ARTICLE

Reconstructing Genetic Ancestry Blocks in Admixed Individuals

Hua Tang, Marc Coram, Pei Wang, Xiaofeng Zhu, and Neil Risch

A chromosome in an individual of recently admixed ancestry resembles a mosaic of chromosomal segments, or ancestry blocks, each derived from a particular ancestral population. We consider the problem of inferring ancestry along the chromosomes in an admixed individual and thereby delineating the ancestry blocks. Using a simple population model, we infer gene-flow history in each individual. Compared with existing methods, which are based on a hidden Markov model, the Markov–hidden Markov model (MHMM) we propose has the advantage of accounting for the background linkage disequilibrium (LD) that exists in ancestral populations. When there are more than two ancestral groups, we allow each ancestral population to admix at a different time in history. We use simulations to illustrate the accuracy of the inferred ancestry as well as the importance of modeling the background LD; not accounting for background LD between markers may mislead us to false inferences about mixed ancestry in an indigenous population. The MHMM makes it possible to identify genomic blocks of a particular ancestry by use of any high-density single-nucleotide–polymorphism panel. One application of our method is to perform admixture mapping without genotyping special ancestry-informative–marker panels.

The genome of an admixed individual represents a mixture of alleles inherited from multiple ancestral (or parental) populations. If the admixing occurred recently, we can imagine that each chromosome was assembled by stitching together long segments of DNA from a particular ancestral population; as a result, changes in ancestry occur only at the “stitch points.” We refer to these chromosomal segments as “ancestry blocks.” The distribution of block sizes depends on when the indigenous populations came into contact; more-recent gene flow gives rise to longer ancestral chromosome blocks on average. Inferences regarding the ancestry of admixed individuals not only are intriguing to population geneticists and anthropologists but also are becoming essential in gene discovery and characterization studies. Because of the potential confounding due to stratification among the ancestral populations, conventional case-control association studies in admixed groups need to adjust for ancestry structure. Moreover, descendants from matings between reproductively isolated ancestors, admixed populations offer unique opportunities to unravel the genetic and environmental components of a variety of diseases. The idea of using admixed populations to map genetic disease loci can be traced to Rife.[1] The rationale of admixture mapping (or mapping by admixture linkage disequilibrium [MALD]) is that, if one of the ancestral populations carries a risk allele at a higher frequency than the other(s), then affected individuals are expected to share a greater level of ancestry from that population around that disease susceptibility locus, compared with the background ancestry level in the genome or compared with the ancestry sharing among unaffected individuals around the same location. The past decade has seen an emergence of theoretical calculations and methods development supporting the application of the method to gene mapping studies in humans.[2–7] For all current MALD methods, the efficiency of the design depends on the accuracy with which one can infer the ancestry at any chromosome location.

Several approaches have been proposed to estimate ancestry at specific genomic locations[4–6,8]; all of them feature a hidden Markov model (HMM), which offers a succinct and computationally efficient framework.[9] HMMs have been successfully used to model a myriad of biological processes; examples include linkage analysis,[10] sequence alignment,[11] nucleotide evolution,[12] and DNA copy-number alterations.[13] For ancestry inference, an HMM extracts more information than does a single-marker analysis, by combining observed genotypes at neighboring markers. This is because most genetic variation is shared across ancestral populations, and so, typically, a single allele does not allow unambiguous inference regarding ancestry at that location.[14] Additionally, the simple structure of an HMM enables it to be augmented into more-complicated models. Thus, several existing approaches for estimating locus-specific ancestry integrate an HMM into a Markov chain Monte Carlo (MCMC) method, which accounts for uncertainties in model parameters, such as difference in allele frequencies between the true ancestors and their contemporary surrogates.[4,5,8] These extensions allow more-accurate point estimates of ancestry as well as a more comprehensive assessment of sampling variability in the estimates. For the estimation of ancestry blocks, Seldin et al.[15] used the program PHASE[16] to estimate haplotypes in a 60-cM region in Europeans, Africans, and African Amer-

From the Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle (H.T.; P.W.); Department of Statistics, University of Chicago, Chicago (M.C.); Department of Preventive Medicine and Epidemiology, Loyola University Medical Center, Maywood, IL (X.Z.); and Institute for Human Genetics, University of California, San Francisco (N.R.)

Received January 23, 2006; accepted for publication March 15, 2006; electronically published May 17, 2006.

Address for correspondence and reprints: Dr. Hua Tang, Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue North, M2-B500, Seattle, WA 98109. E-mail: huatang@fhcrc.org

Am. J. Hum. Genet. 2006;79:1–12. 0002-9297/2006/7901-0002$15.00

icans and inferred ancestry of the estimated African American haplotypes. However, our simulations demonstrate that haplotype inference at the level of an entire chromosome is often infeasible by use of autosomal genotypes in unrelated individuals.

As high-throughput genotyping platforms become available, it is now practical to genotype 1,000–500,000 SNPs in an individual in a single experiment. By inference of ancestry at dense locations along a chromosome, these large data sets offer opportunities to reconstruct the ancestry blocks; in other words, we can infer ancestry even at locations between markers. At the same time, however, high-density genotype data pose a major obstacle for HMM-based analytic approaches. The basic assumption of an HMM, which makes it computationally tractable, is that the observed states are independent conditional on the hidden state (see the "Methods" section). In genetic terms, this amounts to requiring the alleles to be independent, given the ancestral state. Clearly, these assumptions are violated when the marker map is dense and linkage disequilibrium (LD) exists within an ancestral population. Several authors have pointed out that this type of LD, referred to as "background LD," poses a problem for HMM-based models.[8,17,18] However, modeling haplotype structure within each ancestral population is computationally intractable.[8]

In this article, we propose an extended model, which we refer to as the "Markov–hidden Markov model" (MHMM), that accounts for background LD without a great sacrifice in computational efficiency. With phased data or the X chromosome in males, our algorithm infers ancestry blocks. If only unphased genotypes are available, we reconstruct diploid ancestry blocks; as we explain below, this means that we infer ancestry blocks up to a permutation of phase. Our simulation illustrates that the genotyping of markers at a density comparable to Affymetrix's 100K SNP chip allows accurate inference of diploid ancestry blocks; at this density, however, background LD must be accounted for. We envision that the MHMM will prove useful in a variety of analyses of high-density SNP genotype data. In the area of disease association studies, our approach makes it possible to perform admixture mapping by use of any high-density genotyping platform. In the "Discussion" section, we explain why this is important.

Methods

This section describes the population model and statistical methods for estimating ancestry along a chromosome.

Data and Biological Model

We assume that each admixed individual is genotyped at T linked biallelic SNPs on a chromosome and that the recombination distance between consecutive markers, d_t, $t = \{2,3,\ldots,T\}$ (in Morgans), is known without error. Further, we assume individuals representing each of N ancestral populations have been genotyped at the corresponding marker loci, and, on the basis of these genotypes, we infer ancestral allele frequencies. The importance of including these individuals is discussed by Tang et al.[19] We present methods for both phased and unphased data. However, to facilitate the exposition, we lay out the conceptual framework assuming genotypes are phased—that is, haplotypes are available. Our method for phased data may apply in a few special situations, such as in studying the X chromosome in males. Additionally, when samples are analyzed from parents-offspring trios, in which all individuals are genotyped, a majority of marker loci can be phased unambiguously. Markers at which both the parents and the child are heterozygous cannot be phased with certainty; however, chromosomal phase can often be inferred with high confidence on the basis of genotypes at neighboring markers.

Our primary goal is to recover the unobservable ancestry along the chromosomes. As described above, in an individual with recent admixture, we can imagine his or her genome as a mosaic of ancestry blocks. Since the resolution of admixture analyses depends on the length of these ancestral chromosome blocks,[4] we are also interested in examining the variation in block sizes among individuals. For an admixed population with more than two ancestral populations, we expect the distribution of block size to differ depending on the ancestral state, because the indigenous populations may have come into contact at different times. As we will explain below, one important parameter in our model is $\tau = \{\tau_1,\ldots,\tau_N\}$, where the inverse of τ_i reflects the average length of chromosome blocks derived from ancestral population i. We estimate τ for each individual. If, in a person's genealogy, gene flow from each ancestral population occurs in a single generation, then $\hat{\tau}$ is an estimate of the time (in generations) since admixing.[8] Since gene flow may have occurred over many generations continuously, one should be cautious about equating τ with the admixing time. Nonetheless, this parameter provides some information regarding average time of gene flow.

