fastSTRUCTURE: Variational Inference of Population Structure in Large SNP Data Sets

Anil Raj,*,[1] Matthew Stephens,[†] and Jonathan K. Pritchard*,[‡]
*Department of Genetics, [‡]Department of Biology, Howard Hughes Medical Institute, Stanford University, Stanford, California 94305, and [†]Departments of Statistics and Human Genetics, University of Chicago, Chicago, Illinois 60637

ABSTRACT Tools for estimating population structure from genetic data are now used in a wide variety of applications in population genetics. However, inferring population structure in large modern data sets imposes severe computational challenges. Here, we develop efficient algorithms for approximate inference of the model underlying the STRUCTURE program using a variational Bayesian framework. Variational methods pose the problem of computing relevant posterior distributions as an optimization problem, allowing us to build on recent advances in optimization theory to develop fast inference tools. In addition, we propose useful heuristic scores to identify the number of populations represented in a data set and a new hierarchical prior to detect weak population structure in the data. We test the variational algorithms on simulated data and illustrate using genotype data from the CEPH–Human Genome Diversity Panel. The variational algorithms are almost two orders of magnitude faster than STRUCTURE and achieve accuracies comparable to those of ADMIXTURE. Furthermore, our results show that the heuristic scores for choosing model complexity provide a reasonable range of values for the number of populations represented in the data, with minimal bias toward detecting structure when it is very weak. Our algorithm, fastSTRUCTURE, is freely available online at http://pritchardlab.stanford.edu/structure.html.

IDENTIFYING the degree of admixture in individuals and inferring the population of origin of specific loci in these individuals is relevant for a variety of problems in population genetics. Examples include correcting for population stratification in genetic association studies (Pritchard and Donnelly 2001; Price *et al.* 2006), conservation genetics (Pearse and Crandall 2004; Randi 2008), and studying the ancestry and migration patterns of natural populations (Rosenberg *et al.* 2002; Reich *et al.* 2009; Catchen *et al.* 2013). With decreasing costs in sequencing and genotyping technologies, there is an increasing need for fast and accurate tools to infer population structure from very large genetic data sets.

Principal components analysis (PCA)-based methods for analyzing population structure, like EIGENSTRAT (Price *et al.* 2006) and SMARTPCA (Patterson *et al.* 2006), construct low-dimensional projections of the data that maximally retain the variance-covariance structure among the sample genotypes. The availability of fast and efficient algorithms for singular value decomposition has enabled PCA-based methods to become a popular choice for analyzing structure in genetic data sets. However, while these low-dimensional projections allow for straightforward visualization of the underlying population structure, it is not always straightforward to derive and interpret estimates for global ancestry of sample individuals from their projection coordinates (Novembre and Stephens 2008). In contrast, model-based approaches like STRUCTURE (Pritchard *et al.* 2000) propose an explicit generative model for the data based on the assumptions of Hardy-Weinberg equilibrium between alleles and linkage equilibrium between genotyped loci. Global ancestry estimates are then computed directly from posterior distributions of the model parameters, as done in STRUCTURE, or maximum-likelihood estimates of model parameters, as done in FRAPPE (Tang *et al.* 2005) and ADMIXTURE (Alexander *et al.* 2009).

STRUCTURE (Pritchard *et al.* 2000; Falush *et al.* 2003; Hubisz *et al.* 2009) takes a Bayesian approach to estimate global ancestry by sampling from the posterior distribution over global ancestry parameters using a Gibbs sampler that appropriately accounts for the conditional independence relationships between latent variables and model parameters.

doi: 10.1534/genetics.114.164350
Manuscript received December 2, 2013; accepted for publication March 25, 2014; published Early Online April 2, 2014.
Available freely online through the author-supported open access option.
Supporting information is available online at http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.114.164350/-/DC1.
[1]Corresponding author: Stanford University, 300 Pasteur Dr., Alway Bldg., M337, Stanford, CA 94305. E-mail: rajanil@stanford.edu

However, even well-designed sampling schemes need to generate a large number of posterior samples to resolve convergence and mixing issues and yield accurate estimates of ancestry proportions, greatly increasing the time complexity of inference for large genotype data sets. To provide faster estimation, FRAPPE and ADMIXTURE both use a maximum-likelihood approach. FRAPPE computes maximum-likelihood estimates of the parameters of the same model using an expectation-maximization algorithm, while ADMIXTURE computes the same estimates using a sequential quadratic programming algorithm with a quasi-Newton acceleration scheme. Our goal in this article is to adapt a popular approximate inference framework to greatly speed up inference of population structure while achieving accuracies comparable to STRUCTURE and ADMIXTURE.

Variational Bayesian inference aims to repose the problem of inference as an optimization problem rather than a sampling problem. Variational methods, originally used for approximating intractable integrals, have been used for a wide variety of applications in complex networks (Hofman and Wiggins 2008), machine learning (Jordan *et al.* 1998; Blei *et al.* 2003), and Bayesian variable selection (Logsdon *et al.* 2010; Carbonetto and Stephens 2012). Variational Bayesian techniques approximate the log-marginal likelihood of the data by proposing a family of tractable parametric posterior distributions (variational distribution) over hidden variables in the model; the goal is then to find the optimal member of this family that best approximates the marginal likelihood of the data (see *Models and Methods* for more details). Thus, a single optimization problem gives us both approximate analytical forms for the posterior distributions over unknown variables and an approximate estimate of the intractable marginal likelihood; the latter can be used to measure the support in the data for each model, and hence to compare models involving different numbers of populations. Some commonly used optimization algorithms for variational inference include the variational expectation-maximization algorithm (Beal 2003), collapsed variational inference (Teh *et al.* 2007), and stochastic gradient descent (Sato 2001).

In *Models and Methods*, we briefly describe the model underlying STRUCTURE and detail the framework for variational Bayesian inference that we use to infer the underlying ancestry proportions. We then propose a more flexible prior distribution over a subset of hidden parameters in the model and demonstrate that estimation of these hyperparameters using an empirical Bayesian framework improves the accuracy of global ancestry estimates when the underlying population structure is more difficult to resolve. Finally, we describe a scheme to accelerate computation of the optimal variational distributions and describe a set of scores to help evaluate the accuracy of the results and to help compare models involving different numbers of populations. In *Applications*, we compare the accuracy and time complexity of variational inference with those of STRUCTURE and ADMIXTURE on simulated genotype data sets and demonstrate the results of variational inference on a large data set genotyped in the Human Genome Diversity Panel.

Models and Methods

We now briefly describe our generative model for population structure followed by a detailed description of the variational framework used for model inference.

Variational inference

Suppose we have N diploid individuals genotyped at L biallelic loci. A population is represented by a set of allele frequencies at the L loci, $P_k \in [0, 1]^L$, $k \in \{1, \ldots, K\}$, where K denotes the number of populations. The allele being represented at each locus can be chosen arbitrarily. Allowing for admixed individuals in the sample, we assume each individual to be represented by a K-vector of admixture proportions, $Q_n \in [0, 1]^K, \sum_k Q_{nk} = 1, n \in \{1, \ldots, N\}$. Conditioned on Q_n, the population assignments of the two copies of a locus, $Z^a_{nl}, Z^b_{nl} \in \{0, 1\}^K$, $\sum_k Z^a_{nlk} = \sum_k Z^b_{nlk} = 1$, are assumed to be drawn from a multinomial distribution parametrized by Q_n. Conditioned on population assignments, the genotype at each locus G_{nl} is the sum of two independent Bernoulli-distributed random variables, each representing the allelic state of each copy of a locus and parameterized by population-specific allele frequencies. The generative process for the sampled genotypes can now be formalized as

- $p\left(Z^i_{nl}|Q_n\right) = \text{multinomial}(Q_n), i \in \{a, b\}, \forall n, l,$
- $p\left(G_{nl} = 0|[\![Z^a_{nl}]\!] = k, [\![Z^b_{nl}]\!] = k', P_{l\cdot}\right) = (1 - P_{lk})(1 - P_{lk'}),$
- $p\left(G_{nl} = 1|[\![Z^a_{nl}]\!] = k, [\![Z^b_{nl}]\!] = k', P_{l\cdot}\right) = P_{lk}(1 - P_{lk'}) + P_{lk'}(1 - P_{lk}),$
- $p\left(G_{nl} = 2|[\![Z^a_{nl}]\!] = k, [\![Z^b_{nl}]\!] = k', P_{l\cdot}\right) = P_{lk}P_{lk'},$

where $[\![Z]\!]$ denotes the nonzero indices of the vector Z.

Given the set of sampled genotypes, we can either compute the maximum-likelihood estimates of the parameters P and Q of the model (Tang *et al.* 2005; Alexander *et al.* 2009) or sample from the posterior distributions over the unobserved random variables Z^a, Z^b, P, and Q (Pritchard *et al.* 2000) to compute relevant moments of these variables.

Variational Bayesian (VB) inference formulates the problem of computing posterior distributions (and their relevant moments) into an optimization problem. The central aim is to find an element of a tractable family of probability distributions, called variational distributions, that is closest

to the true intractable posterior distribution of interest. A natural choice of distance on probability spaces is the Kullback–Leibler (KL) divergence, defined for a pair of probability distributions $q(x)$ and $p(x)$ as

$$D_{\mathrm{kl}}\big(q(x)\|\,p(x)\big) = \int q(x)\log\frac{q(x)}{p(x)}\,dx. \tag{1}$$

Given the asymmetry of the KL divergence, VB inference chooses $p(x)$ to be the intractable posterior and $q(x)$ to be the variational distribution; this choice allows us to compute expectations with respect to the tractable variational distribution, often exactly.

An approximation to the true intractable posterior distribution can be computed by minimizing the KL divergence between the true posterior and variational distribution. We will restrict our optimization over a variational family that explicitly assumes independence between the latent variables (Z^a, Z^b) and parameters (P, Q); this restriction to a space of fully factorizable distributions is commonly called the *mean field approximation* in the statistical physics (Kadanoff 2009) and machine-learning literature (Jordan *et al.* 1998)). Since this assumption is certainly not true when inferring population structure, the true posterior will not be a member of the variational family and we will be able to find only the fully factorizable variational distribution that best approximates the true posterior. Nevertheless, this approximation significantly simplifies the optimization problem. Furthermore, we observe empirically that this approximation achieves reasonably accurate estimates of lower-order moments (*e.g.*, posterior mean and variance) when the true posterior is replaced by the variational distributions (*e.g.*, when computing prediction error on held-out entries of the genotype matrix). The variational family we choose here is

$$\begin{aligned} q\big(Z^a, Z^b, Q, P\big) &\approx q\big(Z^a, Z^b\big)q(Q,P) \\ &= \prod_{n,l} q\big(Z^a_{nl}\big)q\big(Z^b_{nl}\big)\cdot\prod_n q(Q_n)\cdot\prod_{lk} q(P_{lk}), \end{aligned} \tag{2}$$

where each factor can then be written as

$$\begin{aligned} q\big(Z^a_{nl}\big) &= \mathrm{multinomial}\big(\tilde{Z}^a_{nl}\big) \\ q\big(Z^b_{nl}\big) &= \mathrm{multinomial}\big(\tilde{Z}^b_{nl}\big) \\ q\big(Q_n\big) &= \mathrm{Dirichlet}\big(\tilde{Q}_n\big) \\ q\big(P_{lk}\big) &= \mathrm{Beta}\big(\tilde{P}^u_{lk}, \tilde{P}^v_{lk}\big). \end{aligned} \tag{3}$$

$\tilde{Z}^a_{nl}$, $\tilde{Z}^b_{nl}$, $\tilde{Q}_n$, $\tilde{P}^u_{lk}$, and $\tilde{P}^v_{lk}$ are the parameters of the variational distributions (variational parameters). The choice of the variational family is restricted only by the tractability of computing expectations with respect to the variational distributions; here, we choose parametric distributions that are conjugate to the distributions in the likelihood function.

In addition, the KL divergence (Equation 1) quantifies the tightness of a lower bound to the log-marginal likelihood of the data (Beal 2003). Specifically, for any variational distribution $q(Z^a, Z^b, P, Q)$, we have

$$\begin{aligned} \log p(G|K) = \mathcal{E}\big[q\big(Z^a, Z^b, Q, P\big)\big] \\ + D_{kl}\big(q\big(Z^a, Z^b, Q, P\big)\|p\big(Z^a, Z^b, Q, P|G\big)\big), \end{aligned} \tag{4}$$

where $\mathcal{E}$ is a lower bound to the log-marginal likelihood of the data, $\log p(G|K)$. Thus, minimizing the KL divergence is equivalent to maximizing the log-marginal likelihood lower bound (LLBO) of the data:

$$\begin{aligned} q^* &= \arg\min_q\ D_{kl}\big(q\big(Z^a, Z^b, Q, P\big)\big\|\,p\big(Z^a, Z^b, Q, P|G\big)\big) \\ &= \arg\min_q\ (\log p(G|K) - \mathcal{E}[q]) \\ &= \arg\max_q \mathcal{E}[q]. \end{aligned} \tag{5}$$

The LLBO of the observed genotypes can be written as

$$\begin{aligned} \mathcal{E} &= \sum_{Z^a,Z^b}\int q\big(Z^a, Z^b, Q, P\big)\ \log\frac{p\big(G, Z^a, Z^b, Q, P\big)}{q\big(Z^a, Z^b, Q, P\big)}\,dQ\,dP \\ &= \sum_{Z^a,Z^b}\int q\big(Z^a, Z^b, P\big)\log p\big(G|Z^a, Z^b, P\big)dP \\ &\quad + \sum_{Z^a,Z^b}\int q\big(Z^a, Z^b, Q\big)\log p\big(Z^a, Z^b|Q\big)dQ \\ &\quad + D_{kl}(q(Q)\|p(Q)) + D_{kl}(q(P)\|\,p(P)), \end{aligned} \tag{6}$$

where $p(Q)$ is the prior on the admixture proportions and $p(P)$ is the prior on the allele frequencies. The LLBO of the data in terms of the variational parameters is specified in Appendix A. The LLBO depends on the model, and particularly on the number of populations K. Using simulations, we assess the utility of the LLBO as a heuristic to help select appropriate values for K.

