MOLECULAR ECOLOGY
RESOURCES
Molecular Ecology Resources (2009) **9**, 1322–1332 doi: 10.1111/j.1755-0998.2009.02591.x

TECHNICAL ADVANCES

# Inferring weak population structure with the assistance of sample group information

MELISSA J. HUBISZ,*† DANIEL FALUSH,‡ MATTHEW STEPHENS*§ and JONATHAN K. PRITCHARD*¶
*Department of Human Genetics, §Department of Statistics, and ¶Howard Hughes Medical Institute, University of Chicago, Chicago, IL 60637, USA, †Department of Biological Statistics and Computational Biology, Cornell University, Ithaca, NY 14853, USA, ‡Environmental Research Institute, Department of Microbiology, University College Cork, Cork, Ireland

## Abstract

**Genetic clustering algorithms require a certain amount of data to produce informative results. In the common situation that individuals are sampled at several locations, we show how sample group information can be used to achieve better results when the amount of data is limited. New models are developed for the STRUCTURE program, both for the cases of admixture and no admixture. These models work by modifying the prior distribution for each individual's population assignment. The new prior distributions allow the proportion of individuals assigned to a particular cluster to vary by location. The models are tested on simulated data, and illustrated using microsatellite data from the CEPH Human Genome Diversity Panel. We demonstrate that the new models allow structure to be detected at lower levels of divergence, or with less data, than the original STRUCTURE models or principal components methods, and that they are not biased towards detecting structure when it is not present. These models are implemented in a new version of STRUCTURE which is freely available online at http://pritch.bsd.uchicago.edu/structure.html.**

*Keywords*: admixture, divergence, population structure, prior distribution

*Received 22 July 2008; accepted 12 January 2009*

## Introduction

Clustering algorithms for genetic data have become an important tool in a number of fields including conservation and population genetics (Dawson & Belkhir 2001; Corander *et al*. 2003; Purcell & Sham 2004; Corander & Marttinen 2006; Francois *et al*. 2006; Patterson *et al*. 2006). Such methods are often used to understand the structure of populations, as well as to identify migrant or admixed individuals. They are also used to detect cryptic population structure, as undetected structure may lead to false positives when searching for disease-associated markers in case-control studies.

STRUCTURE is a Bayesian, model-based algorithm that is widely used for clustering genetic data (Pritchard *et al*. 2000; Falush *et al*. 2003; Falush *et al*. 2007). Given the number of clusters ($K$) and assuming Hardy–Weinberg and linkage equilibrium within clusters, STRUCTURE estimates allele frequencies in each cluster and population memberships for every individual. In the simplest, 'no-admixture' model, it assumes that each individual belongs to a single cluster, whereas in the more general 'admixture model', it estimates admixture proportions for each individual. It uses Markov chain Monte Carlo (MCMC) to integrate over the parameter space and make cluster assignments. Although the value of $K$ must be provided to the algorithm, a heuristic method for selecting $K$ is often used, which is based on comparing penalized log likelihoods over independent runs with differing numbers of clusters.

When the data contain relatively little information about population structure, STRUCTURE sometimes produces results that are difficult to interpret. For example, the samples may have come from several distinct populations, and perhaps $F_{ST}$ values calculated between the samples from some pairs of the labelled populations are significantly different

Correspondence: Melissa J. Hubisz, Fax: 607-255-4698;
E-mail: mjhubisz@cornell.edu

from zero, and yet the results indicate no evidence of structure. Or, the population assignments made by the algorithm may hint that there is indeed structure, and yet the highest penalized log likelihood is provided by the model with just one cluster. When such situations arise, it is unclear whether one should conclude that the data are homogeneous after all, or that the amount of data collected is insufficient to make a convincing case for structure.

Although such results may be discouraging, it is worth noting that in a sense, STRUCTURE aims to solve a rather difficult problem. There is an enormous number of ways that $N$ individuals can be partitioned into $K$ populations. The basic STRUCTURE models assume that all partitions of the $N$ individuals into $K$ populations are equally likely, *a priori*. This means that any *particular* clustering solution is highly unlikely, *a priori*, and it takes a considerable amount of statistical evidence to provide strong support for any particular partition. This explains why there can be data sets with significant $F_{ST}$ values between samples of individuals collected at different locations, and yet STRUCTURE does not provide a clear indication of population structure.

In this paper, we extend the basic models to allow STRUCTURE to make use of information about sampling locations, when the data indicate that this information would be helpful. In effect, we place much more prior weight on clustering outcomes that are correlated with the sampling locations. The new models allow much better performance on some data sets where there are too few loci or individuals, or not enough divergence, for the standard STRUCTURE models to perform well. Our approach could also be used in settings where individuals can be classified into discrete groups on the basis of a phenotypic characteristic. The new models have the desirable properties that (i) they do not tend to find structure when none is present; (ii) they are able to ignore the sampling information when the ancestry of individuals is uncorrelated with sampling locations; and (iii) the old and new models give essentially the same answers when the signal of population structure is very strong. Hence, we recommend using the new models in most situations where the amount of available data is limited, especially when the standard STRUCTURE models do not provide a clear signal of structure.

The idea of using sampling locations to help infer population structure has also been considered elsewhere. One approach was taken by Corander *et al*. (2003), and implemented in the program BAPS. BAPS allows the user to pre-specify a set of sample groups; all individuals in the same sample group are assumed to have the same ancestry. The authors have shown that the use of sample group information can greatly improve power to detect structure when the amount of data is limited (Corander *et al*. 2003; Corander & Marttinen 2006). Once the allele frequencies are estimated, migrants and admixture events can be detected in an additional step that does not take the sampling groups into account. By contrast, the methods that we develop here allow for a more flexible relationship between sample groups and ancestry, allowing for the possibility that sample group information might be partially (or even not at all) informative about genetic population structure, and providing simultaneous estimation of allele frequencies and ancestry.

A second type of approach to using location information makes use of spatially explicit models. For example, Wasser *et al*. (2004) used elephant samples from known locations across Africa to estimate the geographical origin of poached ivory. Their method, implemented in SCAT, assumes that allele frequencies vary smoothly across the region of study. Another type of approach has been implemented in the program GENELAND (Francois *et al*. 2006; Guillot *et al*. 2008), and in a recent version of BAPS (Corander *et al*. 2008). The methodologies of the two programs are somewhat different, but they both use a coloured tessellation to model the distribution of the population clusters across space. These spatially explicit methods differ from the models discussed here in that we do not consider the specific geographical coordinates for each individual, but instead simply group together individuals collected at the same sampling location. This allows us to make fewer assumptions about the geographical structure of populations, while still offering improved performance in the common scenario that individuals are sampled at a modest number of distinct locations.

Our new methods are also substantially different from the 'Model with prior population information' introduced in the original STRUCTURE paper (Pritchard *et al*. 2000). That earlier model was designed for the situation in which there is both *strong* evidence of population structure and in which the sampling locations correspond almost exactly to the inferred clusters. That model allows a user to test whether a small number of individuals might be migrants from a different location than where they were sampled and is only useful for highly informative data. In contrast, the new models presented in this paper help to provide useful inference in settings where the data are not highly informative, and in this case it will usually not be possible to identify migrants with any confidence.

