Genetics: Published Articles Ahead of Print, published on May 29, 2012 as 10.1534/genetics.112.140228

Robust estimation of local genetic ancestry in admixed populations using a non-parametric Bayesian approach

Kyung-Ah Sohn*, Zoubin Ghahramani †, Eric P. Xing*

*School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213, USA

†Department of Engineering, University of Cambridge, Cambridge CB2 1PZ, UK

Running Head: Robust estimation of local genetic ancestry in admixed populations

Keywords: local ancestry, admixture, infinite Hidden Markov model, Dirichlet process, Hierarchical Dirichlet process

Corresponding Author:

Eric P. Xing

School of Computer Science

Carnegie Mellon University

5000 Forbes ave, Pittsburgh, PA 15213

412-268-2559 (ph.)

412-268-3431 (fax)

epxing@cs.cmu.edu

Abstract

We present a new haplotype-based approach for inferring local genetic ancestry of individuals in an admixed population. Most existing approaches for local ancestry estimation ignore the latent genetic relatedness between ancestral populations and treat them as independent. In this paper, we exploit such information by building an inheritance model that describes both the ancestral populations and the admixed population jointly in a unified framework. Based on an assumption that the common hypothetical founder haplotypes give rise to both the ancestral and admixed population haplotypes, we employ an infinite hidden Markov model to characterize each ancestral population and further extend it to generate the admixed population. Through an effective utilization of the population structural information under a principled nonparametric Bayesian framework, the resulting model is significantly less sensitive to the choice and the amount of training data for ancestral populations than state-of-the-arts algorithms. We also improve the robustness under deviation from common modeling assumptions by incorporating population-specific scale parameters that allow variable recombination rates in different populations. Our method is applicable to an admixed population from an arbitrary number of ancestral populations and also performs competitively in terms of spurious ancestry proportions under general multi-way admixture assumption. We validate the proposed method by simulation under various admixing scenarios and present empirical analysis results on worldwide distributed dataset from Human Genome Diversity Project.

INTRODUCTION

The problem of inferring genetic ancestries in a population has been widely investigated for various applications such as disease gene mapping and population history inference. For example, the inferred ancestry information has been used in correcting the confounding effect by population stratification in association studies (Wang *et al.* 2010; Price *et al.* 2006). The examination of loci that have elevated probabilities of a specific ancestry has also given critical clues in selecting out potential causal variants of certain diseases in admixture mapping (Zhu *et al.* 2011; Cheng *et al.* 2010; Cheng *et al.* 2009). Broadly, two different problem settings have been commonly considered for ancestral structure analysis (Alexander *et al.* 2009), one on the 'global ancestry' that considers the average proportion of each contributing population across the genome in an 'un-supervised' way (i.e., ancestral labeling of the study population is unknown) (Alexander *et al.* 2009; Patterson *et al.* 2006; Falush *et al.* 2003); and the other on the 'local ancestry' that is more concerned with a locus-by-locus ancestry given reference population data (Price *et al.* 2009; Pasaniuc *et al.* 2009; Tang *et al.* 2006). In this paper, we consider the problem of estimating the local ancestry in an admixed population. As a common scenario of this problem, consider the decomposition of chromosomes of modern African Americans into blocks that have either African or European ancestry given the reference population data close to ancient African and European populations, which we call *ancestral populations*. The populations of CEU and YRI are the most typical choices for such ancestral population data when an *admixed population* of African Americans is considered. We present a new haplotype-based method for local ancestry estimation that can deal with an arbitrary number of ancestral populations in a non-parametric Bayesian framework.

A natural approach to this problem involves a Hidden Markov Model (HMM) that traces the ancestry of each individual along the markers on a chromosome. Most previous approaches using HMM can be largely categorized into two families depending on how they encode the ancestral population. The first family of methods uses a population-specific allele

frequency profile to characterize each ancestral population. Such an allele frequency profile has been typically used as the latent component that generates population data in traditional admixture studies for global ancestry estimation (Falush *et al.* 2003; Pritchard *et al.* 2000; Huelsenbeck and Andolfatto 2007; Alexander *et al.* 2009). When adopted in the local ancestry estimation problem as in LAMP (Pasaniuc *et al.* 2009; Sankararaman *et al.* 2008), it has the general advantages of low computational cost and availability of such frequency profiles in representative datasets. However, the correlations between loci are reflected only by the variation in such allele frequencies and not by the actual recombination events at the chromosome level, so it is rather unnatural to model Linkage Disequilibrium (LD) structure between tightly linked SNPs. Therefore, either a subset of markers in low LD has to be selected in a preprocessing step, or a recombination process needs to be indirectly embedded to utilize a denser set of markers (Patterson *et al.* 2004; Tang *et al.* 2006; Pasaniuc *et al.* 2009). The representation power of this family of methods thus tends to diminish when the correlations between markers are not carefully considered (Price *et al.* 2008).

Another family of methods are based on haplotype data that may contain richer information. These methods utilize representative haplotypes taken from each ancestral population data as reference information for the local ancestry estimation (Sundquist *et al.* 2008; Price *et al.* 2009). Each haplotype in an ancestral population, which we call an *ancestral haplotype*, constitutes a hidden state in an HMM and the basic transition mechanism involves traversing among these ancestral haplotypes. Therefore, these approaches provide a more natural way to reflect the underlying admixing process by simulating recombinations at a real chromosome level. However, the inference result can be rather sensitive to the size and the choice of such ancestral haplotype data because the admixed haplotype is directly compared with the ancestral haplotypes. Moreover, few existing methods make use of the genetic relatedness between ancestral populations resultant from ancient population history and therefore the populations have been typically treated as independent. To improve the

robustness and the accuracy in light of these issues, HAPMIX (Price *et al.* 2009) introduces a 'miscopying' parameter that allows a small possibility for an allele to be copied from population 2 even when it is assumed to be originated from ancestral haplotype in population 1. In this way, it prevents unnecessary transitions among ancestral populations during inference and the allelic information in one population can be naturally borrowed by another population. However, this method is limited to two-way admixture that involves only two ancestral populations, and it is not trivial to generalize this model to consider more general demographic scenarios.

We propose a new Bayesian approach for local ancestry estimation that uses the multi-population haplotype data in a more systematic way. Our method is built on the assumption of a common pool of hypothetical *founder haplotypes* from which the ancestral haplotypes in multiple ancestral populations are to be inherited, and from which in turn the individuals in an admixed population are generated as well by the admixing process between ancestral populations. Motivated by the population model called SPECTRUM in (Sohn and Xing 2007), we model the ancestral population data by an infinite hidden Markov model in which the hidden states correspond to the unknown number of hypothetical founder haplotypes. The recombination and mutation events are then modeled with respect to these founders as transition and emission process. For an individual in an admixed population, we extend the hidden state space to a joint space of founder haplotypes and ancestral populations. That is, we incorporate a hidden state variable consisting of two indicator variables at each marker, one for selecting the ancestral population, and the other for selecting the hypothetical founder haplotype. The hidden state variable corresponding to the ancestral population determines the local admixing status and hence defines the local ancestry along the markers. Furthermore, population-specific scale parameters are incorporated to allow variable recombination probabilities in different populations. These scale parameters can be interpreted as being proportional to two major factors that affect the recombination probabilities in the corresponding populations: the effective population size, and the hypothetical time since

the hypothetical era of founder haplotypes. We observe that this parameterization also enhances the robustness of our model under scenarios that deviate from the common modeling assumption that all the populations participate in the admixture simultaneously.

A subtle issue in the proposed representation is how to choose the number of founders and how to construct them efficiently across multiple populations. Naïvely, we may assume K founders per ancestral population, but under this setting, not only one has to employ a non-trivial model selection process to determine K, but also there is in general no correspondence between the K founders in one population and another set of K founders in a different population. This problem would not only result in serious identifiability and multi-modality issue that can severely slow down inference, but also, it will restrict the information sharing across populations and hence compromise the accuracy of ancestry estimation as well. On the other hand, if we are to use one shared set of K founders, the representational power of population-specific HMM can also be limited. A non-parametric Bayesian framework using an infinite hidden Markov model gives a natural solution for this (Teh *et al.* 2010; Beal *et al.* 2002). Under an infinite HMM, an unbounded number of founder haplotypes can be systematically handled to describe a study population. If we employ multiple such infinite HMMs defined over the same set of founders, one infinite HMM per population, then it allows the founders to be shared between populations, while different populations do not have to include all these founders and can have a unique set of founders with its own frequency and recombination patterns among them. The number and the haplotypes of the founders are recovered as a result of posterior inference from data. Under a Dirichlet process prior, the posterior typically yields a parsimonious set of founders. This non-parametric Bayesian framework allows us to exploit the genetic relatedness between populations in a principled way by describing the ancestral populations in terms of a common set of founder haplotypes. In (Sohn and Xing 2009), a similar approach using a hierarchical Dirichlet process has been successfully used for the problem of haplotype inference from multi-population data. However, the recombination process was not explicitly modeled in that work and a rather

heuristic approach was employed to handle the linkage disequilbrium structure.

In our comparative study with two state-of-the-arts methods of LAMP (Pasaniuc *et al.* 2009; Sankararaman *et al.* 2008) and HAPMIX (Price *et al.* 2009), we show that the proposed method, which we call mSpectrum (admixture model based on multiple SPECTRUM representations), enjoys enhanced robustness and accuracy, evidenced by its substantially less sensitivity to the choice and the amount of ancestral population data. In particular, our method shows very competitive performance even when the sample size of the ancestral population data is very small. This highlights the potential usefulness of this method in the analysis involving underrepresented populations of limited data availability. In addition, the compact population characterization by an infinite hidden Markov model improves the model flexility over existing haplotype-based approaches so that it can naturally handle an arbitrary number of ancestral populations instead of only two in HAPMIX. It is also robust even under deviation from the common modeling assumption that multiple populations participate in the admixture at the same time as in (Pasaniuc *et al.* 2009). The performance of our model is superior in terms of the proportion of spuriously estimated ancestries under general multi-way admixture assumption as well.

In the remainder of this paper, we first describe the statistical model and the inference method. Then we validate the proposed method through simulation study and show empirical analysis result using Human Genome Diversity Panel data (Jakobsson *et al.* 2008). A discussion follows and concludes the paper.

METHODS

Problem setting We consider an admixed population in which J ancestral populations have mixed since G generations ago. For example, if we are to recover the local ancestry of individuals in a Latino population (*admixed population*), we can incorporate $J = 3$ populations of ancient African, European, and Native American as our *ancestral populations*. In our problem setting, we assume that the haplotypes composed of single nucleotide polymor-

phisms are given for the ancestral populations and the admixed population. We will recover the pool of hypothetical founder haplotypes and their associations to individuals by statistical inference. The association of admixed individuals to the ancestral populations will be recovered along with their association to the founders, which would lead to the estimation of local ancestry.