Priors

The choice of priors $p(Q_n)$ and $p(P_{lk})$ plays an important role in inference, particularly when the F_{ST} between the underlying populations is small and population structure is difficult to resolve. Typical genotype data sets contain hundreds of thousands of genetic variants typed in several hundreds of samples. Given the small sample sizes in these data relative to underlying population structure, the posterior distribution over population allele frequencies can be difficult to estimate; thus, the prior over P_{lk} plays a more important role in accurate inference than the prior over admixture proportions. Throughout this study, we choose a symmetric Dirichlet prior over admixture proportions; $p(Q_n) = \mathrm{Dirichlet}\ \big(\frac{1}{K}\mathbf{1}_K\big)$.

Depending on the difficulty in resolving structure in a given data set, we suggest using one of three priors over

allele frequencies. A flat beta-prior over population-specific allele frequencies at each locus, $p(P_{lk}) = \text{Beta}(1, 1)$ (referred to as "simple prior" throughout), has the advantage of computational speed but comes with the cost of potentially not resolving subtle structure. For genetic data where structure is difficult to resolve, the *F*-model for population structure (Falush *et al.* 2003) proposes a hierarchical prior, based on a demographic model that allows the allele frequencies of the populations to have a shared underlying pattern at all loci. Assuming a star-shaped genealogy where each of the populations simultaneously split from an ancestral population, the allele frequency at a given locus is generated from a beta distribution centered at the ancestral allele frequency at that locus, with variance parametrized by a population-specific drift from the ancestral population (we refer to this prior as *F*-prior"):

$$p(P_{lk}) = \text{Beta}\left(P_l^A \frac{1-F_k}{F_k}, \left(1-P_l^A\right)\frac{1-F_k}{F_k}\right). \tag{7}$$

Alternatively, we propose a hierarchical prior that is more flexible than the *F*-prior and allows for more tractable inference, particularly when additional priors on the hyperparameters need to be imposed. At a given locus, the population-specific allele frequency is generated by a logistic normal distribution, with the normal distribution having a locus-specific mean and a population-specific variance (we refer to this prior as logistic prior):

$$\begin{aligned} P_{lk} &= \frac{1}{1+\exp^{-R_{lk}}} \\ p(R_{lk}) &= \mathcal{N}(\mu_l, \lambda_k). \end{aligned} \tag{8}$$

Having specified the appropriate prior distributions, the optimal variational parameters can be computed by iteratively minimizing the KL divergence (or, equivalently, maximizing the LLBO) with respect to each variational parameter, keeping the other variational parameters fixed. The LLBO is concave in each parameter; thus, convergence properties of this iterative optimization algorithm, also called the variational Bayesian expectation-maximization algorithm, are similar to those of the expectation-maximization algorithm for maximum-likelihood problems. The update equations for each of the three models are detailed in *Appendix A*. Furthermore, when population structure is difficult to resolve, we propose updating the hyperparameters ((F, P^A) for the *F*-prior and (μ, λ) for the logistic prior) by maximizing the LLBO with respect to these variables; conditional on these hyperparameter values, improved estimates for the variational parameters are then computed by minimizing the KL divergence. Although such a hyperparameter update is based on optimizing a lower bound on the marginal likelihood, it is likely (although not guaranteed) to increase the marginal likelihood of the data, often leading to better inference. A natural extension of this hierarchical prior would be to allow for a full locus-independent variance–covariance matrix (Pickrell and Pritchard 2012). However, we observed in our simulations that estimating the parameters of the full matrix led to worse prediction accuracy on held-out data. Thus, we did not consider this extension in our analyses.

Accelerated variational inference

Similar to the EM algorithm, the convergence of the iterative algorithm for variational inference can be quite slow. Treating the iterative update equations for the set of variational parameters $\tilde{\theta}$ as a deterministic map $\Phi(\tilde{\theta}^{(t)})$, a globally convergent algorithm with improved convergence rates can be derived by adapting the Cauchy–Barzilai–Borwein method for accelerating the convergence of linear fixed-point problems (Raydan and Svaiter 2002) to the nonlinear fixed-point problem given by our deterministic map (Varadhan and Roland 2008). Specifically, given a current estimate of parameters $\tilde{\theta}^{(t)}$, the new estimate can be written as

$$\tilde{\theta}^{(t+1)}(\nu_t) = \tilde{\theta}^{(t)} - 2\nu_t\,\Delta_t + \nu_t^2\,H_t, \tag{9}$$

where $\Delta_t = \Phi\left(\tilde{\theta}^{(t)}\right) - \tilde{\theta}^{(t)}$, $H_t = \Phi\left(\Phi\left(\tilde{\theta}^{(t)}\right)\right) - 2\Phi\left(\tilde{\theta}^{(t)}\right) + \tilde{\theta}^{(t)}$ and $\nu_t = -||\Delta_t||/||H_t||$. Note that the new estimate is a continuous function of ν_t and the standard variational iterative scheme can be obtained from Equation 9 by setting ν_t to -1. Thus, for values of ν_t close to -1, the accelerated algorithm retains the stability and monotonicity of standard EM algorithms while sacrificing a gain in convergence rate. When $\nu_t < -1$, we gain significant improvement in convergence rate, with two potential problems: (a) the LLBO could decrease, *i.e.*, $\mathcal{E}\left(\tilde{\theta}^{(t+1)}\right) < \mathcal{E}\left(\tilde{\theta}^{(t)}\right)$, and (b) the new estimate $\tilde{\theta}^{(t+1)}$ might not satisfy the constraints of the optimization problem. In our experiments, we observe the first problem to occur rarely and we resolve this by simply testing for convergence of the magnitude of difference in LLBO at successive iterations. We resolve the second problem using a simple back-tracking strategy of halving the distance between ν_t and -1: $\nu_t \leftarrow (\nu_t - 1)/2$, until the new estimate $\tilde{\theta}^{(t+1)}$ satisfies the constraints of the optimization problem.

Validation scores

For each simulated data set, we evaluate the accuracy of each algorithm using two metrics: accuracy of the estimated admixture proportions and the prediction error for a subset of entries in the genotype matrix that are held out before estimating the parameters. For a given choice of model complexity *K*, an estimate of the admixture proportions Q^* is taken to be the maximum-likelihood estimate of Q when using ADMIXTURE, the maximum *a posteriori* (MAP) estimate of Q when using STRUCTURE, and the mean of the variational distribution over Q inferred using fastSTRUCTURE. We measure the accuracy of Q^* by computing the Jensen–Shannon (JS) divergence between Q^* and the true admixture proportions. The Jensen–Shannon divergence (JSD) between two probability vectors P and Q is a bounded distance metric defined as

$$\mathrm{JSD}(P\|Q) = \tfrac{1}{2}D_{kl}(P\|M) + \tfrac{1}{2}D_{kl}(Q\|M), \qquad (10)$$

where $M = \frac{1}{2}(P+Q)$, and $0 \leq \mathrm{JSD}(P\|Q) \leq 1$. Note that if the lengths of P and Q are not the same, the smaller vector is extended by appending zero-valued entries. The mean admixture divergence is then defined as the minimum over all permutations of population labels of the mean JS divergence between the true and estimated admixture proportions over all samples, with higher divergence values corresponding to lower accuracy.

We evaluate the prediction accuracy by estimating model parameters (or posterior distributions over them) after holding out a subset $\mathcal{M}$ of the entries in the genotype matrix. For each held-out entry, the expected genotype is estimated by ADMIXTURE from maximum-likelihood parameter estimates as

$$\hat{G}_{nl} = 2\sum_k P^*_{lk} Q^*_{nk}, \qquad (11)$$

where P^*_{lk} is the maximum-likelihood estimate of P_{lk}. The expected genotype given the variational distributions requires integration over the model parameters and is derived in *Appendix B*. Given the expected genotypes for the held-out entries, for a specified model complexity K, the prediction error is quantified by the deviance residuals under the binomial model averaged over all entries:

$$d_K(\hat{G}, G) = \sum_{n,l\in\mathcal{M}} G_{nl}\log\frac{G_{nl}}{\hat{G}_{nl}} + (2-G_{nl})\log\frac{2-G_{nl}}{2-\hat{G}_{nl}}. \qquad (12)$$

Model complexity

ADMIXTURE suggests choosing the value of model complexity K that achieves the smallest value of $d_K(\hat{G}, G)$, *i.e.*, $K^*_{cv} = \mathrm{argmin}_K d_K(\hat{G}, G)$. We propose two additional metrics to select model complexity in the context of variational Bayesian inference. Assuming a uniform prior on K, the optimal model complexity $K^*_{\mathcal{L}}$ is chosen to be the one that maximizes the LLBO, where the LLBO is used as an approximation to the marginal likelihood of the data. However, since the difference between the log-marginal likelihood of the data and the LLBO is difficult to quantify, the trend of LLBO as a function of K cannot be guaranteed to match that of the log-marginal likelihood. Additionally, we propose a useful heuristic to choose K based on the tendency of mean-field variational schemes to populate only those model components that are essential to explain patterns underlying the observed data. Specifically, given an estimate of Q^* obtained from variational inference executed for a choice of K, we compute the ancestry contribution of each model component as the mean admixture proportion over all samples, *i.e.*, $c_k = \frac{1}{N}\sum_n Q^*_{nk}$. The number of relevant model components $K_{\varnothing^c}$ is then the minimum number of populations that have a cumulative ancestry contribution of at least 99.99%,

$$K_{\varnothing^c} = \min\Big\{|S| : S \in \mathcal{P}(\mathcal{K}) and \sum_{k\in S} c_k > 0.9999\Big\}, \qquad (13)$$

where $\mathcal{K} = \{1, \ldots, K\}$ and $\mathcal{P}(\mathcal{K})$ is the power set of $\mathcal{K}$. As K increases, $K_{\varnothing^c}$ tends to approach a limit that can be chosen as the optimal model complexity $K^*_{\varnothing^c}$.

Applications

In this section, we compare the accuracy and runtime performance of the variational inference framework with the results of STRUCTURE and ADMIXTURE both on data sets generated from the *F*-model and on the Human Genome Diversity Panel (HGDP) (Rosenberg *et al.* 2002). We expect the results of ADMIXTURE to match those of FRAPPE (Tang *et al.* 2005) since they both compute maximum-likelihood estimates of the model parameters. However, ADMIXTURE converges faster than FRAPPE, allowing us to compare it with fastSTRUCTURE using thousands of simulations. In general, we observe that fastSTRUCTURE estimates ancestry proportions with accuracies comparable to, and sometimes better than, those estimated by ADMIXTURE even when the underlying population structure is rather weak. Furthermore, fastSTRUCTURE is about 2 orders of magnitude faster than STRUCTURE and has comparable runtimes to that of ADMIXTURE. Finally, fastSTRUCTURE gives us a reasonable range of values for the model complexity required to explain structure underlying the data, without the need for a cross-validation scheme. Below, we highlight the key advantages and disadvantages of variational inference in each problem setting.

Simulated data sets

To evaluate the performance of the different learning algorithms, we generated two groups of simulated genotype data sets, with each genotype matrix consisting of 600 samples and 2500 loci. The first group was used to evaluate the accuracy of the algorithms as a function of strength of the underlying population structure while the second group was used to evaluate accuracy as a function of number of underlying populations. Although the size of each genotype matrix was kept fixed in these simulations, the performance characteristics of the algorithms are expected to be similar if the strength of population structure is kept fixed and the data set size is varied (Patterson *et al.* 2006).