## Methods

We present both a no-admixture model and an admixture model that allow the individuals' sampling locations to inform cluster assignments. In order to understand how these models work, it is useful first to review the original model. We provide a brief description here, and Table 1 provides a brief summary of the key model parameters. For the complete details, see Pritchard *et al*. (2000) and Falush *et al*. (2003).

**Table 1** Summary of STRUCTURE parameters

| STRUCTURE parameters |
|---|
| $K$: number of clusters |
| $N$: number of individuals |
| $L$: number of loci |
| $q_{ij}$: admixture proportion of individual $i$ in cluster $j$ |
| $z_{ilm}$: cluster of origin for locus $l$, individual $i$, copy $m$ |
| $(\alpha_1, \ldots, \alpha_K)$: parameters to Dirichlet distribution which forms a prior for $q_i$ |
| $p_{klj}$: frequency of allele $j$ in locus $l$, cluster $k$ |
| $\lambda$: parameter to Dirichlet distribution which forms a prior for $p_{kl.}$ |
| $F_k$: the amount of drift from ancestral population to cluster $k$ in the model of correlated allele frequencies |
| **New model parameters** |
| $S$: number of sampling locations |
| $r$: parameter which estimates the informativeness of the sampling location data |
| $(\eta_1, \ldots, \eta_K)$: for the no-admixture model, these parameters reflect the relative proportion of individuals assigned to each cluster |
| $(\gamma_{s1}, \ldots, \gamma_{sK})$: for the no-admixture model, these parameters reflect the relative proportion of individuals from location $s$ assigned to each cluster |
| $(\alpha_1^{(g)}, \ldots, \alpha_K^{(g)})$: for the admixture model, these parameters reflect the relative levels of admixture from each cluster over all individuals |
| $(\alpha_{s1}, \ldots, \alpha_{sK})$: for the admixture model, these parameters reflect the relative levels of admixture from each cluster for an individual from location $s$ |

### *Overview of the STRUCTURE algorithm*

Consider a data set consisting of genotypes for $N$ individuals at $L$ loci. We assume that the sampled individuals have ancestry in $K$ discrete clusters, where the clusters correspond to unobserved populations. $K$ is fixed by the user. Each cluster is characterized by a set of allele frequencies at each locus. The three-dimensional vector $P$ contains the allele frequencies in each cluster for each allele at every locus; the allele frequencies are typically unknown in advance. In the no-admixture model, the algorithm assigns each individual to one of the $K$ clusters. The vector $Z$ records these cluster assignments. In the admixture model, each individual is allowed to have partial ancestry in each of the $K$ clusters. The vector $Q$ describes the proportion of each sampled individual's genome that comes from each cluster. As detailed in Table 1, we use the convention that elements within the vectors $P$, $Q$ and $Z$ are indexed by lower-case '$p$', '$q$', and '$z$' with appropriate subscripts. The likelihood of an individual's genotype is determined as the roduct of the relevant frequencies of the individual's alleles across all loci (the loci are assumed to be independent given cluster memberships). Our goal is to estimate $P$, $Q$ and Z from the data.

STRUCTURE uses MCMC to sample from the posterior distribution of the parameters $P$, $Q$, and $Z$. To estimate the appropriate number of clusters ($K$), the algorithm is usually run many times independently, varying the value for $K$. Although there is some debate as to the best method for choosing $K$ (e.g. Evanno *et al*. 2005), here we use the method suggested in the original STRUCTURE paper, which involves comparing mean log likelihoods penalized by one-half of their variance (Pritchard *et al*. 2000). Although a model of linked loci has been developed (Falush *et al*. 2003), the methods in this paper are most useful when there is a scarcity of data. We assume that when only a small number of loci are genotyped, they are likely to be unlinked, and we will not address the linkage model in this paper.

### *No-admixture model with sample group information*

In the original version of STRUCTURE, an individual is *a priori* assumed to be equally likely to come from any of the $K$ clusters. In the no-admixture model, the prior probability that individual $i$ comes from population $k$ (that is, $z_i = k$) is simply given by:

$$\Pr(z_i = k) = \frac{1}{K}.$$

The idea, then, is to modify this prior to take sampling locations into account. We do this by saying that the probability that an individual is assigned to each cluster may vary among the locations:

$$\Pr(z_i = k \mid \gamma) = \gamma_{l_i k}.$$

Here $\gamma_{lk}$ is the prior probability that an individual from location $l$ will be assigned to cluster $k$, and $l_i$ denotes the location where individual $i$ was sampled. The $\gamma_{lk}$ values are estimated from the data, and these parameterize the extent to which each sampling location is informative about

ancestry. If the $\gamma_{lk}$ are all ~$1/K$, then the location information is relatively uninformative, and this model is similar to the original STRUCTURE model. In contrast if, for each location, one value of $\gamma_{lk}$ is estimated to be ~1 and the rest ~0, then the location information will strongly influence the estimated ancestry.

Therefore, while the $\gamma_{lk}$ might help us to improve inference, it is important that they do not overstate the amount of information contained in the location information. To achieve this, we place the following prior structure on $\gamma$:

$$\gamma_{l\cdot} \sim \text{Dirichlet}(\eta_1 r, \eta_2 r, \ldots \eta_K r),$$

where

$$r \sim \text{uniform}(0, r_{MAX}),$$

and

$$\eta \sim \text{Dirichlet}(1, \ldots 1).$$

Here, $\eta$ is a vector of positive real numbers that, roughly speaking, estimates the overall proportion of individuals from each of the $K$ clusters in the entire data set. Then, $r$ parameterizes the extent to which the ancestry proportions at individual locations can deviate from the overall proportions. $r_{MAX}$ is an upper bound for $r$, preset by the user. If $r$ is large (>>1), then all the locations have essentially the same prior ancestry proportions (i.e. approximately equal to $\eta$). In contrast, if $r$ is ~1 or smaller, then the values of $\gamma_{l\cdot}$ may vary substantially across locations, implying that the location data are informative about ancestry. These priors are chosen so that if either there is no evidence for population structure, or the locations are uncorrelated with ancestry, then $r$ will tend to be large, and we will not be misled by the location information.

For the analyses presented here, we set $r_{MAX} = 1000$. This choice of $r_{MAX}$ puts considerable prior mass on large values of $r$, corresponding to the situation where the locations are uninformative. In some circumstances (e.g. with very small data sets, and good prior information that the locations are likely to be informative), a smaller value of $r_{MAX}$ would probably be preferable. We also found that the algorithm converged best if we started $r$ at a small value ($r_{INIT} = 1$ in our simulations). Appendix I gives details about the MCMC updates for the parameters in this model.

### *Admixture model with sample group information*

The new admixture model works similarly, by modifying the prior distribution for $Q$. In the original version of STRUCTURE, the prior distribution for $q_i$, the ancestry of individual $i$, is given by a Dirichlet distribution with parameters $\alpha_1, \ldots, \alpha_K$. Usually, the $\alpha$ parameters are set equal to each other ($\alpha: = \alpha_1 = \alpha_2 = \ldots = \alpha_K$), and are estimated during the MCMC. Small values of $\alpha$ (i.e. near 0) indicate that most individuals have little admixture, whereas large values indicate that most individuals have substantial ancestry from multiple clusters.

