JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 20, Number 3, 2013

Pp. 199–211
DOI: 10.1089/cmb.2012.0088

# Ancestry Inference in Complex Admixtures via Variable-length Markov Chain Linkage Models

JESSE M. RODRIGUEZ,[1,2,*] SIVAN BERCOVICI,[1,*]
MEGAN ELMORE,[1] and SERAFIM BATZOGLOU[1]

## ABSTRACT

**Inferring the ancestral origin of chromosomal segments in admixed individuals is key for genetic applications, ranging from analyzing population demographics and history, to mapping disease genes. Previous methods addressed ancestry inference by using either weak models of linkage disequilibrium, or large models that make explicit use of ancestral haplotypes. In this paper we introduce ALLOY, an efficient method that incorporates generalized, but highly expressive, linkage disequilibrium models. ALLOY applies a factorial hidden Markov model to capture the parallel process producing the maternal and paternal admixed haplotypes, and models the background linkage disequilibrium in the ancestral populations via an inhomogeneous variable-length Markov chain. We test ALLOY in a broad range of scenarios ranging from recent to ancient admixtures with up to four ancestral populations. We show that ALLOY outperforms the previous state of the art, and is robust to uncertainties in model parameters.**

**Key words:** ancestry inference, FHMM, population genetics, VLMC.

## 1. INTRODUCTION

DETERMINING THE ANCESTRAL origin of chromosomal segments in admixed individuals is a problem that has been addressed by several methods (Patterson et al., 2004; Tang et al., 2006; Sundquist et al., 2008; Bercovici and Geiger, 2009; Pasaniuc et al., 2009; Price et al., 2009). The development of these methods was motivated by various applications, such as studying population migration patterns (Jakobsson et al., 2008; Gravel et al., 2011), increasing the statistical power of association studies by accounting for population structure (Pasaniuc et al., 2011), and enhancing admixture-mapping (Winkler et al., 2010; Seldin et al., 2011) for both disease-gene mapping as well as personalized drug therapy applications (Baye and Wilke, 2010). The ability to accurately infer ancestry is important in genome-wide association studies (GWAS). These studies are based on the premise that a homogenous population sample was collected. Population stratification, however, poses a significant challenge in association studies; the existence of different subpopulations within the examined cases and controls can yield many spurious associations originating from the population substructure rather than the disease status. Inferred substructure within the population enables the correction for this effect, consequently improving the statistical power of these studies.

[1]Department of Computer Science, and [2]Biomedical Informatics Program, Stanford University, Stanford, California.
*These authors contributed equally to this work.

A second disease-gene mapping technique that benefits from an accurate inference of ancestry is admixture-mapping (Winkler et al., 2010). This statistically powerful and efficient method identifies genomic regions containing disease susceptibility genes in recently admixed populations, which are populations formed from the merging of several distinct ancestral populations (e.g., African-Americans). The statistical power of admixture-mapping increases as the disease prevalence exhibits a greater difference between the ancestral populations from which the admixed population was formed. Admixed individuals carrying such a disease are expected to show an elevated frequency of the ancestral population with the higher disease risk near the disease gene loci. Hence, the effectiveness of this method relies on the ability to accurately infer the ancestry along the chromosomes of admixed individuals.

The problem of ancestry inference is commonly viewed at one of two levels: (a) at the global scale, predicting an individual's single origin out of several possible homogenous ancestries, or determining an individual's ancestral genomic composition; and (b) at the finer local scale, labeling the different ancestries along the chromosomes of an admixed individual. In the context of local ancestry inference, most previous methods are based on hidden Markov models (HMM), where the hidden states correspond to ancestral populations and generate the observed genotypes. The work of Patterson et al. (2004) employed such an HMM, integrated into a Markov chain Monte Carlo (MCMC), for estimating ancestry along the genome. The method accounted for uncertainties in model parameters such as number of generation since admixture, admixture proportions, and ancestral allele frequencies. For simplicity, the work assumed that, given the ancestry, the sampled markers are in linkage equilibrium (i.e., independent). This assumption was then relaxed in the work by Tang et al. (2006), applying a Markov hidden Markov model (MHMM) to account for the dependencies between neighboring markers as exhibited within the ancestral populations. While the modeled first-order Markovian dependencies accounted for some of the linkage disequilibrium (LD) between markers, the complex nature of the linkage patterns presented an opportunity for more accurate LD models that would yield better performance in inferring local ancestry. The explicit use of ancestral haplotypes, in methods such as HAPAA (Sundquist et al., 2008) and HAPMIX (Price et al., 2009), enabled a more comprehensive account for background LD (i.e., LD within the ancestral population) over longer segments. In these methods, the hidden states corresponded to specific ancestral haplotypes, and the transition between the states corresponded to intra-population mixture and inter-populations admixture processes. While efficient inference algorithms were applied, the model size grew linearly with the number of parental individuals, and the time complexity grew quadratically with the numbers of parental individuals for the case of genotype-based analysis. The time complexity of such an analyses became prohibitively high with more than a modest number of model individuals.

Other work explored window-based techniques, in which a simple ancestral composition was assumed to occur within a window (i.e., at most a single admixture event within an examined segment). LAMP (Sankararaman et al., 2008), and its extension WINPOP (Pasaniuc et al., 2009), used a naïve Bayes approach, assuming markers within a window are independent given ancestry, applying the inference over a sliding window. Although LD was not modeled, the methods demonstrated an accuracy superior to methods that did account for background LD. An additional window-based framework was developed in Bercovici and Geiger (2009), decoupling the admixture process from the background LD model. Chromosomal ancestral profiles were efficiently enumerated using a dynamic-programming (DP) technique, enabling the instantiation of various LD models for the single-ancestry segments from which a profile was composed. Multiple LD models were studied within the framework, showing that higher-order LD models yield an increase in inference accuracy.

In this work we describe ALLOY, a novel local ancestry inference method that enables the incorporation of complex models for linkage disequilibrium in the ancestral populations. ALLOY applies a factorial hidden Markov model (FHMM) to capture the parallel process producing the maternal and paternal admixed haplotypes. We model background LD in ancestral populations via an inhomogeneous variable-length Markov chain (VLMC). The states in our model correspond to ancestral haplotype clusters, which are groups of haplotypes that share local structure within a chromosomal region, as in Browning and Browning (2007). In our method, each ancestral population is described by a separate LD model that locally fits the varying LD complexities along the genome. We provide an inference algorithm that is subcubic in the maximal number of haplotype clusters at any position. This allows ALLOY to scale well when analyzing admixtures of more than two populations or incorporating more elaborate LD models.