For the first group, the samples were drawn from a three-population demographic model as shown in Figure 1A. The edge weights correspond to the parameter F in the model that quantifies the genetic drift of each of the three current populations from an ancestral population. We introduced a scaling factor $r \in [0, 1]$ that quantifies the resolvability of population structure underlying the samples. Scaling F by r reduces the amount of drift of current populations from the ancestral population; thus, structure is difficult to resolve when r is close to 0, while structure is easy to resolve when r is close to 1. For each $r \in \{0.05, 0.10, \ldots, 0.95, 1\}$, we

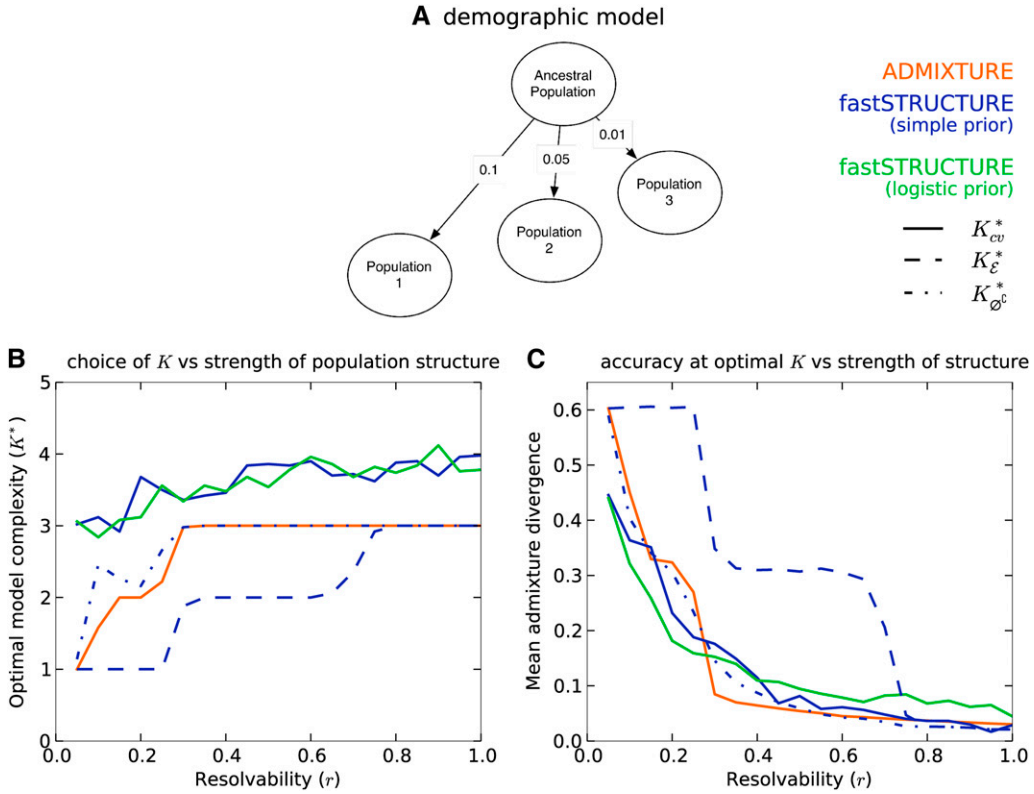

Figure 1 Accuracy of different algorithms as a function of resolvability of population structure. (A) Demographic model underlying the three populations represented in the simulated data sets. The edge weights quantify the amount of drift from the ancestral population. (B and C) Resolvability is a scalar by which the population-specific drifts in the demographic model are multiplied, with higher values of resolvability corresponding to stronger structure. (B) Compares the optimal model complexity given the data, averaged over 50 replicates, inferred by ADMIXTURE (K^*_{cv}), fastSTRUCTURE with simple prior (K^*_{cv}, $K^*_{\mathcal{E}}$, $K^*_{\mathcal{E}^c}$), and fastSTRUCTURE with logistic prior (K^*_{cv}). (C) Compares the accuracy of admixture proportions, averaged over replicates, estimated by each algorithm at the optimal value of K in each replicate.

generated 50 replicate data sets. The ancestral allele frequencies π^A for each data set were drawn from the frequency spectrum computed using the HGDP panel to simulate allele frequencies in natural populations. For each data set, the allele frequency at a given locus for each population was drawn from a beta-distribution with mean π^A_l and variance $rF_k\pi^A_l(1-\pi^A_l)$, and the admixture proportions for each sample were drawn from a symmetric Dirichlet distribution, namely Dirichlet$\left(\frac{1}{10}\mathbf{1}_3\right)$, to simulate small amounts of gene flow between the three populations. Finally, 10% of the samples in each data set, randomly selected, were assigned to one of the three populations with zero admixture.

For the second group, the samples were drawn from a star-shaped demographic model with K_t populations. Each population was assumed to have equal drift from an ancestral population, with the F parameter fixed at either 0.01 to simulate weak structure or 0.04 to simulate strong structure. The ancestral allele frequencies were simulated similar to the first group and 50 replicate data sets were generated for this group for each value of $K_t \in \{1, \ldots, 5\}$. We executed ADMIXTURE and fastSTRUCTURE for each data set with various choices of model complexity: for data sets in the first group, model complexity $K \in \{1, \ldots, 5\}$, and for those in the second group $K \in \{1, \ldots, 8\}$. We executed ADMIXTURE with default parameter settings; with these settings the algorithm terminates when the increase in log likelihood is $<10^{-4}$ and computes prediction error using fivefold cross-validation. fastSTRUCTURE was executed with a convergence criterion of change in the per-genotype log-marginal likelihood lower bound $|\Delta\mathcal{E}| < 10^{-8}$. We held out 20 random disjoint genotype sets, each containing 1% of entries in the genotype matrix and used the mean and standard error of the deviance residuals for these held-out entries as an estimate of the prediction error.

For each group of simulated data sets, we illustrate a comparison of the performance of ADMIXTURE and fastSTRUCTURE with the simple and the logistic prior. When structure was easy to resolve, both the F-prior and the logistic prior returned similar results; however, the logistic prior returned more accurate ancestry estimates when structure was difficult to resolve. Plots including results using the F-prior are shown in Supporting Information, Figure S1, Figure S2, and Figure S3. Since ADMIXTURE uses held-out deviance residuals to choose model complexity, we demonstrate the results of the two algorithms, each using deviance residuals to choose K, using solid lines in Figure 1 and Figure 2. Additionally, in these figures, we also illustrate the performance of fastSTRUCTURE, when using the two alternative metrics to choose model complexity, using blue lines.

Choice of K

One question that arises when applying admixture models in practice is how to select the model complexity, or number of populations, K. It is important to note that in practice there will generally be no "true" value of K, because samples from real populations will never conform exactly to the assumptions of the model. Further, inferred values of K could be influenced by sampling ascertainment schemes (Engelhardt

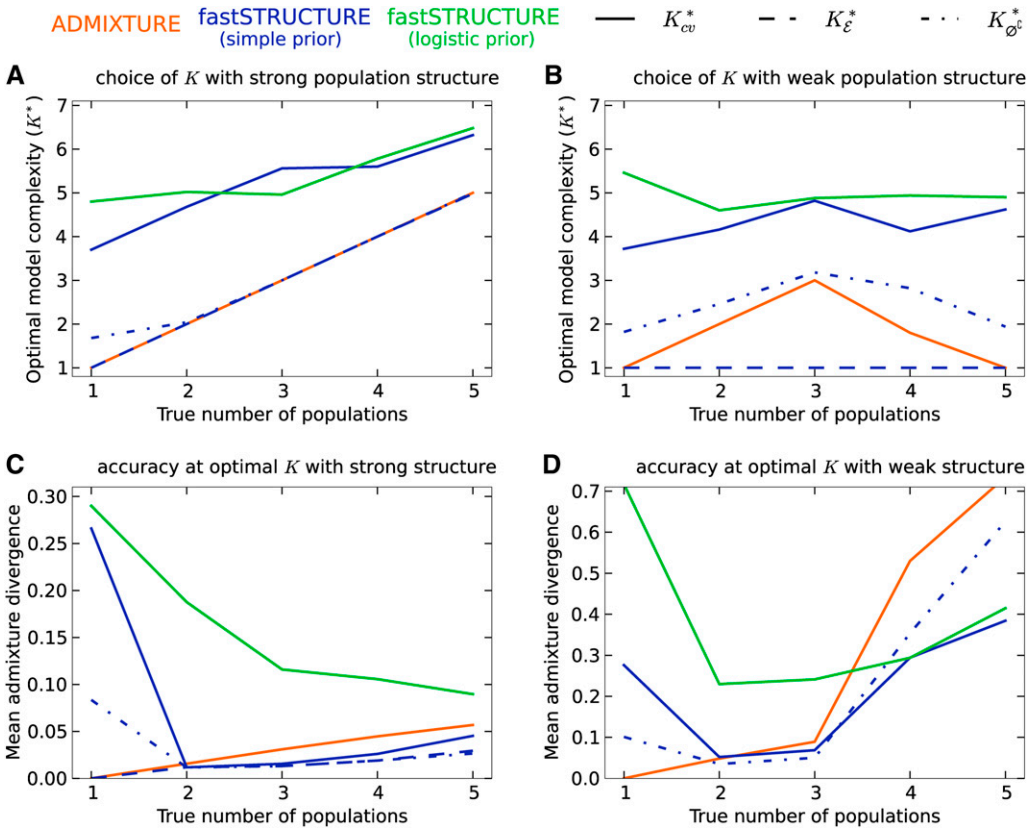

Figure 2 Accuracy of different algorithms as a function of the true number of populations. The demographic model is a star-shaped genealogy with populations having undergone equal amounts of drift. Subfigures A and C correspond to strong structure ($F = 0.04$) and B and D to weak structure ($F = 0.01$). (A and B) Compare the optimal model complexity estimated by the different algorithms using various metrics, averaged over 50 replicates, to the true number of populations represented in the data. Notably, when population structure is weak, both ADMIXTURE and fastSTRUCTURE fail to detect structure when the number of populations is too large. (C and D) Compare the accuracy of admixture proportions estimated by each algorithm at the optimal model complexity for each replicate.

and Stephens 2010) (imagine sampling from g distinct locations in a continuous habitat exhibiting isolation by distance—any automated approach to select K will be influenced by g), and by the number of typed loci (as more loci are typed, more subtle structure can be picked up, and inferred values of K may increase). Nonetheless, it can be helpful to have automated heuristic rules to help guide the analyst in making the appropriate choice for K, even if the resulting inferences need to be carefully interpreted within the context of prior knowledge about the data and sampling scheme. Therefore, we here used simulation to assess several different heuristics for selecting K.

The manual of the ADMIXTURE code proposes choosing model complexity that minimizes the prediction error on held-out data estimated using the mean deviance residuals reported by the algorithm (K^*_{cv}). In Figure 1B, using the first group of simulations, we compare the value of K^*_{cv}, averaged over 50 replicate data sets, between the two algorithms as a function of the resolvability of population structure in the data. We observe that while deviance residuals estimated by ADMIXTURE robustly identify an appropriate model complexity, the value of K identified using deviance residuals computed using the variational parameters from fastSTRUCTURE appear to overestimate the value of K underlying the data. However, on closer inspection, we observe that the difference in prediction errors between large values of K are statistically insignificant (Figure 3, middle). This suggests the following heuristic: select the lowest model complexity above which prediction errors do not vary significantly.

Alternatively, for fastSTRUCTURE with the simple prior, we propose two additional metrics for choosing model complexity: (1) $K^*_{\mathcal{E}}$, value of K that maximizes the LLBO of the entire data set, and (2) $K^*_{\varnothing^c}$, the limiting value, as K increases, of the smallest number of model components that accounts for almost all of the ancestry in the sample. In Figure 1B, we observe that $K^*_{\mathcal{E}}$ has the attractive property of robustly identifying strong structure underlying the data, while $K^*_{\varnothing^c}$ identifies additional model components needed to explain weak structure in the data, with a slight upward bias in complexity when the underlying structure is extremely difficult to resolve. For the second group of simulations, similar to results observed for the first group, when population structure is easy to resolve, ADMIXTURE robustly identifies the correct value of K (shown in Figure 2A). However, for similar reasons as before, the use of prediction error with fastSTRUCTURE tends to systematically overestimate the number of populations underlying the data. In contrast, $K^*_{\mathcal{E}}$ and $K^*_{\varnothing^c}$ match exactly to the true K when population structure is strong. When the underlying population structure is very weak, $K^*_{\mathcal{E}}$ is a severe underestimate of the true K while $K^*_{\varnothing^c}$ slightly overestimates the value of K. Surprisingly, K^*_{cv} estimated using ADMIXTURE and $K^*_{\varnothing^c}$ estimated using fastSTRUCTURE tend to underestimate the number of populations when the true number of populations K_t is large, as shown in Figure 2B.