Overview of admixture model based on founder haplotypes The choice of representation about how to characterize a population is the crucial starting point in admixture modeling. Unlike most previous approaches that use allele frequency profiles (Pasaniuc *et al.* 2009; Sankararaman *et al.* 2008) or representative ancestral haplotypes in their raw forms (Sundquist *et al.* 2008; Price *et al.* 2009), we employ a new haplotype-based method that builds on an assumption of hypothetical founder haplotypes of unknown cardinality. The founder-based population model with explicit recombination modeling has been introduced in (Sohn and Xing 2007) with the application to population structure and recombination analysis. Under this approach, each individual in a population is generated from the hypothetical pool of founders via a series of recombination and mutation. An individual haplotype can then be viewed as a mosaic of the founders whose pattern is determined by the association with founders. This mosaic process could be modeled as a Hidden Markov model in which the founders correspond to the hidden states, the individual haplotypes correspond to the observation sequences, the transition process is modeled by the recombination process, and the emission process by the mutation from founders to the individuals. By employing an infinite hidden Markov model, the number and the haplotypes of the founders can be recovered through posterior inference rather than being pre-specified, and the local inheritance association between the founders and the study individuals can also be derived.

Now we further extend this population model to describe admixture events from an arbitrary number of ancestral populations. When the ancestral populations start to mix and

form an admixed population, each individual haplotype in the admixed population can be decomposed into blocks with distinct ancestry. For each of these blocks, we can trace back the source of the genetic materials to a haplotype in the corresponding ancestral population. Now, recall that this 'ancestral haplotype' is modeled as a mosaic of its founders. This means that each ancestry block in an admixed individual is further dissected into a finer-grained mosaic of founders. Therefore, the admixed inheritance process is a composite process with two different resolutions, one from the founders to ancestral haplotypes, and the other from the ancestral haplotypes to the admixed individuals. A graphical illustration of the proposed model is shown in Figure 1. A variant of the infinite hidden Markov model is employed to make the choice of founders and the ancestral populations at the same time along the chromosome.

[Figure 1 about here.]

Statistical model for generating ancestral and admixed population data We now describe in detail the admixed inheritance model as a generative process of the individual haplotypes in ancestral populations and an admixed population with respect to a set of hypothetical founders.

Transition and emission probabilities For ease of description, we assume that the individuals are haploids. Let individual haplotypes in an admixed population be indexed by i, ancestral populations by j, and the markers by t. And let $H_{it} \in \{0, 1\}$ and $F_{kt} \in \{0, 1\}$ represent the allele of individual i and founder k at marker t, respectively. We introduce a set of hidden state variables $S_{it} = (C_{it}, Z_{it})$ where $C_{it} \in \{1, 2, ...\}$ and $Z_{it} \in \{1, ..., J\}$ represent the indicator variables that select a founder haplotype and an ancestral population, respectively, on an i-th admixed haplotype at marker t. For each ancestral population j, let ν_{jk} be the initial and background probability of founder k, and let $\pi^j_{k'k}$ be the transition probability that determines the probability of switching from copying founder k' to founder k. We also

introduce a set of scale parameters $T_j \in (0, \infty)$ that scale the recombination rate in each population j by T_j. The role of these parameters is to take into account the difference in the hypothetical time since the founder population and also the effective population sizes of different ancestral populations. Let $\eta = (\eta_1, ..., \eta_J)$ denote the global admixing proportion such that η_j is the expected proportion of ancestral population j in the admixed population, let $G \in [0, \infty)$ represent the time since admixture in the admixed population, $\mathbf{r} = (r_1, r_2, ...r_T)$ and $\mathbf{d} = (d_1, \ldots, d_T)$ represent the recombination rate and the physical distance between each neighboring markers, respectively. The final transition probabilities and the emission probabilities are defined as follows:

$$
\begin{aligned}
P(S_{i,0} = (k, j)) &= P(Z_{i,0} = j)P(C_{i,0} = k) = \nu_{jk}\eta_j \\
P(S_{it} = (k, j) \mid S_{i,t-1} = (k', j')) &= (1 - e^{-r_t d_t G})\nu_{jk}\eta_j + \\
&\quad e^{-r_t d_t G} e^{-r_t d_t T_j} I(k = k')I(j = j') + \\
&\quad e^{-r_t d_t G}(1 - e^{-r_t d_t T_j})\pi^j_{k'k} I(j = j') \qquad (1) \\
P(H_{it} \mid S_{it} = (k, j), F_{kt}) &= \delta_k^{I(H_{it} \neq F_{kt})}(1 - \delta_k)^{I(H_{it} = F_{kt})} \qquad (2)
\end{aligned}
$$

where $I(\cdot)$ represents an indicator function such that $I(i = j) = 1$ if $i = j$, and 0 otherwise. We assume a founder-specific mutation parameter δ_k that determines the probability of mutation during the inheritance from a founder k to individuals.

The overall idea underlying this representation is the two-layered inheritance framework, one from the time of hypothetical founders to ancestral populations, and the other from those ancestral populations to the admixed population. If we set $G = 0$ in Equation (1), this two-layered framework is reduced to the model of the first layer that characterizes the ancestral populations with respect to the founder haplotypes. Under the reduced model, each population is associated with its own hidden Markov model parameters and the recombination rate scaled by T_j. Suppose $(C_{i,t-1}, Z_{i,t-1}) = (k', j')$ which means i-th haplotype has inherited from founder k' at marker $t - 1$ in ancestral population j'. At the next marker

t, it either selects a new founder k with probability $(1 - e^{-T_j r_t d_t})\pi^j_{k',k}$ and set $C_{it} = k$, or no recombination takes place with the remaining probability and $C_{it} = C_{i,t-1}$. If we trace the values of C_{it} across all the t, it will decompose the haplotype i into blocks with distinct associated founders. Therefore, each chromosome can be thought of as a mosaic of such founders.

Now, at the second layer which involves the admixture, this sequential process for selecting founders C_{it} occurs within the same ancestral population with probability $e^{-r_t d_t G}$ so that $Z_{it} = Z_{i,t-1}$. Or with probability $(1 - e^{-r_t d_t G})$, a new population j as well as a new founder k is chosen jointly with a probability proportional to the product of admixing proportion η_j and the background probability ν_{jk}. Therefore, haplotypes both in the ancestral populations and in the admixed population are modeled as mosaics of founders determined by the sequence of C_{it}. In addition, each admixed individual i is associated with another resolution of mosaic determined by the sequence of Z_{it} across t. The estimation of local ancestry can be done by tracing the posterior probability of Z_{it} along the markers.

Note that even when no admixing is assumed, we still have the flexibility of choosing a different founder haplotype. This feature helps to control the number of transitions among populations effectively so that the hidden state doesn't need to change excessively. Moreover, although we assume the J populations participate in the admixture simultaneously, the population-specific scale parameters would explicitly allow heterogenous resolution of the genetic mosaics in different ancestral populations to be generated. This greatly improves robustness of the model against the violations of such modeling assumption as well as the accuracy of the ancestry estimation.

The cardinality of the founder space Instead of fixing the number of hypothetical founders by doing statistical model selection, we adapt a more flexible non-parametric approach using an infinite hidden Markov model (iHMM) (Beal *et al.* 2002; Teh *et al.* 2010). Recall that if we consider a finite, say K, hidden states, the transition probabilities will be represented

as a $K \times K$ matrix. Each row k of this matrix sums to one and defines the probabilities of switching from a source state k to all the target states.

Now, if we consider an infinite hidden state space, each row of the transition matrix would be an infinite dimensional vector which sums to one. Dirichlet Process (DP) (Blackwell and MacQueen 1973; Ferguson 1973) has been effectively used to describe such probability distribution. A DP is defined by two parameters: the base measure ('mean' of the DP) and the scale parameter that controls the concentration around the mean. To ensure all the row-specific DPs built on the same state space, another Dirichlet Process is shared as a common base measure at a top level. This model for the hidden Markov transition probabilities actually corresponds to a hierarchical Dirichlet Process (Teh *et al.* 2010). We omit the statistical details of an infinite hidden Markov model formulation in terms of a hierarchical Dirichlet process here (see (Teh *et al.* 2010; Beal *et al.* 2002; Sohn and Xing 2007) for more details). Basically, (k, k')-element of the transition matrix π^j defines the transition probability from state k to state k' in population j, and for a given source state k, the target state index k' can increase as large as needed by the given data. Infinite-dimensional vector of initial probabilities ν_j can be defined in a similar way under the same hierarchical Dirichlet process framework. Since we consider multiple such infinite HMMs for multiple populations, we let the same base measure shared across all the populations. This infinite HMM-based framework leads to a very simple solution to how many founders to consider and how to construct the founder space across multiple populations. The iHMM parameters of our admixture model thus can be summarized as follows:

$$\nu_j \;\sim\; DP(\alpha_0, \beta), \pi_k^j \sim DP(\alpha_0, \beta), \quad \beta \sim GEM(\gamma)$$

where α_0 and γ define the scale parameters for the population-specific DPs and the top level DP, respectively.

Other parameter description We assume Dirichlet distribution prior for the population proportion parameter $\eta \sim Dirichlet(\xi_1, ..., \xi_J)$ and Beta prior for each of the mutation parameters δ_k .

For simplicity of inference, we transform the variables such that r_t and T_j are combined as $g^r_{jt} = r_t T_j$. Similarly, we use the notation $G^r_t := r_t G$. We assume these variables are $i.i.d$ under Gamma prior. Then Equation (1) is transformed as follows:

$$\begin{aligned} P(S_{it} = (k, j) \mid S_{i,t-1} = (k', j')) &= e^{-G^r_t d_t} e^{-g^r_{jt} d_t} I(k = k') I(j = j') + \\ & \quad e^{-G^r_t d_t} (1 - e^{-g^r_{jt} d_t}) I(j = j') \pi^j_{k'k} + \\ & \quad (1 - e^{-G^r_t d_t}) \nu_{jk} \eta_{ij} \end{aligned} \tag{3}$$

In summary, infinite hidden Markov model parameters combined with population genetics parameters are used to capture different characteristics in populations and to describe admixture event from an arbitrary number of ancestral populations. While we assume an infinite number of founders a priori, the posterior inference usually produces a small number of founders and this leads to a compact representation of a population for the admixture analysis.

Posterior Inference To overcome the drawbacks of slow convergence in traditional Gibbs sampling, we employ a variant of beam sampling proposed for infinite HMM (Van Gael *et al.* 2008). Basically, it extends the well-known dynamic programming technique of forward-backward algorithm in a finite state HMM to an infinite state space case. It exploits the property that in an observation sequence of finite length, the number of actually realized hidden states is finite at each iteration step. Therefore, the number of states to be considered in forward-backward algorithm can be adaptively changed over iterations. More specifically, a set of auxiliary variables u are sampled conditional on S such that given $u_1, ..., u_T$, the number of states K having positive forward probabilities is finite. More details of the beam

sampling scheme for the proposed model are described in Supplementary Material .