We demonstrate through simulations that ALLOY accurately infers the position-specific ancestry in a wide range of complex and ancient admixtures. For instance, ALLOY achieves 87% accuracy on a three-population admixture between individuals sampled from Yoruba in Ibadan, Nigeria; Maasai in

Kinyawa, Kenya; and northern and western Europe. Our results represent substantial improvements over previous state of the art. Further, we explore the landscape of background LD models, and find that the highest performance is achieved by LD models that lie between models that assume independence of markers and models that explicitly use the reference haplotypes. Finally, our results demonstrate that as more samples representing the ancestral populations become available, our LD models improve and enable more accurate local ancestry inference.

## 2. METHODS

We consider the problem of local ancestry inference, defined as labeling each genotyped position along the genome of an admixed individual with its ancestry. Here, admixture is assumed to follow the hybrid isolated model (Long, 1991), in which a single past admixture event mixing $K$ ancestral populations with proportions $\pi = (\pi_1, \ldots, \pi_K)$ is followed by $g$ generations of consecutive random mating. For clarity, we assume that a set of $L$ bi-allelic single nucleotide polymorphisms (SNPs) was observed along the genome of an individual; we relax the bi-allelic marker assumption in the Discussion section. Furthermore, at each position, we define a state space of haplotype clusters $A_l$, each of which represents a collection of ancestral haplotypes that share a common local structure (i.e., allelic sequence surrounding a particular location). It immediately follows that each such haplotype cluster $a_l \in A_l$ at location $l$ is mapped to a single allele, denoted by $e(a_l) \in \{0, 1\}$. In our model, each of the $K$ populations is represented by a separate mutuallyexclusive subset of haplotype clusters. We denote by $anc(a_l)$ the ancestry, out of $K$, of a particular haplotype cluster $a_l \in A_l$. We denote by $H_l^m, H_l^p$ the (hidden) haplotype cluster membership drawn from $A_l$ on the maternal and paternal haplotype at position $l$, respectively, and by $G_l \in \{0, 1, 2\}$ the genotype observed at the same marker position, representing the minor allele count. The vectors of haplotype cluster memberships and genotypes across all $L$ marker positions are denoted by $H^{\{m,p\}} = (H_1^{\{m,p\}}, H_2^{\{m,p\}}, \ldots, H_L^{\{m,p\}})$ and $G = (G_1, G_2, \ldots, G_L)$, respectively.

We use a factorial hidden Markov model (FHMM) (Ghahramani et al., 1997) to statistically model the dual mosaic ancestral pattern along the genome of an admixed individual, as depicted in Figure 1. In factorial HMMs, which are equivalent in expressive power to hidden Markov models (HMM) (Rabiner, 1989), the single chain of hidden variables is replaced by a chain of a hidden vector of independent factors. In our application, the FHMM representation allows us to naturally decouple the state space into two parallel dynamic processes generating $H^m$ and $H^p$, pertaining to the presumably independent maternal and paternal admixture processes, and producing the single composed admixed offspring $G$. The decomposition of the state space into independent processes allows efficient inference by leveraging the structure in the compound state transition probabilities. In our model, the values of $H_l = (H_l^m, H_l^p)$ at specific position $l$ are drawn from the Cartesian product $A_l \times A_l$, corresponding to the alleles within specific ancestral haplotypes originating from a restricted prior set of $K$ hypothesized ancestral populations. Note that $A_l$ extends the notion of an allele, which is simply a binary variable, to an allele within an ancestral haplotype; for position $l$, multiple states in $A_l$ may correspond to the same allele.

To infer local ancestry, we first compute the posterior marginals given the sampled genotypes $P(H_l^m, H_l^m|G)$ by applying the forward-backward algorithm

$$P(H_l^m = a_l, H_l^p = a'_l|G) \propto \alpha_l(a_l, a'_l) \cdot \beta_l(a_l, a'_l) \tag{1}$$

where $\alpha_l(a_l, a'_l) = P(G_1, \ldots, G_l, H_l^m = a_l, H_l^p = a'_l)$ and $\beta_l(a_l, a'_l) = P(G_{l+1}, \ldots, G_L|H_l^m = a_l, H_l^p = a'_l)$. A naive recursive computation of $\alpha$ and $\beta$ yields $O(|A_1|^2 + \sum_{l=2}^{L} |A_{l-1}|^2 \cdot |A_l|^2)$ time complexity as the transition from each pair of haplotype cluster memberships to each consecutive pair of haplotype cluster memberships is explicitly assessed. However, the dependency structure of FHMMs allows for a more efficient recursive computation of $\alpha$ and $\beta$, as described in Ghahramani et al. (1997). Specifically, $\alpha$ is computed in the forward direction in three steps as follows

$$\begin{aligned}
\alpha_{l-1}^m(a_l, a'_{l-1}) &= \sum_{a_{l-1} \in A_{l-1}} \alpha_{l-1}(a_{l-1}, a'_{l-1}) \cdot P(H_l^m = a_l|H_{l-1}^m = a_{l-1}) \\
\alpha_{l-1}^p(a_l, a'_l) &= \sum_{a'_{l-1} \in A_{l-1}} \alpha_{l-1}^m(a_l, a'_{l-1}) \cdot P(H_l^p = a'_l|H_{l-1}^p = a'_{l-1}) \\
\alpha_l(a_l, a'_l) &= \alpha_{l-1}^p(a_l, a'_l) \cdot P(G_l|H_l^m = a_l, H_l^p = a'_l),
\end{aligned} \tag{2}$$