The MHMM

Let $\{O_t^f\}_{t=1}^T$ denote a haplotype of observed alleles along a chromosome, say the paternally inherited chromosome of an admixed individual; correspondingly, denote the unobservable ancestral states along this chromosome as $\{Z_t^f\}_t$. The maternally inherited haplotype and its corresponding ancestral states can be similarly defined and are denoted as $\{O_t^m\}_t$ and $\{Z_t^m\}_t$, respectively. Conditional on model parameters, we model the ancestral states along the paternal and the maternal chromosomes as two independent and identical Markov processes. We wish to point out that this model is only approximate. First, because of the constraints imposed by an underlying genealogy, the process along each chromosome is not Markovian.[20] Second, the paternal side of the genealogy and the maternal side of the genealogy may have different levels of admixture, and, therefore, the two processes are not necessarily identical. Finally, we assume that matings are random with respect to ancestry, an assumption that may be violated in some populations. Future work may allow modeling of asymmetric and nonrandom admixing history in a pedigree. For unphased data, we will use the shorthand notation $O_t = \{g^1, g^2\}$ to denote the *unordered* genotypes and $Z_t = \{Z_t^f, Z_t^m\}$ to denote the *ordered* ancestral states combination. Because we analyze each individual independently, we do not need the index for an individual.

The MHMM, with which we propose to model the relationship between the unobservable ancestral states and the observed haplotype along each chromosome, is an example of a Markov-switching model.[21] As illustrated in figure 1*a,* in an HMM, the observed states, O^f, are conditionally independent given the underlying unobservable states, Z^f—that is,

$$\mathbf{P}(O^f_t \mid Z^f_1, \dots, Z^f_t, O^f_1, \dots, O^f_{t-1}) = \mathbf{P}(O^f_t \mid Z^f_t) \ .$$

In contrast, in a Markov-switching model (compare fig. 1*b*), the observed state $O^f_{t^*}$ depends not only on $Z^f_{t^*}$ but also on the past history, $\{Z^f_t\}_{t<t^*}$ and $\{O^f_t\}_{t<t^*}$. Ideally, we would model the background haplotype structure within each ancestral population by allowing $O^f_{t^*}$ to depend on the entire past history. As such a model becomes computationally intractable, we make a compromise and consider only the first-order Markovian dependency along a haplotype. Thus,

$$\mathbf{P}(O^f_t \mid Z^f_1, \dots, Z^f_t, O^f_1, \dots, O^f_{t-1}) = \begin{cases} \mathbf{P}(O^f_t \mid Z^f_t, O^f_{t-1}) & \text{if } Z^f_t = Z^f_{t-1} \\ \mathbf{P}(O^f_t \mid Z^f_t) & \text{otherwise} \end{cases} \ .$$

In other words, if the ancestral state switches between markers $t-1$ and t, the probability of the observed allele depends on only the ancestral allele frequencies at marker t. On the other hand, if the ancestral states do not change between markers $t-1$ and t, then the probability of observing an allele is proportional to the ancestral two-marker haplotype frequency.

As in an HMM, three sets of parameters specify the MHMM: the initial-states distribution (π), the transition matrices ($\mathcal{A} = \{\mathcal{A}_t\}$), and the emission probabilities ($\mathcal{B}_t$). For simplicity, we will denote $\lambda = \{\mathcal{A}, \mathcal{B}, \pi\}$. The initial-states distribution and the transition matrices specify the distribution and conditional distribution of the hidden variables. Falush et al.,[8] for example, adopted the following initial-states distribution and transition probabilities: $\mathbf{P}(Z_1 = i \mid \pi) = \pi_i$, $(i = 1, \dots, N)$, and, for $1 < t \leqslant T$,

$$\begin{aligned} A^{\text{struct}}_{ij}(t) &= \mathbf{P}(Z_t = j \mid Z_{t-1} = i, \tau, \pi) \\ &= \begin{cases} \exp(-d_t\tau) + \pi_j[1 - \exp(-d_t\tau)] & i = j \\ \pi_j[1 - \exp(-d_t\tau)] & \text{otherwise} \end{cases} \ , \end{aligned} \tag{1}$$

where a multinomial probability vector π represents the genomewide average admixture of the individual. Under a simple intermixing model and when d_t is measured in Morgans, τ has the interpretation of the time since admixing.[8] In the "Transition Matrix" section, we discuss how we formulate a transition matrix that allows multiple admixing times.

In an HMM, the emission probability describes the distribution of O^f_t given Z^f_t. A natural choice of emission probabilities at a marker is the allele frequencies in each ancestral population. In the MHMM, we require additionally the joint distribution of alleles at two neighboring markers. The emission probability at marker t is defined by

$$\begin{aligned} \mathcal{B}_t(v,u,j,i) &= \mathbf{P}(O^f_t = v \mid O^f_{t-1} = u, Z^f_t = j, Z^f_{t-1} = i) \\ &= \begin{cases} \tilde{B}_{j,t}(v,u) & \text{if } i = j \\ \breve{B}_{j,t}(v) & \text{otherwise} \end{cases} \ , \end{aligned} \tag{2}$$

where $\breve{B}_{j,t}(v)$ denotes the frequency of allele v in ancestral population j, whereas $\tilde{B}_{j,t}(v,u)$ denotes the probability of observing allele v at marker t, conditioned on observing allele u at marker $t-1$, given that both alleles are derived from ancestral population j.

Efficient computational algorithms have been developed for HMMs and include (1) the forward algorithm, which computes the likelihood of a parameter set given the observed data; (2) the backward algorithm, which, combined with the forward algorithm, estimates the posterior distribution of the hidden state at each observation; (3) the Viterbi algorithm, which searches for the sequence of hidden states that is *jointly* most likely; and (4) the Baum-Welch method, an expectation-maximization (EM)–based algorithm for estimating the model parameters. An excellent tutorial with examples can be found in the work of Rabiner.[22] In the following sections, we explain how to adapt the forward and backward algorithms to compute the likelihood of a parameter set, to estimate the posterior probability of the hidden states in the MHMM, and to sample the sequences of hidden states according to the posterior likelihood.

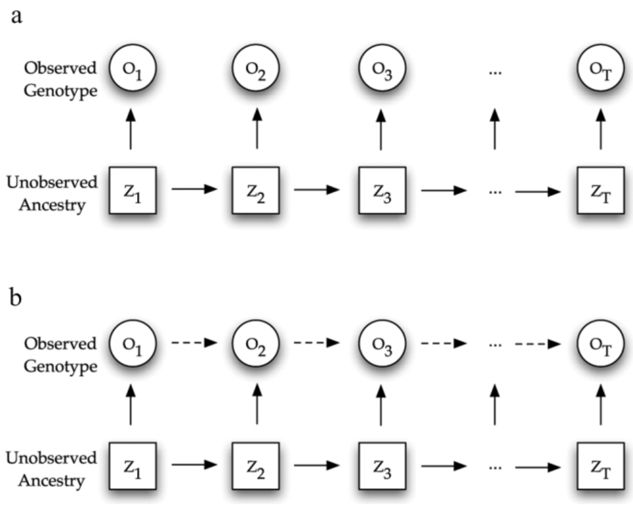

Figure 1. Graphical representation of an HMM (*a*) and an MHMM (*b*).

Likelihood Computation

This section describes modified forward algorithms, which enable us to compute the log likelihood, ℓ, of a parameter set, λ, given genotype data (phased or unphased) on a chromosome:

$$\ell(\lambda \mid O_1, \dots, O_T) = \log[\mathbf{P}(O_1, \dots, O_T, \mid \lambda)] \ . \tag{3}$$

First, let us assume that phase information is available and that, conditional on λ, $\{Z^f_t\}_t$ and $\{Z^m_t\}_t$ are independent:

$$\ell(\lambda \mid O_1, \dots, O_T) = \ell(\lambda \mid O^f_1, \dots, O^f_T) + \ell(\lambda \mid O^m_1, \dots, O^m_T) \ .$$

The forward algorithm for computing $\ell(\lambda \mid O^f_1, \dots, O^f_T)$ closely resembles the corresponding algorithm for an HMM.[23]

Algorithm 1: forward algorithm for phased data.—Define $\alpha^f_t(i) = \mathbf{P}(O^f_1, \dots, O^f_t, Z^f_t = i \mid \lambda)$. These variables are computed inductively in three steps.

1. Initialization.

$$\begin{aligned} \alpha^f_1(i) &= \mathbf{P}(O^f_1 \mid Z^f_1 = i) \\ &= \breve{B}_{i,1}(O^f_1)\pi_i \ . \end{aligned}$$

2. Induction. For $1 < t \leqslant T$,

$$\begin{aligned}\alpha_t^f(j) &= \sum_i \mathbf{P}(O_1^f, \ldots ,O_t^f, Z_t^f = j, Z_{t-1}^f = i \mid \lambda) \\ &= \sum_i \mathbf{P}(O_1^f, \ldots , O_{t-1}^f, Z_{t-1}^f = i \mid \lambda)\mathbf{P}(Z_t^f = j \mid Z_{t-1}^f = i) \\ &\qquad \mathcal{B}_t(O_t^f,O_{t-1}^f,j,i) \\ &= \sum_{i \neq j} \alpha_{t-1}^f(i)\mathcal{A}_{ij}\breve{B}_{j,t}(O_t^f) + \alpha_{t-1}^f(j)\mathcal{A}_{jj}\tilde{B}_{j,t}(O_t^f, O_{t-1}^f) \ , \end{aligned} \tag{4}$$

where $\mathcal{A}_{ij}$ stands for the shorthand notation $\mathcal{A}_{ij}(t-1)$.