For a new data set, we suggest executing fastSTRUCTURE for multiple values of K and estimating $(K^*_{\mathcal{E}}, K^*_{\varnothing^c})$ to obtain

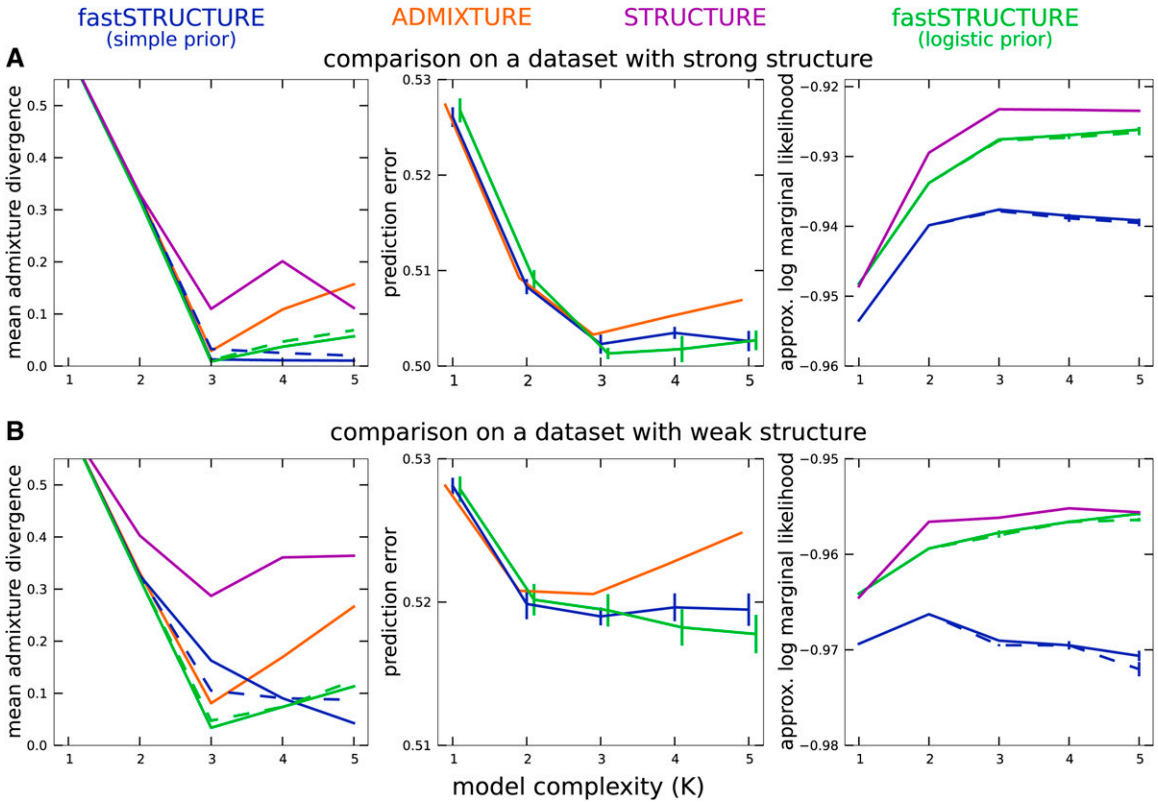

Figure 3 Accuracy of different algorithms as a function of model complexity (K) on two simulated data sets, one in which ancestry is easy to resolve (A; $r = 1$) and one in which ancestry is difficult to resolve: (B; $r = 0.5$) Solid lines correspond to parameter estimates computed with a convergence criterion of $|\Delta\mathcal{E}| < 10^{-8}$, while the dashed lines correspond to a weaker criterion of $|\Delta\mathcal{E}| < 10^{-6}$. (Left) Mean admixture divergence between the true and inferred admixture proportions; (middle) mean binomial deviance of held-out genotype entries. Note that for values of K greater than the optimal value, any change in prediction error lies within the standard error of estimates of prediction error suggesting that we should choose the smallest value of model complexity above which a decrease in prediction error is statistically insignificant. (Right) Approximations to the marginal likelihood of the data computed by STRUCTURE and fastSTRUCTURE.

a reasonable range of values for the number of populations that would explain structure in the data, under the given model. To look for subtle structure in the data, we suggest executing fastSTRUCTURE with the logistic prior with values for values of K similar to those identified by using the simple prior.

Accuracy of ancestry proportions

We evaluated the accuracy of the algorithms by comparing the divergence between the true admixture proportions and the estimated admixture proportions at the optimal model complexity computed using the above metrics for each data set. In Figure 1C, we plot the mean divergence between the true and estimated admixture proportions, over multiple replicates, as a function of resolvability. We observe that the admixture proportions estimated by fastSTRUCTURE at $K^*_{\mathcal{E}}$ have high divergence; however, this is a result of LLBO being too conservative in identifying K. At $K = K^*_{cv}$ and $K = K^*_{\varnothing^c}$, fastSTRUCTURE estimates admixture proportions with accuracies comparable to, and sometimes better than, ADMIXTURE even when the underlying population structure is rather weak. Furthermore, the held-out prediction deviances computed using posterior estimates from variational algorithms are consistently smaller than those estimated by ADMIXTURE (see Figure S3) demonstrating the improved accuracy of variational Bayesian inference schemes over maximum-likelihood methods. Similarly, for the second group of simulated data sets, we observe in Figure 2, C and D, that the accuracy of variational algorithms tends to be comparable to or better than that of ADMIXTURE, particularly when structure is difficult to resolve. When structure is easy to resolve, the increased divergence estimates of fastSTRUCTURE with the logistic prior result from the upward bias in the estimate of K^*_{cv}; this can be improved by using cross-validation more carefully in choosing model complexity.

Visualizing ancestry estimates

Having demonstrated the performance of fastSTRUCTURE on multiple simulated data sets, we now illustrate the performance characteristics and parameter estimates using two specific data sets (selected from the first group of simulated data sets), one with strong population structure ($r = 1$) and one with weak structure ($r = 0.5$). In addition to these algorithms, we executed STRUCTURE for these two data sets using the model of independent allele frequencies to directly compare with the results of fastSTRUCTURE. For each data set, α was kept fixed to $\frac{1}{K}$ for all populations,

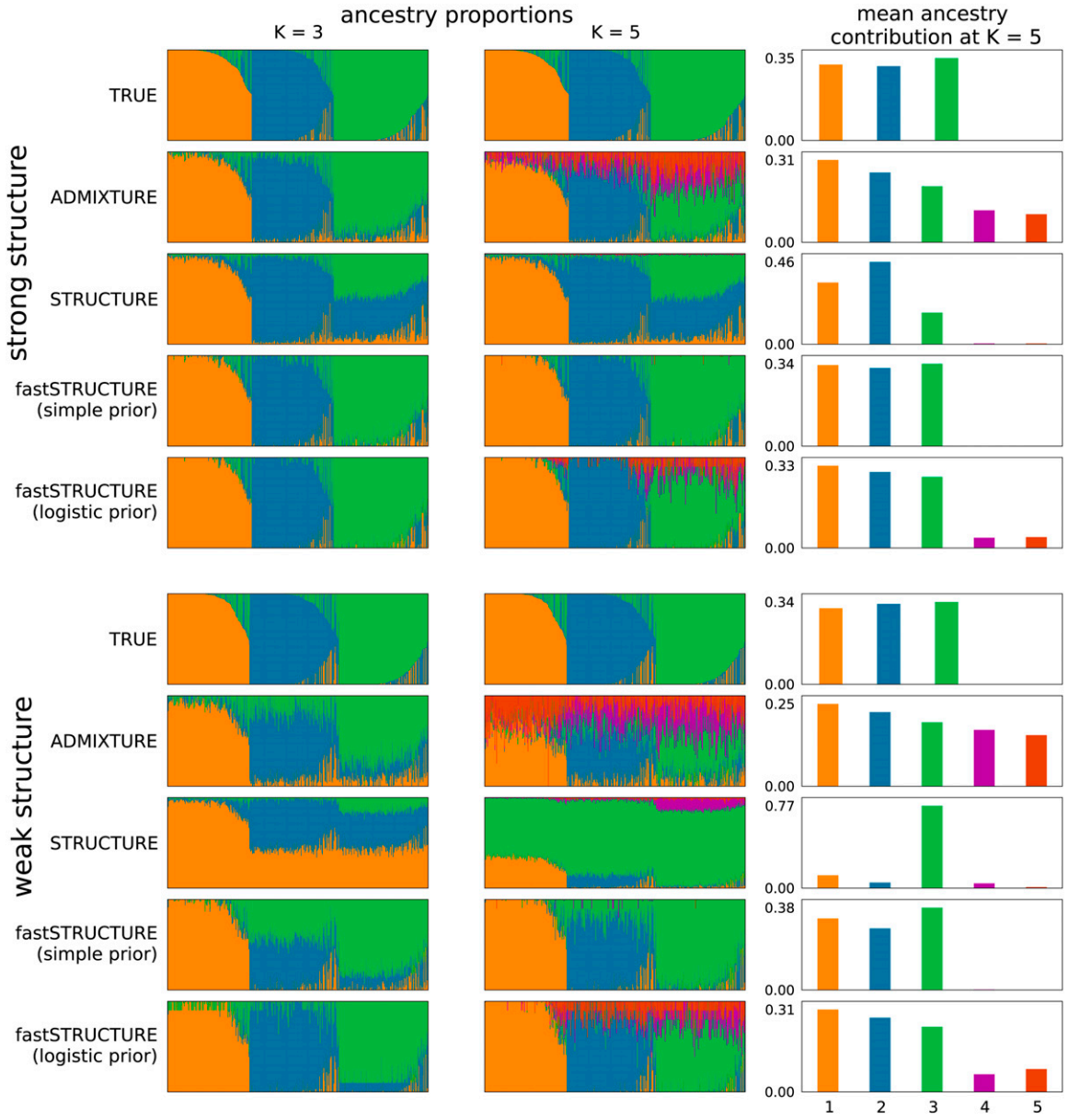

Figure 4 Visualizing ancestry proportions estimated by different algorithms on two simulated data sets, one with strong structure (top, $r = 1$) and one with weak structure (bottom, $r = 0.5$). (Left and middle) Ancestry estimated at model complexity of $K = 3$ and $K = 5$, respectively. Insets illustrate the true ancestry and the ancestry inferred by each algorithm. Each color represents a population and each individual is represented by a vertical line partitioned into colored segments whose lengths represent the admixture proportions from K populations. (Right) Mean ancestry contributions of the model components, when the model complexity $K = 5$.

similar to the prior used for fastSTRUCTURE, and each run consisted of 50,000 burn-in steps and 50,000 MCMC steps. In Figure 3, we illustrate the divergence of admixture estimates and the prediction error on held-out data each as a function of K. For all choices of K greater than or equal to the true value, the accuracy of fastSTRUCTURE, measured using both admixture divergence and prediction error, is generally comparable to or better than that of ADMIXTURE and STRUCTURE, even when the underlying population structure is rather weak. In Figure 3, right, we plot the approximate marginal likelihood of the data, reported by STRUCTURE, and the optimal LLBO, computed by fastSTRUCTURE, each as a function of K. We note that the looseness of the bound between STRUCTURE and fastSTRUCTURE can make the LLBO a less reliable measure to choose model complexity than the approximate marginal likelihood reported by STRUCTURE, particularly when the size of the data set is not sufficient to resolve the underlying population structure.

Figure 4 illustrates the admixture proportions estimated by the different algorithms on both data sets at two values of K, using Distruct plots (Rosenberg 2004). For the larger choice of model complexity, we observe that fastSTRUCTURE with the simple prior uses only those model components that are necessary to explain the data, allowing for automatic

inference of model complexity (Mackay 2003). To better illustrate this property of unsupervised Bayesian inference methods, Figure 4, right, shows the mean contribution of ancestry from each model component to samples in the data set. While ADMIXTURE uses all components of the model to fit the data, STRUCTURE and fastSTRUCTURE assign negligible posterior mass to model components that are not required to capture structure in the data. The number of nonempty model components ($K_{\varnothing^c}$) automatically identifies the model complexity required to explain the data; the optimal model complexity $K^*_{\varnothing^c}$ is then the mode of all values of $K_{\varnothing^c}$ computed for different choices of K. While both STRUCTURE and fastSTRUCTURE tend to use only those model components necessary to explain the data, fastSTRUCTURE is slightly more aggressive in removing model components that seem unnecessary, leading to slightly improved results for fastSTRUCTURE compared to STRUCTURE in Equation 4, when there is strong structure in the data set. This property of fastSTRUCTURE seems useful in identifying global patterns of structure in a data set (*e.g.*, the populations represented in a set of samples); however, it can be an important drawback if one is interested in detecting weak signatures of gene flow from a population to a specific sample in a given data set.

When population structure is difficult to resolve, imposing a logistic prior and estimating its parameters using the data are likely to increase the power to detect weak structure. However, estimation of the hierarchical prior parameters by maximizing the approximate marginal likelihood also makes the model susceptible to overfitting by encouraging a small set of samples to be randomly, and often confidently, assigned to unnecessary components of the model. To correct for this, when using the logistic prior, we suggest estimating the variational parameters with multiple random restarts and using the mean of the parameters corresponding to the top five values of LLBO. To ensure consistent population labels when computing the mean, we permuted the labels for each set of variational parameter estimates to find the permutation with the lowest pairwise Jensen–Shannon divergence between admixture proportions among pairs of restarts. Admixture estimates computed using this scheme show improved robustness against overfitting, as illustrated in Figure 4. Moreover, the pairwise Jensen–Shannon divergence between admixture proportions among all restarts of the variational algorithms can also be used as a measure of the robustness of their results and as a signature of how strongly they overfit the data.