Since the entire inheritance process from founders to ancestral populations and then the admixed population is modeled in a single Bayesian framework, it allows the exact posterior inference by putting the ancestral and admixed population data together in a single series of beam sampling iterations (see A1 in Supplementary Material). However, this is not optimal in terms of time complexity as we often favor to run multiple test sets after we extract reference information about the ancestral populations. Therefore, we split the whole inference process into two phases: 1) training phase where the model parameters about ancestral populations are learned, and 2) ancestry estimation phase that actually recovers the ancestry of admixed individuals.

One caveat of this decomposition is that we may not fully take advantage of the flexibility of the infinite model. This is because we need to constrain the hidden state space somehow as a finite space when the output from the training phase is returned. As an n-th posterior sample from Bayesian inference of the training phase, we get a finite number $K^{(n)}$ of founder haplotypes and the related HMM parameters of $\pi^{(n)}$ and $\nu^{(n)}$ with $g_j^{r(n)}$ for each j. Averaging these results as one training output is not straightforward as $K^{(n)}$ can be different across different n. A plausible approach would be to keep multiple, say N posterior samples $\mathbb{S} = \{\mathbf{F^{(n)}}, \pi^{(\mathbf{n})}, \nu^{(\mathbf{n})}\}_{n=1,\ldots,N}$ and run the ancestry estimation routine N times using each of these parameters in $\mathbb{S}$. Then the N posterior distributions of the ancestry indicator variable Z can be easily averaged to form the final posterior distribution since Z is defined over a fixed number of populations J unlike C or other parameters that depend on K. Note that $g_j^{r(n)}$ does not depend on K, so we can use the posterior mean of $g_j^{r(n)}$ as the final estimate for it. Another practical approach would be to select a single output from the training phase such as a MAP solution, and estimate the local ancestry based on the single set of parameters. Empirically, we observe that the performance degradation by this MAP solution with respect to the first approach is relatively small.

Training phase For an individual in an ancestral population j, we can set the time since admixture G to be zero and the population indicator variables Z to be observed as constant. Then the hidden state variable $S_{it} = (C_{it}, Z_{it})$ can be replaced with a C_{it} indicating the founder and Equation (3) is reduced to the followings :

$$\begin{aligned} P(C_{i0} = k) &= \nu_{Z_{i0}k} \\ P(C_{it} = k \mid C_{i,t-1} = k') &= e^{-g^r_{Z_{i0}t} d_t} I(k = k') + (1 - e^{-g^r_{Z_{i0}t} d_t}) \pi^{Z_{i0}}_{k'k} \end{aligned}$$

We infer the variable C through the Beam sampling algorithm described in A2 in Supplementary Material, and the other variables through the standard Gibbs sampling.

Note that the contribution of transition at each neighboring loci $t-1$ and t to the parameter π and g^r_{jt} is not all equal because of the self-transition probability forced by the recombination model in Equation (3). We handle this by sampling auxiliary binary variables $M_{it} \sim Bernoulli(1 - e^{-g^r_{Z_{i0}t} d_t})$ to indicate whether the jump occurs in the transition or not. The transition probability can be decomposed as follows:

$$P(C_{it} \mid C_{i,t-1}) = P(M_{it} = 0)\delta(C_{it} = C_{i,t-1}) + P(M_{it} = 1)\pi^j_{C_{i,t-1}, C_{it}}$$

Then we sample M_{it} given C_{it} and $C_{i,t-1}$ backward in forward-backward process from

$$P(M_{it} | C_{it} = (k, j), C_{i,t-1}) \propto P(M_{it}) P(C_{it} = k \mid C_{i,t-1} = k', M_{it})$$

Now, π can be sampled as in (Van Gael *et al.* 2008), but conditional on M, which involves the transitions with $M_{it} = 1$ only. g^r_{jt} can also be sampled conditional on M using $P(g^r_{jt} \mid \{C_{:t}, C_{:,t-1}, M_{:t}\}) \propto P(g^r_{jt}) \prod_{i \in Pop\ j} P(C_{i,t} \mid C_{i,t-1}, M_{it})$. The overall sampling procedure is summarized in Algorithm 1.

Algorithm 1 Procedure for training iHMMs in reference populations

Input: Haplotype data H for ancestral populations

Output: N posterior samples of founders and the related HMM parameters $\{\mathbf{F^{(n)}}, \pi^{(\mathbf{n})}, \nu^{(\mathbf{n})}, \mathbf{g^{r(n)}}\}$ for $n = 1, \ldots, N$

1: **repeat**
2: **for** each individual chromosome i **do**
3: Sample the auxiliary variables u_{it} for $t = 0, ..., T-1$.
4: Sample $C_{it} \mid u, H, F$ using the beam sampling algorithm
5: Sample $F_{k,t}$ and δ_k
6: Sample parameters ν, π, β and g^r.
7: **end for**
8: **until** convergence

Ancestry estimation phase As the variables F, g^r, ν, π are returned in the training stage, the unknown variables now are the global admixing proportion η, the generations since admixture G, the mutation rate δ_k of founders, and $S = (C, Z)$ for the admixed individuals. We re-sample δ_k in the ancestry estimation phase instead of getting it from the training step because δ_k can reflect additional information about the admixed population by describing it in terms of the discrepancy between founders and the population. As we now deal with a finite number of hidden states obtained from the training phase, it is not necessary to incorporate the auxiliary variable u to sample S in the ancestry estimation phase. The variables S_{it} thus are sampled through a standard forward-backward algorithm. As in the training stage, the transition probability at each marker can be decomposed into two parts, depending on whether the jump process for admixture occurs or not. We use the similar technique to sample G^r by introducing an auxiliary variable $L_{it} \sim Bernoulli(1 - e^{-G_t^r d_t})$. The overall sampling scheme is summarized in Algorithm 2.

If the time since admixture G, admixing proportion η, and the recombination rate r is assumed to be known as is often the case in admixture analysis, we can omit the second step of parameter sampling (line 5 in Algorithm 2) and re-use δ_k that can be returned from the training stage. Then it is also possible to get an approximate solution by use of a posterior decoding from forward-backward steps in a finite dimensional HMM.

Algorithm 2 Procedure for estimating local ancestry in an admixed individual

Input: Haplotype data H for an admixed population, estimated parameters $\{\mathbf{F^{(n)}}, \pi^{(\mathbf{n})}, \nu^{(\mathbf{n})}, \mathbf{g^{r(n)}}\}$

Output: Posterior distribution of $Z = (Z_{it})$.

for $n = 1, \ldots, N$ **do**
  repeat
    for each individual chromosome i **do**
      Sample $S_{it} = (C_{it}, Z_{it}) \mid H, F$ using the forward-backward algorithm
      Sample δ_k, η, and G^r .
    end for
  until convergence
  Keep S posterior samples of Z
end for
Average $N \cdot S$ posterior samples and return the final posterior distribution of Z

RESULT

Simulation design To validate the proposed method, we simulated admixed haplotypes using the Human Genome Diversity Project (HGDP) data genotyped on Illumina Infinium HumanHap550 BeadChips (Jakobsson *et al.* 2008). Considering previous results that have revealed distinct genetic characteristics across different continents, we selected the following reference populations that would serve as putative ancestral populations: YRI for African, CEU for European, JPT and CHB for East Asia, and Maya for Native American ancestry. Each of the resulting ancestral populations contained 30, 30, 28, and 13 individuals, respectively. In the simulation study, we first focus on chromosome 22 for computational efficiency under diverse types of simulation scenarios.

To take into account the discrepancy between real ancestral populations and those used in training, we generated admixed individuals using populations which are similar but not identical to those used as ancestral populations. For example, individuals in Russian and BantuKenya populations are mixed to simulate an admixed population and then the local ancestries of these individuals are estimated with respect to CEU and YRI populations. A simulation scheme similar to that in (Price *et al.* 2009) was used to generate admixed haplotypes as follows. For each haplotype in an admixed population, we first sample the

ancestry $j \in \{1, ..., J\}$ at the first marker according to the probabilities $\eta = (\eta_1, \ldots, \eta_J)$ and randomly select an ancestral haplotype in the corresponding population j to copy the allele at the first marker. For the following markers, we either assign the same ancestry as the previous marker with probability $\exp(-r_t d_t G)$ and copy the allele of the same ancestral haplotype at the corresponding marker, or with probability $1 - \exp(-r_t d_t G)$, we re-sample the ancestry j' among the J possible populations based on the probabilities η and randomly re-select the ancestral haplotype for allele copy within the selected population j'. We use a constant recombination rate of $r_t = 10^{-8}$ per base pair per generation as in previous studies (Sankararaman *et al.* 2008). Note that our simulation data are not generated under our modeling assumption basing on founder haplotypes, but in more general setting that is commonly considered in previous admixture studies. For each simulation scenario below, we generate 30 admixed individuals per dataset.

The performance is measured as the mean squared error rate of ancestry probabilities along the loci. Specifically, let p_{ijt} denote the probability of ancestry j at a locus t in an individual i. The average error rate of $\sum_{j=1}^{J} \sum_{t=1}^{T} (p_{ijt}^{true} - p_{ijt}^{est})^2 / T$ across all the individuals is reported. We compare our results with the two state-of-the-art methods: LAMP (Pasaniuc *et al.* 2009; Sankararaman *et al.* 2008), the method based on allele frequency profiles as reference information, and HAPMIX (Price *et al.* 2009) that uses representative ancestral haplotypes. These methods appear to outperform other methods such as HAPPA (Sundquist *et al.* 2008), SABER (Tang *et al.* 2006) or ANCESTRYMAP (Patterson *et al.* 2004) in previous studies (Price *et al.* 2009). Since the benchmark algorithms require the parameters for recombination r, the admixture time G, and the population proportion η to be specified as input, we provided the true values of these parameters to all the algorithms in the simulation study. Additionally, each haplotype data for ancestral populations were converted to allele frequency profiles and then LAMP was run with these frequency data as input. For the analysis below, we used the MAP solution as our parameter estimation from the training phase.

Performance on two-way admixture The first simulation scenario considers two-way admixture of ancient European and African populations. We generate admixed individuals using BantuKenya and Russian population data with the admixing proportion of $\eta = (0.5, 0.5)$ and then the local ancestries of the admixed individuals are estimated with respect to YRI and CEU. In Figure 7, we first display the true and the estimated local ancestry probabilities of two sample individuals in an admixed population. The yellow color corresponds to YRI (African) ancestry, and the dark green corresponds to CEU (European) ancestry. The length of the vertical color bar at each chromosomal location along the x-axis is proportional to the corresponding ancestry probability. While all the algorithms produce reasonable results in general, the proposed method denoted by mSpectrum is especially effective in picking out fine details of ancestry changes as can be seen in the example.

[Figure 2 about here.]

