mStruct: Inference of Population Structure in Light of Both Genetic Admixing and Allele Mutations

Suyash Shringarpure * **and Eric P. Xing** †[1]

CMU-ML-08-105

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

*School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA
†School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA

Abstract

Traditional methods for analyzing population structure, such as the *Structure* program, ignore the influence of the effect of allele mutations between the ancestral and current alleles of genetic markers, which can dramatically influence the accuracy of the structural estimation of current populations, and reveal additional information of population evolution such as the the divergence time and migration history of admixed populations. We propose *mStruct*, an admixture of population-specific mixtures of inheritance models, that addresses the task of structure inference and mutation estimation jointly through a hierarchical Bayesian framework, and a variational algorithm for inference. We validated our method on synthetic data, and used it to analyze the HGDP-CEPH cell line panel of microsatellites used in [1] and the HGDP SNP data used in [2]. A comparison of the structural maps of world populations estimated by *mStruct* and *Structure* is presented, and we also report potentially interesting mutation patterns in world populations estimated by *mStruct*.

[1]To whom correspondence should be addressed.
This research was supported by NSF grant CCF-0523757 and DBI-0546594.

Keywords: population structure, graphical models, variational methods, bayesian models

1 Introduction

The deluge of genomic polymorphism data, such as the genome-wide multilocus genotype profiles of variable number of tandem repeats (i.e., microsatellites) and single nucleotide polymorphisms (i.e., SNPs), has fueled the long-standing interest in analyzing patterns of genetic variations to reconstruct the ancestral structures of modern human populations, because such genetic ancestral information can shed light on the evolutionary history of modern populations and provide guidelines for more accurate association studies and other population genetics problems.

Various methods have been proposed to examine the problem of detecting population structure from multi-locus genotype information about a set of indviduals. Pritchard et al [3] proposed a model-based method which uses a statistical methodology known as admixture models to find population structure. This model, and admixture models in general, belong to a more general class of hierarchical Bayesian models known as *mixed membership models* [4], which postulate that genetic markers of each individual are independent identically distributed *iid* [3] or spatially coupled [5] samples from multiple population-specific fixed-dimensional multinomial distributions (known as *allele frequency profiles* [5], or AP) of marker alleles. Under this assumption, the *admixture* model identifies each ancestral population by a specific AP (that defines a unique vector of allele frequencies of each marker in each ancestral population), and displays the fraction of contributions from each AP in a modern individual genome as an *admixing vector* (also known as an *ancestral proportion vector* or *map vector*) in a *structural map* over the study population sample. Figure 1 shows an example of a structural map of four modern populations inferred from a portion of the HapMap multi-population dataset by *Structure*. In this *population structural map*, the *admixing vector* underlying each individual is represented as a thin vertical line which shows the fraction of the individual's genome which originated from each ancestral population, as given by a unique AP. This method has been successfully applied to human genetic data in [1] and has unraveled impressive patterns in the genetic structures of world population.

A recent extension of *Structure*, known as the Structurama [6], relaxes the finite ancestral population assumption in the admixture model by employing a Dirichlet process prior over the ancestral allele frequency profiles, which can automatically estimate the maximum *a posteriori* probable number of ancestral populations. Since the need for manual selection of ancestral population cardinality is often considered to be a drawback of finite admixture models, this extension represents a useful improvement. Anderson et al [7] address the problem of identifying species hybrids into categories using a model-based Bayesian clustering approach implemented in the NewHybrid program. While this problem is not exactly identical to the problem of stratifying the structure of highly admixed populations, it is useful for structural analysis of populations which have recently admixed. The BAPS program developed by Corander et al [8] also uses a Bayesian approach to find the best partition of a set of individuals into sub-populations on the bases of genotypes. Parallel to the aforementioned model-based approaches for genomic structural analysis, direct algebraic eigen-decomposition and dimensionality reduction methods, such as the *Eigensoft* program developed by Patterson et al [9] based on Principal Components Analysis (PCA), offer an alternative approach to explore and visualize the ancestral composition of modern population, and facilitate formal statistical tests for significance of population differentiation. However, unlike the model-based methods such as the *Structure*, where each inferred ancestral population bears a physical meaning as population-specific allele-frequency profiles, the eigen-vectors computed by *Eigensoft* represent the mutually-orthogonal directions in an abstract low-dimensional ancestral space in which population samples can be embedded and visualized; they can be understood as mathematical surrogates of independent genetic sources underlying a population sample, but lack an concrete physical interpretation under a generative genetic inheritance model. Analyses based on *Eigensoft* are usually limited to 2-dimensional ancestral spaces, offering limited power in stratifying highly admixed populations.

This progress notwithstanding, an important aspect of population admixing that is largely missing in the existing methods is the effect of allele mutations between the ancestral and current alleles of genetic markers, which can dramatically influence the accuracy of the structural estimation of current populations, and reveal additional information of population evolution such as the the divergence time and migration history of admixed populations.

Consider for example the *Structure* model- since an AP merely represents the *frequency* of alleles in an ancestral population, rather than the actual allelic content or haplotypes of the alleles themselves, the admixture models developed so far based on AP do not model genetic changes due to mutations from the ancestral alleles. Indeed, a serious pitfall of the model underlying *Structure*, as pointed out in [10], is that there is no mutation model for modern individual alleles with respect to hypothetical common prototypes in the ancestral populations, i.e, every unique allele in the modern population is assumed to have a distinct ancestral proportion, rather than allowing the possibility of it just being a descendent of some common ancestral allele that can also give rise to other closely related alleles at the same locus in the modern population. Thus, while *Structure* aims to provide ancestry information for each individual and each locus, there is no explicit representation of "ancestors" as a physical set of "founding alleles". Therefore, the inferred population structural map emphasizes revealing the contributions of *abstract* population-specific ancestral proportion profiles, which does not directly reflect individual diversity or the extent of genetic changes with respect to the founders. Therefore, *Structure* does not enable inference of the founding genetic patterns, the age of the founding alleles, or the population divergence time [10].

Another important issue in determining population structure is to accurately assess the extent of genetic admixture in different populations. The *Structure* model, like a lot of other models, is based on the assumption of presence of admixture. However, as we shall see later, on real data, it produces results that favor clustering individuals into predominantly one allele frequency profile or the other, thus leading us to conclude that there was little or no admixing between the ancestral human populations. While such a partitioning of individuals would be desirable for clustering them into groups, it does not offer us enough biological insight into the intermixing of the populations.

In this paper, we present *mStruct* (for *structure under mutations*), based on a new model: an admixture of population-specific mixtures of inheritance model (AdMim). Rather than assuming an admixture of multinomial distributions (of different allele frequencies) as in *mStruct*, statistically, AdMim is an *admixture of mixtures* model, which represents each ancestral population as a mixture of ancestral alleles each with its own inheritance process, and each modern individual as an "ancestry vector" (or *map vector*) that represents membership proportions among the ancestral populations. *mStruct* facilitates estimation of both the *structural map* of populations (incorporating mutations) and the mutation parameters of either SNP or microsatellite alleles under various context. A new variational inference algorithm was developed for inference and learning. We compare our method with *Structure* on both synthetic genotype data, and on the microsatellite and SNP genotype data of world populations [1, 2]. Our results show the presence of significant levels of admixture among the founding populations. As our results show, we believe that Structure's inability to allow mutations results in it inferring very low levels of admixture in human data. We also report interesting genetic divergence in world populations revealed by the mutation patterns we estimated.

2 The statistical model

The *mStruct* model differs from the *Structure* model in two main aspects: the representation of ancestral populations, and the generative process for sampling a modern individual. In this section we describe in details the statistical underpinning of these two aspects.

2.1 Representation of Populations

To reveal the genetic composition of each modern individual in terms of contributions from hypothetical ancestral populations via statistical inference on multilocus genotype data, one must first choose an appropriate representation of ancestral populations. We begin with a brief description of a commonly used method, followed by a new method we propose that allows mutations to be captured.

2.1.1 Population-Specific *Allele Frequency* Profiles

Since all markers that are used for population strucure determination are polymorphic in nature, it is not surprising that the most intuitive representation is a set of frequency vectors for all alleles observed at all the loci. For example, we can represent an ancestral population k by a unique set of population-specific

multinomial distributions, $\boldsymbol{\beta}^k \equiv \{\vec{\beta}_i^k \ ; \ i = 1 : I\}$, where $\vec{\beta}_i^k = [\beta_{i,1}^k, \ldots, \beta_{i,L_i'}^k]$ is the vector of multinomial parameters, also known as the *allele frequency profile* [5], or AP, of the allele distribution at locus i in ancestral population k; L_i' denotes the total number of observed marker alleles at locus i, and I denotes the total number of marker loci. This representation, known as *population-specific allele frequency profiles*, is used by the program *Structure*.

