Bayesian Analysis (2004) **1**, Number 1, pp. 1–27

Hidden Markov Dirichlet Process: Modeling Genetic Inference in Open Ancestral Space

Eric P. Xing*
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213
epxing@cs.cmu.edu

Kyung-Ah Sohn
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213
ksohn@cs.cmu.edu

Abstract. The problem of inferring the population structure, linkage disequilibrium pattern, and chromosomal recombination hotspots from genetic polymorphism data is essential for understanding the origin and characteristics of genome variations, with important applications to the genetic analysis of disease propensities and other complex traits. Statistical genetic methodologies developed so far mostly address these problems separately using specialized models ranging from coalescence and admixture models for population structures, to hidden Markov models and renewal processes for recombination; but most of these approaches ignore the inherent uncertainty in the genetic complexity (e.g., the number of genetic founders of a population) of the data and the close statistical and biological relationships among objects studied in these problems. We present a new statistical framework called hidden Markov Dirichlet process (HMDP) to jointly model the genetic recombinations among a possibly infinite number of founders and the coalescence-with-mutation events in the resulting genealogies. The HMDP posits that a haplotype of genetic markers is generated by a sequence of recombination events that select an ancestor for each locus from an unbounded set of founders according to a 1st-order Markov transition process. Conjoining this process with a mutation model, our method accommodates both between-lineage recombination and within-lineage sequence variations, and leads to a compact and natural interpretation of the population structure and inheritance process underlying haplotype data. We have developed an efficient sampling algorithm for HMDP based on a two-level nested Pólya urn scheme, and we present experimental results on joint inference of population structure, linkage disequilibrium, and recombination hotspots based on HMDP. On both simulated and real SNP haplotype data, our method performs competitively or significantly better than extant methods in uncovering the recombination hotspots along chromosomal loci; and in addition it also infers the ancestral genetic patterns and offers a highly accurate map of ancestral compositions of modern populations.

Keywords: Dirichlet Process, Hierarchical DP, hidden Markov model, MCMC, statistical genetics, recombination, population structure, SNP.
* To whom correspondence should be addressed.

1 Introduction

The availability of nearly complete genome sequences for organisms such as humans makes it possible to begin to explore individual differences between DNA sequences,

known as *genetic polymorphisms*, on a genome-wide scale, and to search for associations of such genotypic variations with diseases and other phenotypes. Most human variation that is influenced by genes can be related to a particular kind of genetic polymorphism known as the *single nucleotide polymorphisms*, or SNPs. A SNP refers to the existence of two possible kinds of nucleotides from $\{A, C, G, T\}$ at a single chromosomal *locus* (i.e., a position on the chromosome) in a population; each variant is called an *allele* [1]. A *haplotype* is a list of alleles at contiguous sites in a local region of a single chromosome. Assuming no recombination in this local region, a haplotype is inherited as a unit. But under many realistic biological or genetic scenarios, repeated recombinations between ancestral haplotypes during generations of inheritance may confound the genetic origin of modern haplotypes (Figure 1).

Recombinations between ancestral chromosomes during meiosis play a key role in shaping the patterns of linkage disequilibrium (LD)—the non-random association of alleles at different *loci*—in a population. When a recombination occurs between two loci, it tends to decouple the alleles carried at those loci in its decedents and thus reduce LD; uneven occurrence of recombination events along chromosomal regions during genetic history can lead to "block structures" in molecular genetic polymorphisms such that within each block only low level of diversities are present in a population.

Statistically, for a pair of loci with genetic polymorphic markers, say, X and Y, the LD between these two loci can be characterized by a number of so-called *LD measures*. For example, for bi-allelic markers (i.e., markers that have only two possible states, say "0" and "1"), LD can be measured by the *gametic disequilibrium*, $D = p_{00}p_{11} - p_{01}p_{10}$, where $p_{00} := \text{Prob}(X = 0, Y = 0)$, $p_{11} := \text{Prob}(X = 1, Y = 1)$, $p_{01} := \text{Prob}(X = 0, Y = 1)$, and $p_{10} := \text{Prob}(X = 1, Y = 0)$, are the empirical frequencies of joint allele-state configurations. Another popular LD measure is the p-value for Fisher's exact test over samples of X and Y. When $D = 0$, which means that the two loci of interest are not arranged randomly during inheritance (due to recombination of their host chromosomes at a position between the two loci), they often emerge (e.g., from all possible pairs in a large number of loci being surveyed) as candidates of marker pairs on the chromosome whose locations are physically close so that there is a low probability of having recombination events between them. However, to the best of our knowledge, extant LD-measures remain primarily focused on offering population-level descriptive statistics of the sample, rather than on modeling and inferring the underlying genetic mechanisms and processes that may have generated the data. For example, the pairwise LD measure ignores the global context and overall pattern of the genetic polymorphisms, and thus can not distinguish linkages due to spurious statistical association (e.g., due to problems in sample procedures) from those resulting from true physical proximity, or from genetic coupling due to co-evolution [2]. Such an approach also