The overall performance of each algorithm across all the generated samples are shown in Figure 3. Roughly, we can see that mSpectrum and HAPMIX perform comparably to each other and tend to outperform LAMP in case of two-way admixture. Still, all the three algorithms perform reasonably well as can be seen in the small overall error rates. For example, the average error rates for $G = 10$ were 0.0077, 0.0086, and 0.0116 in mSpectrum, HAPMIX, and LAMP, respectively.

[Figure 3 about here.]

Performance as a function of data size in training set To further evaluate each method in terms of its performance with respect to the training data size, we varied the number of available individual samples per ancestral population. We trained the model using 3, 5, 10, 20, 30 individuals, hence, 6, 10, 20 40, 60 haplotypes, per ancestral population and estimated the ancestries based on each of the trained model. The performance of each algorithm is presented as a function of training data size in Figure 4 for two scenarios:

(a) the two-way admixture scenario from BantuKenya and Russian populations of which the result on the full dataset is shown in Figure 3, and (b) the admixture of YRI and CEU populations where the individuals not contained in the training data are used to generate the admixed individuals. It is clearly seen that the proposed method substantially outperforms the other benchmark algorithms in both cases, especially when the data size is small. Even when only a few ancestral haplotypes are available, it still gives very good estimates of the local ancestries compared to the others. Therefore, our method can be especially useful in the analysis of admixture effect involving non-traditional populations where the amount of available genotypes is still limited. In addition, our method shows greater performance gain over the other two methods when the discrepancy between the training population and the one used in the simulation is large. This implies that the hierarchical structure put on top of the ancestral population data allows more general description of the ancestral populations and hence enhances the accuracy of the ancestry estimation even when the ancestral population used for reference have diverged from the true ancestral populations.

[Figure 4 about here.]

Performance on three-way admixture We now consider the admixture involving more than two ancestral populations. Analogous to the formation of Puerto Rican population (Tang *et al.* 2007), we included CEU, YRI, and Maya as ancestral populations for African, European, and Native American ancestry, and generated an admixed population using Russian, BantuKenya, and Pima with admixing proportion of 0.66, 0.18, and 0.16, respectively.

Figure 5 shows the resulting error rates across different values of G. Since HAPMIX cannot handle more than two ancestral populations directly, we ran it in three different modes such that each run tries to estimate the targeted ancestry versus the other two ancestries as was done in its original paper (Price *et al.* 2009). For this reason, we compare the performance on each ancestry separately. Overall, our method performs significantly better than the other two in most of the analyzed cases.

[Figure 5 about here.]

Robustness under deviation from admixture assumption The modeling assumption that all the ancestral populations participate in the admixing simultaneously does not hold in reality, especially in case of multi-way admixture involving multiple ancestral populations. We test the robustness of each method under deviation from such a modeling assumption by generating admixed haplotypes from three ancestral populations that started to mix at two different time points. More specifically, Russian and BantuKenya populations are mixed for G_1 generations with 50%/50% proportion. Then this admixed population is mixed with the third population of Pima for G_2 generations with 50%/50%, resulting in the overall proportion of 0.5, 0.25, 0.25. We fixed G_2 to be 10 and varied G_1 to be 0, 2, 5, and 10 where $G_1 = 0$ corresponds to the case in which the modeling assumption holds. The result is summarized in Figure 6. In each plot, x-axis corresponds to the values of G_1/G_2 and y-axis shows the error rates. The proposed method resulted in not only the lowest error rates, but also the most stable performance across different values of G_1/G_2. For more quantitative comparison of robustness across different algorithms, we calculated the linear regression coefficient of G_1/G_2 versus the error rates. The resulting slopes were -0.0011, 0.0029, and 0.0074 for mSpectrum, HAPMIX, and LAMP, which again supports the superior robustness of the proposed method.

[Figure 6 about here.]

Performance under four-way admixture assumption When it is unclear how many or which ancestral populations have contributed to the given admixed population due to unknown population history, one needs to run the local ancestry estimation under general assumption of multi-way admixture involving all the candidate ancestral populations. In this case, the proportion of spurious association to non-contributing population is also an important measure for performance comparison in addition to the mean squared error rates

for local ancestry estimation. Or when the contribution of certain population is extremely small, we can test how sensitively each algorithm detects such small portion of ancestries. To examine the behavior of each algorithm under such cases, we let each algorithm assume four ancestral populations of CEU, YRI, Maya and JPTCHB and then estimate the ancestry of admixed haplotypes generated from Russian, BantuKenya, Pima, and Yi populations with admixing proportions of $\eta^{(1)} = (0.2, 0.8, 0, 0)$, and $\eta^{(2)} = (0.8, 0.15, 0.03, 0.02)$.

[Figure 7 about here.]

We first illustrate the true and the estimated local ancestry probabilities of two sample individuals in each of the admixed populations generated using $\eta^{(1)}$ and $\eta^{(2)}$ in Figure 7 (a) and (b). The red color corresponds to YRI ancestry, the black corresponds to CEU, the yellow and the white correspond to Maya and JPT+CHB ancestries, respectively. We find that mSpectrum shows the most accurate and stable result with the least amount of spurious association in both cases.

The global admixing proportion $\hat{\eta}$ computed as the average local ancestry proportion across all the markers and all the individuals in the admixed population is summarized in Figure 8 (a). For the first scenario using $\eta^{(1)}$ that involves only European and African ancestries, the mean proportion of spuriously estimated ancestries is 0.016, 0.016, and 0.051 for mSpectrum, HAPMIX, and LAMP, respectively*. In case of the second scenario using $\eta^{(2)}$ where the true combined proportion of Maya and JPT+CHB populations is 0.036, the estimated proportion of these ancestries in each algorithm is 0.03, 0.13, and 0.23, for mSpectrum, HAPMIX, and LAMP. This result shows that our method is indeed effective in preventing excessive transitions between ancestral populations and hence reducing the

*For HAPMIX, since each ancestry proportion is estimated under two-way admixture assumption of one ancestry versus all the others, the ancestry proportions across all the populations do not necessarily sum to one. While the pie charts and the illustration in Figure 7 show the normalized results, we report the numbers before normalization on the of the pie charts because we find this estimation is more accurate than that after normalization.

proportion of spurious estimations. Figure 8 (b) shows that mSpectrum significantly outperforms HAPMIX or LAMP in terms of the mean squared error rates for the local ancestry estimation as well.

For more detailed comparison of the proportion of spurious ancestries in different methods, in Figure 9, we show the overall distribution of the spuriously estimated ancestry measured over 50 datasets simulated by $\eta^{(1)}$. We find that mSpectrum and HAPMIX estimate similar proportions of spurious JPT+CHB ancestry which is substantially less than that from LAMP. On the other hand, mSpectrum is the most accurate in preventing spurious Maya ancestry.

[Figure 8 about here.]

[Figure 9 about here.]

Sensitivity analysis on model parameters Since the parameters of η and G were assumed to be known in our simulation study in parallel with other methods, we also examine how the performance of mSpectrum is affected by incorrectly specifying these parameters. The performance is shown for the dataset simulated with $G = 10$ and $\eta = (0.5, 0.5)$ with respect to YRI and CEU ancestries in Figure 10. In each plot, x-axis shows the specified parameters where the values are shown in log scale in case of G. We could see that there was almost no effect when η was incorrectly set in the range from 0.2 to 0.8. When we examined the result on G, the algorithm had the general tendency to favor a specified value G smaller than the true value. The effect of mis-specified value of G was minimal when the discrepancy was within a factor of 2. Even in the extreme case such as G varied by a factor of 5, the error still remained within the twice of the error rates when the true value was given.

[Figure 10 about here.]

Empirical analysis of HGDP data To illustrate our method on real data, we applied it to 22 autosomes of the HGDP dataset (Jakobsson *et al.* 2008). Four ancestral populations of YRI, CEU, JPT+CHB, and Maya were chosen as in the simulation study to represent African, European, East Asian, and Native American ancestries. We then recovered the local ancestries in the remaining 28 populations. Since the time since admixture is not available for real data, we let our program estimate the parameters by posterior inference.

The mean ancestry proportion of each population estimated from our algorithm is summarized in Table 1. Overall, the ancestry vector agrees very well with their geographical locations or known history. For example, populations such as Yoruba, Mandenka, BiakaPygmy, or BantuSouthAfrica recovered pure African ancestries, Druze, Basque, Russian and Adygei populations had dominant European ancestries (≥ 0.978), and Pima or Colombian populations resulted in almost pure Native American ancestries (≥ 0.983).

[Table 1 about here.]

More interestingly, the result also identifies the populations that have strong evidence of admixing effect among multiple ancestries. For instance, the proportion of European ancestry in Uygur population was 0.35, that of East Asian ancestry was 0.41, and the remaining proportion of 0.24 in Native American ancestry. Although only one or two populations are selected to serve as each putative ancestral population in our study and hence the interpretation of this result needs to be done carefully, our result largely agrees with the previously reported ancestry proportion in this population. For example, the analysis in (Xu *et al.* 2008; Xu and Jin 2008) claimed that Uygur had roughly 50–60% of European ancestry and 40–50% of East Asian ancestry from the analysis based on two-way admixture. More recent study in (Li *et al.* 2009) showed evidences that the estimation of European ancestry in these studies appear to be biased and suggested a newly estimated proportion of around 30%. Our estimation of East Asian ancestry (41%) is similar to that in (Xu *et al.* 2008) and in addition the estimation of European ancestry (35%) is closer to the more recent result in (Li

et al. 2009) than (Xu *et al.* 2008). Considering its geographical location and the resulting population history, our result suggests that Uygur population has about 35% of European ancestry, 41% of East Asian ancestry, and the remaining proportions of ancestries in other contributing populations that have greater similarity to Native American population.

To further analyze each population data and the behavior of the proposed method, we examined the empirical mutation parameter $\tilde{\delta}$ of each study population computed as an average discrepancy between individuals and corresponding founders within each of the populations. Therefore, $\tilde{\delta}$ can be viewed as reflecting the level of divergence from the founder population. The result is displayed in Figure 11 where the colors of the bars are based on the geographic location of the corresponding population. The ordering of populations by their parameter values almost exactly agrees with the geographic locations out of Africa. That is, all the populations in African continent had the largest values of δ, populations in Eurasia came next, and Oceanian populations were the third. Populations in East Asian region formed the fourth cluster and then Pima and Colombian populations showed the smallest values of δ. It is noteworthy that Yoruba, which appears to be the closest to the training population of YRI, recovers a much larger value of mutation rate δ than all the populations in geographic locations other than African continent. This comes from the nice property of our model that we do not directly use the training haplotype data as our reference, we rather infer the corresponding common founders across all the population data together and then work in a framework dealing with founders and admixed individuals. Otherwise, it would be impossible to obtain such a result because the discrepancy of Yoruba and its reference data would be much smaller than most of the other populations.

[Figure 11 about here.]

