# *mStruct*: A New Admixture Model for Inference of Population Structure in Light of Both Genetic Admixing and Allele Mutations

**Suyash Shringarpure** SUYASH@CS.CMU.EDU
**Eric P. Xing** EPXING@CS.CMU.EDU
School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213 USA

## Abstract

Traditional methods for analyzing population structure, such as the *Structure* program, ignore the influence of mutational effects. 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 (Rosenberg et al., 2002) and the HGDP SNP data used in (Conrad et al., 2006). 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*, which is not possible by *Structure*.

## 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.

One of the state-of-the-art methods for population structure analysis based on multilocus genotype data is the program *Structure*, whose basic form is based on a statistical formalism known as the admixture model (Pritchard et al., 2000). Admixtures are instances of a more general class of hierarchical Bayesian models known as *mixed membership models* (Erosheva et al., 2004), which postulate that genetic markers of each individual are *iid* (Pritchard et al., 2000) or spatially coupled (Falush et al., 2003) samples from multiple population-specific fixed-dimensional multinomial distributions (which we will call *allele frequency profiles* (Falush et al., 2003), or AP) of marker alleles. Under this assumption, the *admixture* model identifies each ancestral population by a specific AP (that defines a unique allele frequency distribution for each ancestral population for each marker) and displays the fraction of contributions from each AP in a modern individual chromosome as a *structural map*. 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*, each individual is represented as a thin vertical line which shows the fraction of the individual's chromosome which originated from each ancestral population, as given by a unique AP. This method has been successfully applied to human genetic data in (Rosenberg et al., 2002) and has unraveled impressive patterns in the genetic structures of world population.

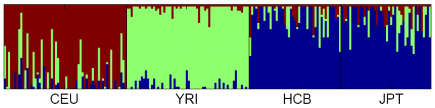


*Figure 1.* Population structural map inferred by *Structure* on HapMap data consisting of 4 populations. The colors represent different populations

However, 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 (Excoffier & Hamilton, 2003), is that there is no

Appearing in *Proceedings of the 25th International Conference on Machine Learning*, Helsinki, Finland, 2008. Copyright 2008 by the author(s)/owner(s).

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 frequency, rather than allowing the possibility of it just being a descendent of some common ancestral allele. 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 allele frequency 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 (Excoffier & Hamilton, 2003).

Another important issue in determining population structure is to look for the presence of admixture, a basic assumption of the *Structure* model. However, as we shall see later, on the HGDP data, it produces results that cluster individuals cleanly into one allele frequency profile or the other, thus leading us to conclude that there was little or no admixture between the human populations. While such a partitioning of individuals would be desirable for clustering them into groups, it does not offer us any biological insight into the intermixing of the populations.

In this paper, we present *mStruct* (for *structure under mutations*), based on an admixture of population-specific mixtures of inheritance model (AdMim). 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 proportion vector" (ancestry vector or *map vector*) that indicates membership proportions among the ancestral populations. By a simple but important extension to the LDA-like (Blei et al., 2003) admixture model used by *Structure*, *mStruct* facilitates estimation of both the *structural map* of populations (incorporating mutations) and the mutation rates of either SNP or microsatellite alleles. 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 (Rosenberg et al., 2002; Conrad et al., 2006). Our results show the presence of significant levels of admixture among the founding populations. We also report interesting genetic divergence in world populations revealed by the mutation patterns we estimated.

## 2. The Statistical Model

### 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. Below, we begin with a brief description of a commonly used method, followed by a new method that we propose.

#### 2.1.1. Population-Specific *Allele Frequency* Profiles

Due to the polymorphic nature of genetic markers, an intuitive statistic to characterize a population is the frequencies of all observed alleles at all 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* (Falush et al., 2003), 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 ancestry proportion profile*, is used by the program *Structure*.

#### 2.1.2. Population-Specific *Mixtures of Ancestral Alleles*

A problem with the population-specific AP profile representation is that it ignores the possibility of mutations underlying the alleles observed in modern populations with respect to their ancestral alleles. To capture this, we propose to represent 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 rate $\delta_i^k$ 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 $i$.

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. 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 $i \in \{1, \dots, I\}$ index the position of a locus in the study genome, $n \in \{1, \dots, N\}$ 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 $\mu^k_{i,l} \in \mu^k_i$ is not directly observable. For each allele founder $\mu^k_{i,l}$, we associate with it an inheritance model $p(\cdot|\mu^k_{i,l}, \delta^k_{i,l})$ from which descendants can be sampled. Then, given specifications of the ancestral population from which $X_{i,n_e}$ is derived (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: $p(x_{i,n_e} = l' \mid Z_{i,n_e} = k) = \sum_{l=1}^{L} \beta^k_{i,l} p(x_{i,n_e}|\mu^k_{i,l}, \delta^k_{i,l})$. Comparing to the counterpart of this function under AP: $p(x_{i,n_e} = l' \mid Z_{i,n_e} = k) = \beta^k_{i,l'}$, 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

The concept of admixture arises when 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., $x_{i,n_e}|z_{i,n_e} = k \sim \text{Multinomial}(\cdot|\vec{\beta}^k_i)$. Using this probability distribution in the general admixture scheme outlined above, we can see that *Structure* essentially implements an *admixture of population-specific allele rates* model. But a serious pitfall of using such a model, as pointed out in (Excoffier & Hamilton, 2003), 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 the mutation of some common ancestral allele at that marker.