In order to modify the prior for $Q$, we now infer a different vector of $\alpha$'s for each location. This is similar in spirit to the new no-admixture model, in that it allows the distribution of cluster assignments to vary by location. If individual $i$ comes from location $l$, then:

$$q_i \sim \text{Dirichlet}(\alpha_{l1}, \ldots, \alpha_{lK}).$$

As for the no-admixture model, it is important to prevent the model from over-fitting the location data when the locations are not truly informative. For this reason, we place the following prior structure on the $\alpha$ values, which has the effect of pulling them towards a set of global values unless the locations are genuinely informative. That is, we define a set of global $\alpha$ values:

$$\alpha_i^{(g)} \sim \text{uniform}(0, \alpha_{MAX}),$$

where $\alpha_i^{(g)}$ denotes the global value of $\alpha$ for the $i$th cluster. Then the local $\alpha$ values for the $l$th location are distributed as where

$$\alpha_{li} \sim \text{gamma}(r * \alpha_i^{(g)}, 1/r),$$

$$r \sim \text{uniform}(0, r_{MAX}).$$

In this model, the global values, $\alpha^{(g)}$, can be thought of as estimating the overall distribution of ancestry. Each is (roughly) proportional to the overall amount of ancestry in cluster $i$. As in the standard STRUCTURE model, the mean of $\alpha^{(g)}$ measures the amount of admixture. The distribution of the local $\alpha$ values is constructed so that each $\alpha_{li}$ has mean $\alpha^{(g)}$ and variance $\alpha_i^{(g)}/r$. Hence, large values of $r$ imply that the local values of $\alpha_{li}$ are very similar to the global values, and the location information has little impact on the model. Conversely, small values of $r$ allow the local values of $\alpha_{li}$ to differ substantially from the global values, implying that the location information is potentially very informative. As in the no-admixture model, the simulations presented here used $r_{MAX} = 1000$, although again we note that smaller values would be appropriate for data sets with strong prior reason to expect structure.

### *Simulations without admixture*

Data were simulated with in-house software using a model of correlated allele frequencies (Nicholson *et al*. 2002) with either two or five populations. It was assumed that each population corresponds perfectly to a sampling location.

All simulated data sets were composed of 100 biallelic loci, to model single nucleotide polymorphisms (SNPs). Each individual had an equal probability of being assigned to each of the populations, and the data sets had 100 and 250 diploid individuals for two and five populations, respectively. $F_{ST}$ was varied in intervals of 0.005, with 50 independent repetitions for each value of $F_{ST}$. Allele frequencies $p_R$ for the root population were simulated from a beta distribution with parameters $\alpha = 0.8$, $\beta = 0.8$. With two populations, the root population was used as population 1, and otherwise a star-like phylogeny of populations was assumed. The allele frequencies for non-root populations were simulated as beta random variables with parameters $\alpha = p_R(1 - F_{ST})/F_{ST}$, $\beta = (1 - p_R)(1 - F_{ST})/F_{ST}$, as suggested by Balding & Nichols (1995).

### *Simulations with admixture*

Data were simulated using a model of independent allele frequencies for $K = 3$, with 100 individuals and a varying number of loci. Each individual had an equal chance of being sampled from each of four locations. The admixture proportions for an individual were drawn from Dirichlet distributions with parameters (10, 0.5, 0.5), (0.5, 10, 0.5), (0.5, 0.5, 10), (0.5, 0.5, 0.5) for locations 1, 2, 3, and 4, respectively. $F_{ST}$ for these simulated data sets was approximately 0.20. An additional set of simulations was performed to demonstrate the behaviour of the admixture model with a large number of sampling locations. Data sets were simulated for $K = 5$ with 100 individuals and 10 microsatellites, for a range of values of $F_{ST}$. Each individual was assigned to one of 25 sampling locations, and population assignments for each individual were highly determined by the sampling location. Specifically, each location was randomly assigned to one of the five clusters, and admixture proportions were drawn from a Dirichlet distribution with parameter 1 for the main cluster, and 0.01 for each other cluster. For example, if a location was assigned to cluster 3, then every individual from that location would have admixture proportions drawn from a Dirichlet distribution with parameters (0.01, 0.01, 1.0, 0.01, 0.01). The microsatellite data were simulated using the correlated allele frequencies model of Falush *et al*. (2003). We assumed that all microsatellites had four possible alleles, and the ancestral allele frequencies were simulated from a Dirichlet distribution with parameters (0.8, 0.8, 0.8, 0.8). For this data set, each STRUCTURE run was repeated four times to ensure proper convergence.

Finally, to illustrate how the results depend on the strength of correlation between location data and population structure, we performed a series of simulations in which we reassigned locations randomly for a fraction $f$ of individuals and re-analysed the data using the new models. This was done for each of the 50 data sets simulated without admixture, assuming five sampling locations, $K = 5$, and $F_{ST} = 0.03$, for values of $f$ in 0, 0.04, ... , 1.0.

For all the above data sets, STRUCTURE was run with each value of $K$ ranging from 1 to $K_T + 1$, where $K_T$ is the true value of $K$ used in the simulation. The estimate for $K$ was then taken as the $K$ with the highest penalized log likelihood as reported by STRUCTURE, which calculates the mean log likelihood minus half of its variance. The model of independent allele frequencies was used for the simulations with admixture in which the number of loci was varied. All other runs used the model of correlated allele frequencies, and estimated a separate $F_{ST}$ for each population. For all runs using the original admixture model, a separate value of $\alpha$ was estimated for each population as well. All runs consisted of 20 000 burn-in steps followed by 10 000 MCMC steps.

### *CEPH Human Genome Diversity Panel (HGDP) microsatellite analysis*

A microsatellite data set consisting of 377 loci genotyped in 1056 individuals from 52 human populations (Rosenberg *et al*. 2002) was downloaded from http://rosenberglab.bioinformatics.med.umich.edu/data/rosenbergEtAl2002/diversitydata.stru. We chose one population from each continent for analysis (Surui from South America, Han from Asia, Basque from Europe, Melanesian from Oceania, and Mandenka from Africa), resulting in a data set with 126 individuals. $F_{ST}$ among populations from different continents is about 7% in this data set (Rosenberg *et al*. 2002). All STRUCTURE analyses were done using the model of correlated allele frequencies, and every run was repeated five times to obtain the run with the highest penalized log-likelihood score. The analysis was repeated 50 times on random subsets of the data for a range of different numbers of loci. Each random subset was created by choosing loci without replacement.

### *Principal components analysis methods*

To provide an additional, and rather different, type of algorithm against which to compare our new methods, we also analysed the simulated data using principal components analysis (PCA). It has been shown (Patterson *et al*. 2006) that the resolution of principal components methods and STRUCTURE are quite similar in many cases. The software package EIGENSOFT was downloaded from http://genepath.med.harvard.edu/~reich/Software.htm and the program SMARTPCA (Patterson *et al*. 2006) was used to analyse the simulated and real data sets. The number of clusters inferred by SMARTPCA was taken as one plus the number of eigenvalues with $p$-value $\leq 0.05$. To get cluster assignments, the $k$-means algorithm (Hartigan & Wong 1979) was applied to the top $K$-1 eigenvectors.