DISCUSSION

Previous admixture studies have suggested that the world populations are not independent of each other, but rather are structured through population admixing history and the re-

sulting gene flow. Most existing approaches for local ancestry estimation have ignored such relatedness and treated the populations as unrelated. We explore this dependency among populations and efficiently utilize it by building a unified model that covers all the ancestral populations and the admixed population together. As shown in our Results, this modeling strategy is especially helpful when only a limited amount of data is available to represent the ancestral populations. Since genetic information in one population can be naturally shared by another population in such a framework, it effectively enhances the robustness of the proposed model regarding the choice of the ancestral population data.

In our comparative study, HAPMIX appears to perform very well when enough amount of data for ancestral populations are given and also for older admixture events. However, this method does not allow one to analyze the admixing effect from more than two ancestral populations. Instead, one ancestry versus all the other ancestries should be estimated. While this setting may be fine for some applications, this constraint limits its applicability to complex admixture scenarios and may compromise its ability to deal with older admixtures.

LAMP has slightly different focus: while its performance was shown to be worse than the other two in general in our simulation study, it can deal with multiple ancestral populations as our model. And computationally this method was significantly faster than the other two haplotype-based methods. LAMP seems to be more suited for very recent admixture case, and its performance tends to drop quite sharply as we consider more ancient admixture events. On the other hand, in a very recent admixture case, LAMP tends to be less sensitive to the amount of training data than HAPMIX as shown in Figure 4. Our approach is more general and of more practical utility in that it can incorporate an arbitrary number of ancestral populations with comparable or superior performance than HAPMIX under various scenarios. In comparison of computation time with HAPMIX, our method requires additional, but off-line computation time for model training, which is linear in the number of individuals and the number of markers. For ancestry estimation phase, we would additionally need a series of MCMC iteration time if we want to estimate the parameters of interest such

as admixture time or mutation rates. As an example running time of our algorithm, it took about 5 minutes to run on a dataset with 30 admixed individuals on chromosome 22.

In the proposed model, we adopted population-specific recombination rates by using a scaling parameter of T_j that explains the different effect population size and the time since the founder population. Although it makes sense to scale the mutation rate by T_j as well in each of the ancestral populations, we found that the performance for the local ancestry estimation did not improve in our experiments. This might be due to statistical reason. During inference, it is observed that the algorithm tends to favor the ancestral population with the smallest mutation rate excessively, so this might have created excessive bias toward such an ancestral population instead of selecting the correct ancestry.

Although our method allows to estimate the admixture time parameter G instead of requiring it as an input when inferring the local ancestry, the parameter estimation result was not very accurate in general. Still, the local ancestry estimation performance was not significantly affected by incorrect estimation of the parameter as implied from our sensitivity analysis in Figure 10 (b). It appears that the likelihood surface from our statistical model is relatively flat over the space of model parameters, so the single optimal point on the model parameter space could not be achieved stably. When we let our program estimate G instead of fixing it in the same scenario considered in Figure 10 (b), the estimate of G averaged over 50 repetitions was around 14 when the true value was $G = 10$. The ancestry estimation accuracy was comparable to the case when we fixed G as 10.

It is worth mentioning some of previous approaches for global ancestry analysis as well to position our method in context. STRUCTURE (Pritchard *et al.* 2000) has been one of the most widely used softwares for admixture analysis, and more recently, other softwares such as EIGENSTRAT (Patterson *et al.* 2006) and ADMIXTURE (Alexander *et al.* 2009) have also gain great popularity especially for their computational efficiency. In global ancestry estimation problems, typically no prior information is provided for the ancestral populations and the ancestries of given individuals are recovered as mean proportions of

each possible ancestry. Therefore, it can be considered as an unsupervised problem. In contrast, local ancestries are mostly estimated based on the given reference information such as allele frequencies or genotypes of putative ancestral populations. There has been more recent work that bridges the gap between these two approaches. For example, LAMP can also run in an 'unsupervised mode' such that it recovers the allele frequency profiles of ancestral populations as well as the local ancestries. Also, ADMIXTURE, which is for the global ancestry estimation, recently added a new feature that the known ancestries of some reference individuals can be exploited (Alexander and Lange 2011). For haplotype-based approaches, this extension is not straightforward in general because one needs to deal with a set of hidden haplotypes that results in a large number of parameters. Regarding this aspect, our model for the local ancestry has the desirable property that it integrates out the ancestral population data during the inference and work with the hypothetical founders and the admixed population data. Therefore, we expect that the extension of the model to an unsupervised case would also be a promising direction to pursue.

In this paper, we assumed that phased haplotype data are given. In practice, a number of softwares are available for haplotype phasing (Li *et al.* 2010; Browning and Browning 2009; Scheet and Stephens 2006), so the phase information can be readily available in processing step. It is also possible to extend our model to deal with unphased genotypes. For example, we may assume that the haplotypes of ancestral populations are given, and then we allow unphased genotypes for admixed individuals, as in the setting considered in (Price *et al.* 2009). The only additional computation then would be one more step in our posterior sampling to recover the phasing of genotypes as well as the hidden states in the ancestry estimation phase.

ACKNOWLEDGMENTS

This material is based upon work supported by a National Science Foundation Career Award to E.P.X. under grant DBI-0546594 and NIH grant 1R01GM087694.

SUPPLEMENTAL MATERIAL

Forward-backward algorithm for the proposed infinite HMM A variant of the beam sampling algorithm for infinite HMM (Van Gael *et al.* 2008) is employed to improve the convergence over standard Gibbs sampling. Specifically, we introduce auxiliary variables u_t for $t = 0, ..., T-1$:

$$\begin{aligned} u_{i0} \mid S_{i0} = (k,j) &\sim \text{Uniform}(0, \nu_{jk}\eta_{ij}) \\ u_{it} \mid S_{it} = (k,j), S_{i,t-1} = (k',j') &\sim \text{Uniform}(0, q_{it}) \qquad \text{for } t = 1, ..., T-1 \end{aligned}$$

where

$$q_{it} = e^{-G_t^r d_t} e^{-g_{jt}^r d_t} I(k=k') I(j=j') + e^{-G_t^r d_t}(1 - e^{-g_{jt}^r d_t}) I(j=j') \pi_{k'k}^j + (1 - e^{-G_t^r d_t}) \nu_{jk}\eta_j$$

For notational convenience, we omit the notation i. Let the forward probabilities be $\alpha_t(k,j) = P(S_t = (k,j) \mid H_{0:t}, u_{0:t})$. Then

$$\begin{aligned} \alpha_0(k,j) &\propto P(S_0 = (k,j), H_0, u_0) \propto P(S_0 = (k,j)) P(u_0 \mid S_0 = (k,j)) P(H_0 \mid C_0 = k) \\ &= I(u_0 < \nu_{jk}\eta_{Z_0}) P(H_0 \mid C_0 = k) \\ \alpha_t(k,j) &\propto \sum_{k',j'} P(S_t = (k,j), S_{t-1} = (k',j'), H_t, u_t \mid H_{0:t-1}, u_{0:t-1}) \\ &\propto P(H_t \mid C_t = k) \sum_{k',j'} P(u_t \mid S_t = (k,j), S_{t-1} = (k',j')) P(S_t = (k,j) \mid S_{t-1} = (k',j')) \alpha_{t-1}(k',j') \\ &\propto P(H_t \mid C_t = k) \sum_{j'=0}^{J-1} \sum_{k'=0}^{\infty} I(u_t < P(S_t = (k,j) \mid S_{t-1} = (k',j'))) \alpha_{t-1}(k',j') \end{aligned} \tag{A1}$$

Given $u_0, ..., u_{T-1}$, the number of states k such that $\alpha_t(k,j) > 0$ for $t = 0, ..., T-1$ is finite: for $t = 0$, the number of k such that $\nu_{jk} > u_0$ is finite for any j since $\sum_k \nu_{jk} = 1$ with $\nu_{jk} \geq 0$, and recursively, we can see the number of k with $\alpha_t(k,j) > 0$ is finite. Therefore, the infinite sum over the previous states in the calculation of forward probability reduces to

a finite sum.

C_{T-1} and Z_{T-1} can be sampled from $\alpha_{T-1}(k, j)$. Then for $t = T-2, ..., 0$, we sample C_t and Z_t using

$$P(C_t, Z_t \mid H_{0:T-1}, u_{0:T-1}, C_{t+1}, Z_{t+1}) \propto P(C_{t+1}, Z_{t+1} \mid C_t, Z_t)\alpha_t(C_t, Z_t)P(u_{t+1} \mid S_t, S_{t+1})$$

If we reduce the model to the training phase, we can treat the variable Z as observed. Therefore, the forward probabilities are written as follows:

$$\begin{aligned}
\alpha_0(k) &\propto P(C_0 = k, H_0, u_0) \propto P(C_0 = k)P(u_0 \mid C_0 = k)P(H_0 \mid C_0 = k) \\
&= I(u_0 < \nu_{Z_0 k}\eta_j)P(H_0 \mid C_0 = k) \\
\alpha_t(k) &\propto \sum_{k'} P(C_t = k, C_{t-1} = k', H_t, u_t \mid H_{0:t-1}, u_{0:t-1}) \\
&\propto P(H_t \mid C_t = k) \sum_{k'} P(u_t \mid C_t = k, C_{t-1} = k')P(C_t = k \mid C_{t-1} = k')\alpha_{t-1}(k') \\
&\propto P(H_t \mid C_t = k) \sum_{k'=0}^{\infty} I(u_t < P(C_t = k \mid C_{t-1} = k'))\alpha_{t-1}(k') \qquad \text{(A2)}
\end{aligned}$$

Once we get the trained parameters, we restrict the model to a finite state space, so we don't need to incorporate the auxiliary variables u, so the standard form of forward-backward probabilities can be used.

LITERATURE CITED

Alexander, D. H. and K. Lange, 2011 Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. BMC bioinformatics **12:** 246.

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

Beal, M. J., Z. Ghahramani, and C. E. Rasmussen, 2002 The infinite hidden Markov model. Advances in Neural Information Processing Systems **14**.

Blackwell, D. and J. B. MacQueen, 1973 Ferguson Distributions Via Polya Urn Schemes.

The Annals of Statistics *1*(2)**:** 363–355.

Browning, B. L. and S. R. Browning, 2009 A Unified Approach to Genotype Imputation and Haplotype-Phase Inference for Large Data Sets of Trios and Unrelated Individuals. The American Journal of Human Genetics *84*(2)**:** 210–223.

Cheng, C.-Y., W. H. L. Kao, N. Patterson, A. Tandon, C. A. Haiman, T. B. Harris, C. Xing, E. M. John, C. B. Ambrosone, F. L. Brancati, J. Coresh, M. F. Press, R. S. Parekh, M. J. Klag, L. A. Meoni, W.-C. Hsueh, L. Fejerman, L. Pawlikowska, M. L. Freedman, L. H. Jandorf, E. V. Bandera, G. L. Ciupak, M. A. Nalls, E. L. Akylbekova, E. S. Orwoll, T. S. Leak, I. Miljkovic, R. Li, G. Ursin, L. Bernstein, K. Ardlie, H. A. Taylor, E. Boerwinckle, J. M. Zmuda, B. E. Henderson, J. G. Wilson, and D. Reich, 2009 Admixture mapping of 15,280 African Americans identifies obesity susceptibility loci on chromosomes 5 and X. PLoS genetics *5*(5)**:** e1000490.