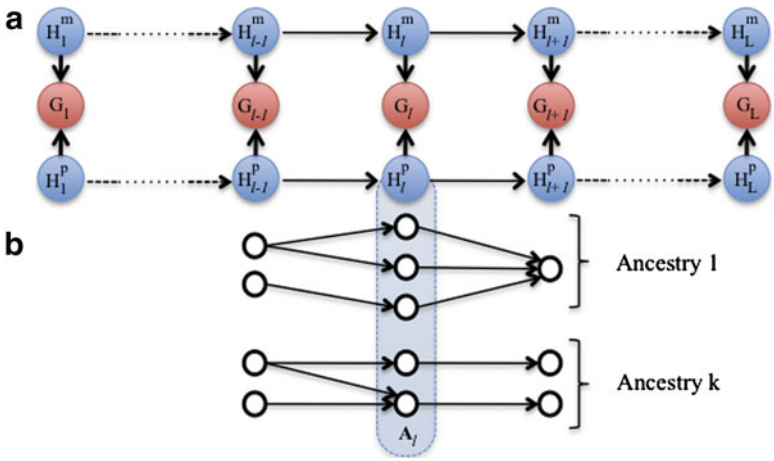


**FIG. 1.** A factorial hidden Markov model capturing the parallel admixture processes generating the maternal and paternal haplotypes and giving rise to the sampled genotypes of the admixed offspring. (**a**) A graphical model depicting the conditional independencies in our model. Each variable in the hidden chains $H_l^{\{m,p\}}$ corresponds to a haplotype cluster membership, and $G_l$ corresponds to the observed genotype at location $l$. (**b**) The state space $A_l$ for a particular location $l$ along the genome. Each ancestry is modeled by an independent set of haplotype cluster membership states and each such state can emit a single allele. Edges in the illustration connecting states correspond to intra-population observed transitions, namely, local haplotypic sequences that were frequent in the corresponding ancestral population. Edges corresponding to admixture transitions, connecting states of different ancestries, are omitted from this illustration for clarity.

namely, advancing on the maternal track, followed by advancing on the paternal track, and finally, incorporating the local observation by multiplying by the emission probability $P(G_l|H_l^m=a_l, H_l^p=a'_l)$. Similarly, $\beta$ is computed in a backward recursion as

$$\begin{aligned}\beta_l^e(a_l,a'_l)&=\beta_l(a_l,a'_l)\cdot P(G_l|H_l^m=a_l,H_l^p=a'_l)\\ \beta_l^m(a_{l-1},a'_l)&=\sum_{a_l\in A_l}P(H_l^m=a_l|H_{l-1}^m=a_{l-1})\cdot\beta_l^e(a_l,a'_l)\\ \beta_{l-1}(a_{l-1},a'_{l-1})&=\sum_{a'_l\in A_l}P(H_l^p=a'_l|H_{l-1}^p=a'_{l-1})\cdot\beta_l^m(a_{l-1},a'_l).\end{aligned}\tag{3}$$

To complete the description, we define $\alpha_1(a_1,a'_1)=P(H_1^m=a_1)\cdot P(H_1^p=a'_1)\cdot P(G_1|H_1^m=a_1,H_1^p=a'_1)$ and $\beta_L(a_L,a'_L)=1$. When computing $\beta$, advancing on the maternal track takes $(|A_{l-1}|\cdot|A_l|)\cdot|A_l|$ time, while advancing on the paternal track takes $(|A_{l-1}|\cdot|A_{l-1}|)\cdot|A_l|$ time, as determined by the size of the corresponding composite state space. Similarly, a single forward step $\alpha_l$ is computed in $|A_l|\cdot|A_{l-1}|\cdot(|A_l|+|A_{l-1}|)$ time. Hence, the time complexity is now reduced to $O(|A_1|^2+\sum_{l=2}^{L}|A_l|\cdot|A_{l-1}|\cdot(|A_l|+|A_{l-1}|))$.

To model genotyping error, the emission probability $P(G_l|H_l^m,H_l^p)$ used in Equations 2 and 3 is defined as follows

$$P(G_l|H_l^m=a_l,H_l^p=a'_l)=\begin{cases}1-2\epsilon, & e(a_l)+e(a'_l)=G_l\\ \epsilon, & \text{otherwise}\end{cases}\tag{4}$$

where $\epsilon$ corresponds to the genotyping error rate.

To increase the numerical stability in the forward-backward computation, scaling is applied. Specifically, $\alpha_l$ and $\beta_l$ are scaled by $s_l=\sum_{a_l,a'_l}\alpha_l(a_l,a'_l)$ as follows

$$\alpha_l^*(a_l, a_l') = \frac{\alpha_l(a_l, a_l')}{s_l} \beta_l^*(a_l, a_l') = \frac{\beta_l(a_l, a_l')}{s_l}. \tag{5}$$

Next, the unordered ancestry pair $\{Z_l^1, Z_l^2\}$ at location $l$ is called by determining the maximal *a posteriori* assignment

$$\{\hat{Z}_l^1, \hat{Z}_l^2\} = \arg\max_{Z_l^1, Z_l^2} \sum_{a_l, a_l' \text{ s.t.}} \alpha_l^*(a_l, a_l') \cdot \beta_l^*(a_l, a_l') \tag{6}$$

$$\{anc(a_l), anc(a_l')\} = \{Z_l^1, Z_l^2\}$$

where, for each $\{Z_l^1, Z_l^2\}$ pair, we sum over all ($a_l$, $a_l'$) haplotype cluster membership pairs that are consistent in their ancestry with the unordered ancestry pair $\{Z_l^1, Z_l^2\}$.

We proceed by describing the transition probabilities $P(H_l|H_{l-1})$. Let $R_l$ be defined as the event in which at least one post-admixture recombination occurred between position $l-1$ and position $l$ since the first population admixture event, and let $\bar{R}_l$ be defined as the complementary event. The transition probability $P(H_l|H_{l-1})$, which captures the process in which an admixed haplotype is generated, mixes the event of intra-ancestral population transition, $P(H_l|H_{l-1}, \bar{R}_l)$, with the event corresponding to the introduction of a new ancestral haplotype, $P(H_l|H_{l-1}, R_l)$, as described by

$$\begin{aligned} P(H_l|H_{l-1}) &= P(R_l) \cdot P(H_l|H_{l-1}, R_l) + P(\bar{R}_l) \cdot P(H_l|H_{l-1}, \bar{R}_l) \\ &= P(R_l) \cdot P_{anc(H_l)}(H_l) + P(\bar{R}_l) \cdot P_{anc(H_l)}(H_l|H_{l-1}) \end{aligned} \tag{7}$$

where $P_{anc(Hl)}(H_l)$ is the position-specific ancestral haplotype cluster prior, and $P_{anc(Hl)}(H_l|H_{l-1})$ models the transition within the ancestral population $anc(H_l)$, capturing the background population-specific LD. Namely, if a post-admixture recombination was introduced ($P(R_l)$), a haplotype $H_l$ is sampled based on the local ancestry prior $P_{anc(Hl)}(H_l)$; if no post-admixture recombination was introduced ($P(\bar{R}_l)$), the next marker is sampled based on the haplotypic structure within population $anc(H_l)$, as defined by $P_{anc(Hl)}(H_l|H_{l-1})$. Assuming the hybrid-isolated model, the probability of post-admixture recombination $P(R_l)$ is approximated via the Haldane function (Haldane, 1919)