Runtime performance

A key advantage of variational Bayesian inference algorithms compared to inference algorithms based on sampling is the dramatic improvement in time complexity of the algorithm. To evaluate the runtimes of the different learning algorithms, we generated from the *F*-model data sets with sample sizes $N \in \{200, 600\}$ and numbers of loci $L \in \{500, 2500\}$, each having three populations with $r = 1$. The time complexity of each of the above algorithms is linear in the number of samples, loci, and populations, *i.e.*, O(*NLK*); in

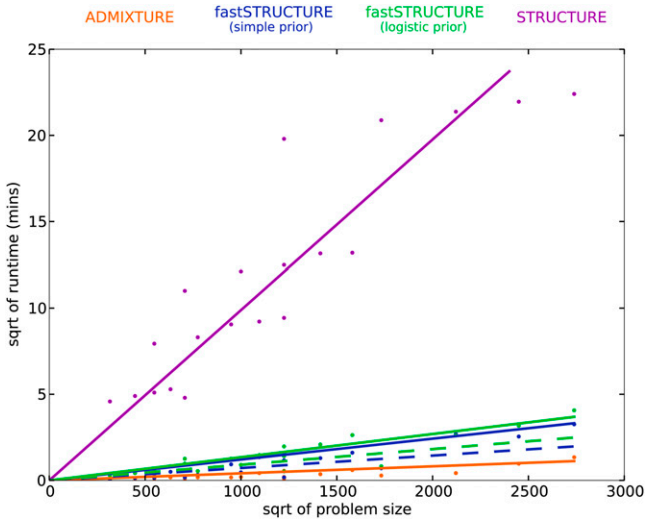

Figure 5 Runtimes of different algorithms on simulated data sets with different number of loci and samples; the square root of runtime (in minutes) is plotted as a function of square root of problem size (defined as $N \times L \times K$). Similar to Figure 3, dashed lines correspond to a weaker convergence criterion than solid lines.

comparison, the time complexity of principal components analysis is quadratic in the number of samples and linear in the number of loci. In Figure 5, the mean runtime of the different algorithms is shown as a function of problem size defined as $N \times L \times K$. The added complexity of the cost function being optimized in fastSTRUCTURE increases its runtime when compared to ADMIXTURE. However, fastSTRUCTURE is about 2 orders of magnitude faster than STRUCTURE, making it suitable for large data sets with hundreds of thousands of genetic variants. For example, using a data set with 1000 samples genotyped at 500,000 loci with $K = 10$, each iteration of our current Python implementation of fastSTRUCTURE with the simple prior takes about 11 min, while each iteration of ADMIXTURE takes ~16 min. Since one would usually like to estimate the variational parameters for multiple values of K for a new data set, a faster algorithm that gives an approximate estimate of ancestry proportions in the sample would be of much utility, particularly to guide an appropriate choice of K. We observe in our simulations that a weaker convergence criterion of $|\Delta\mathcal{E}| < 10^{-6}$ gives us comparably accurate results with much shorter run times, illustrated by the dashed lines in Figure 3 and Figure 5. Based on these observations, we suggest executing multiple random restarts of the algorithm with a weak convergence criterion of $|\Delta\mathcal{E}| < 10^{-5}$ to rapidly obtain reasonably accurate estimates of the variational parameters, prediction errors, and ancestry contributions from relevant model components.

HGDP panel

We now compare the results of ADMIXTURE and fastSTRUCTURE on a large, well-studied data set of genotypes at single nucleotide polymorphisms (SNP) genotyped in the HGDP (Li *et al.* 2008), in which 1048 individuals from 51 different populations were genotyped using Illumina's

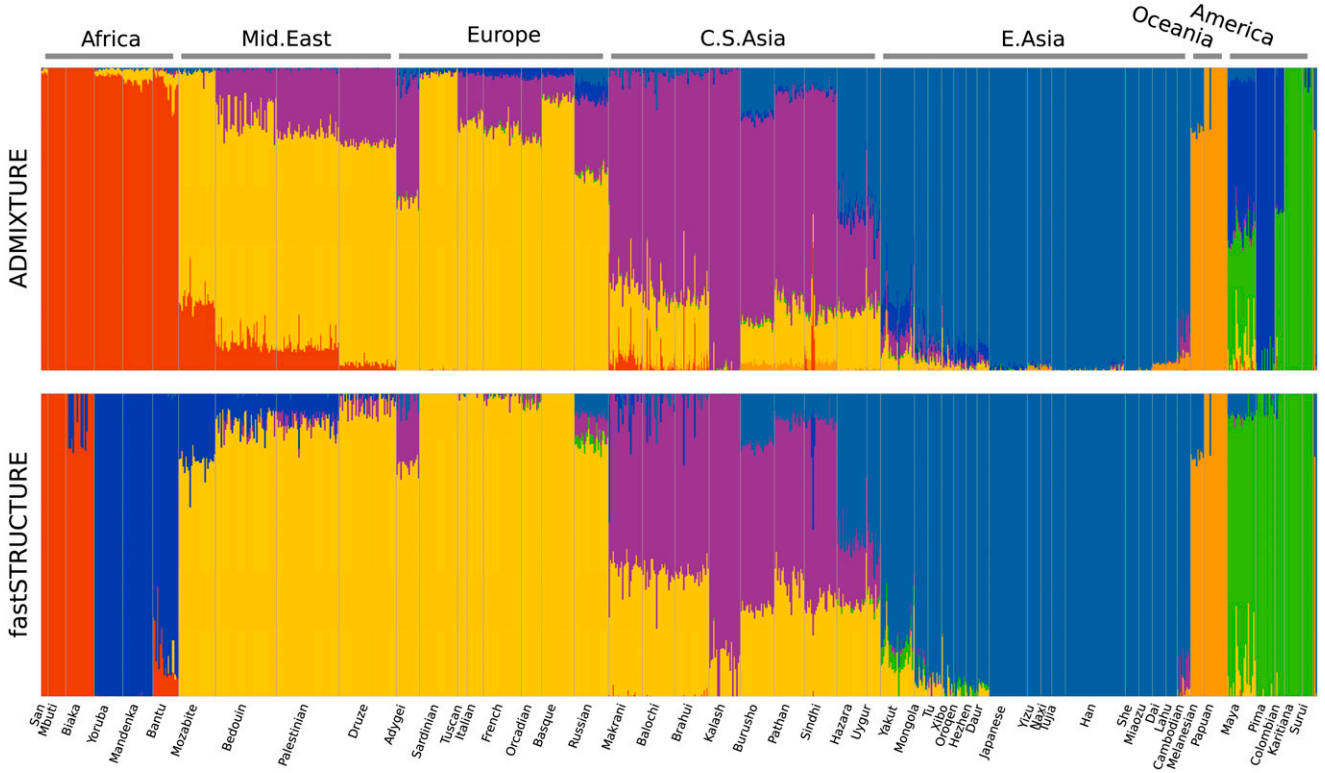

Figure 6 Ancestry proportions inferred by ADMIXTURE and fastSTRUCTURE (with the simple prior) on the HGDP data at $K = 7$ (Li *et al.* 2008). Notably, ADMIXTURE splits the Central and South American populations into two groups while fastSTRUCTURE assigns higher approximate marginal likelihood to a split of sub-Saharan African populations into two groups.

HumanHap650Y platform. We used the set of 938 "unrelated" individuals for the analysis in this article. For the selected set of individuals, we removed SNPs that were monomorphic, had missing genotypes in >5% of the samples, and failed the Hardy–Weinberg Equilibrium (HWE) test at $P < 0.05$ cutoff. To test for violations from HWE, we selected three population groups that have relatively little population structure (East Asia, Europe, Bantu Africa), constructed three large groups of individuals from these populations, and performed a test for HWE for each SNP within each large group. The final data set contained 938 samples with genotypes at 657,143 loci, with 0.1% of the entries in the genotype matrix missing. We executed ADMIXTURE and fastSTRUCTURE using this data set with allowed model complexity $K \in \{5, \ldots, 15\}$. In Figure 6, the ancestry proportions estimated by ADMIXTURE and fastSTRUCTURE at $K = 7$ are shown; this value of K was chosen to compare with results reported using the same data set with FRAPPE (Li *et al.* 2008). In contrast to results reported using FRAPPE, we observe that both ADMIXTURE and fastSTRUCTURE identify the Mozabite, Bedouin, Palestinian, and Druze populations as very closely related to European populations with some gene flow from Central-Asian and African populations; this result was robust over multiple random restarts of each algorithm. Since both ADMIXTURE and FRAPPE maximize the same likelihood function, the slight difference in results is likely due to differences in the modes of the likelihood surface to which the two algorithms converge. A notable difference between ADMIXTURE and fastSTRUCTURE is in their choice of the seventh population—ADMIXTURE splits the Native American populations along a north–south divide while fastSTRUCTURE splits the African populations into central African and south African population groups.

Interestingly, both algorithms strongly suggest the existence of additional weak population structure underlying the data, as shown in Figure 7. ADMIXTURE, using cross-validation, identifies the optimal model complexity to be 11; however, the deviance residuals appear to change very little beyond $K = 7$, suggesting that the model components identified at $K = 7$ explain most of the structure underlying the data. The results of the heuristics implemented in fastSTRUCTURE are largely concordant, with $K^*_{\mathcal{E}} = 7$, $K^*_{\varnothing^c} = 9$ and the lowest cross-validation error obtained at $K^*_{cv} = 10$.

The admixture proportions estimated at the optimal choices of model complexity using the different metrics are shown in Figure 8. The admixture proportions estimated at $K = 7$ and $K = 9$ are remarkably similar, with the Kalash and Karitiana populations being assigned to their own model components at $K = 9$. These results demonstrate the ability of LLBO to identify strong structure underlying the data and that of $K_{\varnothing^c}$ to identify additional weak structure that explain variation in the data. At $K = 10$ (as identified using cross-validation), we observe that only nine of the

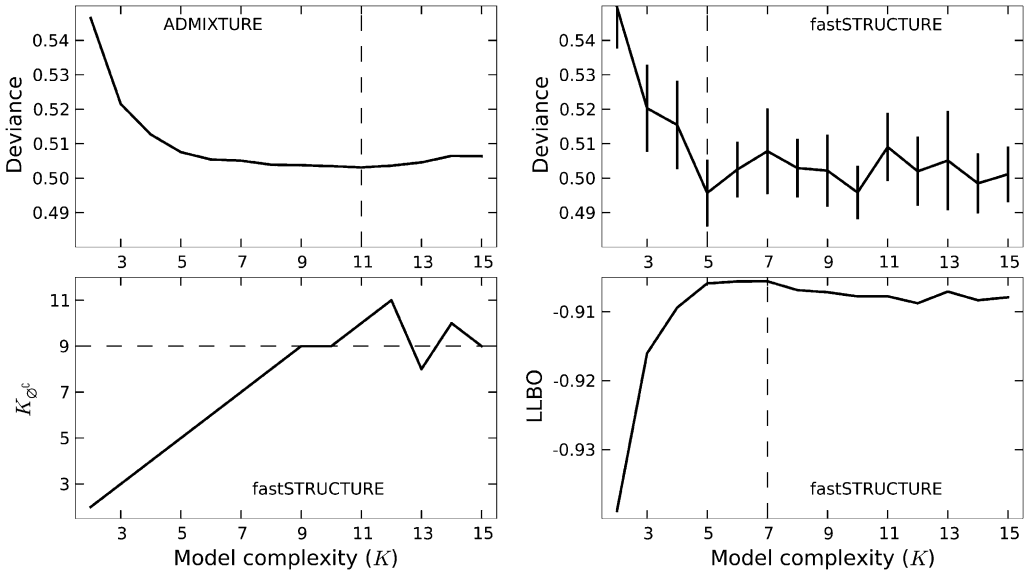

Figure 7 Model choice of ADMIXTURE and fastSTRUCTURE (with the simple prior) on the HGDP data. Optimal value of *K*, identified by ADMIXTURE using deviance residuals, and by fastSTRUCTURE using deviance, $K_{\varnothing^c}$, and LLBO, are shown by a dashed line.

model components are populated. However, the estimated admixture proportions differ crucially with all African populations grouped together, the Melanesian and Papuan populations each assigned to their own groups, and the Middle-Eastern populations represented as predominantly an admixture of Europeans and a Bedouin subpopulation with small amounts of gene flow from Central-Asian populations.

The main contribution of this work is a fast, approximate inference algorithm for one simple admixture model for population structure, used in ADMIXTURE and STRUCTURE. While admixture may not be an exactly correct model for most population data sets, this model often gives key insights into the population structure underlying samples in a new data set and is useful in identifying global patterns of structure in the samples. Exploring model choice, by comparing the goodness-of-fit of different models that capture demographies of varying complexity, is an important future direction.

Discussion

Our analyses on simulated and natural data sets demonstrate that fastSTRUCTURE estimates approximate posterior distributions on ancestry proportions 2 orders of magnitude faster than STRUCTURE, with ancestry estimates and prediction accuracies that are comparable to those of ADMIXTURE. Posing the problem of inference in terms of an optimization problem allows us to draw on powerful tools in convex optimization and plays an important role in the gain in speed achieved by variational inference schemes, when compared to the Gibbs sampling scheme used in STRUCTURE. In addition, the flexible logistic prior enables us to resolve subtle structure underlying a data set. The considerable improvement in runtime with comparable accuracies allows the application of these methods to large genotype data sets that are steadily becoming the norm in studies of population history, genetic association with disease, and conservation biology.