Cheng, C.-Y., D. Reich, T. Y. Wong, R. Klein, B. E. K. Klein, N. Patterson, A. Tandon, M. Li, E. Boerwinkle, A. R. Sharrett, and W. H. L. Kao, 2010 Admixture mapping scans identify a locus affecting retinal vascular caliber in hypertensive African Americans: the Atherosclerosis Risk in Communities (ARIC) study. PLoS genetics *6*(4)**:** e1000908.

Falush, D., M. Stephens, and J. K. Pritchard, 2003 Inference of Population Structure Using Multilocus Genotype Data: Linked Loci and Correlated Allele Frequencies. Genetics *164*(4)**:** 1567–1587.

Ferguson, T. S., 1973 A Bayesian Analysis of Some Nonparametric Problems. The Annals of Statistics *1*(2)**:** 209–230.

Huelsenbeck, J. P. and P. Andolfatto, 2007 Inference of Population Structure Under a Dirichlet Process Model. Genetics *175*(4)**:** 1787–1802.

Jakobsson, M., S. W. Scholz, P. Scheet, J. R. Gibbs, J. M. VanLiere, H.-C. Fung, Z. A. Szpiech, J. H. Degnan, K. Wang, R. Guerreiro, J. M. Bras, J. C. Schymick, D. G. Hernandez, B. J. Traynor, J. Simon-Sanchez, M. Matarin, A. Britton, J. van de Leemput, I. Rafferty, M. Bucan, H. M. Cann, J. A. Hardy, N. A. Rosenberg, and

A. B. Singleton, 2008 Genotype, haplotype and copy-number variation in worldwide human populations. Nature *451*(7181)**:** 998–1003.

Li, H., K. Cho, J. R. Kidd, and K. K. Kidd, 2009 Genetic landscape of Eurasia and "admixture" in Uyghurs. American journal of human genetics *85*(6)**:** 934–7; author reply 937–9.

Li, Y., C. J. Willer, J. Ding, P. Scheet, and G. R. Abecasis, 2010 MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genetic epidemiology *34*(8)**:** 816–834.

Pasaniuc, B., S. Sankararaman, G. Kimmel, and E. Halperin, 2009, June)Inference of locus-specific ancestry in closely related populations. Bioinformatics (Oxford, England) *25*(12)**:** i213–21.

Patterson, N., N. Hattangadi, B. Lane, K. E. Lohmueller, D. A. Hafler, J. R. Oksenberg, S. L. Hauser, M. W. Smith, S. J. O'Brien, D. Altshuler, M. J. Daly, and D. Reich, 2004 Methods for High-Density Admixture Mapping of Disease Genes. The American Journal of Human Genetics *74*(5)**:** 979–1000.

Patterson, N., A. L. Price, and D. Reich, 2006 Population Structure and Eigenanalysis. PLoS genetics *2*(12)**:** e190.

Price, A. L., N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick, and D. Reich, 2006 Principal components analysis corrects for stratification in genome-wide association studies. Nature Genetics *38*(8)**:** 904–909.

Price, A. L., A. Tandon, N. Patterson, K. C. Barnes, N. Rafaels, I. Ruczinski, T. H. Beaty, R. Mathias, D. Reich, and S. Myers, 2009 Sensitive Detection of Chromosomal Segments of Distinct Ancestry in Admixed Populations. PLoS genetics *5*(6)**:** e1000519.

Price, A. L., M. E. Weale, N. Patterson, S. R. Myers, A. C. Need, K. V. Shianna, D. Ge, J. I. Rotter, E. Torres, K. D. Taylor, D. B. Goldstein, and D. Reich, 2008 Long-Range LD Can Confound Genome Scans in Admixed Populations. The American Journal of Human Genetics *83*(1)**:** 132–135.

Pritchard, J. K., M. Stephens, and P. Donnelly, 2000 Inference of population structure

using multilocus genotype data. Genetics *155*(2)**:** 945.

Sankararaman, S., G. Kimmel, E. Halperin, and M. I. Jordan, 2008 On the inference of ancestries in admixed populations. In *RECOMB'08: Proceedings of the 12th annual international conference on Research in computational molecular biology*. Springer-Verlag.

Sankararaman, S., S. Sridhar, G. Kimmel, and E. Halperin, 2008 Estimating Local Ancestry in Admixed Populations. The American Journal of Human Genetics **82:** 290–303.

Scheet, P. and M. Stephens, 2006 A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. American journal of human genetics *78*(4)**:** 629–644.

Sohn, K.-A. and E. P. Xing, 2007 Spectrum: joint bayesian inference of population structure and recombination events. Bioinformatics (Oxford, England) *23*(13)**:** i479–i489.

Sohn, K.-A. and E. P. Xing, 2009 A hierarchical Dirichlet process mixture model for haplotype reconstruction from multi-population data. The Annals of Applied Statistics *3*(2)**:** 791–821.

Sundquist, A., E. Fratkin, C. B. Do, and S. Batzoglou, 2008 Effect of genetic divergence in identifying ancestral origin using HAPAA. Genome Research *18*(4)**:** 676–682.

Tang, H., S. Choudhry, R. Mei, M. Morgan, W. Rodriguez-Cintron, E. G. Burchard, and N. J. Risch, 2007 Recent genetic selection in the ancestral admixture of Puerto Ricans. American journal of human genetics *81*(3)**:** 626–633.

Tang, H., M. Coram, P. Wang, X. Zhu, and N. Risch, 2006 Reconstructing Genetic Ancestry Blocks in Admixed Individuals. The American Journal of Human Genetics **79:** 1–12.

Teh, Y. W., M. I. Jordan, M. J. Beal, and D. M. Blei, 2010 Hierarchical Dirichlet Processes. Journal of the American Statistical Association *101*(476)**:** 1566–1581.

Van Gael, J., Y. Saatci, Y. W. Teh, and Z. Ghahramani, 2008 Beam sampling for the infinite hidden Markov model. In *ICML '08: Proceedings of the 25th international conference on Machine learning*. ACM.

Wang, X., X. Zhu, H. Qin, R. S. Cooper, W. J. Ewens, C. Li, and M. Li, 2010 Adjustment for local ancestry in genetic association analysis of admixed populations. Bioinformatics (Oxford,

England) *27*(5)**:** 670–677.

Xu, S., W. Huang, J. Qian, and L. Jin, 2008 Analysis of genomic admixture in Uyghur and its implication in mapping strategy. American journal of human genetics *82*(4)**:** 883–894.

Xu, S. and L. Jin, 2008 A genome-wide analysis of admixture in Uyghurs and a high-density admixture map for disease-gene discovery. American journal of human genetics *83*(3)**:** 322–336.

Zhu, X., J. H. Young, E. Fox, B. J. Keating, N. Franceschini, S. Kang, B. Tayo, A. Adeyemo, Y. V. Sun, Y. Li, A. Morrison, C. Newton-Cheh, K. Liu, S. K. Ganesh, A. Kutlar, R. S. Vasan, A. Dreisbach, S. Wyatt, J. Polak, W. Palmas, S. Musani, H. Taylor, R. Fabsitz, R. R. Townsend, D. Dries, J. Glessner, C. W. K. Chiang, T. Mosley, S. Kardia, D. Curb, J. N. Hirschhorn, C. Rotimi, A. Reiner, C. Eaton, J. I. Rotter, R. S. Cooper, S. Redline, A. Chakravarti, and D. Levy, 2011 Combined admixture mapping and association analysis identifies a novel blood pressure genetic locus on 5p13: contributions from the CARe consortium. Human Molecular Genetics *20*(11)**:** 2285–2295.

List of Figures

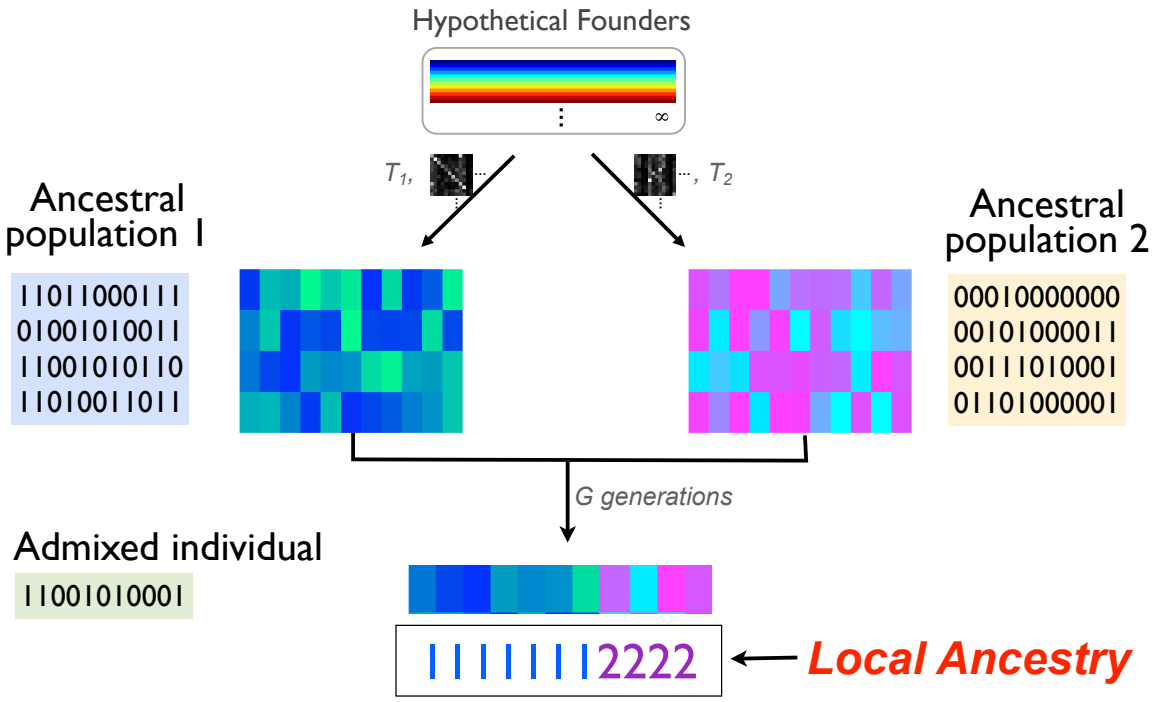
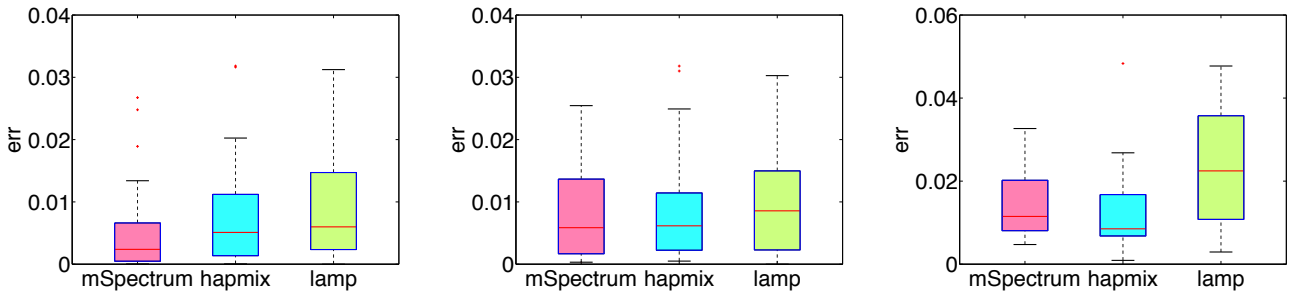