$$P(R_l) = 1 - e^{-\phi(g \cdot d_l)} \tag{8}$$

where $d_l$ is the genetic distance, in Morgans (M), between marker $l-1$ and $l$, and $g+1$ generations are assumed to have passed since the first admixture event. We note that $\varphi(z)$ is defined as a function of the recombination rate $g \cdot d_l$ to enable smoothing; the number of false ancestry changes can be reduced by controlling the probability for recombination (e.g., $\phi(z) = \frac{z}{10}$), overcoming local inaccuracies in ancestry inference due to an imperfect ancestral linkage model. The prior probability of the ancestral haplotype cluster is governed by the mixture proportions $\pi$ and the intra-population haplotype cluster prior $P_{anc(Hl)}(H_l)$, as given by

$$P(H_l) = \pi_{anc(H_l)} \cdot P_{anc(H_l)}(H_l). \tag{9}$$

Finally, we describe the background model we use to capture the ancestral linkage disequilibrium between markers. The range of explored background LD models is illustrated in Figure 2. The most basic models used for ancestry inference assume markers are independent given their corresponding ancestry assignment. An immediate extension that can be captured by our FHMM model incorporates first-order Markovian dependencies to model LD between neighboring markers. However, the model is not limited to first-order dependencies; to capture longer range dependencies between ancestral alleles, the state space $A_l$ from which $H_l$ is drawn can be enriched so as to track ancestral haplotype clusters over a longer range. Specifically, longer range dependencies are effectively translated to additional states that map to specific ancestral local haplotype clusters. Moreover, a different number of states can be introduced at each position, fitting the local ancestral haplotypic complexity. The higher the local complexity is, the more states are used to track dependencies reaching further away. In essence, the model is equivalent to an inhomogeneous VLMC in which regions exhibiting complex LD structures are modeled using longer dependencies (i.e., edges connecting distant nodes in the underlying graphical model). At one extreme, the state space $A_l$ can be constructed assuming a zero-order Markov model (i.e., markers are independent), while at the other extreme, $A_l$ can be extended to have one state per ancestral haplotype instance used in the training phase.

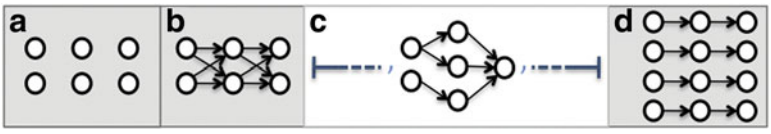


**FIG. 2.** The state space $A_l$ over three consecutive locations in different background LD models, pertaining to the marker dependencies exhibited within a single ancestral population. (**a**) Markers are independent given ancestry. The model contains two states per location, each emitting one of the two possible alleles matching the marginal distribution observed in the ancestral population. (**b**) First order Markovian dependency between adjacent markers. The transition between the neighboring states, which correspond to alleles at specific positions, is derived from the conditional probability estimated from the ancestral population sample. (**c**) Generalized linkage model via haplotype clusters. The number of states at each position correspond to the number of haplotype clusters, each emitting an allele. The local transition probabilities correspond to the Markovian property by which haplotype cluster membership at a given location $l$ is determined by the cluster membership at the previous location $l-1$. (**d**) Explicit use of ancestral haplotypes. For each position, the number of states equals the number of training haplotypes, each emitting a single allele observed in the corresponding haplotype.

An algorithm for fitting inhomogeneous VLMCs was described by Ron et al. (1995), and extended by Browning and Browning (2007), to model haplotypes. We apply Beagle, an implementation of this procedure, to empirically model the local haplotypic structure. Specifically, we determine both the state space of $A_l$ as well as the transition probability through the use of a localized haplotype cluster model described in Browning and Browning (2007). Briefly, given a set of training haplotypes from a single ancestry, the algorithm processes the markers in chromosomal order. With each additional marker considered, nodes, representing some history of allele sequences, are split by considering the subsequent alleles for each such node. Then, nodes at location $l$ are merged based on a Markov criterion roughly guaranteeing that given the cluster membership at position $l$, prior cluster memberships are irrelevant for the prediction of subsequent cluster memberships. Namely, given some parameter $t$, two clusters at position $l$ are merged if the probabilities of allele sequences at markers $l+1, l+2, \ldots, l+t$ resemble each other. For each population *anc*, the procedure yields a weighted directed acyclic graph (DAG), where edges are labeled by alleles, and each training haplotype traces a path through the graph from a root node to a terminal node, defining the weights. For each edge $e_l^i$ at location $l$, the weight $w_l^i$ is defined as the number of haplotypes in the ancestral population sample that pass through the $i^{th}$ cluster. In our model, the state space $A_l^{anc} \subset A_l$ for population *anc* at location $l$ is defined so that each edge $e_l^i$ in the weighted DAG corresponds to the state $a_l^{anc,\,i}$. We denote the source node of each edge $e_l^i$ by $s_l^i$ and its target by $t_l^i$. The prior $P_{anc}(H_l)$ and transition probabilities $P_{anc}(H_l|H_{l-1})$ from Equations 7 and 9, respectively, can be computed as follows

$$P_{anc}(H_l = a_l^{anc,i} | H_{l-1} = a_{l-1}^{anc,j}) = \begin{cases} \frac{w_l^i}{\sum_{k \text{ s.t. } t_{l-1}^k = s_l^i} w_{l-1}^k}, & \text{if} \quad t_{l-1}^j = s_l^i \\ 0, & \text{otherwise} \end{cases} \tag{10}$$

$$P_{anc}(H_l = a_l^{anc,i}) = \frac{w_l^i}{\sum_j w_{l-1}^j}. \tag{11}$$

The process is repeated for each ancestry separately, producing the population-specific $P_{anc(al)}(H_l = a_l)$ and $P_{anc(al)}(H_l = a_l | H_{l-1} = a_{l-1})$.