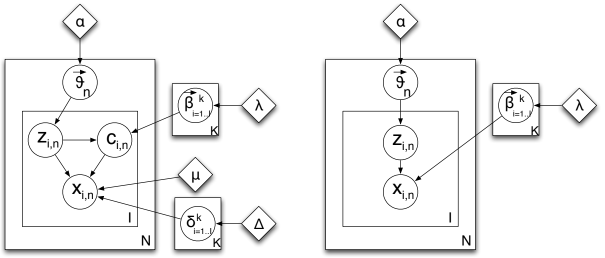
#### 2.2.2. The Proposed Model

We propose to represent each ancestral population by a set of population-specific MAAs. 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. 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 $c_{i,n_e}|z_{i,n_e} = k \sim \text{Multinomial}(\cdot|\vec{\beta}^k_i)$;
- 2.2b: draw the allele $x_{i,n_e}|c_{i,n_e} = l, z_{i,n_e} = k \sim P_m(\cdot|\mu^k_{i,l}, \delta^k_{i,l})$,

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 *Structure* is virtually identical to an LDA model, while *mStruct* is an extended LDA 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( $\mu^k_i = \mu_i$). Also our model assumes Hardy-Weinberg equilibrium within populations. The simplifying assumptions of



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

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

*unlinked 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 the repeats of a small sequences in DNA about 1-4 base pairs in length which are usually represented as integer counts. 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 (Valdes et al., 1993; Lin et al., 2006).

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

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

where $\xi$ is the mutation rate (probability of any mutation), and $\delta$ 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 $\delta$ 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 two-parameter stepwise continuous 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: $P(b|a) \propto \delta^{|b-a|}$. It is not hard to show that normalizing this probability mass function gives us the mutation model as:

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

We can interpret $\delta$ as a variance parameter, the factor by which probability drops as a fuction of the distance between the mutated version $b$ of the allele $a$.

**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 microsatellite mutation models, and use the Bayesian Information Criterion (BIC) to determine the cardinality of the mixture.

**Choice of mutation prior:** In our model, the $\delta$ 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 $\delta$ to improve the log-likelihood, in the absence of any prior distribution on $\delta$. To avoid this situation, and to allow more meaningful and realistic results to emerge from the inference process, we impose on $\delta$ a beta prior that will be biased towards smaller values of $\delta$. 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:

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

In this case, the mutation parameter $\delta$ 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 , 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 $c_{i,l} = \mathcal{I}[c_i = l]$.

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

$$p\left(\mathbf{x}, \mathbf{z}, \mathbf{c}, \vec{\theta} | \alpha, \boldsymbol{\beta}, \mu, \delta\right) = p\left(\vec{\theta}|\alpha\right) \prod_{i=1}^{I} p\left(z_i|\vec{\theta}\right) p\left(c_i|z_i, \vec{\beta}_i^{k=1 \cdot K}\right).$$

The marginal likelihood of the data can be computed by summing/integrating out the latent variables. 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 $\delta$ under AdMim is intractable.(Pritchard et al., 2000) developed an MCMC algorithms for approximate inference for their admixture model underlying *Structure*. We choose to apply a computationally more efficient approximate inference method known as variational inference (Jordan et al., 1999).

### 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 (Xing et al., 2003), we can write the variational distributions of the latent variables as follows:

$$q(\vec{\theta}) \propto \prod_{k=1}^{K} \theta_k^{\alpha_k - 1 + \sum_{i=1}^{I} \langle z_{i,k} \rangle}$$

$$q(c_i) \propto \prod_{l=1}^{L} \left( \prod_{k=1}^{K} \left( \beta_{i,l}^{k} f(x_i|\mu_{i,l}, \delta_i^k) \right)^{\langle z_{i,k} \rangle} \right)^{c_{i,l}}$$

$$q(z_i) \propto \prod_{k=1}^{K} \left( e^{\langle log(\theta_k) \rangle} \left( \prod_{l=1}^{L} \beta_{i,l}^{k} f(x_i|\mu_{i,l}, \delta_i^k)^{\langle c_{i,l} \rangle} \right) \right)^{z_{i,k}} .$$