3. Termination. The likelihood of the parameters can be computed by $\sum_i \alpha_T^f(i)$.

To improve numerical stability, we compute the induction step using a rescaled version of α_t^f that sums to 1 and denote the left-hand side in equation (4) as $\tilde{\alpha}_t^f$. Let $\text{scale}_t^f = \sum_i \tilde{\alpha}_t^f(i)$. It can be shown that the log likelihood is:

$$\ell(\lambda \mid O_1^f, \ldots ,O_T^f) = \sum_{1 \leqslant t \leqslant T} \log(\text{scale}_t^f) \ .$$

To analyze unphased genotype data in a diploid organism, we need to keep track of the phase between consecutive pairs of markers. We introduce a set of variables, X_t. Recall $\{g^1,g^2\}$ denotes the (arbitrarily) ordered pairs of alleles at a marker, and O^m and O^f indicate the maternally and paternally inherited alleles. Then, define

$$X_t = \begin{cases} 1 & \text{if } O_t^m = g^1 \text{ and } O_t^f = g^2 \neq g^1 \\ 0 & \text{if } O_t^m = g^2 \text{ and } O_t^f = g^1 \neq g^2 \text{ or if } O_t \text{ is homozygous} \end{cases} \ .$$

Note that $X_t = 0$ if the genotype at marker t is homozygous ($g^1 = g^2$). Algorithm 1 can be modified to compute the likelihood in equation (3). Define

$$\alpha_t(x,i,j) = \mathbf{P}(O_1, \ldots ,O_t, X_t = x, Z_t^m = i, Z_t^f = j \mid \lambda) \ .$$

These variables are computed in three steps:

1. Initialization.

$$\begin{aligned}\alpha_1(0,i,j) &= \mathbf{P}(g^1 \mid Z_t^m = i)\pi_i P(g^2 \mid Z_t^f = j)\pi_j \\ &= \breve{B}_{i,1}(g_1^1)\breve{B}_{j,1}(g_1^2)\pi_i\pi_j \ ,\end{aligned}$$

and

$$\alpha_1(1,i,j) = \begin{cases}\breve{B}_{i,1}(g_1^2)\breve{B}_{j,1}(g_1^1)\pi_i\pi_j & \text{if } O_1 \text{ is heterozygous} \\ 0 & \text{otherwise}\end{cases} \ .$$

2. Induction. For $1 < t \leqslant T$,

$$\begin{aligned}\alpha_t(0,k,l) &= \sum_i \sum_j \sum_{x \in \{0,1\}} \mathbf{P}(O_1, \ldots ,O_t,X_t = 0, \\ &\qquad Z_t = \{k,l\}, Z_{t-1} = \{i,j\}, X_{t-1} = x) \\ &= \sum_i \sum_j [T_{t,0,0}(i,j,k,l) + T_{t,0,1}(i,j,k,l)] \ ,\end{aligned}$$

where

$$\begin{aligned}T_{t,0,0}(i,j,k,l) &= \mathbf{P}(O_1, \ldots ,O_t, X_t = 0, Z_t, Z_{t-1}, X_{t-1} = 0) \\ &= \mathcal{B}(g_t^1, g_{t-1}^1, k, i)\mathcal{B}(g_t^2, g_{t-1}^2, l, j)\mathcal{A}_{ik}\mathcal{A}_{jl}\alpha_{t-1}(0, i, j) \ ,\end{aligned}$$

and, when O_{t-1} is heterozygous,

$$\begin{aligned}T_{t,0,1}(i,j,k,l) &= \mathbf{P}(O_1, \ldots ,O_t,X_t = 0,Z_t,Z_{t-1},X_{t-1} = 1) \\ &= \mathcal{B}(g_t^1, g_{t-1}^2, k, i)\mathcal{B}(g_t^2, g_{t-1}^1, l, j)\mathcal{A}_{ik}\mathcal{A}_{jl}\alpha_{t-1}(1, i, j) \ .\end{aligned}$$

If O_{t-1} is homozygous, $T_{t,0,1}(i,j,k,l) = 0$. When O_t is heterozygous, we compute $\alpha_t(1,k,l)$ in a similar fashion; otherwise, this term is simply 0.

3. Termination. As in the algorithm for the phased data, we define a scaled α-matrix in the induction for numerical stability,

$$\text{scale}_t = \sum_i \sum_j \sum_{x \in \{0,1\}} \alpha_t(x, i, j) \ ,$$

and compute the log likelihood of the parameter by

$$\ell(\lambda \mid O_1, \ldots ,O_T) = \sum_{1 \leqslant t \leqslant T} \log(\text{scale}_t) \ .$$

In genomewide association studies and admixture mapping studies, genotypes are often available from all chromosomes. Under the assumption that the hidden processes on all chromosomes are generated independently by identical parameters, the log likelihood computed on each chromosome can be summed. The parameter, τ, approximates the average time since admixing and is of particular interest in admixture studies. Assuming other parameters are known without error, we can use a grid search or the Newton-Raphson algorithm to find the maximum-likelihood estimates (MLEs) of τ.

Posterior Probability of Ancestral States

For phased data, we estimate the *marginal* posterior probability that an allele (say, the paternally inherited allele) originates from a specific ancestral population. Our approach to computing these probabilities is an extension of the computation for an HMM.[22] Define

$$\beta_t^f(i) = \begin{cases} 1 & \text{if } t = T \\ \mathbf{P}(O_{t+1}^f, \ldots ,O_T^f \mid Z_t^f = i, O_t^f) & \text{otherwise} \end{cases} \ .$$

We then compute the posterior probability at each allele by

$$\begin{aligned}\gamma_t^f(i) &= \mathbf{P}(Z_t^f = i \mid O^f, \lambda) \\ &\propto \alpha_t^f(i)\beta_t^f(i) \ .\end{aligned}$$

The α^f-matrix is computed using algorithm 1, described in the previous section. Analogously, we modify the backward algorithm to compute the β^f-matrix.

For unphased data, we estimate the posterior probability that a randomly chosen allele at marker t has ancestry from a specific population. Define

$$\beta_t(x, i, j) = \mathbf{P}(O_{t+1}, \ldots, O_T \mid Z_t = \{i, j\}, O_t, X_t = x)$$

and

$$\begin{aligned}\gamma_t(x, i, j) &= \mathbf{P}(Z_t = \{i, j\}, X_t = x \mid O, \lambda) \\ &\propto \alpha_t(x, i, j)\beta_t(x, i, j) \ .\end{aligned}$$

The marginal posterior probability for an allele is computed by

$$\mathbf{P}(\overline{Z_t} = i^* \mid O,\lambda) = \frac{1}{2}\sum_j \sum_{x \in \{0,1\}} \gamma_t(x, i^*, j) + \frac{1}{2}\sum_j \sum_{x \in \{0,1\}} \gamma_t(x, j, i^*) \ .$$

The quantity $\mathbf{P}(\overline{Z_t} = i^* \mid O,\lambda) - \pi_i$ represents the excess ancestry at marker t. Several admixture mapping approaches aim to locate markers at which this quantity deviates from zero in affected individuals but not in healthy controls.[3,5,6]

Posterior Sample of Ancestry Blocks

In HMM literature, the Viterbi algorithm was developed to find the single best-state sequence. In phased data, this is the sequence of ancestral states, which jointly achieves the maximum likelihood given a haplotype. In practice, however, this sequence does not capture all the information; we may want to know, for example, whether there are many other likely sequences of states. For unphased data, an additional complication arises that one cannot unambiguously phase the ancestral states. To see this, suppose the true ancestral sequences along the two haplotypes are $\{ABA\}$ and $\{BBB\}$, where A and B denote the two ancestral populations. By the Markov property, the true ancestral sequences cannot be distinguished from the configuration of $\{ABB\}$ along one haplotype and $\{BBA\}$ on the other. This makes it difficult to study, for example, the length of ancestral chromosome blocks. To overcome this difficulty and to gain additional information about the likelihood surface, we choose to sample ancestral sequences from the posterior distribution; in fact, because we put a noninformative prior on all possible ancestral sequences, the single most likely ancestral sequence configuration selected by the Viterbi algorithm is the posterior mode. In this section, we describe an algorithm for sampling sequences of ancestral states according to the posterior probability of the entire sequence.