[1] In general, an *allele* represents a variant of a SNP, a gene, or some other entity associated with a locus on DNA. In our case (SNPs), the locus harbors a single nucleotide, and therefore the alleles can generally be assumed to be binary, reflecting the fact that "lightning doesn't tend to strike twice in the same place". That is, nucleotide substitutions (i.e., mutations) do not occur to the same locus twice during the inheritance course from a common ancestor. More generally, e.g., in case of *microsatellite* polymorphism, the allele-state can be k-nary, a scenario to which our proposed model also applies.

[2] Co-evolution can occur for DNA sequences that are far apart in the genome if they encode genes or

provides no information regarding the demographical history and ancestral composites of each individual in the study population. In this paper, we propose a new model-based approach to address these issues.

The problem of inferring chromosomal recombination hotspots is essential for understanding the origin and characteristics of genome variations; several combinatorial and statistical approaches have been developed for uncovering optimum block boundaries from single nucleotide polymorphism haplotypes (Daly et al. 2001; Anderson and Novembre 2003; Patil et al. 2001; Zhang et al. 2002). For example, Zhang et al. (2002) proposed a dynamic programming algorithm for partitioning single nucleotide polymorphism (SNP) haplotypes (explained in the sequel) into low-diversity blocks; Daly et al. (2001) and Greenspan and Geiger (2004a) have developed hidden Markov models for locating recombination hotspots in haplotypes; and Anderson and Novembre (2003) proposed a minimum description length (MDL) method for optimal haplotype block finding. Some recent studies resorted to more sophisticated population genetics arguments that more explicitly capture the mechanistic and population genetic foundations underlying recombination and LD pattern formation. For example, Li and Stephens (2003) used a tractable approximation to the recombinational coalescence, via a (latent) genealogy of the population, to capture the conditional dependencies between haplotypes. Rannala and Reeve (2001) also use a coalescence-based model and an MCMC method to integrate over the unknown gene genealogy and coalescence times. These advances have important applications in genetic analysis of disease propensities and other complex traits.

The deluge of SNP data also fuels the long-standing interest of analyzing patterns of genetic variations to reconstruct the evolutionary history and ancestral structures of human populations, using, for example, variants of admixture models on genetic polymorphisms (Pritchard et al. 2000; Rosenberg et al. 2002; Falush et al. 2003). These models are instances of a more general class of hierarchical Bayesian models known as *mixed membership models* (Erosheva et al. 2004), which postulate that genetic markers of each individual are iid (Pritchard et al. 2000) or spatially coupled (Falush et al. 2003) samples from multiple population-specific fixed-dimensional multinomial-distributions of marker alleles. However, the admixture models developed so far do model genetic drift due to mutations from the ancestor allele and therefore do not enable inference of the founding genetic patterns and the age of the founding alleles (Excoffier and Hamilton 2003).

This progress notwithstanding, the statistical methodologies developed so far mostly deal with LD analysis and ancestral inference separately, using specialized models that do not capture the close statistical and genetic relationships of these two problems. Moreover, most of these approaches ignore the inherent uncertainty in the genetic complexity (e.g., the number of genetic founders of a population) of the data and rely on inflexible models built on a pre-fixed, closed genetic space. Recently, we have developed a nonparametric Bayesian framework for modeling genetic polymorphisms based

---

regulatory elements that jointly or corporately perform an indispensable biology function. For example, proteins that form a complex to carry out enzymatic activities usually co-evolve.