The choice of model complexity, or the number of populations required to explain structure in a data set, is a difficult problem associated with the inference of population structure. Unlike in maximum-likelihood estimation, the model parameters have been integrated out in variational inference schemes and optimizing the KL divergence in fastSTRUCTURE does not run the risk of overfitting. The heuristic scores that we have proposed to identify model complexity provide a robust and reasonable range for the number of populations underlying the data set, without the need for a time-consuming cross-validation scheme.

As in the original version of STRUCTURE, the model underlying fastSTRUCTURE does not explicitly account for linkage disequilibrium (LD) between genetic markers. While LD between genotype markers in the genotype data set will lead us to underestimate the variance of the approximate posterior distributions, the improved accuracy in predicting held-out genotypes for the HGDP data set demonstrates that the underestimate due to unmodeled LD and the mean field approximation is not too severe. Furthermore, not accounting for LD appropriately can lead to significant biases in local ancestry estimation, depending on the sample size and population haplotype frequencies. However, we believe global ancestry estimates are likely to incur very little bias due to unmodeled LD. One potential source of bias in global ancestry estimates is due to LD driven by segregating, chromosomal inversions. While genetic variants on inversions on the human genome and those of different model organisms are fairly well characterized and can be easily masked, it is important to identify and remove genetic variants that lie in inversions for nonmodel organisms, to avoid them from biasing global ancestry estimates. One heuristic approach to searching for such large blocks would be to compute a measure of differentiation for each locus between one population and the remaining populations, using the inferred variational posteriors on allele frequencies. Long stretches of the genome that have highly differentiated

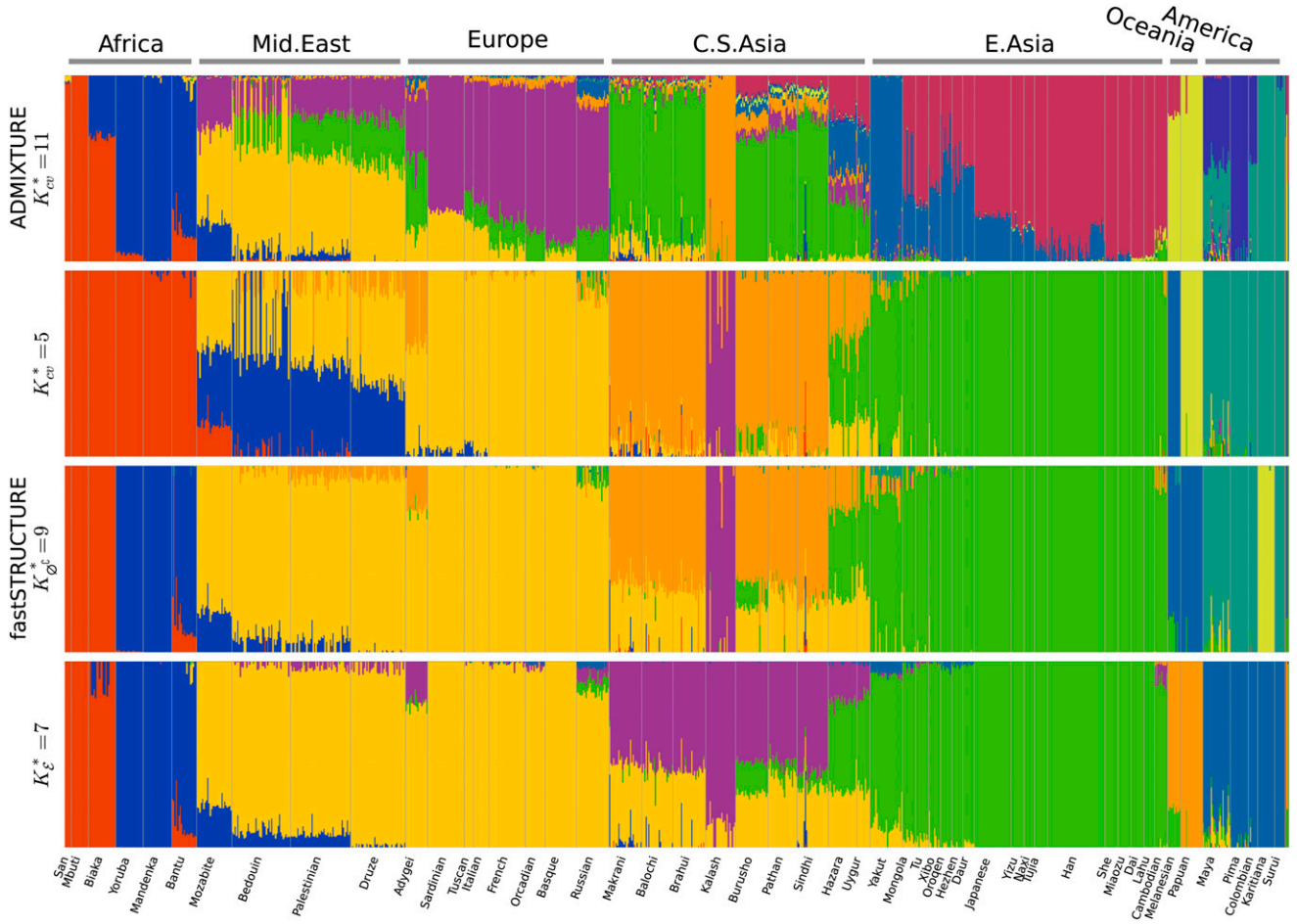

Figure 8 Ancestry proportions inferred by ADMIXTURE and fastSTRUCTURE (with the simple prior) at the optimal choice of K identified by relevant metrics for each algorithm. Notably, the admixture proportions at $K = K^*_{\mathcal{E}}$ and $K = K^*_{\emptyset^c}$ are quite similar, with estimates in the latter case identifying the Kalash and Karitiana as additional separate groups that share very little ancestry with the remaining populations.

genetic variants can then be removed before recomputing ancestry estimates.

In summary, we have presented a variational framework for fast, accurate inference of global ancestry of samples genotyped at a large number of genetic markers. For a new data set, we recommend executing our program, fastSTRUCTURE, for multiple values of K to obtain a reasonable range of values for the appropriate model complexity required to explain structure in the data, as well as ancestry estimates at those model complexities. For improved ancestry estimates and to identify subtle structure, we recommend executing fastSTRUCTURE with the logistic prior at values of K similar to those identified when using the simple prior. Our program is available for download at http://pritchardlab.stanford.edu/structure.html.

Acknowledgments

We thank Tim Flutre, Shyam Gopalakrishnan, and Ida Moltke for fruitful discussions on this project and the editor and two anonymous reviewers for their helpful comments and suggestions. This work was funded by grants from the National Institutes of Health (HG007036,HG002585) and by the Howard Hughes Medical Institute.

Literature Cited

Alexander, D. H., J. Novembre, and K. Lange, 2009 Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 19(9): 1655–1664.

Beal, M. J., 2003 Variational algorithms for approximate Bayesian inference. Ph.D. Thesis, Gatsby Computational Neuroscience Unit, University College London, London.

Blei, D. M., A. Y. Ng, and M. I. Jordan, 2003 Latent dirichlet allocation. J. Mach. Learn. Res. 3: 993–1022.

Carbonetto, P., and M. Stephens, 2012 Scalable variational inference for Bayesian variable selection in regression, and its accuracy in genetic association studies. Bayesian Anal. 7(1): 73–108.

Catchen, J., S. Bassham, T. Wilson, M. Currey, C. O'Brien *et al.*, 2013 The population structure and recent colonization history of Oregon threespine stickleback determined using restriction-site associated DNA-sequencing. Mol. Ecol. 22: 2864–2883.

Engelhardt, B. E., and M. Stephens, 2010 Analysis of population structure: a unifying framework and novel methods based on sparse factor analysis. PLoS Genet. 6(9): e1001117.

Falush, D., M. Stephens, and J. K. Pritchard, 2003 Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics 164: 1567–1587.
Hofman, J. M., and C. H. Wiggins, 2008 Bayesian approach to network modularity. Phys. Rev. Lett. 100(25): 258701.
Hubisz, M. J., D. Falush, M. Stephens, and J. K. Pritchard, 2009 Inferring weak population structure with the assistance of sample group information. Mol. Ecol. Res. 9(5): 1322–1332.
Jordan, M. I., Z. Gharamani, T. S. Jaakkola, and L. K. Saul, 1999 An introduction to variational methods for graphical models. Mach. Learn. 37(2): 183–233.
Kadanoff, L. P., 2009 More is the same: phase transitions and mean field theories. J. Stat. Phys. 137(5–6): 777–797.
Li, J. Z., D. M. Absher, H. Tang, A. M. Southwick, A. M. Casto *et al.*, 2008 Worldwide human relationships inferred from genome-wide patterns of variation. Science 319(5866): 1100–1104.
Logsdon, B. A., G. E. Hoffman, and J. G. Mezey, 2010 A variational Bayes algorithm for fast and accurate multiple locus genome-wide association analysis. BMC Bioinformatics 11(1): 58.
Mackay, D. J., 2003 Information theory, inference and learning algorithms. Cambridge University Press, Cambridge, UK.
Novembre, J., and M. Stephens, 2008 Interpreting principal component analyses of spatial population genetic variation. Nat. Genet. 40(5): 646–649.
Patterson, N., A. L. Price, and D. Reich, 2006 Population structure and eigenanalysis. PLoS Genet. 2(12): e190.
Pearse, D., and K. Crandall, 2004 Beyond FST: analysis of population genetic data for conservation. Conserv. Genet. 5(5): 585–602.
Pickrell, J. K., and J. K. Pritchard, 2012 Inference of population splits and mixtures from genomewide allele frequency data. PLoS Genet. 8(11): e1002967.
Price, A. L., N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick *et al.*, 2006 Principal components analysis corrects for stratification in genomewide association studies. Nat. Genet. 38 (8): 904–909.
Pritchard, J. K., and P. Donnelly, 2001 Case-control studies of association in structured or admixed populations. Theor. Popul. Biol. 60(3): 227–237.
Pritchard, J. K., M. Stephens, and P. Donnelly, 2000 Inference of population structure using multilocus genotype data. Genetics 155: 945–959.
Randi, E., 2008 Detecting hybridization between wild species and their domesticated relatives. Mol. Ecol. 17(1): 285–293.
Raydan, M., and B. F. Svaiter, 2002 Relaxed steepest descent and Cauchy–Barzilai–Borwein method. Comput. Optim. Appl. 21(2): 155–167.
Reich, D., K. Thangaraj, N. Patterson, A. L. Price, and L. Singh, 2009 Reconstructing Indian population history. Nature 461 (7263): 489–494.
Rosenberg, N. A., 2004 DISTRUCT: a program for the graphical display of population structure. Mol. Ecol. Notes 4(1): 137–138.
Rosenberg, N. A., J. K. Pritchard, J. L. Weber, H. M. Cann, K. K. Kidd *et al.*, 2002 Genetic structure of human populations. Science 298(5602): 2381–2385.
Sato, M. A., 2001 Online model selection based on the variational Bayes. Neural Comput. 13(7): 1649–1681.
Tang, H., J. Peng, P. Wang, and N. J. Risch, 2005 Estimation of individual admixture: analytical and study design considerations. Genet. Epidemiol. 28(4): 289–301.
Teh, Y. W., D. Newman, and M. Welling, 2007 A collapsed variational Bayesian inference algorithm for latent Dirichlet allocation. Adv. Neural Inf. Process. Syst. 19: 1353.
Varadhan, R., and C. Roland, 2008 Simple and globally convergent methods for accelerating the convergence of any EM algorithm. Scand. J. Stat. 35(2): 335–353.