As before, we first consider phased data. This algorithm bears close resemblance to the backward Gibbs sampling step in STRUCTURE.[8] To begin, sample Z_T according to the distribution $\mathbf{P}(Z^f_T = j) \propto \alpha^f_T(j)$. Subsequently, iteratively sample Z_t according to

$$\begin{aligned}&\mathbf{P}(Z^f_t = i \mid Z^f_{t+1} = j, \ldots, Z^f_T, O^f_1, \ldots, O^f_T) \\ &\propto \mathbf{P}(O^f_1, \ldots, O^f_t, Z^f_t = i)\mathbf{P}(Z^f_{t+1} = j \mid Z^f_t = i) \\ &\quad \mathbf{P}(O^f_{t+1} \mid O^f_t, Z^f_t = i, Z^f_{t+1} = j) \\ &= \alpha^f_t(i)\mathcal{A}_{ij}\mathcal{B}(O^f_{t+1}, O^f_t, j, i) \ .\end{aligned}$$

For unphased data, we sample

$$\mathbf{P}(Z_T = \{i,j\}, X_T = x) \propto \alpha_T(x,i,j) \ ,$$

and, subsequently,

$$\begin{aligned}&\mathbf{P}(Z_t = \{i,j\}, X_t = x \mid Z_{t+1} = \{k,l\}, \ldots, Z_T, \\ &\quad X_{t+1} = x', \ldots, X_T, O_1, \ldots, O_T) \\ &\propto \alpha_t(x,i,j)\mathcal{A}_{ik}\mathcal{A}_{jl}\mathbf{P}(O_{t+1} \mid O_t, Z_t = \{i,j\}, \\ &\quad Z_{t+1} = \{k,l\}, X_t = x, X_{t+1} = x') \ .\end{aligned} \tag{5}$$

The last term in equation (5) is the emission probability, which depends on the phase indicators, X_t and X_{t+1}, and can be evaluated in a similar fashion as we computed the $T_{x,i,j}$ terms in the modified forward algorithm.

Transition Matrix

The transition matrix models the probability with which the ancestry switches between two consecutive markers. The transition matrix implemented in STRUCTURE[8] models a simple intermixing process, which assumes that all chromosomes in the sampled admixed subjects descended from a mixed group of ancestral chromosomes g generations ago, who have subsequently mated randomly.[24] Under this model, the transition matrix specified in equation (1) has several appealing properties: it guarantees that the stationary distribution of the Markov chain coincides with the genome-average individual admixture (IA); it applies for an arbitrary number of ancestral populations; and, when intermarker distance is measured in Morgans, the parameter τ has an approximate interpretation as the admixing time, g. The transition matrix that represents a continuous gene-flow model has been worked out by Zhu et al.[6] The result, however, applies only to the two-ancestral population case and becomes cumbersome to derive as the number of populations increases.

Here, we extend the transition matrix of Falush et al.[8] to reflect different admixing times for N ($N \geqslant 3$) parental populations. Let τ_n, $n \in 1, \ldots, N$, be the inverse of the expected length of the chromosome blocks that are derived from ancestral population n. Define the N-by-N matrix Q by

$$Q_{i,j} = \begin{cases} \pi_i \dfrac{\tau_i^2}{\sum\limits_{n=1}^{N} \pi_n\tau_n} - \tau_i & \text{if } i = j \\ \pi_j \dfrac{\tau_i\tau_j}{\sum\limits_{n=1}^{N} \pi_n\tau_n} & \text{otherwise} \end{cases} \ .$$

Q_{ij} represents the *instantaneous* rate of transition from ancestral state i to j. Our formulation of the transition rate is based on two observations. First, given the current state i, the waiting time to the first jump (point of recombination that may lead to a change in ancestral state) follows an exponential distribution with an expectation inversely proportional to the number of meioses since admixing (τ_i). Second, holding the stationary distribution, π, constant, the probability of switching into a given state should be inversely related to the expected length of time that the Markov process stays in that state. Therefore, we choose the new state with a probability proportional to $\pi_i\tau_i$. The stationary distribution, taken as the genome-average ancestry, can be estimated jointly with other parameters. However, for high-density genotype data, in which many markers are tightly linked, it is computationally more efficient to estimate the stationary distribution by using a subset of weakly linked markers and existing methods[8,4,19] (X.

Zhu, S. Zhang, H. Tang, and R. Cooper, unpublished data). Therefore, in the simulations below, we assume that individual admixtures are known. Let d be the distance (in Morgans) between two markers. The transition matrix is then computed by matrix exponentiation[25]:

$$\mathcal{A}(d) = \exp(-d \times Q) \ . \qquad (6)$$

It can be shown that $\mathcal{A}$ retains all the appealing features of equation (1) but is more flexible to permit the average length of a chromosome block to depend on its ancestry. In the case $\tau_1 = \tau_2 = \ldots = \tau_N$, matrix $\mathcal{A}$ simplifies to equation (1).

Estimation of Ancestral Haplotype Frequencies

The computation of the forward (α) and the backward (β) matrices requires, for the emission probabilities, both the ancestral allele frequency and two-marker haplotype frequencies, (g_1,g_2). In this section, we explain how to estimate these frequencies.

To estimate ancestral-allele frequencies, we can simply count alleles in each ancestral population. However, because the number of ancestral individuals genotyped is often limited, the sampling variance of these estimates can be large. Incorporating genotypes from the admixed individuals increases the information on those frequencies. For example, STRUCTURE uses a Gibbs step to update the ancestral allele frequency estimates.[8,26] Alternatively, X. Zhu, S. Zhang, H. Tang, and R. Cooper (unpublished data) and Tang et al.[19] suggest updating these frequencies via an EM algorithm.[27] All these methods produce more-accurate allele frequency estimates. Furthermore, several large genotyping projects are underway, including the HapMap project[28] and the ALFRED[29] database, and we expect rapid improvements in the estimates of population-specific allele frequencies.

Similarly, we can estimate the two-marker haplotype frequencies by using the ancestral individuals alone. Various methods have been proposed to estimate haplotype frequencies from unphased population genotype data.[16,30–34] Again, such estimates have large sampling errors because of the limited number of ancestral individuals. The problem is especially prominent when one or both SNPs have rare alleles. For example, within a large ancestry block, observing a single two-marker haplotype in an admixed individual that is absent in the corresponding ancestral population would force an abrupt change in ancestral state. The absence of the allele in the ancestral population may be the result of the sheer paucity of ancestral individuals examined. In theory, as for the allele frequency estimates, we could also improve the haplotype frequency estimates by using either the EM algorithm or a Gibbs sampling method, which would incorporate the genotypes in the admixed individuals. This, however, is computationally expensive. We choose an alternative approach by observing that there is often richer information on ancestral allele frequency than on haplotype frequency. As we explained in the previous paragraph, more-accurate allele frequency estimates either can be computed jointly on ancestral and admixed individuals or may be obtained from external sources. In other words, in the notation illustrated in the tabulation below, we assume the allele frequencies $p_{1\cdot}$, $p_{2\cdot}$, $p_{\cdot 1}$, and $p_{\cdot 2}$ to be known from a larger data set. We then model the observed ancestral haplotype counts n_{11}, n_{12}, n_{21}, and n_{22} as a sample from an underlying multinomial distribution, whose parameter is of interest.

	SNP 2 Allele		
SNP 1 Allele	B	b	
A	n_{11}	n_{12}	$p_{1\cdot}$
a	n_{21}	n_{22}	$p_{2\cdot}$
	$p_{\cdot 1}$	$p_{\cdot 2}$	N

Because we consider the marginal frequencies to be fixed, there is only one unknown parameter in the model, which is the LD parameter $D = \mathbf{P}_{AB} - \mathbf{P}_A\mathbf{P}_B$. Thus, we compute $\tilde{B}(B,A)$ by $\breve{B}(B) + \hat{D}/\breve{B}(A)$, where $\tilde{B}$ and $\breve{B}$ are the conditional frequency and the marginal frequency, respectively, of the B allele defined in equation (2). This is likely to improve the haplotype frequency estimates. Because the estimate of the LD parameter D tends to have an upward bias in small samples,[35] we introduce a shrinkage procedure. We assume that a number, c, of haplotypes have been observed a priori, which falls into the four cells in the tabulation above according to linkage equilibrium. Thus, we seek D that maximizes the likelihood of the multinomial data, $n_{11} + cp_{1\cdot}p_{\cdot 1}$, $n_{12} + cp_{1\cdot}p_{\cdot 2}$, $n_{21} + cp_{2\cdot}p_{\cdot 1}$, and $n_{22} + cp_{2\cdot}p_{\cdot 2}$. In our simulations, we take $c = 5$. For fixed c, the shrinkage becomes negligible as the sample size N increases; for a fixed sample size N, increasing c shrinks D closer toward 0. Note that, if we ignore background LD and let $D = 0$, the MHMM is reduced to a standard HMM.