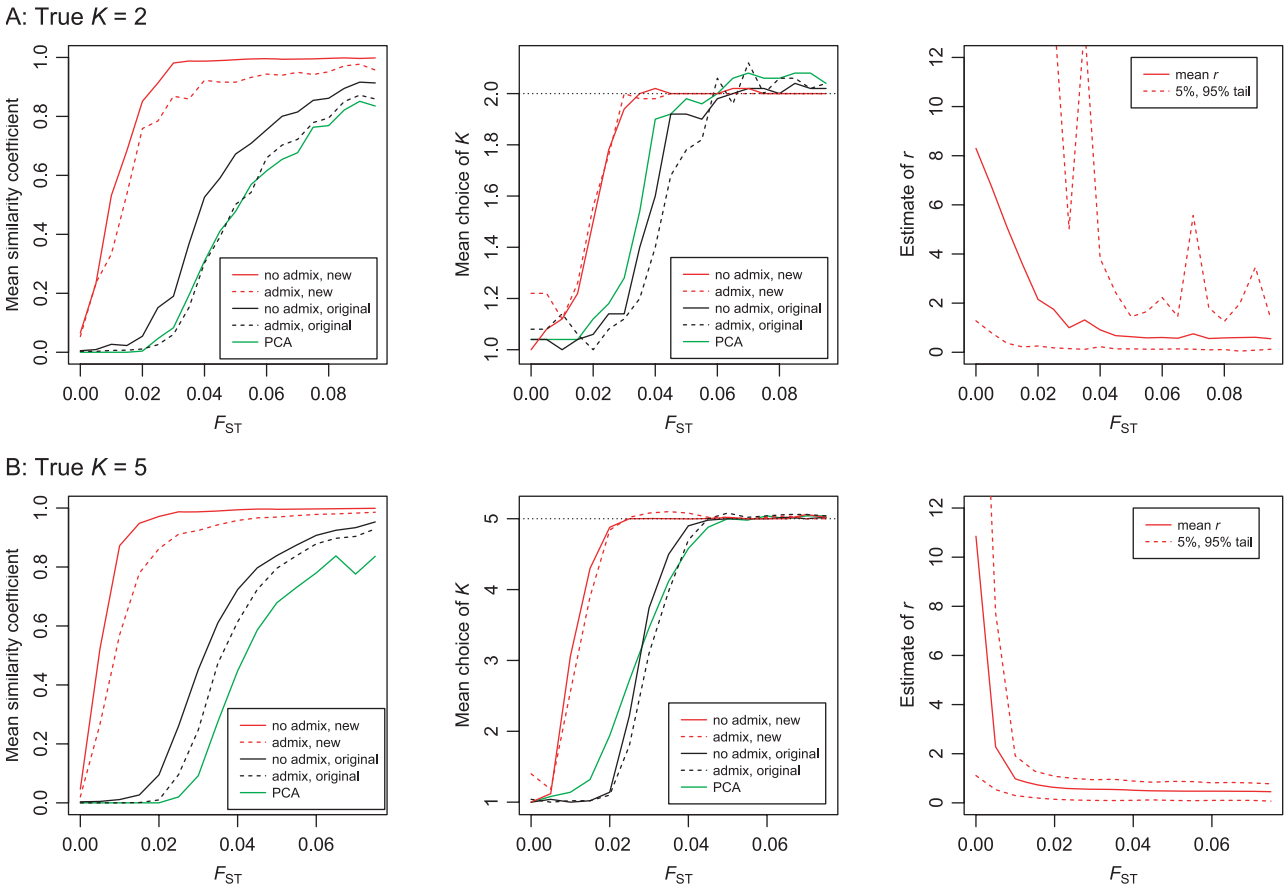



**Fig. 1** Results for simulations without admixture. Data were simulated for $K = 2$ and $K = 5$, as described in the Methods. On the left is plotted the mean similarity coefficient between the true and estimated ancestry, as a function of $F_{ST}$, each averaged over 50 simulated data sets. The middle plots show the average choice of $K$, with the dotted horizontal lines indicating the true value of $K$. On the right, the solid line shows the average estimate of $r$ calculated using the new no-admixture model with sampling locations. The dotted lines show the 5% and 95% tails of the distribution.

### *Similarity score*

To measure the similarity between the true and estimated population assignments, we used an adaptation of the standard Brier similarity score. That is, let $q_{ik}$ be the true fraction of ancestry of individual $i$ in population $k$ and let $\hat{q}_{ik}$ be the corresponding estimate of $q_{ik}$. Then, we define a score $S$ as

$$S = \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K}(q_{ik} - \hat{q}_{ik})^2,$$

where $N$ is the number of individuals. Note that $S$ will be zero when $\hat{Q} = Q$, and can be as large as 2 if there is a complete mismatch between $Q$ and $\hat{Q}$. In practice, the labelling of clusters identified by STRUCTURE is arbitrary, and thus, we computed $S$ for each of the $K!$ possible permutations of the cluster labels, and recorded the minimum of $S$ across permutations (call this $S'$). When the data are completely uninformative, a clustering solution $\hat{Q}^*$ that places a fraction $1/K$ of each individual into each cluster would receive a smaller score (call this $S^*$) than a solution that puts all individuals into a single cluster (provided that true ancestry is not highly skewed towards particular clusters). Finally, to obtain a similarity score which is equal to one when $\hat{Q} = Q$, and zero for any $q$ that performs as poorly as $\hat{Q}^*$, we recorded the similarity score as $1 - \min(S', S^*)/S^*$.

## Results

To evaluate the performance of the new models, we tested them on simulated and real data under a variety of conditions. Together, the examples illustrate the performance of the methods as a function of the amount of divergence among populations and as a function of the number of loci; as well as under a variety of different conditions: variable numbers of loci; variable levels of information in the location data; discrete populations and admixture; and SNPs and microsatellites. The parameter values for the simulations were chosen because they illustrate the