## 3. RESULTS

**Simulation of admixed individuals and training the background LD models**. We evaluated the performance of ALLOY for local ancestry inference. In our experiments, we simulated admixed individuals and trained ALLOY's background model using data from six HapMap (Altshuler et al., 2010) populations: individuals from the Centre d'Etude du Polymorphisme Humain collected in Utah, with ancestry from northern and western Europe (CEU); Han Chinese in Beijing, China (CHB); Japanese in Tokyo, Japan (JPT); Yoruba in Ibadan, Nigeria (YRI); Maasai in Kinyawa, Kenya (MKK); and Tuscans in Italy (Toscani in Italia, TSI). All SNPs present in the HapMap Phase III panel on the first arm of Chromosome 1 were used to expedite the results. We partitioned the HapMap data such that 100 individuals from each population were used as training data, and the remainder were used as test data to evaluate the performance of our method. We used haplotypes from the test set to simulate admixed individuals for six different combinations of ancestral populations: YRI-MKK (YM), CHB-JPT (CJ), YRI-MKK-CEU (YMC), CHB-JPT-CEU (CJC), YRI-MKK-CEU-CHB (YMCC), and CHB-JPT-CEU-YRI (CJCY). Each test data set contained 100 simulated admixed individuals. In this section, we use $g_{sim}$ and $\pi_{sim}$ to denote the parameters used for simulation, and $g$ and $\pi$, to denote the parameters used for inference. Each simulated admixed individual was generated by traversing the set of markers in chromosomal order, generating a pair of admixed maternal and paternal haplotypes in parallel. The initial pair of ancestries and alleles, corresponding to the first marker, was randomly selected based on the prior ancestral admixture proportions $\pi_{sim}$. Alleles were then copied from the ancestral reference haplotypes. With each subsequent marker, the probability for an admixture-related recombination was evaluated via Equation 8. In case of a recombination, a new ancestral source was selected using the $\pi_{sim}$ admixture proportions and the copying process continued.

We used the Beagle package (Browning and Browning, 2007) with default parameters, to phase the training and testing individuals separately. Next, the ancestral background LD model states and parameters were determined through Equations 10 and 11 by examining Beagle's DAG output. To build an efficient background LD model for ALLOY, we selected a subset of ancestry informative markers (AIM), which are genetic variants that carry a population-specific characterizing allele distribution and can be used to efficiently distinguish between genetic segments of different origins. In order to select the set of ancestry informative markers, we used the Shannon Information Content (SIC) criteria (Rosenberg et al., 2003). Namely, for a given set of markers and their corresponding allele distribution in the ancestral populations, we measured the mutual information (MI) $I(X_l;Z)$ between ancestry $Z$ and allele $X_l$ at position $l$. Using the SIC measurement, we followed the marker selection heuristic presented in Tian et al. (2006), choosing a constant number of highly informative markers within a window of fixed size. Specifically, in our simulations, we selected the single most informative marker in windows of 0.05 centimorgans. For the YM, YMC, and YMCC data sets, we used SNPs with the highest MI differentiating the YRI and MKK populations; for the CJ, CJC, and CJCY admixture scenarios, we selected markers with the highest MI when differentiating the CHB and JPT populations. While ALLOY's background LD model was based on a subset of SNPs, inference was performed on all SNPs, calling the ancestry of the excluded SNPs using a nearest marker approach.

**Evaluating ALLOY's accuracy under complex and ancient admixtures.** When performing inference, we modeled the genotyping error rate with $\epsilon = 0.01$, and used $\phi(z) = \frac{z}{10}$ as our smoothing function in Equation 8. We compared the performance of ALLOY to WINPOP (Pasaniuc et al., 2009), a local ancestry

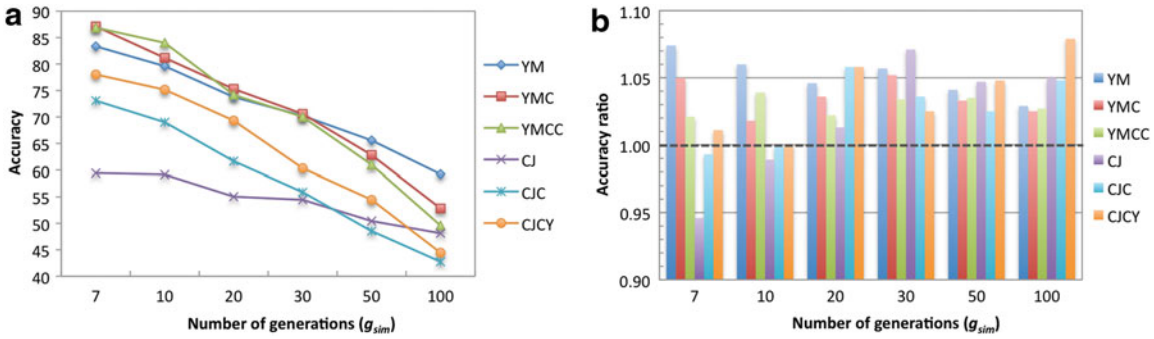


**FIG. 3.** (**a**) The performance of ALLOY based on the number of generations since admixture $g_{sim}$ for various admixture configurations with equal ancestral proportions. ALLOY was run with $g = g_{sim}$ and $\pi = \pi_{sim}$, and the accuracy was conservatively measured as the fraction of markers for which the exact ancestry pair was inferred. (**b**) For the same experiments, ALLOY was compared to WINPOP by measuring the accuracy ratio between them ($\frac{\text{ALLOY's accuracy}}{\text{WINPOP's accuracy}}$). The results clearly demonstrate ALLOY's superior accuracy in the vast majority of tested admixture configurations, with an increase in performance in more than 86% of the tests.

inference platform that has been shown to outperform previous state-of-the-art methods such as SABER (Tang et al., 2006), HAPAA (Sundquist et al., 2008), and HAPMIX (Price et al., 2009). We measured the accuracy of ALLOY and WINPOP when inferring local ancestry of simulated admixed individuals under increasingly complex admixtures, and with a varying number of generations since the first admixture event, ranging from recent admixture ($g = 7$) to more ancient admixture ($g = 100$). Accuracy was conservatively measured as the average fraction of SNPs for which the correct ancestry was inferred. As depicted in Figure 3, our results show that ALLOY's accuracy is greater than WINPOP's in nearly all tested scenarios. Our experiments show that applying WINPOP over the full set of markers achieves a higher performance in comparison to analyzing only a subset of ancestry information markers. WINPOP performs SNP selection prior to inference to confirm with their model assumptions, and hence benefits from the larger initial set of markers. We therefore reported WINPOP results corresponding to an analysis applied on the entire HapMap Phase III set of SNPs rather than the SNP subsets used for training ALLOY.

**Exploring background LD models.** As previously described, the background LD models in ALLOY can capture a wide range of complexities, from simpler models such as those used in AncestryMap (Patterson et al., 2004) and SABER, which model zero- and first-order dependencies between markers, respectively, to more complex explicit haplotype models as used by HAPAA and HAPMIX. More importantly, ALLOY is able to capture models of intermediate complexity. We explored the performance of ALLOY using a range of background LD models with varying complexities. Background models of different complexities were generated by applying Beagle on our training data using different values for Beagle's *scale* parameter, which controls the complexity of the generated DAG underlying ALLOY's model. As *scale* approaches 0, the model approaches the explicit model used in HAPAA, and as the value of *scale* grows, the generated model approaches a zero-order model similar to the one used by AncestryMap. The results, shown in Figure 4, illustrate that the models of intermediate complexity outperform both the more complex as well as the simpler models used by previous methods.