Simulations

Simulation 1.—The first simulation aims to illustrate the advantage of the haplotype frequency estimation procedure described in the previous section. We generated a large haplotype pool by resampling haplotypes of chromosome 22 in the 60 unrelated European parents (CEPH individuals from Utah [CEU]) genotyped in the HapMap project.[28] The observed haplotype frequencies are taken as the underlying truth. Next, we created 50 diploid and unphased individuals by sampling 100 haplotypes from the haplotype pool. We then compare two approaches for estimating the two-marker haplotype frequencies. The naive method uses an EM algorithm and jointly estimates allele frequencies and haplotype frequencies from the 50 individuals. In the second approach, we assume the allele frequencies at both markers are known without error and use the EM algorithm to estimate LD, as described in the previous section. We then compare both estimates with the true sampling frequencies.

Simulation 2.—Next, we examine the importance of modeling background LD, using a combination of simulated and real data. For the simulation, we consider an admixed population with three ancestral populations: two populations admixed 25 generations ago and a third ancestral population introduced 10 generations ago. Underlying ancestral states along the genome were generated according to a Markov chain, the transition matrix of which is given by equation (6). To simulate the observed genotypes, we sample from the phased data produced by the HapMap project. This way, our simulated data incorporates a realistic level of high-order dependency among linked markers, and we have the opportunity to examine whether the MHMM is adequate. The three ancestral populations consist of 120 European chromosomes (CEU), 120 African chromosomes (Yoruba), and 178 East Asian chromosomes (90 Han Chinese and 88 Japanese). We then scan along the simulated ancestry sequence, identifying segments of the genome in which the ancestry does not change. For

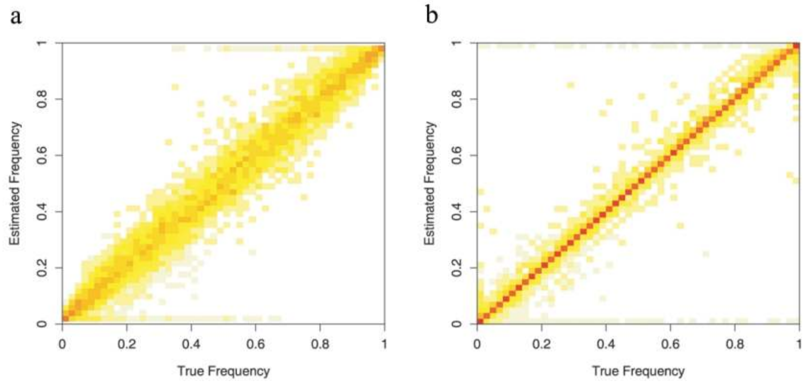

Figure 2. Estimation of two-marker haplotype frequency estimation. Unphased genotype data in 50 individuals were simulated on the basis of chromosome 22 haplotypes of the CEU individuals genotyped in the HapMap project. Each plot can be viewed as a two-dimensional histogram, in which the *X*-axis represents the true haplotype frequency, and the *Y*-axis represents the corresponding estimated frequencies. The intensity at each pixel indicates the height of the histogram, or the number of marker pairs whose true haplotype frequency is at the *X*-coordinate while the estimated haplotype frequency is at the *Y*-coordinate. *a,* Naive haplotype frequency estimates. Both allele frequencies and haplotype frequencies are estimated from a small sample of individuals. *b,* Augmented haplotype frequency estimates. Haplotype frequencies were estimated from same set of individuals as in panel a, but allele frequencies were estimated from a larger sample.

each of these segments, a segment of a haplotype is sampled independently from an individual from the corresponding genomic region and ancestral population. Markers are chosen at a density comparable to that in the Affymetrix 100K SNP chip, with an average spacing of 30 kb. In our analysis, we eliminated any marker that was either in complete LD with its left neighbor or within 10-kb distance to its left neighbor; dropping such markers reduces computation time without losing much ancestry information. The ancestral allele frequencies are estimated under both the HMM and the MHMM, by use of the unphased HapMap genotypes. The two-marker haplotype frequencies are inferred from the same ancestral individuals. MLEs of admixing times, τ, are computed by evaluating the likelihood, over a dense grid, by use of the modified forward algorithm. Similarly, we compute the MLEs under the HMM. Posterior ancestry estimates are obtained according to both the HMM and the MHMM. Under the MHMM, we also obtained 10 posterior samples of ancestry sequences.

Simulation 3.—We hypothesize that, as the markers become more densely located, the impact of background LD becomes more prominent. To test this hypothesis and to understand the adequacy of the MHMM for analyzing denser marker sets, we randomly sampled 100K markers from a Han Chinese individual genotyped by the HapMap project. This individual is removed from the ancestral individuals when ancestral allele and haplotype frequencies are estimated. Posterior mean ancestry was estimated assuming IA proportions of (1/3,1/3,1/3) and $\tau =$ (25,25,25). The experiment was repeated for a randomly sampled panel of 500K markers and for the complete set of HapMap markers.

Simulation 4.—As we discussed in the "Transition Matrix" section, the admixing model from which our method is derived represents a simplification of the historical process. Therefore, the final simulation provides an example illustrating how our proposed ancestry-block-reconstruction approach performs when the data-generating mechanism deviates from the assumed model. In this simulation, we assume that admixing occurred 25 generations ago in the paternal lineage with ancestry proportions of 0.4, 0.4, and 0.2, whereas, in the maternal lineage, admixing occurred 2 generations ago with ancestry proportions of 0.75, 0.125, and 0.125. All other parameters are the same as in simulation 2. We obtained the posterior ancestry estimates, assuming various parameter values of τ and π.

Results

Simulation 1

Although inferring haplotype frequencies on the basis of a small number of ancestral individuals produces large sampling errors, the estimates are substantially better when we incorporate external information about allele frequencies at each marker (fig. 2). Each plot can be thought of as a two-dimensional histogram, in which the *X*-axis represents the true haplotype frequency and the *Y*-axis represents the corresponding estimated frequencies. The intensity at each pixel indicates the height of the histogram, or the number of marker pairs whose true haplotype frequency is at the *X*-coordinate while the estimated haplotype-frequency is at the *Y*-coordinate. If the estimated frequencies entirely coincide with the true values, we will see red pixels on the diagonal and white elsewhere. On the other hand, if the estimated frequencies bear no relationship to the truth, all pixels will show the same color intensity. Clearly, the estimated frequencies clusters more tightly around the true values in figure 2*b* (allele frequencies known) than they do in figure 2*a* (allele frequencies unknown).

Simulation 2

Estimating model parameter, τ.—Figure 3 shows the distribution of the MLE of admixing time. Under the MHMM,

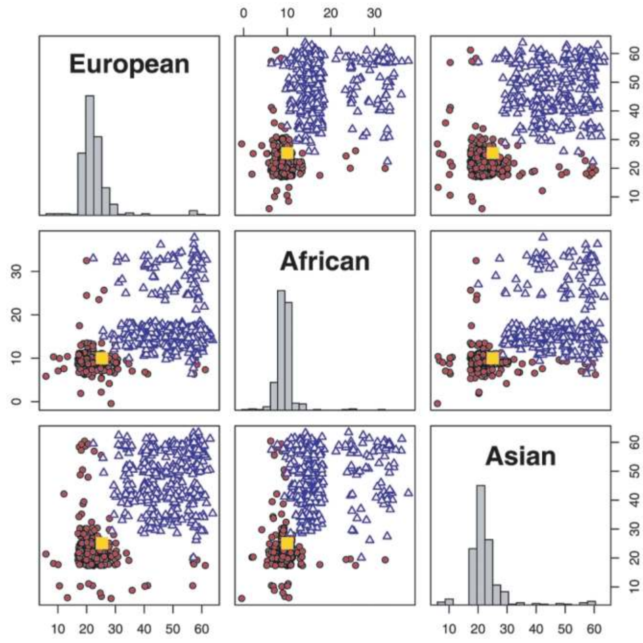

Figure 3. Estimated admixing time, τ, of 400 simulated individuals. Red circles represent the MLE under the MHMM; blue triangles represent the MLE under the HMM by use of the same genotype data. True times are 25, 10, and 25, indicated with a yellow square. Some jitter is added to the MLEs to aid visualization.

the mean estimated admixing times are 23.3, 9.6, and 23.2 generations, respectively, compared with the true parameter values of 25, 10, and 25 generations. In contrast, in ignoring the background LD, an HMM substantially overestimates the times, with mean estimates of 47.5, 17.6, and 43.7 generations, respectively. Note that the comparison is between the MHMM and an HMM algorithm we implemented, which resembles the MHMM in all respects except that it does not account for the background LD. This HMM algorithm we implemented is similar to the core component used in programs such as STRUCTURE,[8] ADMIXMAP,[4] and ANCESTRYMAP[5] but differs in two important aspects. First, these latter programs may have somewhat different parameter estimates, since they iteratively update all model parameters through MCMC algorithms. Second, as we explain in the "Transition Matrix" section, all these programs use only a single τ for all ancestral populations. Because of computational challenges and because our primary goal is to investigate the importance of accounting for background LD, we have not analyzed the simulated data with the use of MCMC-based programs.