2.1.2 Population-Specific *Mixtures of Ancestral Alleles*

Being a vector of observed allele frequencies, an AP does not enable us to allow the possibility of mutations, i.e, there is no way of representing that two observed alleles might have been derived from a single ancestral allele by two different mutations. This possibility can be represented in a population by a genetically more realistic statistical model known as the *population-specific mixtures of ancestral alleles (MAA)*. For each locus i, an MAA for ancestral population k is a triple $\{\mu_i^k, \delta_i^k, \vec{\beta}_i^k\}$ consisting of a set of *ancestral* (or founder) alleles $\mu_i^k = (\mu_{i,1}^k, \ldots, \mu_{i,L_i}^k)$, which can differ from their descendent alleles in the modern population; a mutation parameter δ_i^k associated with the locus, which can be further generalized to be allele-specific if necessary; and an AP $\vec{\beta}_i^k$ which now represents the frequencies of the *ancestral* alleles. Here L_i denotes the total number of ancestral alleles at loci i. By explictly linking a mutation model with the population, we can now capture mutation events as described above.

An MAA is strictly more expressive than an AP, because the incorporation of a mutation model helps to capture details about the population structure which an AP cannot; and the MAA reduces to the AP when the mutation rates become zero and the founders are identical to their descendents. MAA is also arguably more realistic because it allows mutation rates to be different for different founder alleles even within the same ancestral population, as is commonly the case with many genetic markers. For example, the mutation rates for microsatellite alleles are believed to be dependent on their length (number of repeats). As we show shortly, with an MAA, one can examine the mutation parameters corresponding to each ancestral population via Bayesian inference from genotype data; this might enable us to infer the age of alleles, and also estimate population divergence times.

Let $i \in \{1, \ldots, I\}$ index the position of a locus in the study genome, $n \in \{1, \ldots, N\}$ index an individual in the study population, and $e \in \{0, 1\}$ index the two possible parental origin of an allele (in this study we do not require strict phase information of the two alleles, so the index e is merely used to indicate diploid data). Under an MAA specific to an ancestral population k, the correspondence between a marker allele X_{i,n_e} and a founder $\mu_{i,l}^k \in \mu_i^k$ is not directly observable. For each allele founder $\mu_{i,l}^k$, we associate with it an inheritance model $p(\cdot|\mu_{i,l}^k, \delta_{i,l}^k)$ from which descendants can be sampled. Then, given specifications of the ancestral population from which X_{i,n_e} is derived from, which is denoted by hidden indicator variable Z_{i,n_e}, the conditional distribution of X_{i,n_e} under MAA follows a mixture of population-specific inheritance model: $P(x_{i,n_e} = l' \mid z_{i,n_e} = k) = \sum_{l=1}^{L} \beta_{i,l}^k P(x_{i,n_e}|\mu_{i,l}^k, \delta_{i,l}^k)$. Comparing to the counterpart of this function under AP: $P(x_{i,n_e} = l' \mid z_{i,n_e} = k) = \beta_{i,l'}^k$, we can see that the latter cannot explicitly model allele diversities in terms of molecular evolution from the founders.

2.2 A New Admixture Model for Population Structure

Admixtures are useful for modeling objects (e.g., human beings) each comprising multiple instances of some attributes (e.g., marker alleles), each of which comes from a (possibly different) source distribution $P_k(\cdot|\Theta_k)$, according to an individual-specific *admixing coefficient vector* (a.k.a. *map vector*) $\vec{\theta}$. The *map vector* represents the normalized contribution from each of the source distributions $\{P_k \ ; \ k = 1 : K\}$ to the study object. For example, for every individual, the alleles at all marker loci may be inherited from founders in different ancestral populations, each represented by a unique distribution of founding alleles and the way they can be inherited. Formally, this scenario can be captured in the following generative process:

1. For each individual n, draw the admixing vector: $\vec{\theta_n} \sim P(\cdot|\alpha)$, where $P(\cdot|\alpha)$ is a pre-chosen map prior.
2. For each marker allele $x_{i,n_e} \in \mathbf{x}_n$
   - 2.1: draw the latent *ancestral-population-origin* indicator $z_{i,n_e} \sim \text{Multinomial}(\cdot|\vec{\theta_n})$;
   - 2.2: draw the allele $x_{i,n_e}|z_{i,n_e} = k \sim P_k(\cdot|\Theta_k)$.

As discussed in the previous section, an ancestral population can be either represented as an AP or as an MAA. These two different representations lead to two different probability distributions for $P_k(\cdot|\Theta_k)$ in the last sampling step above, and thereby two different admixtures of very different characteristics.

2.2.1 The existing model

In *Structure*, the ancestral populations are represented by a set of population-specific APs. Thus the distribution $P_k(\cdot|\Theta_k)$ from which an observed allele can be sampled is a multinomial distribution defined by the rates of all observed alleles in the ancestral population, i.e., $x_{i,n_e}|z_{i,n_e} = k \sim \text{Multinomial}(\cdot|\vec{\beta}_i^k)$. Using this probability distribution in the general admixture scheme outlined above, we can see that *Structure* essentially implements an *admixture of population-specific allele frequency profiles* model. This model has been successfully applied to human genetic data in [1] and has unraveled impressive patterns in the genetic structures of world population; and it has been generalized to allow linked loci and correlated allele frequencies in [5]. But a serious pitfall of using such a model, as pointed out in [10], is that there is no error model for individual alleles with respect to the common prototypes, i.e, every unique measurement at a particular allele is assumed to be a new allele, rather than allowing the possibility of it just being derived from some common ancestral allele at that marker as a result of a mutation.

2.2.2 The proposed model

We propose to represent each ancestral population by a set of population-specific MAAs. Recall that in an MAA for each locus we define a finite set of founders with prototypical alleles $\mu_i^k = (\mu_{i,1}^k, \ldots, \mu_{i,L_i}^k)$ that can be different from the alleles observed in a modern population; each founder is associated with a unique frequency $\beta_{i,l}^k$, and a unique (if desired) mutation model from the prototype allele parameterized by rate $\delta_{i,l}^k$. Under this representation, now the distribution $P_k(\cdot|\Theta_k)$ from which an observed allele can be sampled becomes a mixture of inheritance models each defined on a specific founder; and the ensuing sampling module to be plugged into the general admixture scheme outlined above (to replace step 2.2) becomes a two-step generative process:

- 2.2a: draw the latent founder indicator $c_{i,n_e}|z_{i,n_e} = k \sim \text{Multinomial}(\cdot|\vec{\beta}_i^k)$;
- 2.2b: draw the allele $x_{i,n_e}|c_{i,n_e} = l, z_{i,n_e} = k \sim P_m(\cdot|\mu_{i,l}^k, \delta_{i,l}^k)$,

where $P_m()$ is a mutation model that can be flexibly defined based on whether the genetic markers are microsatellites or single nucleotide polymorphisms. We call this model an *admixture of population-specific inheritance models* (AdMim), while the previous model is technically only an admixture of population specific allele frequency profiles. Figure 2(a) shows a graphical model the overall generative scheme for AdMim, in comparison with the admixture of population-specific allele rates discussed earlier. From the figure, we can clearly see that *mStruct* is an extended *Structure* model which allows noisy observations.

For simplicity of presentation, in the model described above we assume that for a particular individual, the genetic markers at each locus are conditionally *iid* samples from a set of population-specific fixed-dimensional mixture of inheritance models, and that the set of founder alleles at a particular locus is the same for all ancestral populations ($\mu_i^k = \mu_i$). We shall also assume that the mutation parameters for each population at any locus are independent of the alleles at that locus ($\delta_{i,l}^k = \delta_i^k$). Also our model assumes Hardy-Weinberg equilibrium within populations. The simplifying assumptions of *unliked loci, no linkage disequilibrium between loci within populations* can be easily removed by incorporating Markovian dependencies over ancestral indicators Z_{i,n_e} and Z_{i+1,n_e} of adjacent loci, and over other parameters such as the allele frequencies $\vec{\beta}_i^k$ in exactly the same way as in *Structure*. We can also introduce Markovian dependencies over mutation rates at adjacent loci, which might be desirable to better reflect the dynamics of molecular evolution in the genome. We defer such extensions to a later paper.

2.3 Mutation model

As described above, our model is applicable to almost all kinds of genetic markers by plugging in an appropriate allele mutation model (i.e., inheritance model) $P_m()$. We now discuss two mutation models, for microsatellites and SNPs, respectively.

2.3.1 Microsatellite mutation model

Microsatellites are a class of tandem repeat loci that involve a base DNA unit of 1-4 base pairs in length. They are special in that microsatellite DNA has significantly high mutation rates as compared to other DNA, with mutation rates as high as 10^{-3} or 10^{-4} [11, 12]. The large amount of variation present in microsatellite DNA makes it ideal for differentiating between closely related populations. Microsatellite loci have been used before DNA fingerprinting [13], linkage analysis [14], and in the reconstruction of human phylogeny [15]. By applying theoretical models of microsatellite evolution to data, questions such as time of divergence of two populations can be attempted to be addressed [16, 17].