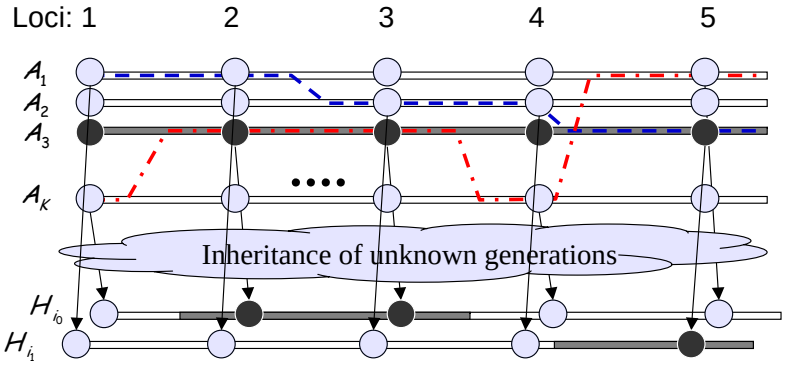

Figure 1: An illustration of a hidden Markov Dirichlet process for haplotype recombination and inheritance. Note that the total number of ancestors is unknown.

on the Dirichlet process (DP) mixtures and extensions, which attempts to allow more flexible control over the number of genetic founders than has been provided by the statistical methods proposed thus far (Xing et al. 2004) . In this paper, we leverage on this approach and present a unified framework to model complex genetic inheritance processes that allows recombinations among possibly infinite founding alleles and coalescence-with-mutation events in the resulting genealogies.

We assume that individual chromosomes in a modern population are originated from an unknown number of ancestral haplotypes via biased random recombinations and mutations (Figure 1). The recombinations between the ancestors follow a state-transition process we refer to as hidden Markov Dirichlet process (originated from the infinite HMM by Beal et al. (2002)), which travels in an open ancestor space, with nonstationary recombination rates depending on the genetic distances between SNP loci. Our model draws inspiration from the HMM proposed in Greenspan and Geiger (2004b), but we employ a two-level Pólya urn scheme akin to the hierarchical DP (Teh et al. 2006) to accommodate an open ancestor space, and allow full posterior inference of the recombination sites, mutation rates, haplotype origin, ancestor patterns, etc., conditioning on phased SNP data, rather than estimating them using information theoretic or maximum likelihood principles. On both simulated and real genetic data, our model and algorithm show competitive or superior performance on a number of genetic inference tasks over the state-of-the-art parametric methods.

The remainder of this paper is presented as follows. In section 2, we formulate the problem, and present details of the proposed model. In section 3, we describe a block Gibbs sampling algorithm for posterior inference of the latent variables. In section 4, we present experimental results on a simulated data haplotype data set, and on two published real data sets, one from a single population, and the other from two populations. We conclude with a brief discussion in section 6. A short version of this manuscript was presented earlier in Sohn and Xing (2006), but the current version offers more details on the biological background, the model specifications, and the experimental results.

2 The Statistical Model

Sequentially choosing recombination targets from a set of ancestral chromosomes can be modeled as a hidden Markov process (Niu et al. 2002; Greenspan and Geiger 2004b), in which the hidden states correspond to the index of the candidate chromosomes, the transition probabilities correspond to the recombination rates between the recombining chromosome pairs, and the emission model corresponds to a mutation process that passes the chosen chromosome region in the ancestors to the descents. When the number of ancestral chromosomes is not known, it is natural to consider an HMM whose state space is countably infinite (Beal et al. 2002; Teh et al. 2006). In this section, we describe such an infinite HMM formalism, which we would like to call *hidden Markov Dirichlet process*, for modeling recombination in an open ancestral space.

2.1 Dirichlet Process mixtures

For self-containedness, we begin with a quick overview of the fundamentals of the Dirichlet process and its connection to the coalescent process in population genetics, followed by a brief recapitulation of the basic Dirichlet process mixture model we proposed in Xing et al. (2004) for haplytope inheritance without recombination.

As mentioned earlier, a *haplotype* refers to the joint allele configuration of a contiguous list of SNPs located on a chromosome. Under a well-known genetic model known as *coalescence with infinite-many-alleles (IMA) mutations* (but without recombination), one can treat a haplotype from a modern individual as a descendent of a most recent common ancestor (MRCA) of unknown haplotype via random mutations that alter the allelic states of some SNPs (Kingman 1982). Hoppe (1984) observed that a coalescent process in an infinite population leads to a partition of the population at every generation that can be succinctly captured by the following Pólya urn scheme.

Consider an urn that at the outset contains a ball of a single color. At each step we either draw a ball from the urn and replace it with two balls of the same color, or we are given a ball of a new color which we place in the urn. One can see that such a scheme leads to a partition of the balls according to their color. Mapping each ball to a haploid individual [3] and each color to a possible haplotype, this partition is equivalent to the one resulting from the *coalescence-with-mutation* process (Hoppe 1984), and the probability distribution of the resulting *allele spectrum*—the numbers of colors (resp. haplotypes) with every possible number of representative balls (resp. decedents)—is captured by the well-known Ewens' sampling formula (Tavare and Ewens 1998).