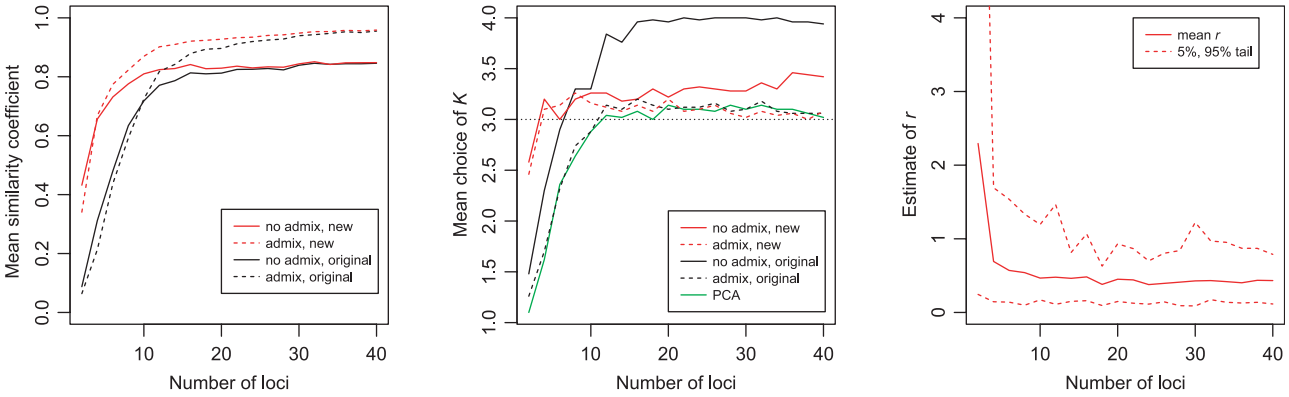



**Fig. 2** Results for simulations with admixture. Data were simulated with $K = 3$, as described in the Methods. On the left is the mean similarity coefficient over 50 simulated data sets as a function of the number of loci. In the middle is the mean estimate of $K$, with the dotted horizontal line indicating the true value of $K$. The right plot shows the average estimate of $r$ calculated using the new admixture model with sampling locations, with the dotted lines giving the 5% and 95% tails of the distribution.

differences between the new and original models; for larger or more informative data sets, the differences between the new and old models tend to be small, and in some contexts, we prefer the original STRUCTURE models (see below for further discussion).

The first set of simulations (Fig. 1) considered a setting in which individuals are sampled from either two or five different sampling locations, and where each sampling location consists of a distinct non-admixed population. As expected, all the methods struggle to assign individuals accurately to populations at low divergence ($F_{ST}$ near 0), and provide accurate assignments at high divergence. However, there is a range of $F_{ST}$ values for which the new models perform much better than the existing methods: both in terms of making more accurate cluster assignments (similarity coefficient), and in choosing the correct value of $K$ at lower divergence levels. Importantly, all of the models predict just one cluster when $F_{ST} = 0.0$, suggesting that the new models do not bias the algorithm towards finding structure when it is not present.

Figure 1 also plots values of the tuning parameter, $r$, which measures the amount of information contained in the location information. Recall that $r>>1$ implies that the location labels are uninformative about ancestry, while small values of $r$ allow the ancestry proportions to vary substantially among locations. Notice that when $F_{ST}$ is near 0, the mean estimate of $r$ is considerably larger than 1, consistent with the estimates of $K$ near 1. As the amount of information in the data increases the estimate of $r$ quickly decreases, indicating that the sampling groups are contributing information. At $F_{ST} = 0$, one might have expected that the posterior mean of $r$ should be approximately $r_{MAX}/2$, since in this case $r_{MAX}$ was set to be very large. The fact that $r$ is much smaller than $r_{MAX}/2$ suggests that $r$ has not fully explored its posterior range during the course of the MCMC run length used here (recall that $r$ was initialized at 1). However this should not be a serious concern as the model is relatively insensitive to the precise value of $r$ when $r$ is considerably larger than 1, and in practice, we would recommend a smaller value of $r_{MAX}$ for most applications.

A second set of simulations was performed with admixture (Fig. 2). In this case, we set $K = 3$ and simulated four sampling locations with different mixtures of ancestry coefficients. We set $F_{ST} = 0.20$ and varied the number of genotyped loci. The plot of similarity coefficients shows that again the new models substantially improve the ancestry estimates when the data sets are small, even providing some information with just one genotyped locus. The old and new models become more similar as the number of genotyped loci increases. We have observed that these new methods tend to improve estimation of admixture coefficients for all the individuals in these data sets, including individuals who are outliers within their sampling group. This indicates that the new methods are not simply working by grouping the individuals in the same location together; instead, the location information also improves the estimation of allele frequencies, leading to more accurate parameter estimation.

To assess the behaviour of the new model when there are many sampling locations, we also simulated data with 100 individuals sampled from across 25 sampling locations, with $K = 5$. The simulations were set up so that individuals from the same sampling location generally drew most of their ancestry from the same cluster. Figure 3 shows the performance over a range of values of $F_{ST}$. Even with a relatively small number of individuals per group, the new models still benefit from using the location information, compared to the original models, although the advantage appears to be smaller than when larger numbers of individuals are sampled in each location. We also found that for these data sets, the estimation of $K$ was a little erratic for small values of $F_{ST}$. In particular, both models frequently estimated $K > 1$ even when $F_{ST} = 0$ (implying that there is no real population structure, so that we would want to estimate

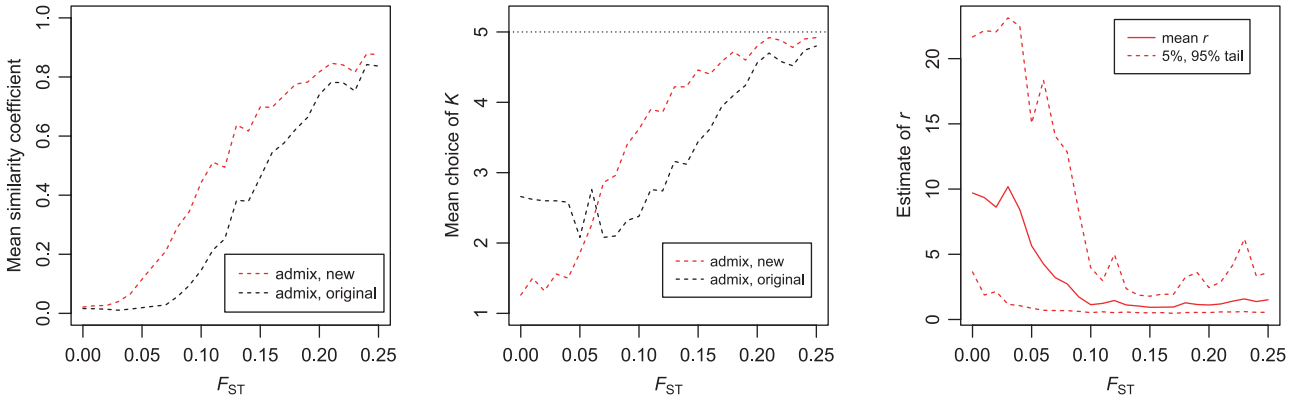
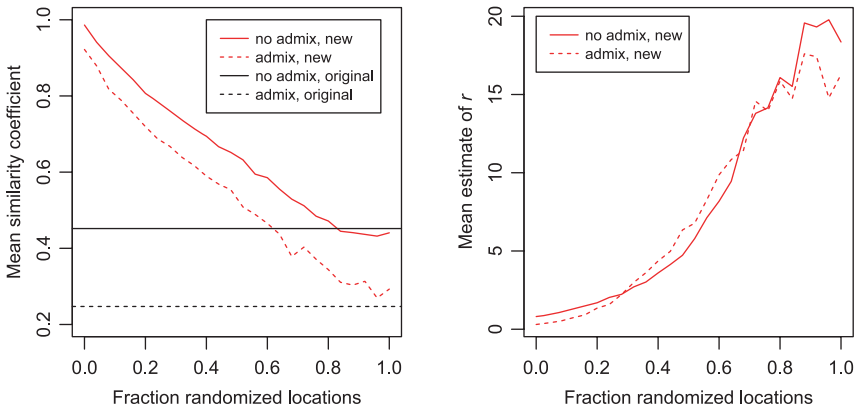



**Fig. 3** Results for simulations with admixture, using 25 sampling locations with an average of 4 individuals per location, and $K = 5$. See Figs 2 and 3 for descriptions of the plots. Each data point is an average over 50 simulated data sets for a given value of $F_{ST}$.