The choice of a suitable microsatellite mutation model is important, for both computational and interpretation purposes. Below we discuss the mutation model that we use and the biological interpretation of the parameters of the mutation model. We begin with a stepwise mutation model for microsatellites widely used in forensic analysis [18, 19].

This model defines a conditional distribution of a progeny allele b given its progenitor allele a, both of which take continuous values:

$$p(b|a) = \frac{1}{2}\xi(1-\delta)\delta^{|b-a|-1}, \tag{1}$$

where ξ is the mutation rate (probability of any mutation), and δ is the factor by which mutation decreases as distance between the two alleles increases. Although this mutation distribution is not stationary (i.e. it does not ensure allele frequencies to be constant over the generations), it is simple and commonly used in forensic inference. To some degree δ can be regarded as a parameter that controls the probability of unit-distance mutation, as can be seen from the following identity: $p(b+1|a)/p(b|a) = \delta$.

In practice, the alleles for almost all microsatellite markers are represented by discrete integer counts. Also the two-parameter stepwise mutation model described above complicates the inference process. We propose a discrete microsatellite mutation model that is a simplification of Eq.1, but captures its main idea. We posit that: $P(b|a) \propto \delta^{|b-a|}$. Since $b \in [1, \infty)$, the normalization constant of this distribution is:

$$\begin{aligned}
\sum_{b=1}^{\infty} P(b|a) &= \sum_{b=1}^{a} \delta^{a-b} + \sum_{b=a+1}^{\infty} \delta^{b-a} \\
&= \frac{1-\delta^a}{1-\delta} + \frac{\delta}{1-\delta} \\
&= \frac{1+\delta-\delta^a}{1-\delta},
\end{aligned}$$

which gives the mutation model as

$$P(b|a) = \frac{1-\delta}{1-\delta^a+\delta}\delta^{|b-a|}. \tag{2}$$

We can interpret δ as a variance parameter, the factor by which probability drops as a fuction of the distance between the mutated version b of the allele a. Figure 3 shows the discrete pdf for various values of δ.

Determination of founder set at each locus: According to our model assumptions, there can be a different number of founder alleles at each locus. This number is typically smaller than the number of alleles observed at each marker since the founder alleles are "ancestral". To estimate the appropriate number and allele states of founders, we fit finite mixtures (of fixed size, corresponding to the desired number of ancestral alleles) of microsatellite mutation models over all the measurements at a particular marker for all individuals. We use the Bayesian Information Criterion (BIC) to determine the best number and state-estimation of founder alleles to use at each locus, since information criteria tend to favor smaller number of founder alleles which fit the observed data well.

For each locus, we fit many different finite-sized mixtures of mutation distributions, with the size varying from 1 to the number of observed alleles at the locus. For each mixture size, the likelihood is optimized and a BIC value is computed. The number of founder alleles is chosen to be the size of the mixture that has the best (minimum) BIC value. It is important to note that this can be performed as a pre-processing step

before the actual inference or estimation procedures since we assumed that the set of founder alleles at each locus was the same for all populations.

Choice of mutation prior: In our model, the δ parameter, as explained above, is a population-specific parameter that controls the probability of stepwise mutations. Being a parameter that controls the variance of the mutation distribution, there is a possibility that inference on the model will encourage higher values of δ to improve the log-likelihood, in the absence of any prior distribution on δ. To avoid this situation, and to allow more meaningful and realistic results to emerge from the inference process, we impose on δ a beta prior that will be biased towards smaller values of δ. The beta prior will be a fixed one and will not be among the parameters we estimate.

2.3.2 SNP mutation model

SNPs, or single nucleotide polymorphisms, represent the largest class of individual differences in DNA. In general, there is a well-defined correlation between the age of the mutation producing a SNP allele and the frequency of the allele. For SNPs, we use a simple pointwise mutation model, rather than more complex block models. Thus, the observations in SNP data are only binary in nature (0/1). So, given the observed allele b, we say that the probability of it being derived from the founder allele a is given by:

$$P(b|a) = \delta^{\mathcal{I}[b=a]} \times (1-\delta)^{\mathcal{I}[b\neq a]}; \quad a, b \in \{0, 1\}. \tag{3}$$

In this case, the mutation parameter δ is the probability that the observed allele is not identical to the founder allele, but derived from it due to a mutation.

2.4 Inference and Parameter Estimation

2.4.1 Probability distribution on the model

For notational convenience, we will ignore the diploid nature of observations in the analysis that follows. With the understanding that the analysis is carried out for the n^{th} individual, we will drop the subscript n. Also, we overload the indicator variables z_i and c_i to be both, arrays with only one element equal to 1 and the rest equal to 0, as well as scalars with a value equal to the index at which the array forms have 1s. In other words: $z_i \in \{1, \ldots, K\}, c_i \in \{1, \ldots, L\}, z_{i,k} = \mathcal{I}[z_i = k]$, and $c_{i,l} = \mathcal{I}[c_i = l]$, where $\mathcal{I}[\cdot]$ denotes an indicator function that equals to 1 when its argument is true and 0 otherwise.

The joint probability distribution of the the data and the relevant variables under the AdMim model can then be written as:

$$\begin{aligned} & P\left(\mathbf{x}, \mathbf{z}, \mathbf{c}, \vec{\theta} | \alpha, \boldsymbol{\beta}, \mu, \delta\right) \\ = \; & p\left(\vec{\theta}|\alpha\right) \prod_{i=1}^{I} P\left(z_i|\vec{\theta}\right) P\left(c_i|z_i, \vec{\beta}_i^{k=1:K}\right) P\left(x_i|c_i, z_i, \mu_i, \delta_i^{k=1:K}\right) \end{aligned}$$

The marginal likelihood of the data can be computed by summing/integrating out the latent variables.

$$\begin{aligned} P\left(x|\alpha, \boldsymbol{\beta}, \mu, \delta\right) = & \frac{\Gamma\left(\sum_{k=1}^{K} \alpha_k\right)}{\prod_{k=1}^{K} \Gamma\left(\alpha_k\right)} \int \left(\prod_{k=1}^{K} \theta_k^{\alpha_k - 1}\right) \ldots \\ & \times \prod_{i=1}^{I} \sum_{k=1}^{K} \left(\prod_{k=1}^{K} \theta_k^{z_{i,k}}\right) \sum_{i=1}^{I} \prod_{k=1}^{K} \prod_{l=1}^{L_i} \left(\beta_{i,l}^{k}\right)^{c_{i,l} z_{i,k}} \times P\left(x_i|\mu_{i,l}, \delta_i^k\right)^{c_{i,l} z_{i,k}} d\vec{\theta} \end{aligned}$$

However, a closed-form solution to this summation/integration is not possible, and indeed exact inference on hidden variables such as the map vector $\vec{\theta}$, and estimation of model parameters such as the mutation rates δ under AdMim is intractable. [3] developed an MCMC algorithms for approximate inference for their admixture model underlying *Structure*. While it is straightforward to implement a similar MCMC scheme for AdMim, we choose to apply a computationally efficient approximate inference method known as variational inference [20].

2.5 Variational Inference

We use a mean-field approximation for performing inference on the model. This approximation method approximates an intractable joint posterior $p()$ of the all hidden variables in the model by a product of marginal distributions $q() = \prod q_i()$, each over only a single hidden variable. The optimal parameterization of $q_i()$ for each variable is obtained by minimizing the Kullback-Leibler divergence between the variational approximation q and the true joint posterior p. Using results from the the Generalised Mean Field theory [21], we can write the variational distributions of the latent variables as follows:

$$
\begin{aligned}
q(\vec{\theta}) &\propto \prod_{k=1}^{K} \theta_k^{\alpha_k - 1 + \sum_{i=1}^{I} \langle z_{i,k} \rangle} \\
q(c_i) &\propto \prod_{l=1}^{L} \left(\prod_{k=1}^{K} \left(\beta_{i,l}^k f(x_i | \mu_{i,l}, \delta_i^k) \right)^{\langle z_{i,k} \rangle} \right)^{c_{i,l}} \\
q(z_i) &\propto \prod_{k=1}^{K} \left(e^{\langle \log(\theta_k) \rangle} \left(\prod_{l=1}^{L} \beta_{i,l}^k f(x_i | \mu_{i,l}, \delta_i^k)^{\langle c_{i,l} \rangle} \right) \right)^{z_{i,k}}.
\end{aligned}
$$