Figure 1: Graphical illustration of the proposed model

Figure 2: True and estimated local ancestries of two sample individuals in an admixed population from African and European populations. The x-axis corresponds to chromosomal position and the y-axis corresponds to the ancestry probability (yellow: African, dark grean: European)



$G = 5$ $G = 10$ $G = 20$

Figure 3: Boxplot for mean squared error rates of ancestry estimation for two-way admixture of African and European populations since G generations ago

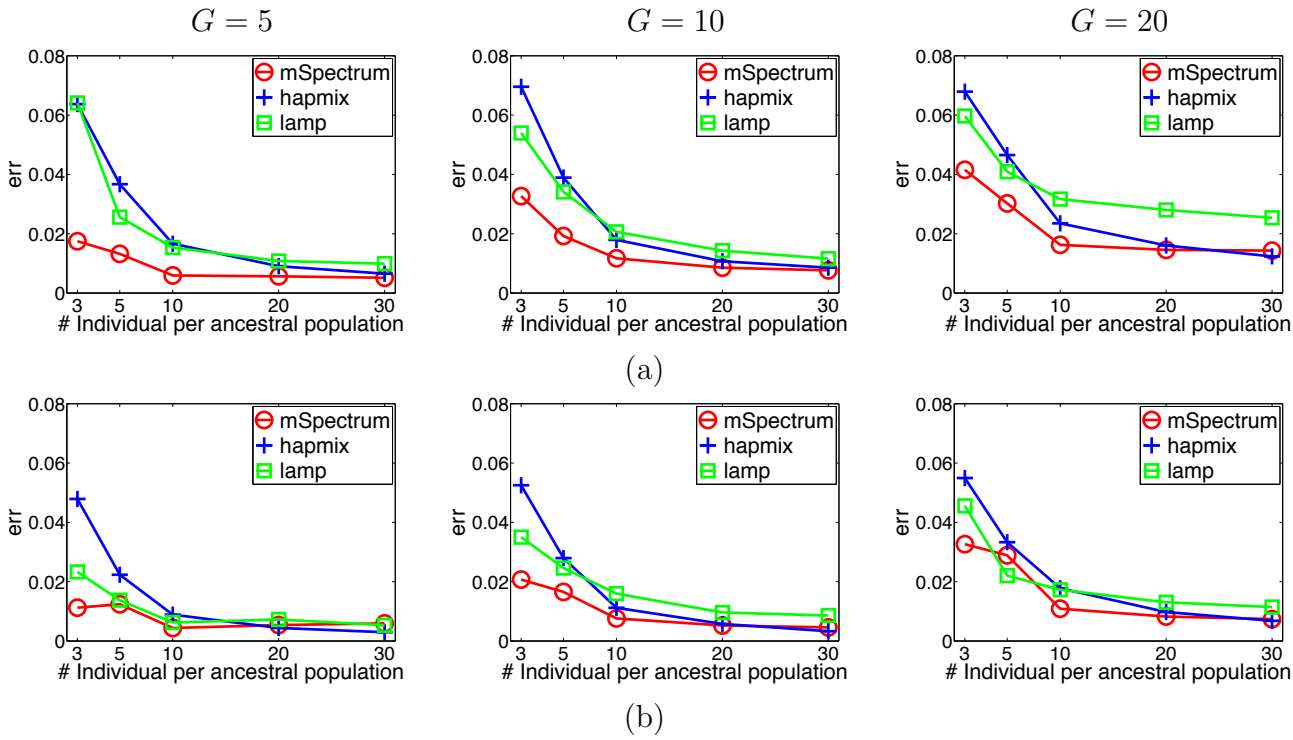


Figure 4: Error rate as a function of the number of individuals per train population. Two-way admixture of African and European populations since G generations ago using (a): Russian and BantuKenya populations, (b): CEU and YRI populations.

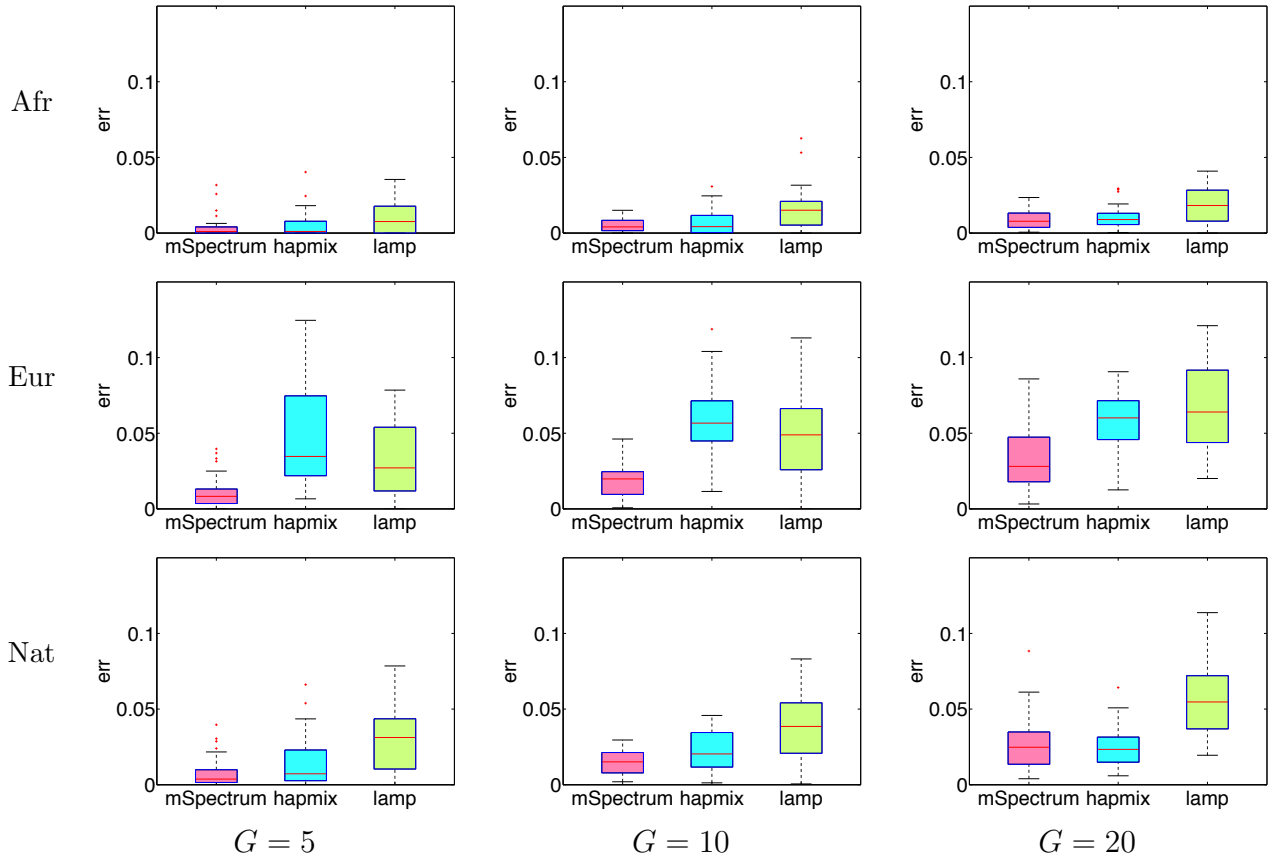
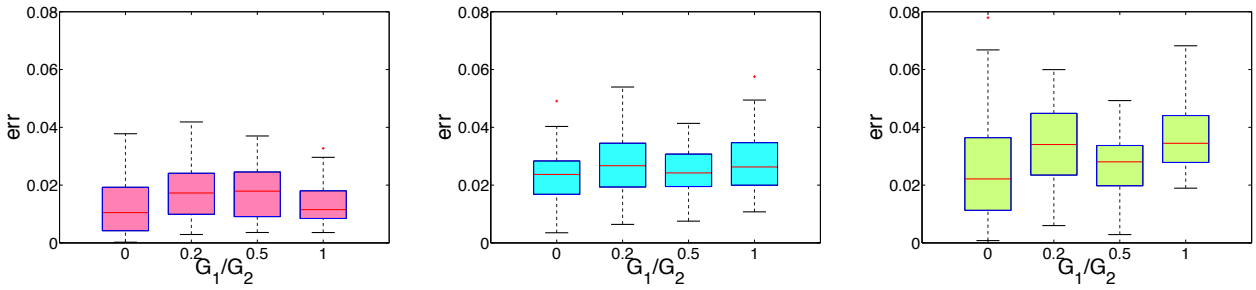


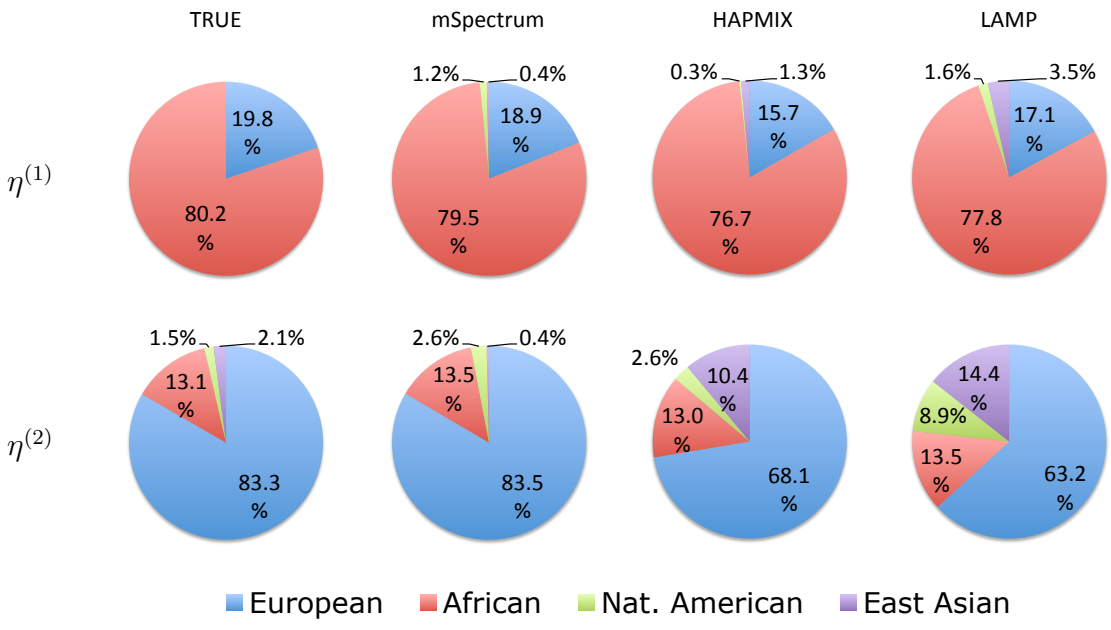
Figure 5: Boxplot for mean squared error rates of ancestry estimation. Three-way admixture of African, European, and Native American populations since G generations ago. Since HAPMIX is applicable to only two-way admixture case and was run to estimate each ancestry versus the other two, we report the error rate on each ancestry separately.