A few points in figure 3 appear to have poor estimates under the MHMM. Upon inspection, we find that the likelihood surface of the time parameters are very flat in these individuals. In most cases, the genomewide average ancestry from one population is close to 0 or 1. In the former case, few segments in the person's genome are derived from the corresponding ancestral population; in the latter case, there are few transitions in the underlying ancestral states. Therefore, parameter estimates for an individual with a low level of admixture can be unreliable.

Inferring ancestry of an admixed individual.—Figure 4 shows the posterior mean estimates of the ancestry on chromosome 22 in a simulated individual. The *X*-axis represents the physical locations of the SNP markers. The *Y*-axis is the probability that a randomly sampled allele at that locus has an ancestry from a specific population (blue = European, red = African, and yellow = Asian). The true ancestry is delineated in the top panel; both paternal and maternal copies of the chromosome are largely Asian (yellow), with one chromosome having a small African ancestry block (red) and the other chromosome having a European ancestry block (blue). The middle panel shows the MHMM estimates, and the bottom panel shows the HMM estimates. The MHMM appears to produce more-accurate ancestry estimates than the HMM. For each of the 400 simulated admixed individuals, we compared the mean squared error (MSE) of the posterior estimates produced by the HMM and MHMM. The MSE of the *n*th individual is a sum over all markers:

$$\mathrm{MSE}_n = \sum_t (\hat{p}_{t,1} - p_{t,1})^2 + (\hat{p}_{t,2} - p_{t,2})^2 + (\hat{p}_{t,3} - p_{t,3})^2 \ ,$$

where $(\hat{p}_{t,1}, \hat{p}_{t,2}, \hat{p}_{t,3})$ denote the posterior mean estimates of ancestry at marker t and $p_{t,i}$ represents the true ancestry composition—for example, if one allele at the marker originates from population 1 and the other allele from population 3, then we take $(p_1,p_2,p_3) = (1/2,0,1/2)$. Figure 5 presents a histogram of the MSE reduction by use of the MHMM, compared with use of the HMM—that is, $(\mathrm{MSE}_n^{\mathrm{HMM}} - \mathrm{MSE}_n^{\mathrm{MHMM}})/\mathrm{MSE}_n^{\mathrm{HMM}}$. The reduction appears to be quite striking, ranging from 15% to >70%.

Reconstructing ancestry blocks.—Of 10 posterior samples obtained for this region under the MHMM, all correctly identified the presence of the European and the African blocks, although there is slight ambiguity with respect to the precise locations at which ancestry changes. Posterior samples of the ancestry sequences under the HMM appear more variable, with some samples identifying a spurious European block of ~ 3.3×10^7 bp or ~ 4.2×10^7 bp. However, we wish to point out that, when analyzing unphased genotype data, neither the MHMM nor the HMM resolves the phase of these ancestry blocks; in other words, we cannot distinguish the true block configuration in figure 4 from the one in which both the European (blue) and African (red) blocks resides on one chromosome, while the other chromosome is entirely Asian (yellow). The posterior sampling algorithm described in the "Posterior Probability of Ancestral States" section would choose the two-phase configuration with equal probability; thus, we construct diploid ancestry blocks. Of course, for phased data or X-chromosome data in males, we can construct ancestry blocks with no phase ambiguity.

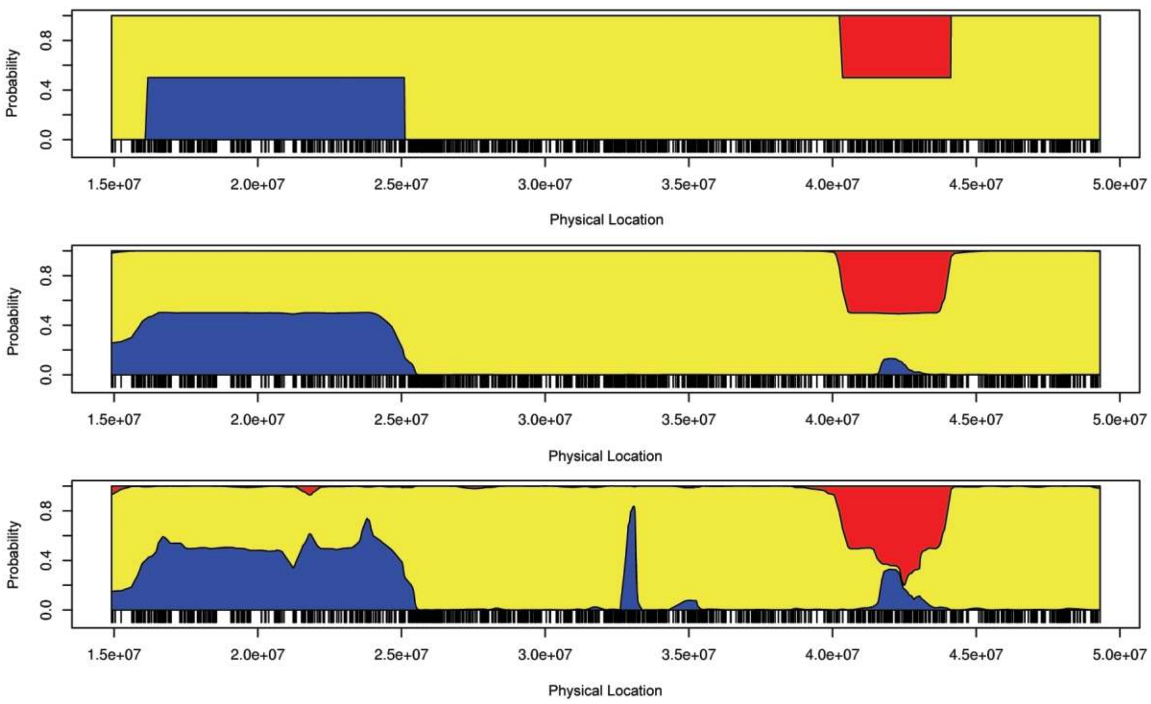

Figure 4. Ancestry for a simulated admixed individual. The *Y*-axis represents the posterior probability that one allele is derived from a specific ancestry; the *X*-axis indicates the physical locations of the markers. *Top,* True ancestral states. *Middle,* MHMM estimates. *Bottom,* HMM estimates.

Simulation 3: Inferring Ancestry of an Indigenous Individual

Figure 6 shows the posterior mean estimates of ancestry for chromosome 22 in a Han Chinese individual from Beijing. The intermarker spacing is 30 kb, 6 kb, and 3 kb for the three rows. The MHMM (left column) estimates predominantly Asian ancestry, as we would expect. This held even when we used all HapMap SNPs and therefore expected the background LD to be quite strong. In contrast, ignoring background LD, the HMM (right column) mistakenly identifies several regions as having European ancestry or African ancestry. Furthermore, the unexpected ancestry switches occur increasingly often as the markers become more densely located. Thus, not accounting for background LD between markers may mislead us to false inferences about mixed ancestry in an indigenous population.

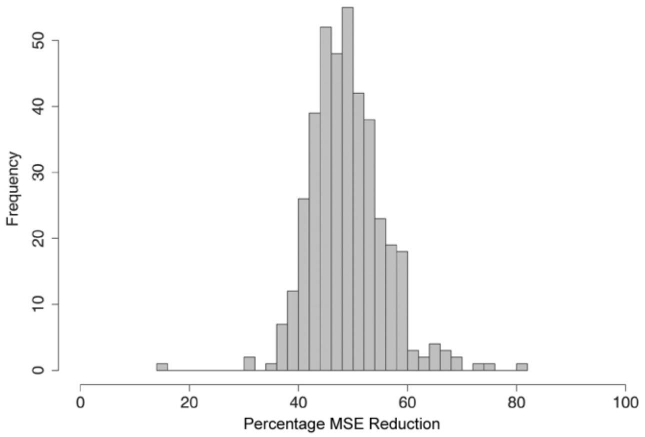

Figure 5. Comparison of percentage reduction in MSE. Percentage reduction for individual n is defined as $(\mathrm{MSE}_n^{\mathrm{HMM}} - \mathrm{MSE}_n^{\mathrm{MHMM}})/\mathrm{MSE}_n^{\mathrm{HMM}}$.