**Fig. 4** Effect of varying the amount of information contained in the location data. The simulations assumed 250 individuals, five sampling locations, $K = 5$, $F_{ST} = 0.03$, and no admixture. The $x$-axis shows the fraction of individuals whose location data were randomized. For all other individuals, the location number matched the true population number.




$K = 1$). We believe that STRUCTURE may be struggling with the relatively small data sets simulated in this case (100 individuals with 10 microsatellites; for example, compare this to Fig. 1A, which includes 100 individuals genotyped at 100 SNPs). In the plot shown in Fig. 3, the new model seems to perform better than the original model at estimating small $K$ when $F_{ST} = 0$, but this does not seem to be a general property of the new model. For example, when we analysed the same data using the ONEFST model in STRUCTURE, both models overestimated $K$ in the case where $F_{ST} = 0$.

We also investigated the performance of the new models as the correlation between locations and clusters changes. The left plot in Fig. 4 shows the effect of similarity coefficients as the fraction of individuals with randomly assigned locations is increased. The horizontal lines show the average performance of the original STRUCTURE models on the same data. As expected, the performance of the new models is best when the locations correspond perfectly to the underlying structure. However, even when the locations are completely random, the new models perform almost identically to the old models. This implies that there is little cost to using the new models, even when the location data are potentially uninformative. The right plot in Fig. 4 shows that the value of $r$ estimated by STRUCTURE seems to be a good indicator of the usefulness of the location data.

Finally, we illustrate the new methods with a simple application to microsatellite data from the Human Genome Diversity Panel (Rosenberg *et al.* 2002). We selected a set of 126 individuals representing five populations on five different continents. Figure 5A shows the average results of choosing subsets of the microsatellites at random. We see that the new models almost always estimate $K = 5$ with as few as 6 random loci, whereas 16 or more loci are required to make the same estimate when the sampling location data are not used. Also, the new models substantially improve the accuracy of the estimated admixture proportions, when the 'true' ancestry proportions are estimated using all 377 microsatellites. Figure 5B shows some example results, using the first 2, 6, and 10 microsatellites, respectively, from the data set (in a single random order), compared to the complete data set. It is clear that with 2 and 6 microsatellites, the new models have much more success at separating the continental groups than do the original models.

Once the data set increases to 10 microsatellites or more, the differences among the results become quite subtle. However, for the complete data set of 377 loci, there is a slight but noteworthy difference between results from the

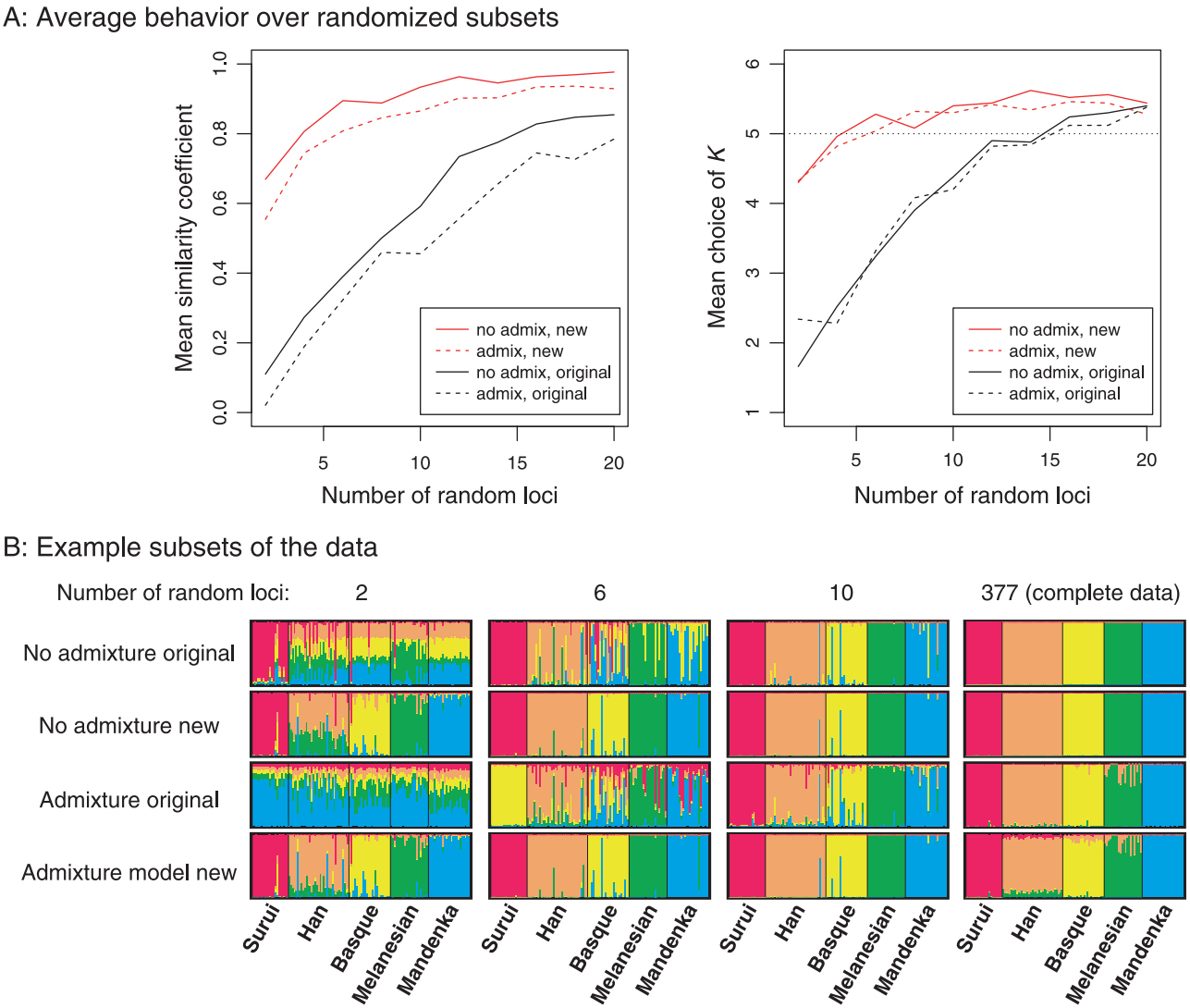



**Fig. 5** Analysis of five populations from the Human Genome Diversity Panel microsatellite data set. In Fig. 5A, the mean similarity coefficient and choice of *K* are plotted, averaged over 50 runs using a number of randomly chosen microsatellites, shown on the *x*-axis. Figure 5B shows Structure results for the first 2, 6, and 10 loci, as well as the entire data set.

new and original admixture models (Fig. 4B). Unlike the original admixture model, the new admixture model estimates that all the Han Chinese individuals contain a small amount of ancestry from both the Melanesians and the Surui. Since it is implausible that there has been recent gene flow of this magnitude from Native Americans and Oceanians into the Chinese population, this argues that the new prior model is subtly shifting the performance of the method on this highly informative data set.