In the distributions above, the '⟨⟩' are used to indicate the expected values of the enclosed random variables. A close inspection of the above formulae reveals that these variational distributions have the form $q(\vec{\theta}) \sim$ Dirichlet$(\gamma_1, \ldots, \gamma_K)$, $q(z_i) \sim$ Multinomial$(\rho_{i,1}, \ldots, \rho_{i,K})$, and $q(c_i) \sim$ 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:

$$\gamma_k = \alpha_k + \sum_{i=1}^{I} \langle z_{i,k} \rangle$$

$$\rho_{i,k} = \frac{e^{\langle log(\theta_k) \rangle} \left( \prod_{l=1}^{L} \beta_{i,l}^{k} f(x_i|\mu_{i,l}, \delta_i^k)^{\langle c_{i,l} \rangle} \right)}{\sum_{k=1}^{K} \left( e^{\langle log(\theta_k) \rangle} \left( \prod_{l=1}^{L} \beta_{i,l}^{k} f(x_i|\mu_{i,l}, \delta_i^k)^{\langle c_{i,l} \rangle} \right) \right)}$$

$$\xi_{i,k} = \frac{\prod_{k=1}^{K} \left( \beta_{i,l}^{k} f(x_i|\mu_{i,l}, \delta_i^k) \right)^{\langle z_{i,k} \rangle}}{\sum_{k=1}^{K} \left( \prod_{k=1}^{K} \left( \beta_{i,l}^{k} f(x_i|\mu_{i,l}, \delta_i^k) \right)^{\langle z_{i,k} \rangle} \right)}$$

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 (Xing et al., 2003).

## 3. Hyperparameter Estimation

The parameters of our model, i.e., $\{\mu, \delta, \boldsymbol{\beta}\}$, and the Dirichlet hyperparameter $\alpha$, 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. Due to space limits, details of this empirical Bayes estimation scheme are available in the extended version.

## 4. Experiments and Results

We validated our model on a synthetic microsatellite dataset where the simulated values of the hidden map vector $\theta_n$ 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 2.1).

### 4.1. Validations on Synthetic Data

We simulated 20 microsatellite genotype datasets using the AdMim generative process described in section 2.2, with 100 diploid individuals from 2 ances-

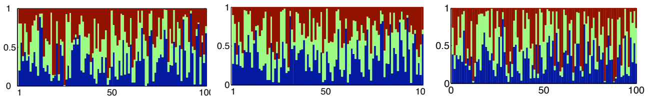


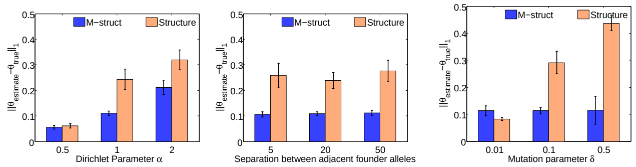
*Figure 3.* Ancestry spectra for a 3-population simulated dataset. First panel shows the true ancestry proportion vectors. Middle panel shows the estimate by *mStruct*. Right panel shows the estimate from *Structure* .

tral 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 $\theta_n$ were sampled from a symmetric beta distribution with parameter $\alpha$, 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. Due to space limitations, we only report two experiments below; the additional results on accuracy of ancestral alleles recovery and the ancestral-allele frequency estimation are available in the full paper.

### 4.1.1. Accuracy of Population Map Estimate

The map vector $\theta_n$ reflects the proportions of contributions from different ancestral population to the maker-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 $\theta_n$ 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 4 shows the accuracy of the Map estimate by *mStruct* on synthetic datasets simulated with different properties, in comparison with that of *Structure*. Fig. 4(a) shows that, under different degrees of biases of population admixing induced by the Beta prior of $\theta_n$, *mStruct* consistently outperforms *Structure*. Specifically, as the value of the Beta prior hyperparameter $\alpha$ 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 performs better than *Structure* as the separations between founding alleles decreases (Fig.4(b)), which tends to increasingly confound the ancestral origins of modern alleles. Finally, Fig. 4(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.



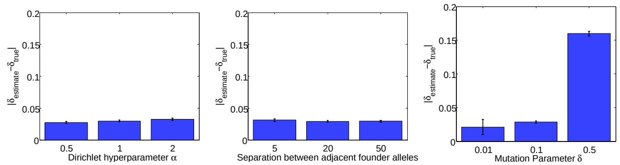
(a) (b) (c)

*Figure 4.* Accuracy of $\theta$ est. under different conditions.

### 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 only report highlights of mutation rate estimation.

**Mutation parameter estimation:** We evaluate the performance at recovery of $\delta^k$'s by a simple distance measure, (L1 distance measure), between the true and inferred values. We expect that using the beta prior described earlier improves the recovery of the population-specific mutation parameters. As shown in Figure 5, the estimates of $\delta^k$'s are robust and remain low-bias under different degree of admixing (due to changing $\alpha$) and different ancestor dispersion (due to changing distances among the $\mu^k$'s). The accuracy decreases as the value of the mutation parameter itself increases, but remains respectable, as shown in Figure. 5(c).