**Measuring robustness to inaccuracies in model parameters.** Our method assumes that the admixture parameters, such as the number of generations $g$ and the admixture proportions $\pi$ are given. When applied on real data, however, the true values for these parameters are unknown. We examined the robustness of ALLOY to inaccuracies in model parameters. Specifically, we measured the impact of misspecified admixture proportion $\pi$ on the accuracy of inference. To test for robustness, we simulated a YM mixture with $\pi_{sim} = (0.5,0.5)$ and $g_{sim} = 30$, and evaluated ALLOY's performance varying $\pi$ between (0.05,0.95) and (0.95,0.05) during inference. Our results indicate that ALLOY's performance is robust to inaccuracies in $\pi$, yielding the highest accuracy when $\pi = \pi_{sim}$, slightly reducing the accuracy by 0.0029 to its lowest value at the two extremes [i.e., $\pi = (0.95,0.05)$ and $\pi = (0.05, 0.95)$]. We further evaluated ALLOY's performance when $g$ was misspecified. A YM mixture was simulated with $g_{sim} = 20$, $\pi_{sim} = (0.5,0.5)$. When $g = g_{sim}$, ALLOY achieved 73.77% accuracy; for misspecified values of $g$ between 10 and 40, accuracy ranged from 73.41% to 73.88%, respectively.

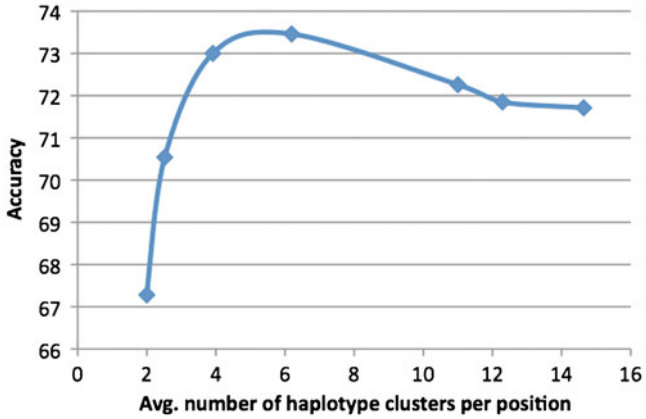


**FIG. 4.** Ancestry inference accuracy as a function of model complexity, as measured by the average number of haplotype clusters under a certain background LD model. The models range from a simplistic assumption of independence on the left, to more explicit models on the right. The plot illustrates that both oversimplification, corresponding to the LD models used in AncsetryMap and SABER, and overspecification, corresponding to the models leaning toward those used in HAPAA and HAPMIX, yield reduced performance in comparison with a more generalizing local haplotypic model.

Additionally, we explored the sensitivity of ALLOY's performance to different genotyping error rates $\epsilon$. When simulating a YM mixture with $g_{sim} = 20, \pi_{sim} = (0.5, 0.5)$ as above, and performing inference with values of $\epsilon \leq 0.025$, ALLOY's accuracy was at least 73.20%. Assuming a 5% genotyping error rate ($\epsilon = 0.05$), the accuracy decreased by less than 1%.

**Evaluating model accuracy under varying amounts of training data.** Currently, the amount of available genotype data is limited by the number of individuals genotyped and the density of SNPs measured. However, the number of genotyped individuals and the SNP density of genotyping technologies are expected to greatly increase in the near future. To evaluate the effect of training set size on ALLOY's performance, we trained our background LD model on sets of individuals with increasing size. Specifically, we derived a model for the YRI and MKK ancestral populations using subsets of the individuals of varying size and evaluated the inference accuracy. The results, shown in Figure 5a, emphasize the importance of training set size to the improved performance, suggesting that as more samples are collected and genotyped, more accurate background models could be derived, yielding a higher level of accuracy.

We further evaluated the performance of ALLOY with respect to the number of SNPs used during training. We generated subsets of informative SNPs of various sizes by using different window sizes during the AIM selection phase. To evaluate the importance of using AIMs, we also selected random SNP subsets of matching sizes. We simulated individuals from a YM admixed population with $g_{sim} = 30$ generations of admixture, evaluating ALLOY's accuracy when trained using the different SNP subset. Figure 5b shows

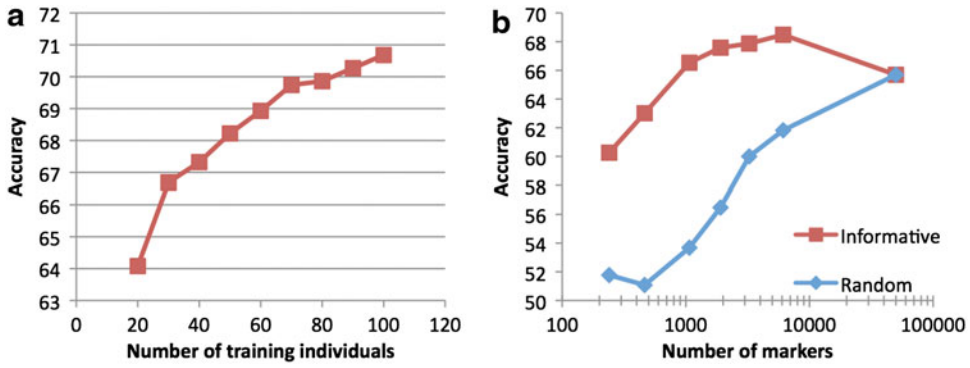


**FIG. 5.** (**a**) Local ancestry inference accuracy as a function of training set size. A various number of individuals were used as representatives of the ancestral populations in the computation of the background LD model, demonstrating an increased performance as more samples are used in the training phase. (**b**) The accuracy of inferring ancestry as a function of the number of markers used. The plot illustrates the significance of using ancestry-informative markers in comparison to a randomly chosen set, as for all tested resolutions, the use of the informative set yielded an improved performance. The results also indicate that the addition of noninformative markers reduces performance (demonstrated by the right-most data point) as these are assumed to interfere with the construction of an effective background LD model.

that ALLOY's accuracy increases as more SNPs are used. The results further demonstrate that ALLOY's performance is significantly higher with informative SNPs compared to random ones. The rightmost point in Figure 5b corresponds to ALLOY's performance when all SNPs are used. These results indicate that using excessive uninformative markers can reduce accuracy in comparison to a model based on informative markers.

