m}_{kl}[t]^2} \right) + (2 - x_{dl}) \zeta_{dl}[k] \log(1 - \widetilde{m}_{kl}[t]) \\
&= \frac{T}{2} - \frac{1}{2} \sum_{t=1}^{T} (\log \sigma_k^2[t] - \log \widetilde{v}_{kl}[t]) - \frac{1}{2} \sum_{t=1}^{T} \frac{1}{\sigma_k^2[t]} (\widetilde{m}_{kl}[t] - \widetilde{m}_{kl}[t-1])^2 - \frac{\widetilde{v}_{kl}[t]}{\sigma_k^2[t]} - \frac{\widetilde{v}_{kl}[t-1]}{\sigma_k^2[t-1]} && (36) \\
&\quad + \sum_{t=1}^{T} \sum_{l=1}^{L} \sum_{d:t_d=t} x_{dl} \phi_{dl}[k] \left( \log \widetilde{m}_{kl}[t] - \frac{\widetilde{v}_[kl][t]}{2 \widetilde{m}_{kl}[t]^2} \right) + (2 - x_{dl}) \zeta_{dl}[k] \log(1 - \widetilde{m}_{kl}[t])
\end{aligned}
$$

where we define $\widetilde{v}_{kl}[0] = 0$, $\sigma_k^2[0] = 1$, and $m_{kl}[0] = \widetilde{m}_{kl}[0] = \beta_{kl}[0]$ for notational convenience. Taking partial derivatives with respect to the pseudo-outputs gives us

$$
\begin{aligned}
\frac{\partial L_*}{\partial \hat{\beta}_{kl}[s]} &= - \sum_{t=1}^{T} \frac{1}{\sigma_k^2[t]} (\widetilde{m}_{kl}[t] - \widetilde{m}_{kl}[t-1]) \left( \frac{\partial \widetilde{m}_{kl}[t]}{\partial \hat{\beta}_{kl}[s]} - \frac{\partial \widetilde{m}_{kl}[t-1]}{\partial \hat{\beta}_{kl}[s]} \right) && (37) \\
&\quad + \frac{\partial \widetilde{m}_{kl}[t]}{\partial \hat{\beta}_{kl}[s]} \sum_{d:t_d=t} x_{dl} \phi_{dl}[k] \left( \frac{1}{\widetilde{m}_{kl}[t]} + \frac{\widetilde{v}_k[t]}{\widetilde{m}_{kl}[t]^3} \right) + (2 - x_{dl}) \zeta_{dl}[k] \frac{1}{(\widetilde{m}_{kl}[t] - 1)}
\end{aligned}
$$

The full algorithm iterates between optimizing the local parameters $\hat{\boldsymbol{\beta}}_{kl}$, $\phi_{dl}[k]$, and $\zeta_{dl}[k]$ for each locus in each individual in each population using (21), (22), and (37), then updating global parameters $\hat{\boldsymbol{\theta}}_d$ according to (26) until convergence.

### A.4 Inference Algorithm

We can perform stochastic variational inference through a slight modification to the coordinate ascent algorithm presented above [4, 13]. Briefly, stochastic variational inference computes noisy estimates of the optimal global parameters by stochastically subsampling data points, and using the optimal local parameters to update the global parameters. The optimal global parameters are a weighted average of the previous global parameters, with the newly computed global parameters. Following [11], the $n+1$ stochastic variational inference update for the global parameters $\hat{\boldsymbol{\theta}}_d$ is

$$
\hat{\theta}_{dk}^{n+1} = (1 - \epsilon_n) \hat{\theta}_{dk}^{n} + \alpha_k + \epsilon_n L \left( x_{dl} \phi_{dl}[k] + (2 - x_{dl}) \zeta_{dl}[k] \right) \qquad (38)
$$

where $\epsilon_n$ is the step size for iteration $n$ and $L$ is the number of loci. Provided the step size meets certain criteria the algorithm is guaranteed to converge. See [13] or [4] for more details. We picked a step size of $\epsilon_n = (1+n)^{-0.5}$ for the first 10000 iterations, and $\epsilon_n = (n - 7825)^{-0.6}$ for the remaining iterations.