Communicating editor: M. K. Uyenoyama

Appendix A

Given the parametric forms for the variational distributions and a choice of prior for the fastSTRUCTURE model, the per-genotype LLBO is given as

$$\begin{aligned}\mathcal{E} = \frac{1}{\mathcal{G}}\sum_{n,l}\delta(\mathbf{G}_{nl})\Bigg\{\sum_k\left(\mathbf{E}\left[Z^a_{nlk}\right]+\mathbf{E}\left[Z^b_{nlk}\right]\right)\left(\mathbb{I}[\mathbf{G}_{nl}=0]\mathbf{E}\left[\log(1-P)_{lk}\right]+\mathbb{I}[\mathbf{G}_{nl}=2]\,\mathbf{E}[\log P_{lk}]+\mathbf{E}[\log Q_{nk}]\right)\\ +\mathbb{I}[\mathbf{G}_{nl}=1]\sum_k\left(\mathbf{E}\left[Z^a_{nlk}\right]\mathbf{E}[\log P_{lk}]+\mathbf{E}\left[Z^b_{nlk}\right]\mathbf{E}[\log(1-P_{lk})]\right)-\mathbf{E}\left[\log Z^a_{nl}\right]-\mathbf{E}\left[\log Z^b_{nl}\right]\Bigg\}\\ +\sum_{l,k}\log\frac{\mathrm{B}(\tilde{P}^u_{lk},\tilde{P}^v_{lk})}{\mathrm{B}(\beta,\gamma)}+(\beta-\tilde{P}^u_{lk})\mathbf{E}[\log P_{lk}]+(\gamma-\tilde{P}^v_{lk})\mathbf{E}[\log(1-P_{lk})]\\ +\sum_n\Bigg\{\sum_k(\alpha_k-\tilde{Q}_{nk})\mathbf{E}[\log Q_{nk}]+\log\Gamma(\alpha_k)-\log\Gamma(\tilde{Q}_{nk})\Bigg\}+\log\Gamma(\tilde{Q}_{no})-\log\Gamma(\alpha_o),\end{aligned} \tag{A1}$$

where $\mathbf{E}[\cdot]$ is the expectation taken with respect to the appropriate variational distribution, $\mathrm{B}(\cdot)$ is the beta function, $\Gamma(\cdot)$ is the gamma function, $\{\alpha, \beta, \gamma\}$ are the hyperparameters in the model, $\delta(\cdot)$ is an indicator variable that takes the value of zero if the genotype is missing, $\mathcal{G}$ is the number of observed entries in the genotype matrix, $\alpha_o = \sum_k \alpha_k$, and $\tilde{Q}_{no} = \sum_k \tilde{Q}_{nk}$. Maximizing this lower bound for each variational parameter, keeping the other parameters fixed, gives us the following update equations:

$(\tilde{Z}^a, \tilde{Z}^b)$:

$$\tilde{Z}^a_{nlk} \propto \exp\left\{\Psi^a_{\mathbf{G}_{nl}} - \psi(\tilde{P}^u_{lk}+\tilde{P}^v_{lk}) + \psi(\tilde{Q}_{nk}) - \psi(\tilde{Q}_{no})\right\} \tag{A2}$$

$$\tilde{Z}^b_{nlk} \propto \exp\left\{\Psi^b_{\mathbf{G}_{nl}} - \psi\left(\tilde{P}^u_{lk}+\tilde{P}^v_{lk}\right) + \psi(\tilde{Q}_{nk}) - \psi(\tilde{Q}_{no})\right\}, \tag{A3}$$

where

$$\Psi^a_{\mathbf{G}_{nl}} = \mathbb{I}[\mathbf{G}_{nl}=0]\psi\left(\tilde{P}^v_{lk}\right)+\mathbb{I}[\mathbf{G}_{nl}=1]\psi\left(\tilde{P}^u_{lk}\right)+\mathbb{I}[\mathbf{G}_{nl}=2]\psi\left(\tilde{P}^u_{lk}\right) \tag{A4}$$

$$\Psi^b_{\mathbf{G}_{nl}} = \mathbb{I}[\mathbf{G}_{nl}=0]\psi\left(\tilde{P}^v_{lk}\right)+\mathbb{I}[\mathbf{G}_{nl}=1]\psi\left(\tilde{P}^v_{lk}\right)+\mathbb{I}[\mathbf{G}_{nl}=2]\psi\left(\tilde{P}^u_{lk}\right) \tag{A5}$$

$\tilde{Q}$:

$$\tilde{Q}_{nk} = \alpha_k + \sum_l \delta(\mathbf{G}_{nl})\left(\tilde{Z}^a_{nlk}+\tilde{Z}^b_{nlk}\right) \tag{A6}$$

$(\tilde{P}^u, \tilde{P}^v)$:

$$\tilde{P}^u_{lk} = \beta + \sum_n\left(\mathbb{I}[\mathbf{G}_{nl}=1]\;\tilde{Z}^a_{nlk}+\mathbb{I}[\mathbf{G}_{nl}=2]\left(\tilde{Z}^a_{nlk}+\tilde{Z}^b_{nlk}\right)\right) \tag{A7}$$

$$\tilde{P}^v_{lk} = \gamma + \sum_n\left(\mathbb{I}[\mathbf{G}_{nl}=1]\;\tilde{Z}^b_{nlk}+\mathbb{I}[\mathbf{G}_{nl}=0]\left(\tilde{Z}^a_{nlk}+\tilde{Z}^b_{nlk}\right)\right). \tag{A8}$$

In the above update equations, $\psi(\cdot)$ is the digamma function. When the *F*-prior is used, the LLBO and the update equations remain exactly the same, after replacing β with $\pi^A_l([1-F_k]/F_k)$ and γ with $(1-\pi^A_l)([1-F_k]/F_k)$. In this case, the LLBO is also maximized with respect to the hyperparameter F using the L-BFGS-B algorithm, a quasi-Newton code for bound-constrained optimization.

When the logistic prior is used, a straightforward maximization of the LLBO no longer gives us explicit update equations for $\tilde{P}^u_{lk}$ and $\tilde{P}^v_{lk}$. One alternative is to use a constrained optimization solver, like L-BFGS-B; however, the large number of variational parameters to be optimized greatly increases the per-iteration computational cost of the inference algorithm. Instead, we propose update equations for $\tilde{P}^u_{lk}$ and $\tilde{P}^v_{lk}$ to have a similar form as those obtained with the simple prior,

$$\tilde{P}^u_{lk} = \beta_{lk} + \sum_n \mathbb{I}[\mathbf{G}_{nl} = 1]\tilde{Z}^a_{nlk} + \mathbb{I}[\mathbf{G}_{nl} = 2]\left(\tilde{Z}^a_{nlk} + \tilde{Z}^b_{nlk}\right) \tag{A9}$$

$$\tilde{P}^v_{lk} = \gamma_{lk} + \sum_n \mathbb{I}[\mathbf{G}_{nl} = 1]\tilde{Z}^b_{nlk} + \mathbb{I}[\mathbf{G}_{nl} = 0]\left(\tilde{Z}^a_{nlk} + \tilde{Z}^b_{nlk}\right), \tag{A10}$$

where β_{lk} and γ_{lk} implicitly depend on $\tilde{P}^u_{lk}$ and $\tilde{P}^v_{lk}$ as follows:

$$\begin{aligned}(\psi'(\tilde{P}^u_{lk}) - \psi'(\tilde{P}^u_{lk} + \tilde{P}^v_{lk}))\beta_{lk} - \psi'(\tilde{P}^u_{lk} + \tilde{P}^v_{lk})\,\gamma_{lk} = &-\lambda_k\psi'(\tilde{P}^u_{lk})\left(\psi(\tilde{P}^u_{lk}) - \psi(\tilde{P}^v_{lk}) - \mu_l\right) - \frac{1}{2}\lambda_k\psi''(\tilde{P}^u_{lk}) - \psi'(\tilde{P}^u_{lk} + \tilde{P}^v_{lk})\beta_{lk} \\ &+\left(\psi'\left(\tilde{P}^u_{lk}\right) - \psi'\left(\tilde{P}^u_{lk} + \tilde{P}^v_{lk}\right)\right)\gamma_{lk} = \lambda_k\psi'\left(\tilde{P}^v_{lk}\right)\left(\psi\left(\tilde{P}^u_{lk}\right) - \psi\left(\tilde{P}^v_{lk}\right) - \mu_l\right) \\ &-\frac{1}{2}\lambda_k\psi''\left(\tilde{P}^v_{lk}\right).\end{aligned} \tag{A11}$$

The optimal values for $\tilde{P}^u_{lk}$ and $\tilde{P}^v_{lk}$ can be obtained by iterating between the two sets of equations to convergence. Thus, when the logistic prior is used, the algorithm is implemented as a nested iterative scheme where for each update of all the variational parameters, an iterative scheme computes the update for $(\tilde{P}^u, \tilde{P}^v)$. Finally, the optimal value of the hyperparameter μ is obtained straightforwardly as

$$\mu_l = \sum_k \lambda_k\left(\psi\left(\tilde{P}^u_{lk}\right) - \psi\left(\tilde{P}^v_{lk}\right)\right)\Big/\sum_k \lambda_k \tag{A12}$$

while the optimal λ is computed using a constrained optimization solver.

Appendix B

Given the observed genotypes $\mathbf{G}$, the probability of the unobserved genotype $\mathbf{G}^{\text{hid}}_{nl}$ for the nth sample at the lth locus is given as

$$p(\mathbf{G}^{\text{hid}}_{nl}|\mathbf{G}) = \int p(\mathbf{G}^{\text{hid}}_{nl}|P, Q)p(P, Q|\mathbf{G})\,dQ\,dP. \tag{B1}$$

Replacing the posterior $p(P, Q|\mathbf{G})$ with the optimal variational posterior distribution, we obtain

$$p(\mathbf{G}^{\text{hid}}_{nl} = 0) \approx \int p(\mathbf{G}^{\text{hid}}_{nl} = 0|P, Q)q(P)q(Q)\,dQ\,dP \tag{B2}$$

$$= \sum_{k,k'} \int Q_{nk}Q_{nk'}(1 - P_{lk})(1 - P_{lk'})q(P)q(Q)\,dQ\,dP \tag{B3}$$

$$= \sum_{k\neq k'} \mathbf{E}[Q_{nk}Q_{nk'}](1 - \mathbf{E}[P_{lk}])(1 - \mathbf{E}[P_{lk'}]) \tag{B4}$$

$$+ \sum_{k=k'} \mathbf{E}\left[Q^2_{nk}\right]\mathbf{E}\left[(1-P_{lk})^2\right] \tag{B5}$$

$$p(\mathbf{G}^{\text{hid}}_{nl} = 1) \approx \int p(\mathbf{G}^{\text{hid}}_{nl} = 1|P, Q)q(P)q(Q)\,dQ\,dP \tag{B6}$$

$$= 2\sum_{k,k'} \int Q_{nk} Q_{nk'} P_{lk}(1 - P_{lk'}) q(P) q(Q)\, dQ\, dP \tag{B7}$$

$$= \sum_{k \neq k'} \mathbf{E}\left[Q_{nk} Q_{nk'}\right] \mathbf{E}[P_{lk}]\left(1 - \mathbf{E}\left[P_{lk'}\right]\right) \tag{B8}$$

$$+ \sum_{k=k'} \mathbf{E}\left[Q_{nk}^2\right] \mathbf{E}\left[P_{lk}(1 - P_{lk})\right] \tag{B9}$$

$$p(\mathbf{G}_{nl}^{\text{hid}} = 2) \approx \int p(\mathbf{G}_{nl}^{\text{hid}} = 2 | P, Q) q(P) q(Q)\, dQ\, dP \tag{B10}$$

$$= \sum_{k,k'} \int Q_{nk} Q_{nk'} P_{lk} P_{lk'} q(P) q(Q)\, dQ\, dP \tag{B11}$$

$$= \sum_{k \neq k'} \mathbf{E}\left[Q_{nk} Q_{nk'}\right] \mathbf{E}[P_{lk}] \mathbf{E}\left[P_{lk'}\right] \tag{B12}$$

$$+ \sum_{k=k'} \mathbf{E}\left[Q_{nk}^2\right] \mathbf{E}\left[P_{lk}^2\right], \tag{B13}$$

where

$$\mathbf{E}\left[Q_{nk} Q_{nk'}\right] = \frac{\tilde{Q}_{nk} \tilde{Q}_{nk'}}{\tilde{Q}_{no}(\tilde{Q}_{no} + 1)} \tag{B14}$$

$$\mathbf{E}\left[Q_{nk}^2\right] = \frac{\tilde{Q}_{nk}(\tilde{Q}_{nk} + 1)}{\tilde{Q}_{no}(\tilde{Q}_{no} + 1)} \tag{B15}$$

$$\mathbf{E}[P_{lk}] = \frac{\tilde{P}_{lk}^u}{\tilde{P}_{lk}^u + \tilde{P}_{lk}^v} \tag{B16}$$

$$\mathbf{E}\left[P_{lk}^2\right] = \frac{\tilde{P}_{lk}^u(\tilde{P}_{lk}^u + 1)}{(\tilde{P}_{lk}^u + \tilde{P}_{lk}^v)(\tilde{P}_{lk}^u + \tilde{P}_{lk}^v + 1)} \tag{B17}$$

$$\mathbf{E}[P_{lk}(1 - P_{lk})] = \frac{\tilde{P}_{lk}^u \tilde{P}_{lk}^v}{(\tilde{P}_{lk}^u + \tilde{P}_{lk}^v)(\tilde{P}_{lk}^u + \tilde{P}_{lk}^v + 1)} \tag{B18}$$

$$\mathbf{E}\left[(1 - P_{lk})^2\right] = \frac{\tilde{P}_{lk}^v(\tilde{P}_{lk}^v + 1)}{(\tilde{P}_{lk}^u + \tilde{P}_{lk}^v)(\tilde{P}_{lk}^u + \tilde{P}_{lk}^v + 1)}. \tag{B19}$$

$$\tag{B20}$$

The expected genotype can then be straightforwardly computed from these genotype probabilities.