Simulation 4: Robustness to Model Deviation

We simulated ancestry blocks and genotypes in an individual with asymmetric admixing history in the paternal and maternal lineages. The top panel in figure 7 depicts the true ancestry blocks: the paternal chromosome (upper strand) consists of European, African, and Asian blocks, each relatively short, and reflects a longer time since admixing; in contrast, the maternal chromosome is entirely European, reflecting a history of more recent admixing. Subsequent panels in figure 7 present posterior ancestry estimates with various values of the parameter τ. Although the ancestry was simulated using unequal ancestry proportions in the paternal and the maternal chromosomes, we assumed an IA of (1/3,1/3,1/3) in performing the MHMM analyses. Despite the erroneous assumptions about the model and parameter values, the posterior ancestry estimates captured the major blocks accurately. Although this demonstrates the robustness of the MHMM in an example that deviates substantially from the generating model, more-comprehensive insights will be ob-

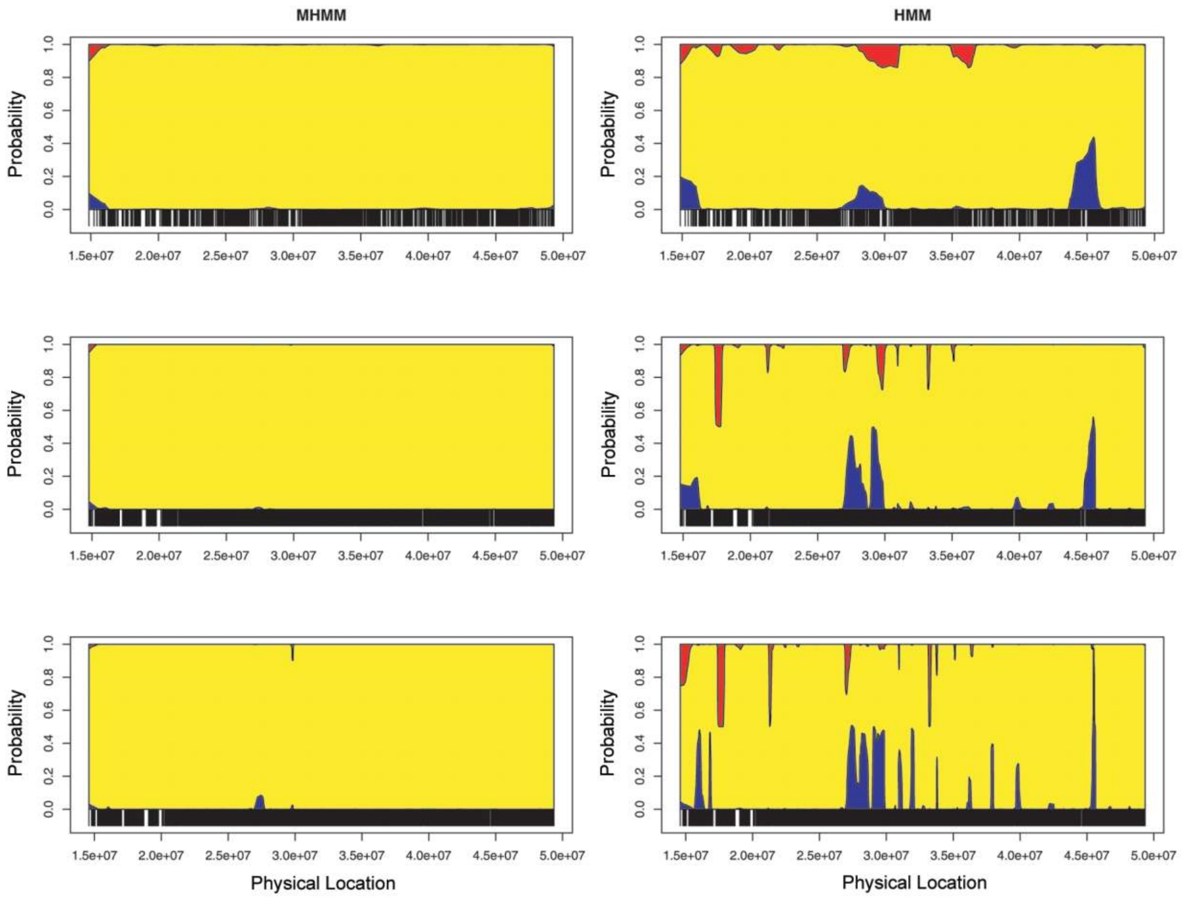

Figure 6. Estimated ancestry for a Han Chinese individual from Beijing. The *Y*-axis represents the posterior probability that one allele is derived from a specific ancestry; the *X*-axis indicates the physical locations of markers. Markers were sampled at an average spacing of 30 kb (*top panels*), 6 kb (*middle panels*), and 3 kb (*bottom panels*), which approximated the density of a 100K SNP chip, approximated the density of a 500K SNP chip, and used all HapMap SNPs, respectively. *Left panels,* MHMM correctly infers Asian ancestry (*yellow*) at most markers. *Right panels,* HMM assigns considerable probability of European ancestry (*blue*) or African ancestry (*red*) in several regions.

tained through analysis of real genetic data, which are rapidly accumulating.

Discussion

Ancestry inference, whether for mapping disease loci or for conducting gene-association studies, is a critical component of genetic analysis in an admixed population. LD between tightly linked markers within ancestral populations complicates such analyses. One option to circumvent the background LD problem is to eliminate markers that are in LD in each ancestral population. Toward this end, a panel of ancestry-informative markers (AIMs) has been developed for admixture mapping in African Americans. Such a map does not exist for other admixed populations but may become available in the near future. However, as Patterson et al.[5] recognize, admixture mapping cannot replace genotype- or haplotype-based association analyses. First, there is considerable risk in genotyping a large number of AIMs, which are tailored for one special design. The superiority of admixture mapping over conventional association approaches hinges on the assumption that the frequency of the risk allele differs greatly between ancestral populations. While this may sometimes be the case, genetic differentiation between ancestral populations will generally not be sufficiently large.[14] Furthermore, in the event that admixture mapping is not successful, the researchers cannot use the genotype data for conventional analyses, because the AIMs are chosen to eliminate background LD and thus are very far apart.

The estimates of the parameter τ shed light on aspects of admixing history. For example, in the simulation example we presented, $\hat{\tau}_1$ and $\hat{\tau}_3$ are generally greater than $\hat{\tau}_2$, conveying that ancestral population 2 (African) admixed more recently than the other two populations. However, we warn against equating $\hat{\tau}$ with the actual time of admixing. The transition matrix we adopted represents a compromise between realism and model complexity. Although we generalize the transition matrix of Falush et