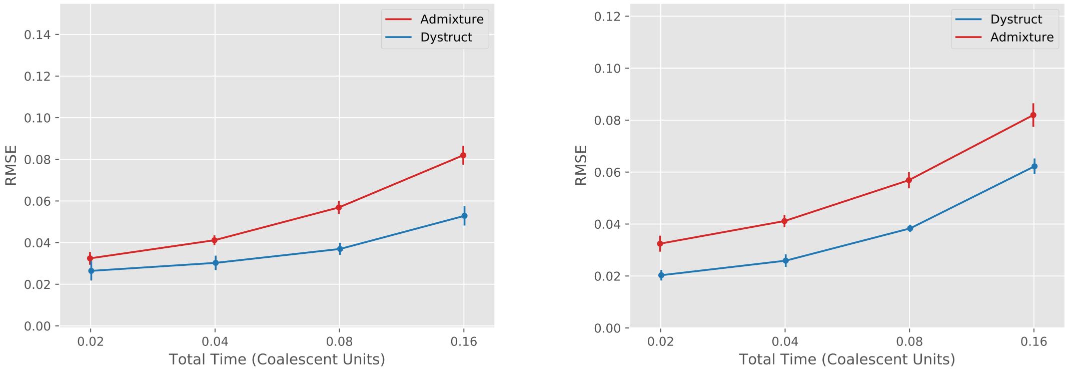
**Algorithm 1** Dystruct inference algorithm

```
Input: Genotypes $x_{1:D,1:L}$; Sample Times $t_d$; Population Size $N_k = N$ for all populations.
while $\hat{\theta}_d$ have not converged do
    Pick $l \sim \text{Uniform}(1, L)$
    while $\phi_{dl}$ and $\zeta_{dl}$ have not converged do
        Update auxiliary parameters $\phi_{dl}$ and $\zeta_{dl}$ for $d = 1, 2, ..., D$ according to (21) and (22).
        Update allele frequency parameters $\hat{\beta}_{kl}$ for $k = 1, 2, ...K$ using the numerical optimization routine described
        in section A.3.
    end while
    Update global parameters $\hat{\theta}_d$ for $d = 1, 2, ..., D$ according to (38)
end while
```

### A.5 Extensions to missing data

The above algorithm only holds for complete data. A small modification is required for missing data, where not every sample has an observed genotype at every locus. Rather than a single global step size $\epsilon_t$, we maintain a step size for every individual $\epsilon_{n_d}$ where $n_d$ is the number of iterations for individual $d$. When a locus is subsampled, we only update global ancestry estimates for individuals with observed genotypes at that locus, and the step size for those individuals. We further replace the parameter $L$ with $L_d$, the number of loci for each observed in each individual.