GENETICS

Supporting Information
http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.114.164350/-/DC1

fastSTRUCTURE: Variational Inference of Population Structure in Large SNP Data Sets

Anil Raj, Matthew Stephens, and Jonathan K. Pritchard

DOI: 10.1534/genetics.114.164350

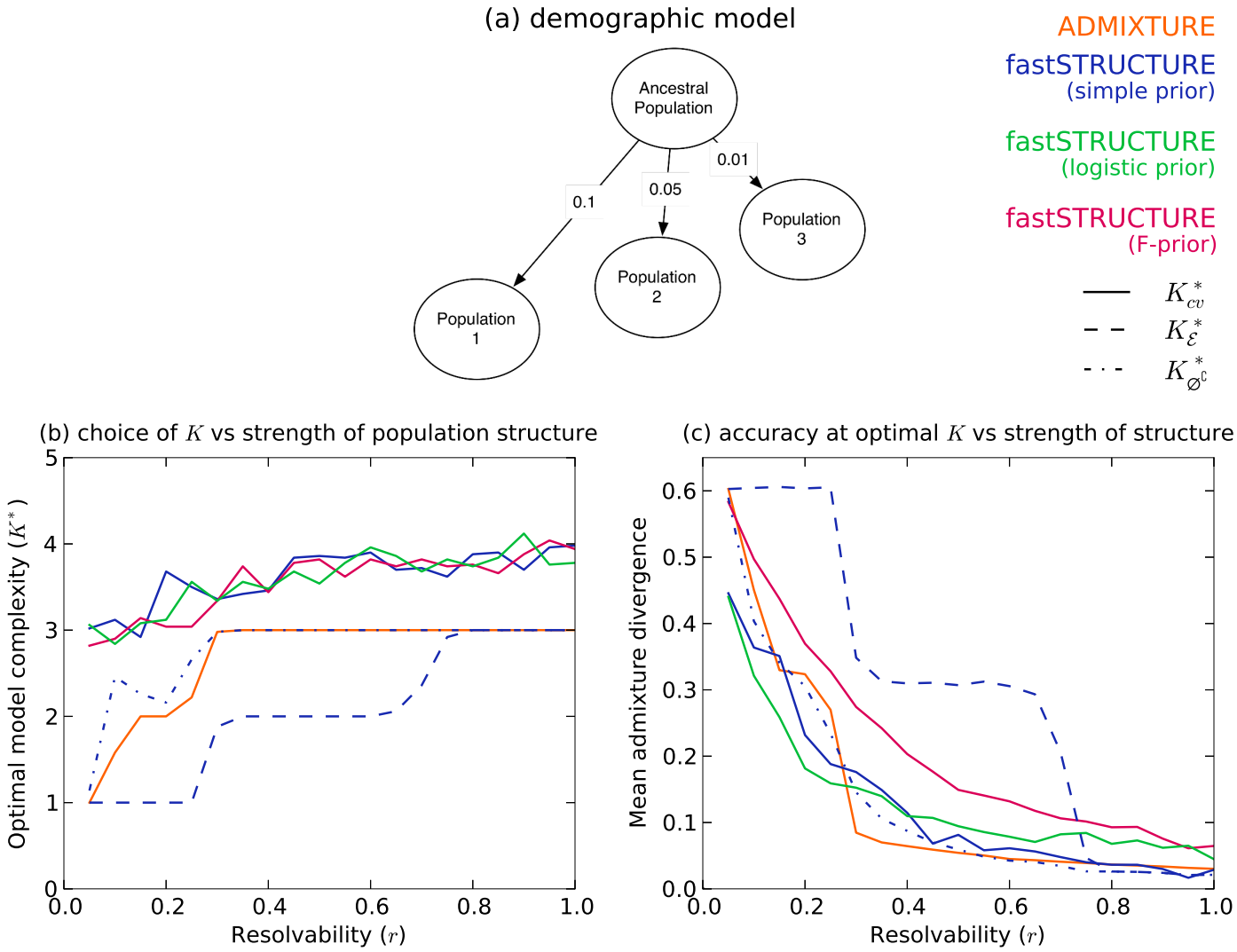

Figure S1: Accuracy of different algorithms as a function of resolvability of population structure. This figure is similar to Figure 1 in the main text, with results using the F-prior included. Subfigure (a) illustrates the demographic model underlying the three populations represented in the simulated datasets. Subfigure (b) compares the optimal model complexity inferred by ADMIXTURE (K^*_{cv}), fastSTRUCTURE with simple prior ($K^*_{cv}, K^*_{\mathcal{E}}, K^*_{\mathcal{C}}$), fastSTRUCTURE with F-prior (K^*_{cv}), and fastSTRUCTURE with logistic prior (K^*_{cv}). Subfigure (c) compares the accuracy of admixture proportions estimated by each algorithm at the optimal value of K in each replicate.

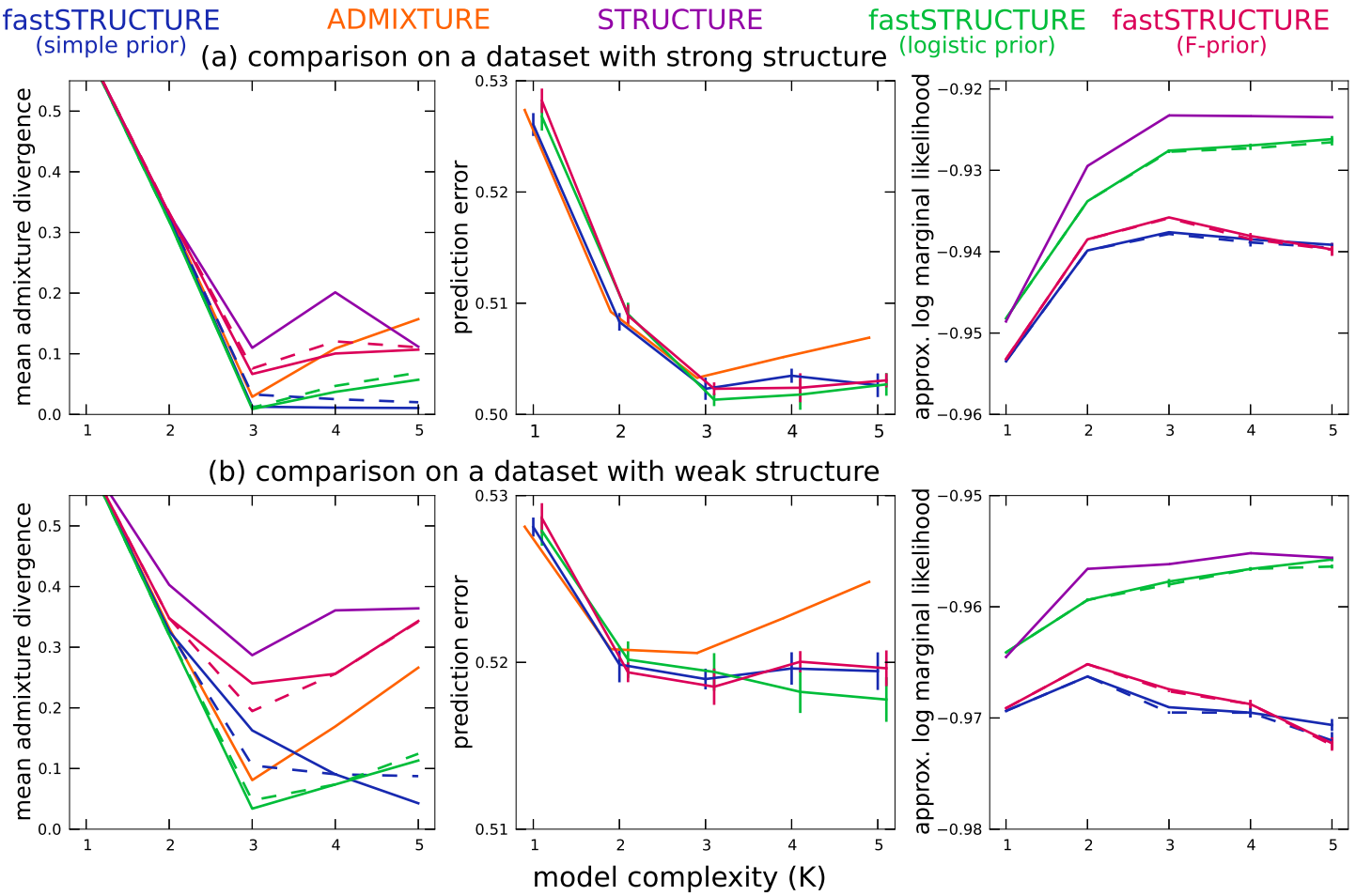

Figure S2: Accuracy of different algorithms as a function of model complexity (K) on two simulated data sets, one in which ancestry is easy to resolve (top panel; $r = 1$) and one in which ancestry is difficult to resolve (bottom panel; $r = 0.5$). This figure is similar to Figure 3 in the main text, with results using the F-prior included. Solid lines correspond to parameter estimates computed with a convergence criterion of $|\Delta\mathcal{E}| < 10^{-8}$, while the dashed lines correspond to a weaker criterion of $|\Delta\mathcal{E}| < 10^{-6}$. The left panel of subfigures shows the mean admixture divergence, the middle panel shows the mean binomial deviance of held-out genotype entries, and the right panel shows the approximations to the marginal likelihood of the data computed by STRUCTURE and fastSTRUCTURE.

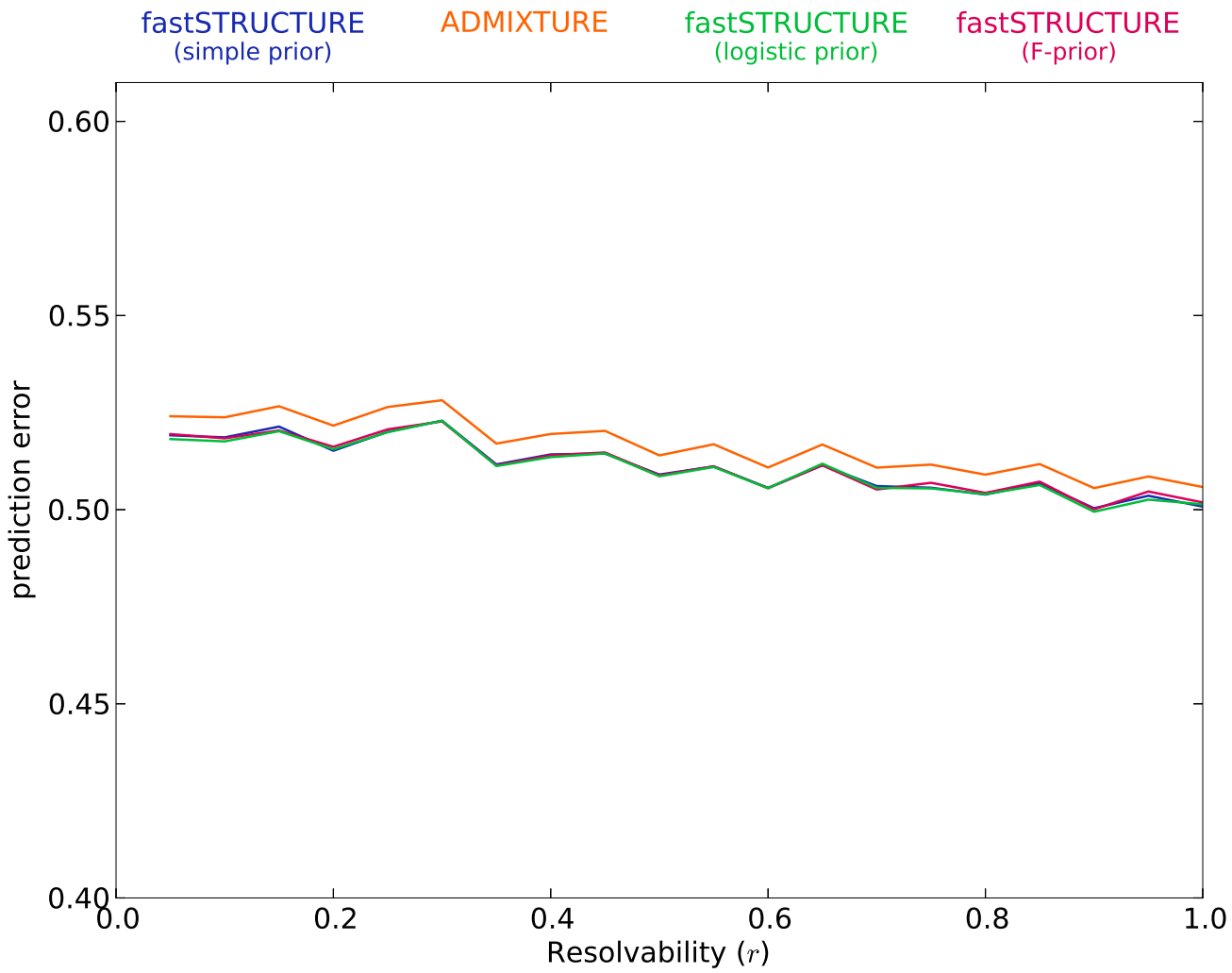

Figure S3: Prediction error of different algorithms as a function of resolvability of population structure.