In the distributions above, the '$\langle \cdot \rangle$' are used to indicate the expected values of the enclosed random variables. A close inspection of the above formulas reveals that these variational distributions have the form $q(\vec{\theta}) \sim \text{Dirichlet}(\gamma_1, \ldots, \gamma_K)$, $q(z_i) \sim \text{Multinomial}(\rho_{i,1}, \ldots, \rho_{i,K})$, and $q(c_i) \sim \text{Multinomial}(\xi_{i,1}, \ldots, \xi_{i,L})$, respectively, where the parameters $\gamma_k, \rho_{i,k}$ and $\xi_{i,l}$ are given by the following equations:

$$
\begin{aligned}
\gamma_k &= \alpha_k + \sum_{i=1} \langle z_{i,k} \rangle \\
\rho_{i,k} &= \frac{e^{\langle \log(\theta_k) \rangle} \left(\prod_{l=1}^{L} \beta_{i,l}^k f(x_i | \mu_{i,l}, \delta_i^k)^{\langle c_{i,l} \rangle} \right)}{\sum_{k=1}^{K} \left(\theta_k \left(\prod_{l=1}^{L} \beta_{ikl} f(x_i | \mu_{il}, \delta_{ik})^{\langle c_{il} \rangle} \right) \right)} \\
\xi_{i,l} &= \frac{\prod_{k=1}^{K} \left(\beta_{i,l}^k f(x_i | \mu_{i,l}, \delta_i^k) \right)^{\langle z_{i,k} \rangle}}{\sum_{k=1}^{K} \left(\prod_{k=1}^{K} \left(\beta_{i,l}^k f(x_i | \mu_{i,l}, \delta_{i,k}) \right)^{\langle z_{i,k} \rangle} \right)}
\end{aligned}
$$

and they have the properties: $\langle \log(\theta_k) \rangle = \gamma_k$, $\langle z_{i,k} \rangle = \rho_{i,k}$ and $\langle c_{i,l} \rangle = \xi_{i,l}$, which suggest that they can be computed via fixed point iterations. It can be shown that this iteration will converge to a local optimum, similar to what happens in an EM algorithm. Empirically, a near global optimal can be obtained by multiple random restarts of the fixed point iteration. Typically, such a mean-field variational inference converges much faster then sampling [21]. Upon convergence, we can easily compute an estimate of the map vector $\vec{\theta}$ for each individual from $\{\langle \log(\theta_k) \rangle \ ; \ k = 1 : K\}$.

3 Hyperparameter Estimation

The hyperparameters of our model, i.e., $\{\mu, \delta, \boldsymbol{\beta}\}$, and the Dirichlet hyperparameter α, can be estimated by maximizing the lower bound on the log-likelihood as a function of the current values of the hyperparameters, via a variational EM algorithm. For the hyperparameter estimation, we perform empirical Bayes estimation using the variational Expectation Maximization algorithm described in [22]. The variational inference described in Section 2.5 provides us with a tractable lower bound on the log-likelihood as a function of the current values of the hyperparameters. We can thus maximize it with respect to the hyperparameters. If we alternately carry out variational inference with fixed hyperparameters, followed by a maximization of the lower bound with respect to the hyperparameters for fixed values of the variational parameters, we can get an empirical Bayes estimate of the hyperparameters. The derivation, details of which we will not show here, leads to the following iterative algorithm:

1. (*E-step*) For each individual, find the optimizing values of the variational parameters $(\gamma^n, \rho^n, \xi^n; n \in 1, \ldots, N)$ using the variational updates described above.

2. (*M-step*) Maximize the resulting variational lower bound on the likelihood with respect to the model parameters, namely $\alpha, \beta, \mu, \delta$.

The two steps are repeated until the lower bound on the log-likelihood converges.

3.1 Bayes estimates of hyperparameters

Denote the original set of hyperparameters by

$$\mathbb{H} = \{\alpha, \boldsymbol{\beta}, \mu, \delta\} \tag{4}$$

and the variational parameters for the n^{th} individual by

$$\mathbb{V}^n = \{\gamma^n, \rho^n, \xi^n\} \tag{5}$$

The variational lower bound to the log-likelihood for the n^{th} individual is given by:

$$\begin{aligned} L_n(\mathbb{H}, \mathbb{V}^n) &= \mathbb{E}_q[\log p(x_n, \vec{\theta}_n, z_{.,n}, c_{.,n}; \mathbb{H})] \\ &\quad - \mathbb{E}_q[\log q(\vec{\theta}_n, z_{.,n}, c_{.,n}; \mathbb{H}, \mathbb{V}^n)] \end{aligned} \tag{6}$$

The subscripts indicate the n^{th} individual. As described earlier, we partition the variational approximation as:

$$q(\vec{\theta}_n, z_{.,n}, c_{.,n}; \mathbb{H}, \mathbb{V}) = q(\vec{\theta}_n) \prod_{i=1}^{I} q(z_{i,n}) q(c_{i,n}) \tag{7}$$

So we can expand Equation 6 as

$$\begin{aligned} L^n(\mathbb{H}, \mathbb{V}_i) &= \mathbb{E}_q[\log p(\vec{\theta}_n; \alpha)] + \mathbb{E}_q[\log p(z_{.,n}|\vec{\theta}_n)] + \mathbb{E}_q[\log p(c_{.,n}|z_{.,n})] \\ &\quad + \mathbb{E}_q[\log p(x_n|c_{.}, n, z_{.,n}, \boldsymbol{\beta})] - \mathbb{E}_q[\log q(\vec{\theta}_n)] - \mathbb{E}_q[\log q(z_{.}, n)] - \mathbb{E}_q[\log q(c_{.}, n)] \end{aligned} \tag{8}$$

The lower bound to the total data log-likelihood is

$$L(\mathbb{H}, \mathbb{V}) = \sum_{n=1}^{N} L^n(\mathbb{H}, \mathbb{V}^n)$$

which, on substituting from Equation 8 becomes

$$\begin{aligned} L(\mathbb{H}, \mathbb{V}) &= \sum_{n=1}^{N} \mathbb{E}_q[\log p(\vec{\theta}_n; \alpha)] + \sum_{n=1}^{N} \mathbb{E}_q[\log p(z_{.,n}|\vec{\theta}_n)] \\ &\quad + \sum_{n=1}^{N} \mathbb{E}_q[\log p(c_{.,n}|z_{.,n})] + \sum_{n=1}^{N} \mathbb{E}_q[\log p(x_n|c_{.,n}, z_{.,n}, \boldsymbol{\beta})] \\ &\quad - \sum_{n=1}^{N} \mathbb{E}_q[\log q(\vec{\theta}_n)] - \sum_{n=1}^{N} \mathbb{E}_q[\log q(z_{.,n})] \\ &\quad - \sum_{n=1}^{N} \mathbb{E}_q[\log q(c_{.,n})] \end{aligned} \tag{9}$$

Simplifying each term in Equation 9, we get

$$
\begin{aligned}
L(\mathbb{H},\mathbb{V}) &= N\log\Gamma\left(\sum_{k=1}^{K}\alpha_k\right) - N\sum_{k=1}^{K}\log\Gamma\left(\alpha_k\right) + \sum_{n=1}^{N}\sum_{k=1}^{K}\left(\alpha_k-1\right)\left[\psi\left(\gamma_k^n\right) - \psi\left(\sum_{k=1}^{K}\gamma_k^n\right)\right] \\
&+ \sum_{n=1}^{N}\sum_{i=1}^{I}\sum_{k=1}^{K}\rho_{i,k}^n\left[\psi\left(\gamma_k^n\right) - \psi\left(\sum_{k=1}^{K}\gamma_k^n\right)\right] \\
&+ \sum_{n=1}^{N}\sum_{i=1}^{I}\sum_{k=1}^{K}\sum_{l=1}^{L_i}\xi_{i,l}^n\rho_{i,k}^n\log\beta_{il}^k \\
&+ \sum_{n=1}^{N}\sum_{i=1}^{I}\sum_{k=1}^{K}\sum_{l=1}^{L_i}\xi_{i,l}^n\rho_{i,k}^n\left[\log\left(1-\delta_i^k\right) + \left|x_{i,n}-\mu_{i,l}\right|\log\delta_i^k - \log\left(1+\delta_i^k-\left(\delta_i^k\right)^{\mu_{i,l}}\right)\right] \\
&- \sum_{n=1}^{N}\left[\log\Gamma\left(\sum_{k=1}^{K}\gamma_k^n\right) - \sum_{k=1}^{K}\log\Gamma\left(\gamma_k^n\right)\sum_{k=1}^{K}\left(\gamma_k^n-1\right)\left[\psi\left(\gamma_k^n\right) - \psi\left(\sum_{k=1}^{K}\gamma_k^n\right)\right]\right] \\
&- \sum_{n=1}^{N}\sum_{i=1}^{I}\sum_{l=1}^{L_i}\xi_{i,l}^n\log\xi_{i,l}^n \\
&- \sum_{n=1}^{N}\sum_{i=1}^{I}\sum_{k=1}^{K}\rho_{i,k}^n\log\rho_{i,k}^n
\end{aligned}
\tag{10}
$$

Each line in Equation 10 corresponds to an expectation term in Equation 9. In the following subsections, we will briefly describe how the maximum-likelihood estimates of the hyperparameters were obtained from the variational lower bound.