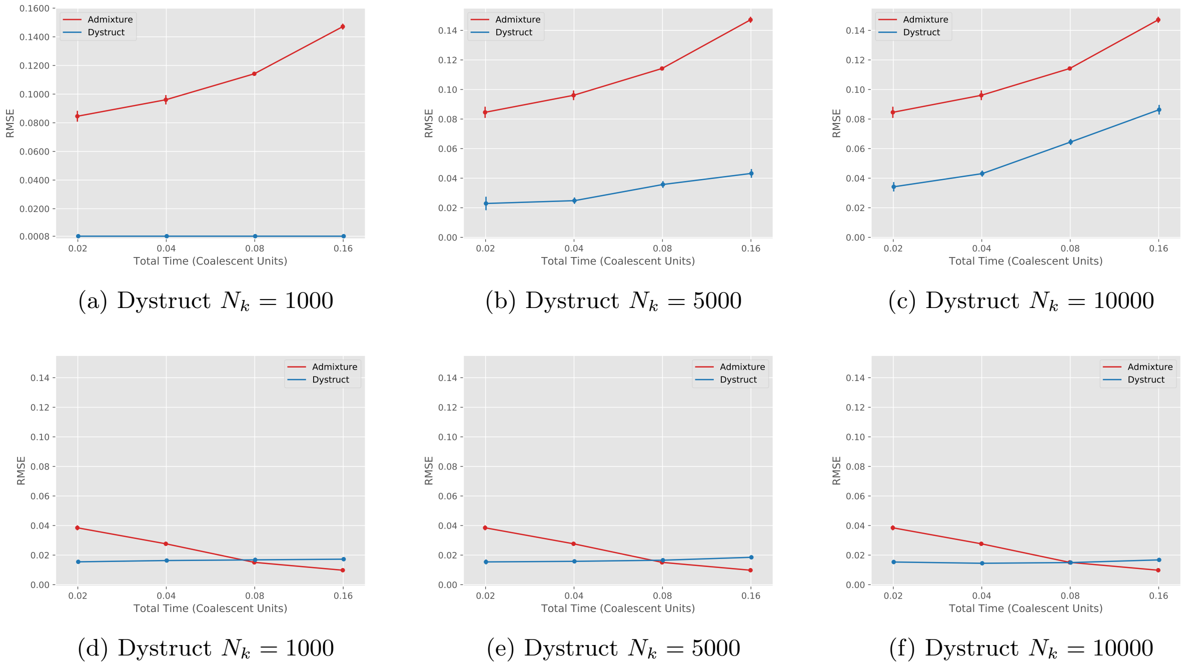
## B Supplementary Figures



(a) Dystruct $N_k = 2500$

(b) Dystruct $N_k = 5000$

Fig. 8: RMSE for the *baseline* simulation scenarios across several effective population sizes provided to Dystruct. True $N_k = 2500$.



(a) Dystruct $N_k = 1000$

(b) Dystruct $N_k = 5000$

(c) Dystruct $N_k = 10000$

(d) Dystruct $N_k = 1000$

(e) Dystruct $N_k = 5000$

(f) Dystruct $N_k = 10000$

Fig. 9: RMSE for the *historical* simulation scenarios across several effective population sizes provided to Dystruct. True $N_k = 2500$. (a) (b) (c) RMSE on ancient samples. (d) (e) (f) RMSE on modern samples.

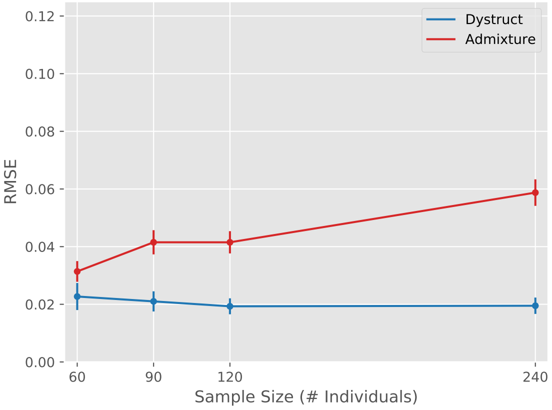


Fig. 10: Increasing the number of sampled individuals for the baseline scenario at $\tau = 0.02, N_k = 5000$.