(a) (b) (c)

*Figure 5.* Accuracy of microsatellite mutation para. est.

## 4.2. Empirical Analysis of Real Datasets

The HGDP-CEPH cell line panel (Cann et al., 2002; Cavalli-Sforza, 2005) used in (Rosenberg et al., 2002)contains genotype information from 1056 indviduals from 52 populations at 377 autosomal microsatellite loci, along with geographical and population labels. The HGDP SNP data (Conrad et al., 2006) 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 (Rosenberg et al., 2002), the choice of the total number of ancestral populations is left to the user, and here we only show results of $K = 4$ due to space limitations.

#### 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 6). 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 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 actual evolutionary history of populations. In contrast, the structure map produced by *mStruct* for microsatellites suggests that all populations share a common ancestral population with a unique extra component 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 SNP mutation model.This issue will be explored in more detail in the full version of the paper.

#### 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 6 (lower subpanels 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 6 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 (e.g., 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 7. The BIC curves for both SNPs and microsatellites suggest K=4 as the best fit for the data.

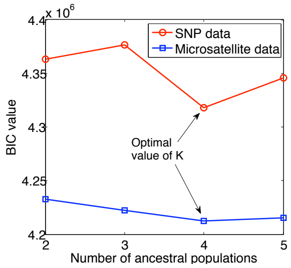
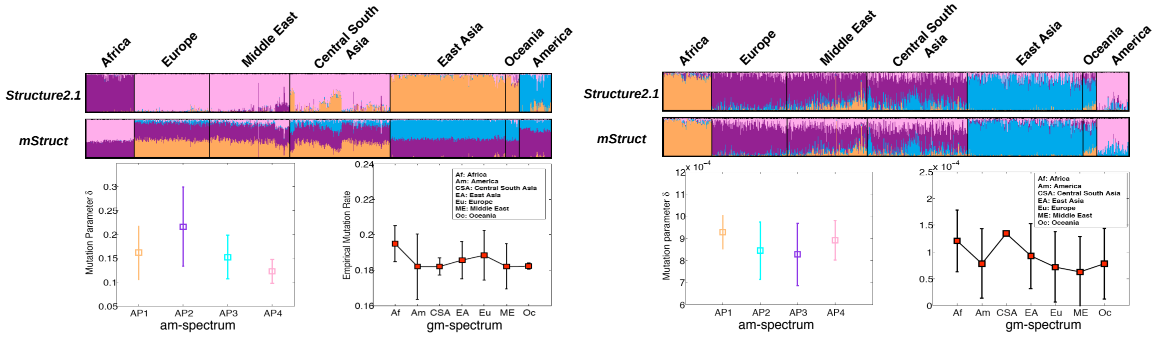


*Figure 7.* BIC scores for K=2 to 5.



(a) from Microsatellite data

(b) from SNP data

*Figure 6.* Structural maps, mutation spectrums, from the HGDP data via *mStruct* and *Structure*.

## 5. Discussions

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. 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 (Falush et al., 2003). 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 such as (Sohn & Xing, 2007). Our model might be described as a noisy-channel version of the LDA model, where the observations are modified instances of the original alleles. It is not hard to imagine applications of this model to other tasks such as image modeling or IR tasks involving noisy data, with minor changes in the distributions from which the observations are sampled.

## Acknowledgements

This research was supported in part by NSF Grant CCF-0523757 and DBI-0546594.

## References

Blei, D., Ng, A., & Jordan, M. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Cann, H., de Toma, C., Cazes, L., Legrand, M., Morel, V., Piouffre, L., Bodmer, J., Bodmer, W., Bonne-Tamir, B., Cambon-Thomsen, A., et al. (2002). A Human Genome Diversity Cell Line Panel. *Science*, *296*, 261–262.

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

Conrad, D., Jakobsson, M., Coop, G., Wen, X., Wall, J., Rosenberg, N., & Pritchard, J. (2006). A worldwide survey of haplotype variation and linkage disequilibrium in the human genome. *Nat Genet*, *38*, 1251–1260.

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

Excoffier, L., & Hamilton, G. (2003). Comment on Genetic Structure of Human Populations. *Science*, *300*, 1877–1877.

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.

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

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

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

Rosenberg, N., Pritchard, J., Weber, J., Cann, H., Kidd, K., Zhivotovsky, L., & Feldman, M. (2002). Genetic Structure of Human Populations. *Science*, *298*, 2381–2385.

Sohn, K., & Xing, E. (2007). Spectrum: joint bayesian inference of population structure and recombination events. *Bioinformatics*, *23*, i479.

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

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.*