3.2 Estimating ancestral allele frequency profiles β

Since β is a table of probability distributions, the values of its elements are constrained by the equality $\sum_{l=1}^{L_i}\beta_{i,l}^k = 1$ for all combinations of $\{i,k\}$. So to find the optimal values of β satisfying this constraint while maximizing the variational lower bound, we introduce lagrange multipliers $\nu_{i,k}$. The new objective function to maximize is then given by:

$$
L_{new}(\mathbb{H},\mathbb{V}) = L(\mathbb{H},\mathbb{V}) + \sum_{i=1}^{I}\sum_{k=1}^{K}\nu_{i,k}\left(\sum_{l=1}^{L_i}\beta_{i,l}^k - 1\right) \tag{11}
$$

Maximizing this objective function gives:

$$
\beta_{i,l}^k = \frac{\sum_{n=1}^{N}\xi_{i,l}^n\rho_{i,k}^n}{\sum_{l=1}^{L_i}\sum_{n=1}^{N}\xi_{i,l}^n\rho_{i,k}^n} \tag{12}
$$

We use a uniform dirichlet prior λ on each multinomial $\vec{\beta}_i^k$. Under this prior, it is not difficult to show that the estimate of $\beta_{i,l}^k$ changes to

$$
\beta_{i,l}^k = \frac{\lambda + \sum_{n=1}^{N}\xi_{i,l}^n\rho_{i,k}^n}{\lambda * L_i + \sum_{l=1}^{L_i}\sum_{n=1}^{N}\xi_{i,l}^n\rho_{i,k}^n} \tag{13}
$$

3.3 Estimating the Dirichlet prior on populations α

For estimating α we use the method described by Minka in [23]. This gives a Newton-Raphson iteration for α that does not involve inversion of the Hessian, and hence is reasonably fast. The update equation we get is:

$$
\alpha^{new} = \alpha^{old} - \left(\mathbf{H^{-1}g}\right) \tag{14}
$$

where

$$\begin{aligned}
\left(\mathbf{H^{-1}g}\right)_k &= \frac{g_k - b}{q_{kk}} \\
g_k &= N\psi\left(\sum_k \alpha_k\right) - N\sum_k \psi\left(\alpha_k\right) \\
&\quad + \sum_i \sum_k \left[\psi\left(\gamma_k^n\right) - \psi\left(\sum_k \gamma_k^n\right)\right] \\
q_{jk} &= -N\psi^{'}\left(\alpha_k\right)\delta\left(j-k\right) \\
b &= \frac{\sum_j g_j/q_{jj}}{1/z - \sum_j 1/q_{jj}} \\
z &= N\psi^{'}\left(\sum_k \alpha_k\right)
\end{aligned}$$

3.4 Estimating the ancestral alleles μ and the cumulative mutation parameters δ

For finding the optimal values of μ and δ, we use simple gradient ascent with line search. μ are actually discrete variables, however, as an approximation, we will assume them to be continuous in the optimization and round off the result to the nearest integer. The gradient of the variational lower bound with respect to μ_{il} is given by

$$\frac{\partial L}{\partial \mu_{il}} = \sum_{n=1}^{N}\sum_{k=1}^{K} \xi_{il}^n \rho_{ik}^n \log \delta_{ik} \left[\text{sign}\left(x_m^n - \mu_{il}\right) + \frac{\delta_{ik}^{\mu_{il}}}{1+\delta_{ik}-\delta_{ik}^{\mu_{il}}}\right] \tag{15}$$

The gradient with respect to δ_{ik} is given by

$$\frac{\partial L}{\partial \delta_{ik}} = \sum_{n=1}^{N}\sum_{l=1}^{L_i} \xi_{il}^n \rho_{ik}^n \left[\frac{|x_m^n - \mu_{il}|}{\delta_{ik}} - \frac{1}{1-\delta_{ik}} - \frac{1-\mu_{il}\delta_{ik}^{\mu_{il}-1}}{1+\delta_{ik}-\delta_{ik}^{\mu_{il}}}\right] \tag{16}$$

Since the values of δ are constrained to be in $[0,1]$, we use the logit transformation to create a mapping from $[0,1]$ to $\mathbb{R}$. This gives us the equations:

$$\begin{aligned}
\sigma_{ik} &= \log\left(\frac{\delta_{ik}}{1-\delta_{ik}}\right) \\
\delta_{ik} &= \text{sigmoid}\left(\sigma_{ik}\right) \\
\frac{\partial L}{\partial \sigma_{ik}} &= \frac{\partial L}{\partial \delta_{ik}}\frac{\partial \delta_{ik}}{\partial \sigma_{ik}} \\
&= \frac{\partial L}{\partial \delta_{ik}} \times \delta_{ik}\left(1-\delta_{ik}\right)
\end{aligned}$$

We can then perform gradient ascent on each μ and δ separately, and repeat this a number of times, to obtain values that increase the lower bound. To constrain values of the mutation parameter δ to allow meaningful interpretation, we use a β prior on it with a small expected value(around 0.1).

While the gradient methods developed are useful for small datasets, they are inefficient on larger datasets and increase the time required for estimation. Hence we look at a couple of small approximations that help speed up the hyperparameter estimation. A careful look at the results produced indicates that once the founder alleles have been picked initially by fitting a mixture of mutation distributions individually at each

locus, the later gradient descent on μ only makes very minor changes in their values, if any at all. So, to improve the speed of the algorithm, we do not perform gradient descent on the founder alleles μ but fix them after initialization. We show below an approximation for estimating the mutation parameter δ.
For the estimation of the mutation parameter (δ), the only relevant term in the likelihood lower bound is the term:

$$\begin{aligned} L(\delta_i^k) = & \sum_{n=1}^{N} \sum_{l=1}^{L_i} \xi_{i,l}^n \rho_{i,k}^n \times \log f(x_i^n; \mu_{i,l}, \delta_i^k) \\ & + \frac{{\delta_i^k}^{\Delta_1 - 1} \left(1 - \delta_i^k\right)^{\Delta_2 - 1}}{B\left(\Delta_1, \Delta_2\right)} \\ & + (\text{Terms not involving } \delta_i^k) \end{aligned} \tag{17}$$

And for the mutation distribution, we use the discrete distribution whose pdf is

$$f(x|\mu, \delta) = \frac{(1-\delta)\delta^{|x-\mu|}}{1 + \delta - \delta^{\mu}} \tag{18}$$

3.4.1 Approximation

We will assume δ to be small in equation 18. So we can ignore the term exponential in μ in the denominator, reducing it to only $(1 + \delta)$. The expansion of $(1 + \delta)^{-1}$ is given by

$$\frac{1}{1+\delta} = 1 - \delta + \delta^2 - \delta^3 + \ldots \tag{19}$$

$$\geq 1 - \delta \tag{20}$$

This gives us a lower bound to the mutation distribution to be

$$f_{lb}(x|\mu, \delta) = (1-\delta)^2 \delta^{|x-\mu|} \tag{21}$$

It is not hard to show that using this form for the mutation distribution allows a closed form MLE for δ. This approximation gives us a lower bound to the likelihood that is not as tight as the variational lower bound. However, it offers a significant improvement in time complexity due to the the existence of a closed form solution, thus avoiding the need for slow gradient-based methods.

4 Experiments and Results

We validated our model on a synthetic microsatellite dataset where the simulated values of the hidden map vector θ_n of each individual and the population parameters $\{\mu^k, \delta^k, \vec{\beta}^k\}$ of each ancestral population are known as ground truth. The goal is to assess the performance of *mStruct* in terms of accuracy and consistency of the estimated map vectors and population parameters, and test of the correctness of the inference and estimation algorithms we developed. We also conduct empirical analysis using *mStruct* of two real datasets: the HGDP-CEPH cell line panel of microsatellite loci and the HGDP SNP data, in comparison with the *Structure* program (version 2.2).

4.1 Validations on Synthetic Data

We simulated 20 microsatellite genotype dataset using the AdMim generative process described in section 2.2, with 100 diploid individuals from 2 ancestral populations, at 50 genotype loci. Each locus has 4 founding alleles, separated by adjustable distances; the mutation parameter at each locus for both populations had default value 0.1, but can be varied to simulate different degrees of divergence. The founding allele frequencies, $\vec{\beta}_i^k$, were drawn from a flat Dirichlet prior with parameter 1. The map vectors θ_n were sampled from a symmetric beta distribution with parameter α, allowing different levels of admixing. We examine the accuracies of several estimates of interest under a number of different simulation conditions, and for each condition we report the statistics of the accuracies across 20 iid synthetic datasets.

4.1.1 Accuracy of population Map estimate

The map vector θ_n reflects the proportions of contributions from different ancestral population to the marker-alleles of each individual. The display of the map vectors of all individuals in a study population gives a *Map* of population structure (see, e.g., Fig. 1 in the introduction), which has been the main output of the *Structure* program . We compare the accuracy of the estimated θ_n w.r.t. the ground truth recorded during the simulation in terms of their L1 distances.