## 4. DISCUSSION

ALLOY represents the LD structure of each population with a highly expressive model that lies between the simpler first-order Markov hidden Markov model in SABER, and the explicit-haplotype model in HAPAA and HAPMIX. The first advantage of this approach is its improved accuracy compared to either extreme, as shown in Figure 4. Additionally, our inference algorithm has higher computational efficiency than explicit-haplotype models. In this work, we derive the population-specific LD structures by generating haplotype clusters through the Beagle package. We translate the produced DAG into prior and transition probabilities that define the parameters of our factorial hidden Markov model. In future work, alternatives to Beagle can be used for modeling LD; for instance, one can develop ancestry-aware methods that produce LD models that emphasize the structural differences between ancestral populations.

In our experiments, we assumed a hybrid isolated (HI) model (Long, 1991) for simulating admixed individuals. However, other models, such as the continuous gene flow (CGF) model (Long, 1991), may better correspond with population migration and admixture patterns, and as such will more accurately fit the ancestral mosaic patterns observed in admixed populations. ALLOY assumes an HI admixture model. To evaluate the robustness of ALLOY to misspecification of admixture models, we measured ALLOY's accuracy under the scenario where the admixed individuals were simulated using a CGF model. ALLOY achieved 86.0% accuracy on a YM mixture with $\pi = (0.5,0.5)$, $g = 10$, and a generational donor contribution rate $\alpha = 0.01$ from both ancestries. These results indicate an approximately 8% increase in accuracy compared to the results achieved when inferring the local ancestry of simulated admixed individuals generated using the HI model and the same values for $g$ and $\pi$. The increase in accuracy can be attributed to the fact that CGF generates longer ancestral tracts in comparison to the HI model with the same admixture parameters, and the fact that longer tracts are easier to predict correctly. To explore our model under a more challenging scenario, we further simulated admixed individuals from a YM mixture using the CGF model with an adjusted $g$ such that the average ancestral tract length was equal to the average length under an HI model with the same parameters. ALLOY achieved 81.0% accuracy, which is comparable to our previous result for the HI model (79.6% accuracy). We concluded that ALLOY is robust to such differences in the underlying admixture model and can support more realistic admixture models.

ALLOY assumes that the admixture parameters are given. In particular, the number of generations since admixture $g$, and the relative proportions of the ancestral populations $\pi$, are required. We showed through simulations that our method is robust to inaccuracies in the estimation of the admixture proportion. Nonetheless, $\pi$ can be estimated by direct examination of the sampled individuals' genotype likelihood. Alternatively, given a set of individuals representing a particular admixed population, demographic parameters such as admixture time $g$ and ancestral proportions $\pi$ can be derived as a post-processing step. For instance, as suggested in Pool and Nielsen (2009) and Henn et al. (2012), the length of ancestral tracts can be used to infer changes in migration patterns. In particular, as our method has been shown to be robust to inaccuracies in $\pi$ and $g$, as well as to misspecified admixture models, we can first apply ALLOY to accurately infer the individuals' ancestral mosaic. Then, statistics over the inferred ancestral tracts, such as their length and number, can be sequentially used in combination with a variety of admixture models to compute the maximum likelihood estimate for parameters such as the time of migration and nature of admixture. To infer these parameters, the method presented in Pool and Nielsen (2009) examined the distribution of tracts larger than a given threshold, as shorter tracts cannot be reliably inferred. By leveraging the structure stemming from the ancestral linkage disequilibrium, ALLOY can accurately infer shorter ancestral tracts, enabling the observation of more distant admixture events and historical changes in migration rate. We also note that the flexibility of our FHMM enables different admixture times and proportions to be incorporated separately for the maternal and paternal haplotypes. Hence, pedigrees exhibiting very recent complex admixture at the grandparental level can be explicitly modeled. For example, the parameters of our method can be tuned to accurately infer the ancestry of an admixed individual

that has one African-American and one Chinese parent. Finally, our model assumes a single genetic map is given, capturing the genetic distance between neighboring SNPs that is shared between all ancestral populations. Previous work showed that more accurate recombination rates can be inferred using admixed populations by observing the ancestral switch points among admixed individuals (Hinch et al., 2011; Wegmann et al., 2011). As with the methods used to infer admixture parameters, the ancestral mosaic of admixed individuals is first inferred; then, the rate of ancestral switches per position is estimated. While such methods can be used to infer more accurate maps, our experiments have shown that inaccuracies in the estimation of these recombination rates do not have a significant effect on ALLOY's ability to infer local ancestry under the examined scenarios.

In the Methods section, we described an inference algorithm with a time complexity that depends on the local ancestral LD structure rather than the number of ancestral haplotypes used when training the background model. Specifically, the algorithm's time complexity is $O(L \cdot C^3)$, where $C$ is an upper bound on the number of states in a single position (i.e., $C = \max_l |A_l|$). In our implementation of ALLOY, we reduced the time complexity by rearranging the calculations corresponding to the transition probabilities in the forward and backward computations, described by Equations 2 and 3, respectively. In particular, transitioning between states corresponding to an admixture recombination event can be collapsed into a single term. For instance, when transitioning between states corresponding to different ancestries in the forward iteration, Equation 7 is reduced to the term $P(R_l) \cdot P(H_l)$. Hence, $\alpha^m_{l-1}(a_l, a'_{l-1})$ can be rewritten as

$$\alpha^m_{l-1}(a_l, a'_{l-1}) = P(R_l) \cdot P(a_l) + P(\bar{R}_l) \cdot \sum_{a_{l-1} \in A^{anc(a_l)}_{l-1}} P_{anc(a_l)}(a_l | a_{l-1}).$$

When such an optimization is applied, the time complexity is reduced to $O(L \cdot C^2 \cdot C_K)$, where $C_K$ is an upper bound on the number of states corresponding to a single population (i.e., $C_K = \max_{l,k} |A^k_l|$). We note that this implementation of ALLOY has a practical running time, completing a single experiment as described in the Results section in approximately one minute.