# References

1. David H Alexander, John Novembre, and Kenneth Lange. Fast model-based estimation of ancestry in unrelated individuals. *Genome research*, 19(9):1655–1664, 2009.
2. Morten E Allentoft, Martin Sikora, Karl-Göran Sjögren, Simon Rasmussen, Morten Rasmussen, Jesper Stenderup, Peter B Damgaard, Hannes Schroeder, Torbjörn Ahlström, Lasse Vinner, et al. Population genomics of bronze age eurasia. *Nature*, 522(7555):167–172, 2015.
3. David M Blei. Probabilistic topic models. *Communications of the ACM*, 55(4):77–84, 2012.
4. David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 2017.
5. David M Blei and John D Lafferty. Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*, pages 113–120. ACM, 2006.
6. David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine learning research*, 3:993–1022, 2003.
7. Luigi L Cavalli-Sforza and Anthony WF Edwards. Phylogenetic analysis: models and estimation procedures. *Evolution*, 21(3):550–570, 1967.
8. Qiaomei Fu, Heng Li, Priya Moorjani, Flora Jay, Sergey M Slepchenko, Aleksei A Bondarev, Philip LF Johnson, Ayinuer Aximu-Petri, Kay Prüfer, Cesare de Filippo, et al. Genome sequence of a 45,000-year-old modern human from western siberia. *Nature*, 514(7523):445–449, 2014.
9. Qiaomei Fu, Cosimo Posth, Mateja Hajdinjak, Martin Petr, Swapan Mallick, Daniel Fernandes, Anja Furtwängler, Wolfgang Haak, Matthias Meyer, Alissa Mittnik, et al. The genetic history of ice age europe. *Nature*, 2016.
10. Cristina Gamba, Eppie R Jones, Matthew D Teasdale, Russell L McLaughlin, Gloria Gonzalez-Fortes, Valeria Mattiangeli, László Domboróczki, Ivett Kővári, Ildikó Pap, Alexandra Anders, et al. Genome flux and stasis in a five millennium transect of european prehistory. *Nature communications*, 5:5257, 2014.
11. Prem Gopalan, Wei Hao, David M Blei, and John D Storey. Scaling probabilistic models of genetic variation to millions of humans. *Nat Genet*, 48(12), 2016.
12. Wolfgang Haak, Iosif Lazaridis, Nick Patterson, Nadin Rohland, Swapan Mallick, Bastien Llamas, Guido Brandt, Susanne Nordenfelt, Eadaoin Harney, Kristin Stewardson, et al. Massive migration from the steppe was a source for indo-european languages in europe. *Nature*, 522(7555):207–211, 2015.
13. Matthew D Hoffman, David M Blei, Chong Wang, and John William Paisley. Stochastic variational inference. *Journal of Machine Learning Research*, 14(1):1303–1347, 2013.
14. Michael I Jordan, Zoubin Ghahramani, Tommi S Jaakkola, and Lawrence K Saul. An introduction to variational methods for graphical models. *Machine learning*, 37(2):183–233, 1999.
15. Andreas Keller, Angela Graefen, Markus Ball, Mark Matzas, Valesca Boisguerin, Frank Maixner, Petra Leidinger, Christina Backes, Rabab Khairat, Michael Forster, et al. New insights into the tyrolean iceman's origin and phenotype as inferred by whole-genome sequencing. *Nature communications*, 3:698, 2012.
16. Iosif Lazaridis, Nick Patterson, Alissa Mittnik, Gabriel Renaud, Swapan Mallick, Karola Kirsanow, Peter H Sudmant, Joshua G Schraiber, Sergi Castellano, Mark Lipson, et al. Ancient human genomes suggest three ancestral populations for present-day europeans. *Nature*, 513(7518):409–413, 2014.
17. Mark Lipson, Po-Ru Loh, Alex Levin, David Reich, Nick Patterson, and Bonnie Berger. Efficient moment-based inference of admixture parameters and sources of gene flow. *Molecular biology and evolution*, 30(8):1788–1802, 2013.
18. Rasmus Nielsen, Joshua M Akey, Mattias Jakobsson, Jonathan K Pritchard, Sarah Tishkoff, and Eske Willerslev. Tracing the peopling of the world through genomics. *Nature*, 541(7637):302–310, 2017.
19. Inigo Olalde, Morten E Allentoft, Federico Sánchez-Quinto, Gabriel Santpere, Charleston WK Chiang, Michael DeGiorgio, Javier Prado-Martinez, Juan Antonio Rodríguez, Simon Rasmussen, Javier Quilez, et al. Derived immune and ancestral pigmentation alleles in a 7,000-year-old mesolithic european. *Nature*, 507(7491):225–228, 2014.
20. Nick Patterson, Priya Moorjani, Yontao Luo, Swapan Mallick, Nadin Rohland, Yiping Zhan, Teri Genschoreck, Teresa Webster, and David Reich. Ancient admixture in human history. *Genetics*, 192(3):1065–1093, 2012.
21. Benjamin M Peter. Admixture, population structure, and f-statistics. *Genetics*, 202(4):1485–1501, 2016.
22. Jonathan K Pritchard, Matthew Stephens, and Peter Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–959, 2000.
23. Kay Prüfer, Fernando Racimo, Nick Patterson, Flora Jay, Sriram Sankararaman, Susanna Sawyer, Anja Heinze, Gabriel Renaud, Peter H Sudmant, Cesare De Filippo, et al. The complete genome sequence of a neanderthal from the altai mountains. *Nature*, 505(7481):43–49, 2014.
24. Shaun Purcell, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel AR Ferreira, David Bender, Julian Maller, Pamela Sklar, Paul IW De Bakker, Mark J Daly, et al. Plink: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, 81(3):559–575, 2007.