Figure 3 shows an example of this comparison, and we can see that *mStruct* is visually more accurate than *Structure*. Figure 5 shows the accuracy of the Map estimate by *mStruct* on synthetic datasets simulated with different properties, in comparison with that of *Structure*. Fig. 5(a) shows that, under different degrees of biases of population admixing induced by the Dirichlet prior of θ_n, *mStruct* consistently performs better than *Structure*, which shows the consistency and correctness of our inference and parameter estimation procedures. Specifically, as the value of the Beta prior hyperparameter α increases, fewer individuals tend to belong completely to only one population, and more and more individuals become highly admixed. As the figure shows, the performance of both methods degrades as we progress toward this end; however, the severity of degradation of *mStruct* is much less than that of *Structure*. *mStruct* remains robust and dominates *Structure* constantly as the separations between founding alleles decreases (Fig.5(b)), which tends to increasingly confound the ancestral origins of modern alleles. Finally, Fig. 5(c) shows how the presence of mutations affects the performance of both methods. At very low values of the mutation parameters, the performances of both models are comparable; but as the mutation parameter increases in magnitude, the performance of *Structure* degrades significantly. On the other hand, the decrease in accuracy for *mStruct* is hardly noticeable. This shows that our model is resistant to the confounding effect of large mutations.

4.1.2 Accuracy of parameter estimation

An important aspect of guarantee and utility we desire for our model and inference algorithm is that it should offer consistent estimates of the population parameters $\{\mu^k, \delta^k, \vec{\beta}^k\}$ underlying the composition of the ancestral population and their inheritance processes. These estimates offer important insight of the evolutionary history and dynamics of modern population genotype data. We have extensively investigated the robustness and accuracy of all these estimates. Due to space limitations, here we report highlights of mutation rate estimation.

Mutation parameter estimation: We evaluate the performance at recovery of δ^k's by a simple distance measure, such as the L1 distance measure, between the true and inferred values. As explained earlier, the use of a beta prior on each parameter avoids the situation where the variation parameters can be used to accommodate errors into the model if not properly biased initially. So we expect that using the prior improves the recovery of the population-specific mutation parameters. As shown in Figure 6, the estimates of δ^k's are robust and remain low-bias under different degree of admixing (due to changing α) and different ancestor dispersion (due to changing distances among the μ^k's). The accuracy decreases as the value of the mutation parameter itself increases, but remains respectable, as shown in Figure. 6(c).

4.2 Empirical Analysis of Real Datasets

The HGDP-CEPH cell line panel [24, 25] used in [1] contains genotype information from 1056 indviduals from 52 populations at 377 autosomal microsatellite loci, along with geographical and population labels. The HGDP SNP data [2] contains the SNPs genotypes at 2834 loci of 927 unrelated individuals that overlap with the HGDP-CEPH data. To make results for both types of data comparable, we chose the set of only those individuals present in both datasets. As in [1], the choice of the total number of ancestral populations can be left to the user; we tried K ranging from 2 to 5, and we applied BIC to decide the Bayes optimal number of ancestral populations within this range to be $K = 4$. Below, we present the structural analysis under four ancestral populations.

4.2.1 Structural maps from HGDP data

We compare the structural maps inferred from both the microsatellite and the SNP data using *mStruct* and *Structure* (top panels in Figure 7 and Figure 8). The structural maps produced by both programs are quite

similar in the case of SNPs, but are very different for microsatellites. The most obvious difference between the maps produced by both programs is the degree of admixing that the individuals in the program are assigned. *Structure* assigns each geographical population to a distinct ancestral allele frequency profile. Thus, it seems to predict very little admixing effect in modern human populations. While useful for clustering, this might result in loss of potentially useful information about the actual evolutionary history of populations. In contrast, the structure map produced by *mStruct* from microsatellite data suggests that all populations share a common ancestral population with a unique extra component (represented by the magenta color in Figure 7) that characterizes their particular genotypes. It is interesting to note that clustering individuals by the ancestry proportion vectors due to *mStruct* will produce exactly the same clustering partitions as that due to *Structure*. The structural maps produced in the case of SNP data are quite similar for both softwares, with results from *mStruct* again predicting more admixture than *Structure*. It is also interesting to see that the ancestry proportions for European and Middle Eastern regions are more distinct from each other in *mStruct* than in *Structure*, allowing for better separation of the two geographical regions. A possible cause for the inconsistency between the results produced by *mStruct* for SNP data and microsatellite data could be the large difference between their mutation rates, or due to the choice of a simplistic mutation model that might not be powerful enough to capture mutations.

4.2.2 Analysis of the mutation spectrums

Now we report a preliminary analysis of the evolutionary dynamics reflected by the estimated mutation spectrums of different ancestral populations (denoted "am-spectrum"), and of different modern geographical populations (denoted "gm-spectrum"), which is not possible by *Structure*. For the am-spectrum, we compute the mean mutation rates over all loci and founding alleles for each ancestral population as estimated by *mStruct*. We estimate the gm-spectrum as follows: for every individual, a mutation rate is computed as the per-locus number of observed alleles that are attributed to mutations, weighted by the mutation rate corresponding to the ancestral allele chosen for that locus. This can be computed by observing the population-indicator (Z) and the allele-indicator (C) for each individual. We then compute the population mutation rates by averaging mutation rates of all individuals having the same geographical label.

As shown in the gm-spectrums in Figure 7 and Figure 8 (lower sub-panels on the right), the mutation rates for African populations are indeed higher than those of other modern populations. This indicates that they diverged earlier, a common hypothesis of human migration. Other trends in the gm-spectrums also reveal interesting insights, which we do not have space to discuss. The am-spectrums of SNP data in Figure 8 suggest that the founder ancestral population that dominates modern African populations has a higher mutation rate than the other ancestral population, indicating that is the older of the two ancestral populations. The mutation estimates are largely consistent for both microsatellites and SNPs in comparative order, but vastly different in numerical values.

4.3 Model selection

As with all probabilistic models, we face a tradeoff between model complexity and the log-likelihood value that the model achieves. In our case, complexity is controlled by the number of ancestral populations we pick, K. Unlike non-parametric or infinite dimensional models(Dirichlet processes etc.), for models of fixed dimension, it is not clear in general as to what value of K gives us the best balance between model complexity and log-likelihood. In such cases, different information criteria are often used to determine the optimal model complexity. To determine what number of ancestral populations fit the HGDP SNP and microsatellite data best, we computed BIC scores for K=2 to K=5 for both kinds of data separately. The results are shown in Figure 9. From the BIC curves for both SNP and microsatellite data, we can see that the curves suggest K=4 as the best fit for the data.

5 Discussions

The task of estimating the genetic contributions of ancestral populations, i.e., structural map estimation, in each modern individual is an important problem in population genetics. Due to the relatively high rates of mutation in markers such as microsatellites and SNPs, multilocus genotype data usually harbor a large

amount of variations, which allows differentiation even between populations that have close evolutionary relationships. However, to our knowledge, none of the existing methods is able to take advantage of this property to compare how marker mutation rates vary with population and locus, while at the same time exploiting such information for population structural estimation. Traditionally, population structure estimation and mutation spectrum estimation have been performed as separate tasks.

We have developed *mStruct*, which allows estimation of genetic contributions of ancestral populations in each modern individual in light of both population admixture and allele mutation. The variational inference algorithm that we developed allows tractable approximate inference on the model. The ancestral proportions of each individual enable representing population structure in a way that is both visually easy to interpret, as well as amenable to further computational analysis.

The statistical modeling differences between *mStruct* and *Structure* provide an interesting insight into the possible reasons which lead to *mStruct* inferring higher levels of admixture than *Structure*. In Structure's representation of population, every microsatellite allele is considered to be a separate element of the population, even though they might be very similar. In the inheritance model representation, such alleles are considered to be derived from a single ancestral allele. This gives extra similarity to the individuals possessing these alleles. This is probably the main reason that the inferred levels of admixture are higher in *mStruct* and *Structure*.

Another parameter that would also affect inferred levels of admixture is the δ parameter which determines the variance of the mutation distributions. Higher values of δ (tending to 1) lead to significantly higher levels of inferred admixture. If a strong prior is not used, the δ values tend towards 1 in the initial few steps of the variational EM algorithm. This seems to happen due to the initial imprecise assignments for the z and c indicator variables. However, the region of high δ values is a region of low log-likelihood in the parameter space and the EM quickly finds a local optimum which is undesirable due to the low log-likelihood of that region of the parameter space.

In conjunction with geographical location, the inferred ancestry proportions could be used to detect migrations, sub-populations etc. quite easily. Moreover, the ability to estimate population and locus specific mutation rates also allows us to substantiate evolutionary dynamics claims based on high/low mutation rates in certain geographical population, or on high/low mutation rates at certain loci in the genome. While the estimates of mutation rates that *mStruct* provides are not on an absolute scale, the comparison of their relative magnitudes is certainly informative.

As of now, there remain a number of possible extensions to the methodology we presented so far. It would be instructive to see the impact of allowing linked loci as in [5]. We have not yet addressed the issue of the most suitable choice of mutation process, but instead have chosen one that is reasonable and computationally tractable. It would be interesting to combine *mStruct* with the nonparametric Bayesian models based on the Dirichlet processes as in programs such as Spectrum [26] and Structurama [6].