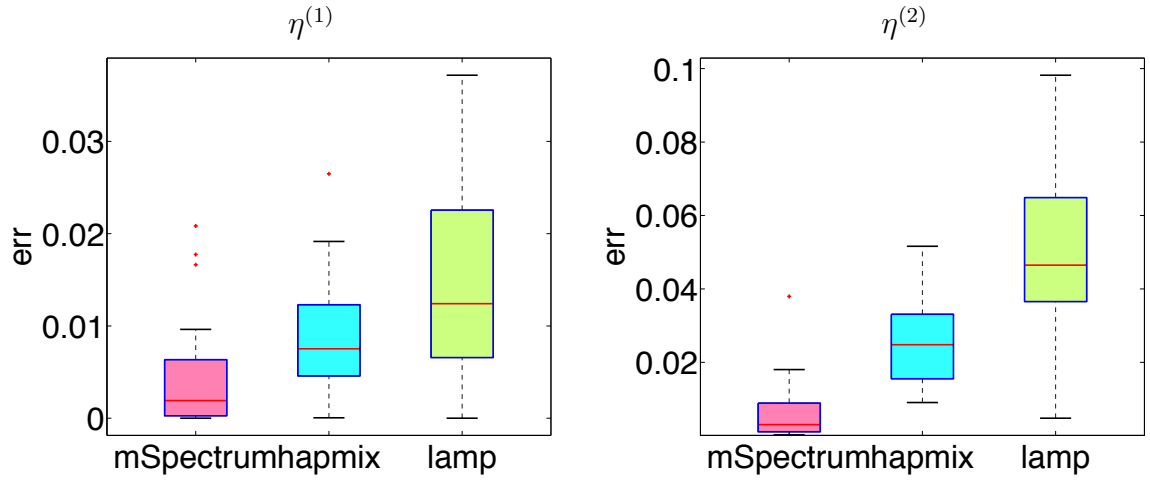
mSpectrum HAPMIX LAMP

Figure 6: Robustness under deviation from the modeling assumption. The x-axis represents the ratio G_1/G_2, where G_1 denotes the number of generations for which the first two populations had mixed and G_2 means the additional number of generations since the third population joined and have further mixed together.

Figure 7: True and estimated local ancestries of two sample individuals in an admixed population from African and European populations when the ancestry is estimated with respect to four ancestral populations of YRI (red), CEU (black), Maya (yellow), and JPT+CHB (white). The x-axis corresponds to chromosomal position and the the length of each colored vertical bar is proportional to the corresponding ancestry probability)



(a)



(b)

Figure 8: Performance under four-way admixture assumption when the admixed population is generated with admixing proportions of $\eta^{(1)} = (0.2, 0.8, 0, 0)$ and $\eta^{(2)} = (0.8, 0.15, 0.03, 0.02)$ using Russian, BantuKenya, Pima, and Yi populations. We show two performance measures of (a) the true and the empirical η estimated as the average local ancestry proportions across individuals and markers, and (b) the mean squared error rates of local ancestry estimation.

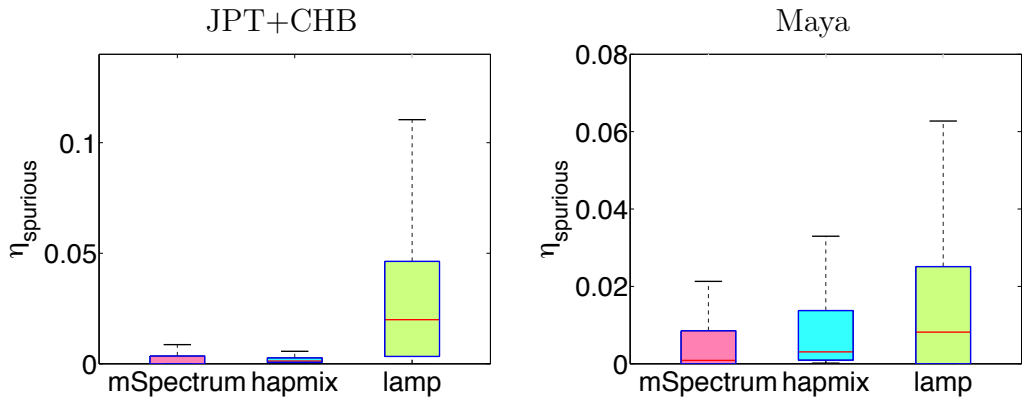
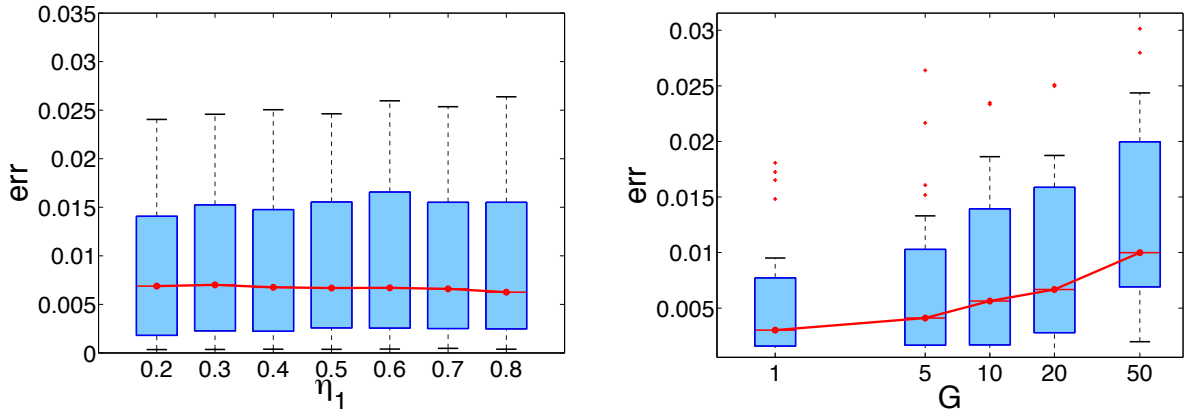


Figure 9: Proportions of spuriously estimated ancestry proportions under four-way admixture assumption when the admixed population is generated using Russian and BantuKenya populations only, computed over 50 datasets each of which containing 30 individuals.



(a) (b)

Figure 10: Sensitivity analysis: boxplot for error rates as a function of specified parameter values (a) η_1 and (b) G when the true values are $\eta_{true} = (0.5, 0.5), G_{true} = 10$.

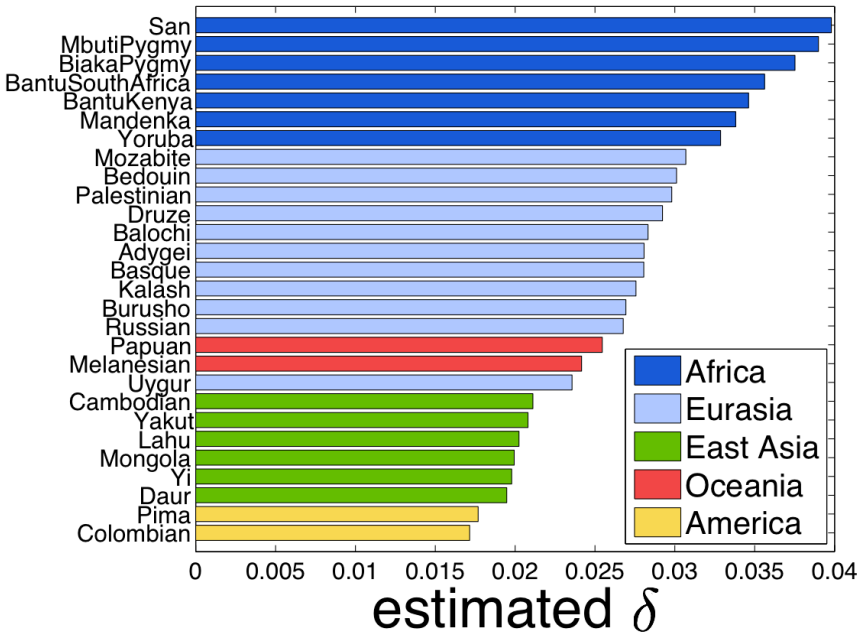


Figure 11: The empirical mutation rate δ of each HGDP population computed as the average discrepancy between individuals and their founders.

List of Tables

Table 1: Estimated ancestry proportions of populations in HGDP dataset.

	African	European	East Asian	Native Amer
Yoruba	1.000	0.000	0.000	0.000
Mandenka	1.000	0.000	0.000	0.000
BiakaPygmy	1.000	0.000	0.000	0.000
BantuSouthAfrica	1.000	0.000	0.000	0.000
San	0.999	0.001	0.000	0.000
MbutiPygmy	0.999	0.000	0.000	0.001
BantuKenya	0.998	0.001	0.000	0.000
Mozabite	0.141	0.818	0.013	0.028
Bedouin	0.035	0.941	0.006	0.018
Palestinian	0.013	0.966	0.006	0.015
Basque	0.000	0.998	0.000	0.001
Russian	0.000	0.990	0.003	0.007
Druze	0.002	0.989	0.002	0.006
Adygei	0.000	0.978	0.008	0.014
Kalash	0.000	0.930	0.027	0.043
Balochi	0.015	0.888	0.031	0.066
Burusho	0.000	0.741	0.088	0.170
Uygur	0.000	0.348	0.414	0.239
Yakut	0.000	0.045	0.848	0.106
Mongola	0.000	0.006	0.960	0.034
Daur	0.000	0.004	0.972	0.024
Cambodian	0.000	0.004	0.977	0.019
Lahu	0.000	0.000	0.987	0.013
Yi	0.000	0.001	0.991	0.009
Melanesian	0.001	0.039	0.821	0.140
Papuan	0.002	0.081	0.733	0.185
Pima	0.001	0.012	0.004	0.983
Colombian	0.002	0.001	0.001	0.996