## Discussion

The new models presented in this study are designed to help detect population structure and to produce more accurate ancestry estimates for data sets with low information content. Our simulation studies suggest that the models can help considerably in such cases. As the information content in the data increases, the results become similar to those obtained using the original models. In general, our simulations show that the new models provide an appropriate balance between the potential value of incorporating location information into the inference, while still remaining reasonably robust when there is no population structure. Moreover, the new models are able to ignore the sampling information when there is clear evidence of population structure, but the structure is uncorrelated with sampling locations.

For these reasons, we feel that it will often be beneficial to use the new models for analysing small- or medium-sized data sets, such as are currently typical in studies of molecular

ecology or conservation genetics. However, we would still encourage users to run the original models as well, and to check that substantial differences between results from the new and old models seem biologically sensible. We also suggest that the value of $r$ can be a useful indicator of whether the location information is relevant to the model: values of $r$ near or below 1 imply that the ancestry proportions differ substantially between sampling locations.

However, we also caution that the new models are not a panacea. For example, STRUCTURE sometimes overestimates the number of clusters: for example when there is inbreeding or relatedness among some individuals. Moreover, the number of clusters is not well-defined in settings where the allele frequencies vary smoothly across the landscape (Wasser *et al*. 2004). The new models are likely to be affected similarly by these issues. Finally, for very informative data sets, the new and old models should provide very similar results. However, in one example (the HGDP data, described above), we noted slight differences between results with the old and new priors. Given this, and the fact that there is now a great deal of accumulated experience with the standard STRUCTURE models, we recommend that the standard models should continue to be the default for data sets in which the data are highly informative.

Finally, we remind users that the new models serve a very different purpose from an existing model in STRUCTURE that also uses location information (obtained in the software by setting USEPOPINFO = 1) (Pritchard *et al*. 2000). That model was designed for identifying migrant individuals in data that are *highly informative*, in contrast to the goal here of detecting very weak population structure.

The models presented here have been implemented in a forthcoming version of STRUCTURE, version 2.3. The use of the new models will be described in detail in the next release of the STRUCTURE manual. The new software and documentation will be available online at http://pritch.bsd.uchicago.edu/structure.html.

## Acknowledgements

This work was supported by a National Institutes of Health Genetics and Regulation Training Grant (M.J.H.), a Packard Foundation grant (J.K.P.), and a Science Foundation of Ireland (D.F., grant no. 05/FE1/B882). J.K.P. is an investigator of the Howard Hughes Medical Institute. We thank Jukka Corander, three anonymous reviewers, and the editor, Jared Strasburg, for helpful comments, and Tim Wootton for a conversation that helped to stimulate this project.

## References

Balding DJ, Nichols RA (1995) A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. *Genetica*, **96**, 3–12.

Corander J, Marttinen P (2006) Bayesian identification of admixture events using multilocus molecular markers. *Molecular Ecology*, **15**, 2833–2843.

Corander J, Siren J, Arjas E (2008) Bayesian spatial modeling of genetic population structure. *Computational Statistics*, **23**, 111–129.

Corander J, Waldmann P, Sillanpaa MJ (2003) Bayesian analysis of genetic differentiation between populations. *Genetics*, **163**, 367–374.

Dawson KJ, Belkhir K (2001) A Bayesian approach to the identification of panmictic populations and the assignment of individuals. *Genetics Research*, **78**, 59–77.

Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software Structure: a simulation study. *Molecular Ecology*, **14**, 2611–2620.

Falush D, Stephens M, Pritchard JK (2003) Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics*, **164**, 1567–1587.

Falush D, Stephens M, Pritchard JK (2007) Inference of population structure using multilocus genotype data: dominant markers and null alleles. *Molecular Ecology Notes*, **7**, 574–578.

Francois O, Ancelet S, Guillot G (2006) Bayesian clustering using hidden Markov random fields in spatial population genetics. *Genetics*, **174**, 805–816.

Guillot G, Santos F, Estoup A (2008) Analysing georeferenced population genetics data with GENELAND: a new algorithm to deal with null alleles and a friendly graphical user interface. *Bioinformatics*, **24**, 1406–1407.

Hartigan JA, Wong MA (1979) A K-means clustering algorithm. *Applied Statistics*, **28**, 100–108.

Nicholson G, Smith AV, Jónsson F *et al*. (2002) Assessing population differentiation and isolation from single-nucleotide polymorphism data. *Journal of the Royal Statistical Society B*, **64**, 695–715.

Patterson N, Price AL, Reich D (2006) Population structure and eigen analysis. *Public Library of Science, Genetics* **2**, e190.

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics*, **155**, 945–959.

Purcell S, Sham P (2004) Properties of structured association approaches to detecting population stratification. *Human Heredity*, **58**, 93–107.

Rosenberg NA, Pritchard JK, Weber JL *et al*. (2002) Genetic structure of human populations. *Science*, **298**, 2381–2385.

Wasser SK, Shedlock AM, Comstock K *et al*. (2004) Assigning African elephant DNA to geographic region of origin: applications to the ivory trade. *Proceedings of the National Academy of Sciences, USA*, **101**, 14847–14852.

## Appendix: MCMC updates

### *No admixture model with sample groups*

To sample from $\Pr(P, Z, r, \eta, \gamma \mid X)$, the algorithm proceeds as follows:

1. Sample $P^{(m)}$ from $\Pr(P \mid Z^{(m-1)}, \gamma^{(m-1)}, \eta^{(m-1)}, r^{(m-1)}, X)$.
2. Sample $Z^{(m)}$ from $\Pr(Z \mid P^{(m)}, \gamma^{(m-1)}, \eta^{(m-1)}, r^{(m-1)}, X)$.
3. Update $r$ using a Metropolis–Hastings step.
4. Update $\eta$ using a Metropolis–Hastings step.
5. Update $\gamma$ using a Metropolis–Hastings step.

Because the new models have only modified the prior for Z, $\Pr(P \mid Z^{(m-1)}, \gamma^{(m-1)}, \eta^{(m-1)}, r^{(m-1)}, X)$ does not depend on $\gamma$, $\eta$, or $r$, and step 1 does not need to be modified from the original STRUCTURE algorithm.

For step 2, we note that since $\eta$ and $r$ form a prior for $\gamma$, $\Pr(Z \mid P^{(m)}, \gamma^{(m-1)}, \eta^{(m-1)}, r^{(m-1)}, X)$ is equivalent to $\Pr(Z \mid P^{(m)}, \gamma^{(m-1)}, X)$. Then, for each individual $i$ from location $l_i$ we can sample $z_i$ based on the distribution:

$$\Pr(z_i = k \mid X, P, \gamma) = \frac{\Pr(z_i = k \mid \gamma)\Pr(X \mid P, z_i = k)}{\sum_{k'=1}^{K} \Pr(z_i = k' \mid \gamma)\Pr(X \mid P, z_i = k')},$$

where $\Pr(z_i = k \mid \gamma) = \gamma_{l_i k}$, and $\Pr(X \mid P, Z_i = k)$ is a product of allele frequencies in cluster $k$ corresponding to the genotype data. The exact expression is defined in the appendix of Pritchard *et al*. (2000).