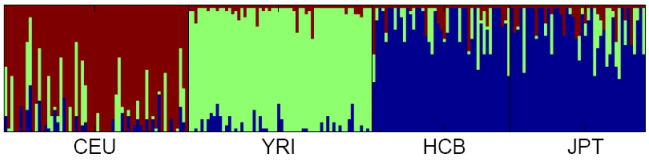
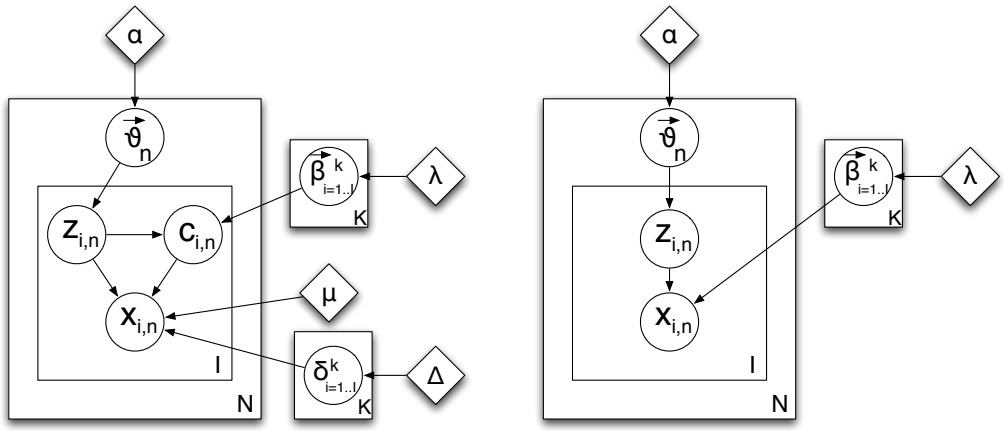


Figure 1: Population structural map inferred by *Structure* on HapMap data consisting of 4 populations.



(a) *mStruct* (b) *Structure 2.1*

Figure 2: Graphical Models: the circles represent random variables and diamonds represent hyperparameters.

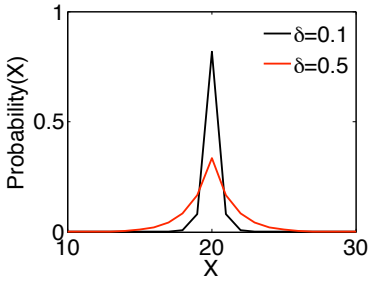


Figure 3: Discrete pdf for two values of mutation parameter.

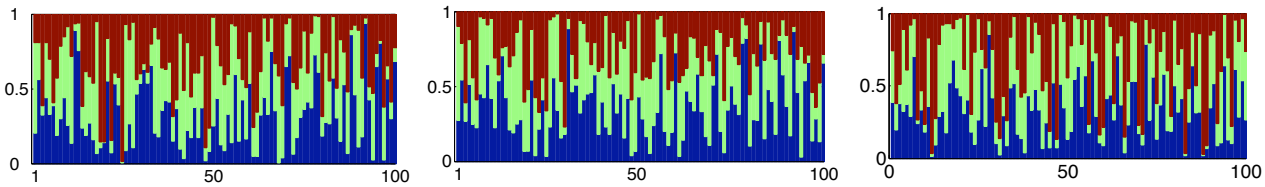
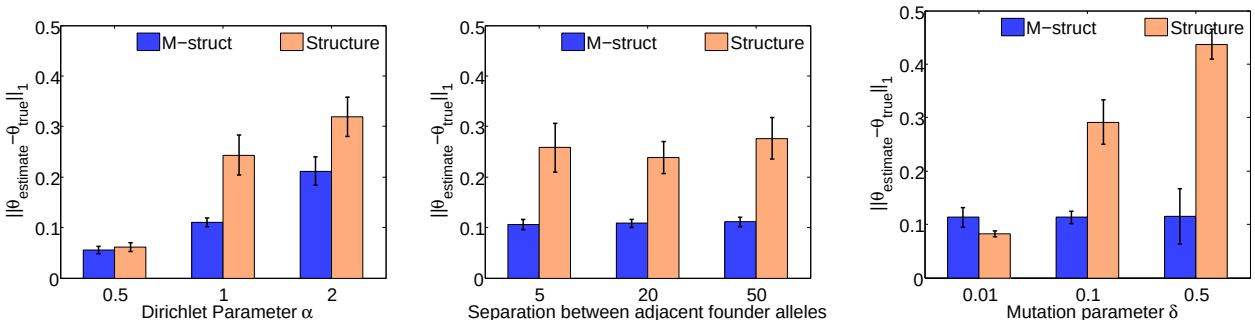
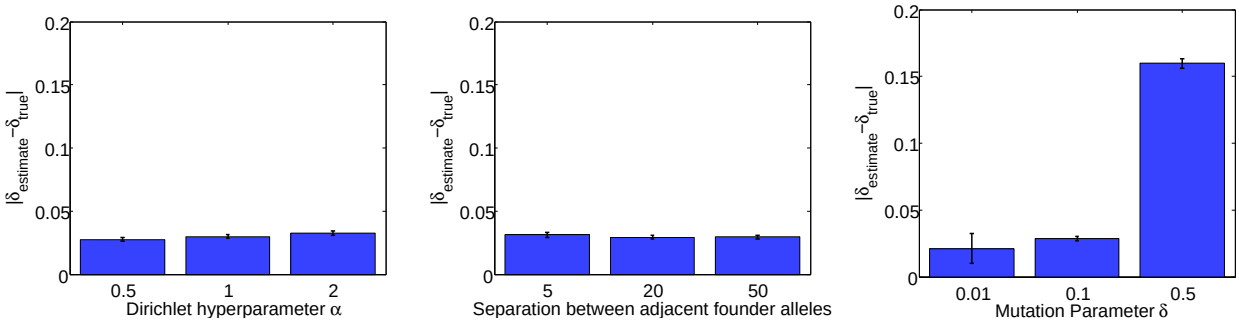


Figure 4: Ancestry structure map for a 3-population simulated dataset. First panel shows the true ancestry proportion vectors θ. Middle panel shows the estimates by *mStruct*. Right panel shows the estimates from *Structure*.



(a) (b) (c)

Figure 5: Accuracy of ancestry proportion vector θ estimation under different conditions.



(a) (b) (c)

Figure 6: Accuracy of microsatellite mutation parameter estimation

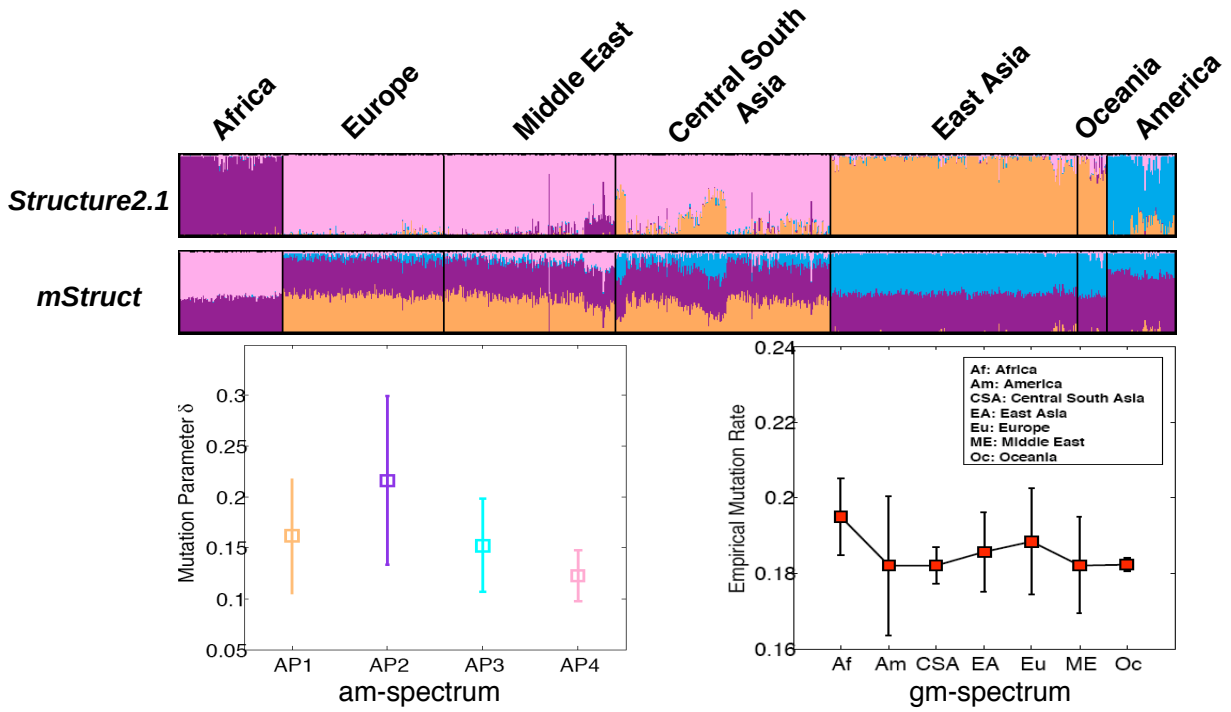


Figure 7: Ancestry structure maps inferred from microsatellite portion of the HGDP dataset, using *mStruct* and *Structure* with 4 ancestral population. The colors represent different ancestral populations.