25. Maanasa Raghavan, Pontus Skoglund, Kelly E Graf, Mait Metspalu, Anders Albrechtsen, Ida Moltke, Simon Rasmussen, Thomas W Stafford Jr, Ludovic Orlando, Ene Metspalu, et al. Upper palaeolithic siberian genome reveals dual ancestry of native americans. *Nature*, 505(7481):87–91, 2014.
26. Anil Raj, Matthew Stephens, and Jonathan K Pritchard. *fastSTRUCTURE*: variational inference of population structure in large SNP data sets. *Genetics*, 197(2):573–589, 2014.
27. Morten Rasmussen, Yingrui Li, Stinus Lindgreen, Jakob Skou Pedersen, Anders Albrechtsen, Ida Moltke, Mait Metspalu, Ene Metspalu, Toomas Kivisild, Ramneek Gupta, et al. Ancient human genome sequence of an extinct palaeo-eskimo. *Nature*, 463(7282):757–762, 2010.
28. David Reich, Richard E Green, Martin Kircher, Johannes Krause, Nick Patterson, Eric Y Durand, Bence Viola, Adrian W Briggs, Udo Stenzel, Philip LF Johnson, et al. Genetic history of an archaic hominin group from denisova cave in siberia. *Nature*, 468(7327):1053–1060, 2010.
29. Carina M Schlebusch, Helena Malmström, Torsten Günther, Per Sjödin, Alexandra Coutinho, Hanna Edlund, Arielle R Munters, Maryna Steyn, Himla Soodyall, Marlize Lombard, et al. Ancient genomes from southern africa pushes modern human divergence beyond 260,000 years ago. *bioRxiv*, page 145409, 2017.
30. Andaine Seguin-Orlando, Thorfinn S Korneliussen, Martin Sikora, Anna-Sapfo Malaspinas, Andrea Manica, Ida Moltke, Anders Albrechtsen, Amy Ko, Ashot Margaryan, Vyacheslav Moiseyev, et al. Genomic structure in europeans dating back at least 36,200 years. *Science*, 346(6213):1113–1118, 2014.
31. Pontus Skoglund, Helena Malmström, Ayça Omrak, Maanasa Raghavan, Cristina Valdiosera, Torsten Günther, Per Hall, Kristiina Tambets, Jüri Parik, Karl-Göran Sjögren, et al. Genomic diversity and admixture differs for stone-age scandinavian foragers and farmers. *Science*, 344(6185):747–750, 2014.
32. Pontus Skoglund, Helena Malmström, Maanasa Raghavan, Jan Storå, Per Hall, Eske Willerslev, M Thomas P Gilbert, Anders Götherström, and Mattias Jakobsson. Origins and genetic legacy of neolithic farmers and hunter-gatherers in europe. *Science*, 336(6080):466–469, 2012.
33. Martin J Wainwright and Michael I Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305, 2008.