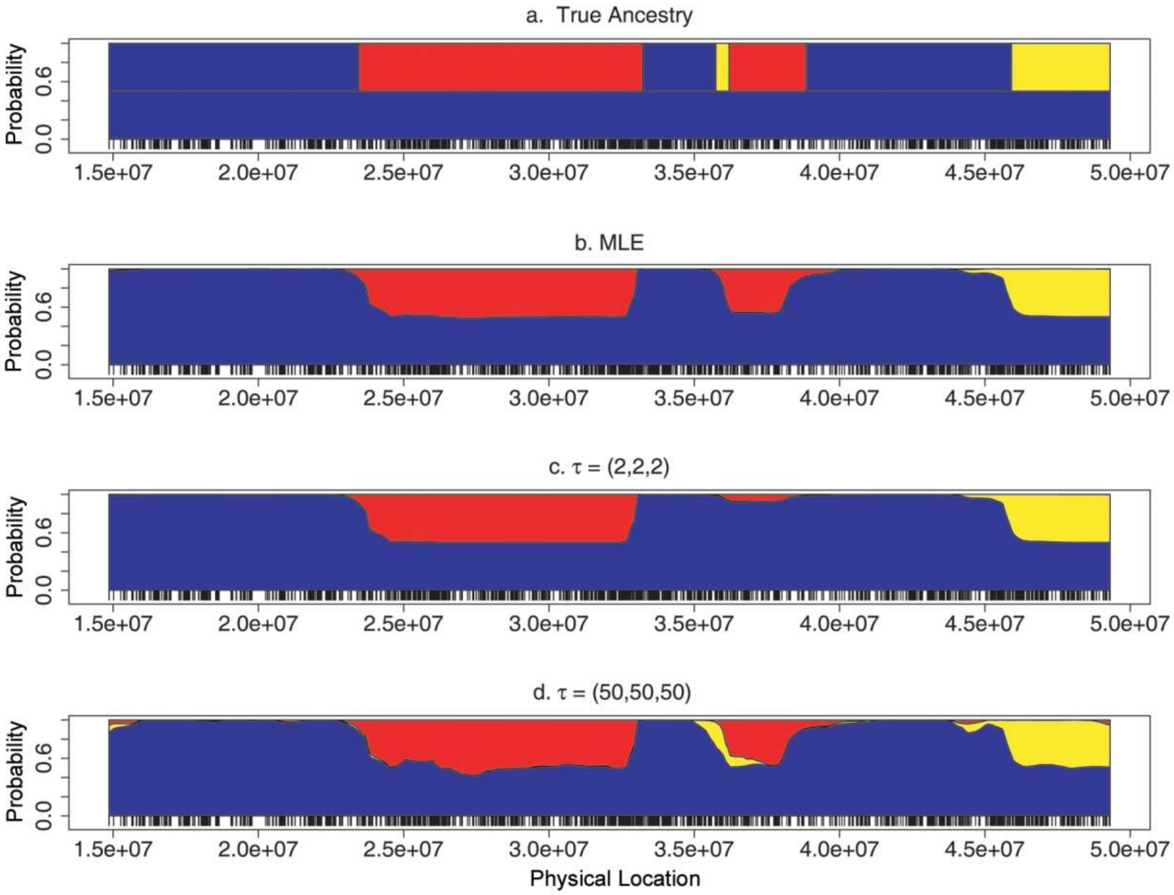

Figure 7. Estimated ancestry for a simulated individual with asymmetric admixing history. The *Y*-axis represents the posterior probability that one allele is derived from a specific ancestry; the *X*-axis indicates the physical locations of markers. *a,* True ancestry along the paternal and the maternal chromosomes. The paternal chromosome was generated assuming $\tau = (25,25,25)$ and $\pi = (0.4,0.4,0.2)$, whereas the maternal chromosome was generated assuming $\tau = (2,2,2)$ and $\pi = (0.75,0.125,0.125)$. *b,* Posterior ancestry estimates at the MLE of τ. *c,* Posterior ancestry estimates under the assumption $\tau = (2,2,2)$. *d,* Posterior ancestry estimates under the assumption $\tau = (50,50,50)$.

al.[8] to allow different admixing times, it nonetheless represents a great simplification of the historical process of admixing, in which the gene flow from each ancestral population may have occurred continuously or intermittently over many generations.

Having to estimate the two-marker haplotype frequencies substantially enlarges the parameter space of the MHMM compared with an HMM. The estimate can be particularly unreliable when the ancestral information is sparse or inaccurate or when one of the alleles is rare. Thus, a potential weakness of the MHMM, compared with an HMM, is its requirement for richer genetic information on the ancestral populations. Fortunately, high-density SNP platforms are becoming more available and less expensive.

In this article, we propose a computationally tractable model for inferring admixing times and delineating ancestry along admixed chromosomes, which also accounts for background LD in ancestral populations. This approach opens up the possibility that admixture analyses, including MALD and candidate-gene association studies, can be performed using the existing high-density genotype platform, even if the marker panel has not been preselected to be ancestry informative. The simulation results we presented demonstrate the importance of accounting for background LD, both for estimating model parameters and for estimating underlying ancestry. We find it encouraging that the MHMM appears to adequately account for background LD, even for very dense marker panels. The MHMM is implemented in a program, SABER, which will be available online.

Acknowledgments

This research was supported by National Institutes of Health grants GM073059 (to H.T.) and HG003054 (to X.Z.). We thank E. Burchard, S. Choudhry, H. Li, R. Olshen, E. Ziv, and the anonymous reviewers for helpful discussions and comments.

Web Resource

The URL for data presented herein is as follows:

SABER, http://www.fhcrc.org/science/labs/tang/

References

1. Rife D (1954) Populations of hybrid origin as source material for the detection of linkage. Am J Hum Genet 6:26–33
2. McKeigue P (1998) Mapping genes that underlie ethnic differences in disease risk: methods for detecting linkage in admixed populations, by conditioning on parental admixture. Am J Hum Genet 63:241–251
3. Montana G, Pritchard J (2004) Statistical tests for admixture mapping with case-control and cases-only data. Am J Hum Genet 75:771–789
4. Hoggart C, Shriver M, Kittles R, Clayton D, McKeigue P (2004) Design and analysis of admixture mapping studies. Am J Hum Genet 74:965–978
5. Patterson N, Hattangadi N, Lane B, Lohmueller K, Hafler D, Oksenberg J, Hauser S, Smith M, O'Brien S, Altshuler D, Daly M, Reich D (2004) Methods for high-density admixture mapping of disease genes. Am J Hum Genet 74:979–1000
6. Zhu X, Cooper R, Elston R (2004) Linkage analysis of a complex disease through use of admixed populations. Am J Hum Genet 74:1136–1153
7. Zhu X, Luke A, Cooper R, Quertermous T, Hanis C, Mosley T, Gu C, Tang H, Rao D, Risch N, Weder A (2005) Admixture mapping for hypertension loci with genome-scan markers. Nat Genet 37:177–181
8. Falush D, Stephens M, Pritchard J (2003) Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics 164:1567–1587
9. Baum LE, Petrie T (1966) Statistical inference for probabilistic functions of finite state Markov chains. Ann Math Stat 37: 1554–1563
10. Lander E, Green P (1987) Construction of multilocus genetic linkage maps in humans. Proc Natl Acad Sci USA 84:2363–2367
11. Hughey R, Krogh A (1996) Hidden Markov models for sequence analysis: extension and analysis of the basic method. Comput Appl Biosci 12:95–107
12. Felsenstein J, Churchill G (1996) A hidden Markov model approach to variation among sites in rate of evolution. Mol Biol Evol 13:93–104
13. Fridlyand J, Snijders AM, Pinkel D, Albertson DG, Jain AN (2004) Hidden Markov models approach to the analysis of array CGH data. J Multivariate Anal 90:132–153
14. Rosenberg N, Pritchard J, Weber J, Cann H, Kidd K, Zhivotovsky L, Feldman M (2002) Genetic structure of human populations. Science 298:2381–2385
15. Seldin M, Morii T, Collins-Schramm H, Chima B, Kittles R, Criswell L, Li H (2004) Putative ancestral origins of chromosomal segments in individual African Americans: implications for admixture mapping. Genome Res 14:1076–1084
16. Stephens M, Smith N, Donnelly P (2001) A new statistical method for haplotype reconstruction from population data. Am J Hum Genet 68:978–989
17. Reich D, Patterson N (2005) Will admixture mapping work to find disease genes? Philos Trans R Soc Lond B Biol Sci 360: 1605–1607
18. McKeigue P (2000) Multipoint admixture mapping [letter]. Genet Epidemiol 19:464–467
19. Tang H, Peng J, Wang P, Risch N (2005) Estimation of individual admixture: analytical and study design considerations. Genet Epidemiol 28:289–301
20. McPeek M, Sun L (2000) Statistical tests for detection of misspecified relationships by use of genome-screen data. Am J Hum Genet 66:1076–1094
21. Cappé O, Moulines E, Rydén T (2005) Inference in hidden Markov models. Springer, New York
22. Rabiner L (1989) A tutorial on hidden Markov models and selected applications in speech recognition. Proc IEEE 77: 257–286
23. Baum L, Petrie T, Soules G, Weiss N (1970) A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. Ann Math Stat 41:164–171
24. Long J (1991) The genetic structure of admixed populations. Genetics 127:417–428
25. Karlin S, Taylor HM (1975) A first course in stochastic processes. Academic Press, London, p 152
26. Pritchard J, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155:945–959
27. Dempster A, Laird N, Rubin D (1977) Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc Ser B 39:1–38
28. International HapMap Consortium (2005) A haplotype map of the human genome. Nature 437:1299–1320
29. Rajeevan H, Osier M, Cheung K, Deng H, Druskin L, Heinzen R, Kidd J, Stein S, Pakstis A, Tosches N, Yeh C, Miller P, Kidd K (2003) Inference of population structure using multilocus genotype data. Nucleic Acids Res 31:270–271
30. Excoffier L, Slatkin M (1995) Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. Mol Biol Evol 12:921–927
31. Hawley M, Kidd K (1995) Haplo: a program using the EM algorithm to estimate the frequencies of multi-site haplotypes. J Hered 86:409–411
32. Long J, Williams R, Urbanek M (1995) An E-M algorithm and testing strategy for multiple-locus haplotypes. Am J Hum Genet 56:799–810
33. Clark A (1990) Inference of haplotypes from PCR-amplified samples of diploid populations. Mol Biol Evol 7:111–122
34. Fallin D, Schork N (2000) Accuracy of haplotype frequency estimation for biallelic loci, via the expectation-maximization algorithm for unphased diploid genotype data. Am J Hum Genet 67:947–959
35. Weiss K, Clark A (2002) Linkage disequilibrium and the mapping of complex human traits. Trends Genet 18:19–24