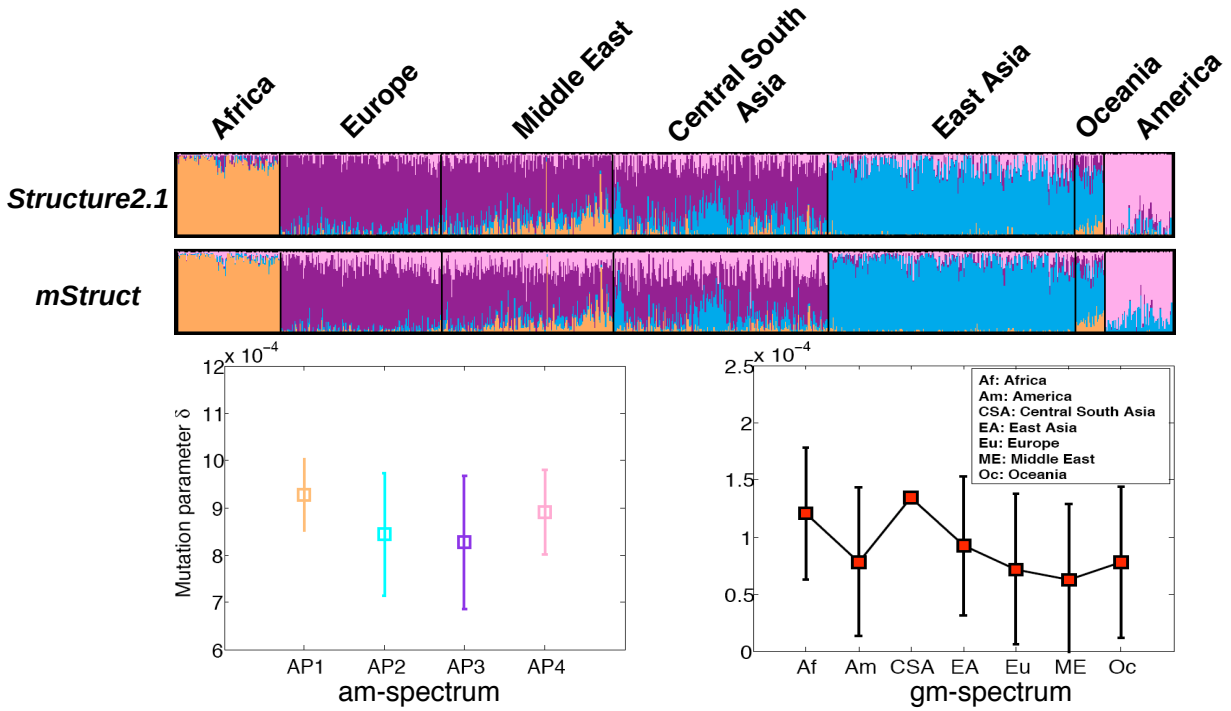


Figure 8: Ancestry structure maps inferred from SNPs portion of the HGDP dataset, using *mStruct* and *Structure* with 4 ancestral population.

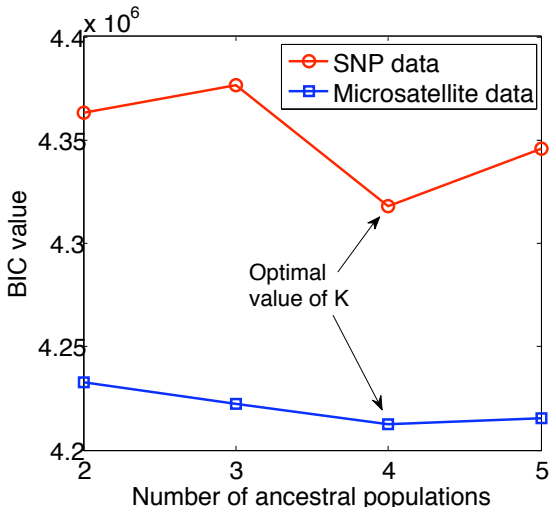


Figure 9: Model selection with BIC score for the HGDP data with *mStruct* on SNP and microsatellite data

References

[1] Rosenberg N, Pritchard J, Weber J, Cann H, Kidd K, et al. (2002) Genetic Structure of Human Populations. Science 298:2381–2385.

[2] Conrad D, Jakobsson M, Coop G, Wen X, Wall J, et al. (2006) A worldwide survey of haplotype variation and linkage disequilibrium in the human genome. NatureGenetics 38:1251–1260.

[3] Pritchard J, Stephens M, Donnelly P (2000) Inference of Population Structure Using Multilocus Genotype Data. Genetics 155:945–959.

[4] Erosheva E, Fienberg S, Lafferty J (2004) Mixed-membership models of scientific publications. Proceedings of the National Academy of Sciences 101:5220–5227.

[5] Falush D, Stephens M, Pritchard J (2003) Inference of Population Structure Using Multilocus Genotype Data Linked Loci and Correlated Allele Frequencies. Genetics 164:1567–1587.

[6] Huelsenbeck J, Andolfatto P (2007) Inference of Population Structure Under a Dirichlet Process Model. Genetics 175:1787–1802.

[7] Anderson E, Thompson E (2002) A Model-Based Method for Identifying Species Hybrids Using Multilocus Genetic Data. Genetics 160:1217–1229.

[8] Corander J, Waldmann P, Sillanpaa M (2003) Bayesian Analysis of Genetic Differentiation Between Populations. Genetics 163:367–374.

[9] Patterson N, Price A, Reich D (2006) Population structure and eigenanalysis. PLoS Genet 2:e190.

[10] Excoffier L, Hamilton G (2003) Comment on Genetic Structure of Human Populations. Science 300:1877–1877.

[11] Kelly R, Gibbs M, Collick A, Jeffreys A (1991) Spontaneous Mutation at the Hypervariable Mouse Minisatellite Locus Ms6-hm: Flanking DNA Sequence and Analysis of Germline and Early Somatic Mutation Events. Proceedings: Biological Sciences 245:235–245.

[12] Henderson S, Petes T (1992) Instability of simple sequence DNA in Saccharomyces cerevisiae. Molecular and Cellular Biology 12:2749–2757.

[13] Queller D, Strassmann J, Hughes C (1993) Microsatellites and kinship. Trends in Ecology & Evolution 8:285–288.

[14] Dietrich W, Katz H, Lincoln S, Shin H, Friedman J, et al. (1992) A Genetic Map of the Mouse Suitable for Typing Intraspecific Crosses. Genetics 131:423–447.

[15] Bowcock A, Ruiz-Linares A, Tomfohrde J, Minch E, Kidd J, et al. (1994) High resolution of human evolutionary trees with polymorphic microsatellites. Nature 368:455–457.

[16] Pisani D, Poling L, Lyons-Weiler M, Hedges S (2004) The colonization of land by animals: molecular phylogeny and divergence times among arthropods. BMC Biology .

[17] Zhivotovsky L, Underhill P, Cinnioglu C, Kayser M, Morar B, et al. (2004) The effective mutation rate at Y chromosome short tandem repeats, with application to human population-divergence time. American journal of human genetics 74:50–61.

[18] Valdes A, Slatkin M, Freimer N (1993) Allele Frequencies at Microsatellite Loci: The Stepwise Mutation Model Revisited. Genetics 133:737–749.

[19] Lin T, Myers E, Xing E (2006) Interpreting anonymous DNA samples from mass disasters–probabilistic forensic inference using genetic markers. Bioinformatics 22:e298.

[20] Jordan M, Ghahramani Z, Jaakkola T, Saul L (1999) An Introduction to Variational Methods for Graphical Models. Machine Learning 37:183–233.

[21] Xing E, Jordan M, Russell S (2003) A generalized mean field algorithm for variational inference in exponential families. Uncertainty in Artificial Intelligence (UAI2003) Morgan Kaufmann Publishers .

[22] Blei D, Ng A, Jordan M (2003) Latent Dirichlet allocation. Journal of Machine Learning Research 3:993–1022.

[23] Minka T (2000) Estimating a Dirichlet distribution .

[24] Cann H, de Toma C, Cazes L, Legrand M, Morel V, et al. (2002) A Human Genome Diversity Cell Line Panel. Science 296:261–262.

[25] Cavalli-Sforza L (2005) The Human Genome Diversity Project: past, present and future. Nat Rev Genet 6:333–40.

[26] Sohn K, Xing E (2007) Spectrum: joint bayesian inference of population structure and recombination events. Bioinformatics 23:i479–i489.