Our simulations experimented with SNP markers that were found to be polymorphic in 1,184 individuals sampled from 11 populations in the third phase of the HapMap project (Altshuler et al., 2010). However, additional variation exists in these populations beyond the SNPs assayed in this data set. In particular, rare SNPs, which have been found to exhibit little sharing among diverged populations (Gravel et al., 2011) and can therefore act as highly informative markers for ancestry inference, are likely to be missing from the panel. Therefore, as additional rare SNPs are discovered and sampled, we expect the accuracy of ALLOY to improve. We further note that the spectrum of human genetic variation ranges beyond SNPs. For instance, copy-number variations (CNV) and other structural variations constitute a large fraction of the total human genomic variation (Alkan et al., 2011). As with SNPs, rare CNVs are useful for separating ancestries and have been shown to be more abundant than rare SNPs ( Jakobsson et al., 2008). Our model is not limited to bi-allelic SNPs and supports the incorporation of markers of higher variability, such as CNVs, by adjusting Equation 4. The construction of the variable-length Markov chain linkage-models, either through Beagle or other methods, can be extended to take such additional genetic variation into account.

ALLOY is a novel method for inferring the local ancestry of admixed individuals, which is an essential task for various applications in human genetics. We have shown that our approach has higher accuracy than the previous state of the art and that its VLMC-based LD model plays a crucial role in its superior performance. Our method is applicable to ancient and complex admixtures and is capable of separately modeling the maternal and paternal histories. We expect that as the genetic variation of worldwide populations is extensively sampled, ALLOY will be able to better characterize the particular histories of examined individuals. ALLOY is publicly and freely available online.

## ACKNOWLEDGMENTS

We thank Chuong B. Do for helpful discussions. This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE-1147470. This publication was made possible by Grant Number 5RC2HG005570-02 from the National Institutes of Health (NIH). Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the NIH. This material is based upon work supported by the National Science Foundation under

Grant No. 0640211. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

## DISCLOSURE STATEMENT

The authors declare that no competing financial interests exist.

## REFERENCES

Alkan, C., Coe, B.P., and Eichler, E.E. 2011. Genome structural variation discovery and genotyping. *Nature reviews. Genetics* 12, 363–376.

Altshuler, D.M., Gibbs, R.A., Peltonen, L., 2010. et al. Integrating common and rare genetic variation in diverse human populations. *Nature* 467, 52–58.

Baye, T.M., and Wilke, R.A. 2010 Mapping genes that predict treatment outcome in admixed populations. *The Pharmacogenomics Journal* 10, 465–477.

Bercovici, S., and Geiger, D. 2009. Inferring ancestries efficiently in admixed populations with linkage disequilibrium. *Journal of Computational Biology : A Journal of Computational Molecular Cell Biology* 16, 1141–50.

Browning S.R., and Browning, B.L. 2007. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *The American Journal of Human Genetics* 81, 1084–1097, 2007.

Ghahramani, Z., Jordan, M.I., and Smyth, P. 1997. Factorial hidden Markov models. In *Machine Learning*. MIT Press, Cambridge, MA.

Gravel, S., Henn, B.M., Gutenkunst, R.N., et al. 2011. Demographic history and rare allele sharing among human populations. *Proceedings of the National Academy of Sciences* 108, 11983–11988.

Haldane J.B.S. 1919. The combination of linkage values, and the calculation of distance between the loci of linked factors. *J Genet* 8, 299–309, 1919.

Henn, B.M., Botigué, L.R., Gravel, S. et al. 2012. Genomic Ancestry of North Africans supports back-to-Africa migrations. *PLoS Genet* 8, e1002397.

Hinch, A.G., Tandon, A., Patterson, N., et al. 2011. The landscape of recombination in African Americans. *Nature* 476, 170–175.

Jakobsson, M., Scholz, S.W., Scheet, P., et al. 2008. Genotype, haplotype and copy-number variation in worldwide human populations. *Nature* 451, 998–1003.

Long, J.C. 1991. The genetic structure of admixed population. *Genetics* 127, 417–428.

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

Pasaniuc, B., Zaitlen, N., Lettre, G., et al. 2011. Enhanced statistical tests for GWAS in admixed populations: assessment using African Americans from CARe and a Breast Cancer Consortium. *PLoS genetics* 7, e1001371.

Patterson, N., Hattangadi, N., Lane, B. et al. 2004. Methods for high-density admixture mapping of disease genes. *American Journal of Human Genetics* 74, 979–1000.

Pool, J.E., and Nielsen, R. 2009. Inference of historical changes in migration rate from the lengths of migrant tracts. *Genetics* 181, 711–719.

Price, A.L., Tandon, A., Patterson, N., et al. 2009. Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. *PLoS Genet* 5, e1000519.

Rabiner, L.R. 1989. A tutorial on hidden Markov models and selected applications in speech recognition. In *Proceedings of the IEEE*, 257–286.

Ron, D., Singer, Y., and Tishby, N., 1995. On the learnability and usage of acyclic probabilistic finite automata. In *Journal of Computer and System Sciences* 31–40.

Rosenberg, N.A., Li, L.M., Ward, R., and Pritchard, J.K. 2003. Informativeness of genetic markers for inference of ancestry. *The American Journal of Human Genetics* 73, 1402–1422.

Sankararaman, S., Sridhar, S., Kimmel, G., and Halperin, E. 2008. Estimating local ancestry in admixed populations. *Journal of Human Genetics* February, 290–303.

Seldin, M.F., Pasaniuc, B., and Price, A.L. 2011. New approaches to disease mapping in admixed populations. *Nature reviews. Genetics* 12, 523–8.

Sundquist, A., Fratkin, E., Do, C.B., and Batzoglou, S. 2008. Effect of genetic divergence in identifying ancestral origin using HAPAA. *Genome research* 18, 676–82.

Tang, H., Coram, M., Wang, P., et al. 2006. Reconstructing genetic ancestry blocks in admixed individuals. *American Journal of Human Genetics* 79, 1–12.

Tian, C., Hinds, D.A., Shigeta, R., et al. 2006. A genomewide single-nucleotide polymorphism panel with high ancestry information for african american admixture mapping. *The American Journal of Human Genetics* 79, 640–649.

Wegmann, D., Kessner, D.E., Veeramah, K.R., et al. 2011. Recombination rates in admixed individuals identified by ancestry-based inference. *Nature Genetics* 43, 847–853.

Winkler, C.A., Nelson, G.W., and Smith, M.W. Admixture mapping comes of age. *Annual Review of Genomics and Human Genetics* 11, 65–89.

Address correspondence to:
*Jesse M. Rodriguez*
*Stanford University*
*Biomedical Informatics Program*
*318 Campus Drive Room S260*
*Stanford, CA 94305*

*E-mail:* jesserod@cs.stanford.edu