For step 3, $r'$ is simulated from a uniform distribution in $(r^{(m-1)} - r_\varepsilon, r^{(m-1)} + r_\varepsilon)$. $r'$ is rejected if it is not in the range $(0, r_{MAX})$. Otherwise, it is accepted with the probability:

$$\prod_{l=1}^{S} \frac{f(\gamma_{l\cdot} \mid r', \eta)}{f(\gamma_{l\cdot} \mid r, \eta)}$$

where $l = 1$ ... S indicates the sampling locations, and where $f(\gamma_{l\cdot} \mid r, \eta)$ is given by the Dirichlet distribution:

$$f(\gamma_{l\cdot} \mid r, \eta) = \frac{\Gamma(\sum_{k=1}^{K} r\eta_k)}{\prod_{k=1}^{K} \Gamma(r\eta_k)} \prod_{k=1}^{k} \gamma_{lk}^{r\eta_k - 1}.$$

If $r'$ is accepted, than $r^{(m)}$ is set to $r'$, otherwise $r^{(m)}$ is set to $r^{(m-1)}$.

In all the analyses in this manuscript, $r_\varepsilon$ was set to 0.1.

For step 4, two clusters, $i$ and $j$, are chosen at random so that $i \neq j$. A random number $\varepsilon$ is simulated randomly in the range $(0, \varepsilon_{MAX})$. Then, is set to $\eta_i^{(m-1)} + \varepsilon$, and is set to $\eta_j^{(m-1)} - \varepsilon$. All other elements $\eta_k'$ are set to $\eta_k^{(m-1)}$ for $k$ not equal to $i$ or $j$. The update is rejected if either or $\eta_j'$ is not in the range (0, 1). In this way, the elements of the $\eta'$ vector are guaranteed to sum to 1, given that the elements of $\eta^{(m-1)}$ sum to 1. Then, $\eta'$ is accepted with the probability:

$$\prod_{l=1}^{S} \frac{f(\gamma_{l\cdot} \mid r, \eta')}{f(\gamma_{l\cdot} \mid r, \eta)}$$

If $\eta'$ is accepted, $\eta^{(m)}$ is set to $\eta'$. Otherwise, $\eta^{(m)}$ is set to $\eta^{(m-1)}$. For all analysis in this paper, $\varepsilon_{MAX}$ was set to 0.025.

For step 5, each vector $\gamma_{l\cdot}$ is updated in turn, for each location $l$. A $\gamma_{l\cdot}'$ is generated in exactly the same manner as $\eta'$, and is rejected if any of the elements are not in the range (0,1). Then, $\gamma_{l\cdot}'$ is accepted with the probability:

$$\frac{f(\gamma_{l\cdot}' \mid r, \eta)}{f(\gamma_{l\cdot} \mid r, \eta)} \prod_{i=1}^{N} \left[ \frac{g(z_i \mid \gamma')}{g(z_i \mid \gamma)} \right]^{I(l_i = l)}$$

Here, $I(l_i = l)$ is the indicator function which equals 1 if individual $i$ comes from location $l$, and zero otherwise, and $g(z_i \mid \gamma)$ is the probability of observing a particular value of $z_i$, given $\gamma$. If $\gamma_{\cdot}'$ is accepted, $\gamma_{l\cdot}^{(m)}$ is set to $\gamma'$, otherwise $\gamma_{l}^{(m)}$ is set to $\gamma_{l\cdot}^{(m-1)}$.

### *Admixture model with sample groups*

To sample from $\Pr(Z, Q, P, \alpha, r \mid X)$, the algorithm proceeds as follows:

1. Sample $P^{(m)}$ from $\Pr(P \mid Z^{(m-1)}, Q^{(m-1)}, \alpha^{(m-1)}, r^{(m-1)}, X)$.
2. Sample $Q^{(m)}$ from $\Pr(Q \mid P^{(m)}, Z^{(m-1)}, \alpha^{(m-1)}, r^{(m-1)}, X)$.
3. Sample $Z^{(m)}$ from $\Pr(Z \mid P^{(m)}, Q^{(m)}, \alpha^{(m-1)}, r^{(m-1)}, X)$.
4. Update $r$ using a Metropolis–Hastings step.
5. Update $\alpha$ using a Metropolis–Hastings step.

The new admixture model only affects the prior for $Q$, and therefore steps 1 and 3 do not need to be modified from the original algorithm. To perform step 2, the admixture proportions for individual $i$ from location $l$ have a distribution given by:

$$q_i \sim \text{Dirichlet}(\alpha_{l1} + n_{i1}, \alpha_{l2} + n_{i1}, \ldots, \alpha_{lK} + n_{iK})$$

where $n_{ik}$ is the total number of copies of each locus assigned to population $k$ in individual $i$.

For step 4, $r'$ is simulated from a uniform distribution in $(r^{(m-1)} - r_\varepsilon, r^{(m-1)} + r_\varepsilon)$, where $r_\varepsilon$ is the same as in the new no-admixture model. $r'$ is rejected if it is not in the range $(0, r_{MAX})$. Otherwise, it is accepted with the probability:

$$\prod_{l=1}^{S} \prod_{k=1}^{K} \frac{h(\alpha_{lk} \mid r', \alpha_k^{(g)})}{h(\alpha_{lk} \mid r, \alpha_k^{(g)})},$$

where $h(\alpha_{lk} \mid r, )$ is given by the Gamma distribution with parameters $r$, $1/r$.

Step 5 is achieved by independently updating every element of the $\alpha$ vector. First each element of $\alpha^{(g)}$ is updated. $\alpha_k^{(g)\prime}$ is simulated from a normal distribution with mean $\alpha_k^{(g)(m-1)}$ and standard deviation $\sigma_\alpha$. It is rejected if it is outside the range $(0, \alpha_{MAX})$. Otherwise, it is accepted with the probability:

$$\prod_{l=1}^{S} \frac{h(\alpha_{lk} \mid r, \alpha_k^{(g)\prime})}{h(\alpha_{lk} \mid r, \alpha_k^{(g)})}.$$

Finally, to update each element of $\alpha_{lk}$, an $\alpha_{lk}'$ is simulated from a normal distribution with mean $\alpha_{lk}^{(m-1)}$ and standard deviation $\sigma_\alpha$. It is accepted with the probability:

$$\frac{h(\alpha_{lk}' \mid r, \alpha_k^{(g)})}{h(\alpha_{lk} \mid r, \alpha_k^{(g)})} \prod_{i=1}^{N} \left[ \frac{f(q_i \mid 1, \alpha_{l\cdot}')}{f(q_i \mid 1, \alpha_{l\cdot})} \right]^{I(l_i = l)}.$$

For all the analysis in this paper, $\sigma_\alpha$ was set to 0.025.