Letting parameter α define the probabilities of the two types of draws in the aforementioned Pólya urn scheme, and viewing each (distinct) color as a sample from Q_0, and each ball as a sample from Q [4], Blackwell and MacQueen (1973) showed that

[3] A haploid individual refers to an individual with only one haplotype — a simplifying assumption often used on population genetics when the paternal and maternal haplotypes of a diploid individual are inherited independently.

[4] Here we deviate from the conventional notations in the statistics literature (e.g., Neal (2000); Escobar and West (2002); Ishwaran and James (2001)) and use Q and Q_0, instead of G and G_0 (or

this Pólya urn model yields samples whose distributions are those of Q_0 the marginal probabilities under the *Dirichlet process* (Ferguson 1973). Formally, a random probability measure Q is generated by a DP if for any measurable partition $B_1, \ldots, B_k$ of the sample space, the vector of random probabilities $Q(B_i)$ follows a Dirichlet distribution: $(Q(B_1), \ldots, Q(B_k)) \sim \text{Dir}(\alpha Q_0(B_1), \ldots, \alpha Q_0(B_k))$, where α denotes a *scaling parameter* and Q_0 denotes a *base measure*. The Pólya urn model makes explicit that the association of data points to colors defines a "clustering" of the data. Specifically, having observed n values $(\phi_1, \ldots, \phi_n)$ sampled from a Dirichlet process $DP(\alpha, Q_0)$, the probability of the $(n+1)$th value is given by:

$$\phi_{n+1}|\phi_1, \ldots, \phi_n, \alpha, Q \;\sim\; \sum_{i=1}^{n} \frac{1}{n+\alpha}\delta_{\phi_i}(\cdot) + \frac{\alpha}{n+\alpha}Q_0(\cdot), \tag{1}$$

where $\delta_{\phi_i}(\cdot)$ denotes a point mass at value ϕ_i. Another very useful representation of DP is the stick-breaking construction by Sethuraman (1994). This construction is based on independent sequences of independent random samples $\{\pi'_{k,i}\}_{i=1}^{\infty}$ and $\{\phi_i\}_{i=1}^{\infty}$ generated in the following way: $\pi'_i|\alpha, Q_0 \sim \text{Beta}(1, \alpha)$ and $\phi_i|\alpha, Q_0 \sim Q_0$, where $\text{Beta}(a, b)$ is the Beta distribution with parameter a and b. Let $\pi_i = \pi'_i \prod_{l=1}^{k-1}(1-\pi'_l)$ (analogous to a process of repetitively breaking a stick at fraction π'_l), Sethuraman showed that the random measure arising from $DP(\alpha, Q_0)$ admits the representation $Q = \sum_{i=1}^{\infty} \pi_i \delta_{\phi_i}$. The ϕ_i's can be understood as the *locations* of samples in their space, and the π_i's are the *weights* of these samples.

The discrete nature of the DP, as obviated from the stick-breaking construction, is well suited for the problem of placing priors on mixture components in mixture modeling. In the context of mixture models, one can associate mixture component centroids (e.g., haplotype founders, as explained in the sequel) with colors in the Pólya urn model and thereby define a "clustering" of the (possibly noisy) data (e.g., modern haplotypes that are "recognizable" variants of their corresponding founders). This mixture model is known as a DP mixture (Antoniak 1973; Escobar and West 2002) (also known as "infinite" mixture model in machine learning community). Note that a DP mixture requires no prior specification of the number of components, which is typically unknown in genetic demography and general data clustering problems. It is important to emphasize that here DP is used as a *prior distribution* of mixture components. Multiplying this prior by a likelihood that relates the mixture components to the actual data yields a *posterior distribution* of the mixture components, and the design of the likelihood function is completely up to the modeler based on specific problems. MCMC algorithms have been developed to sample from the posterior associated with DP priors (Escobar and West 2002; Neal 2000; Ishwaran and James 2001). This nonparametric Bayesian formalism forms the technical foundation of the haplotype modeling and inference algorithms to be developed in this paper.

Back to haplotype modeling, a straightforward statistical genetics argument shows that the distribution of haplotypes can be formulated as a mixture model, where the

H), to denote the random probability measure under DP and the base measure of DP, because in the genetic context, G and H have been used to denote the genotype and haplotype of polymorphic markers (Pritchard et al. 2000; Stephens et al. 2001; Li and Stephens 2003; Xing et al. 2004).

set of mixture components corresponds to the pool of ancestor haplotypes, or *founders*, of the population (Excoffier and Slatkin 1995; Niu et al. 2002; Kimmel and Shamir 2004). Crucially, however, the size of this pool is unknown; indeed, knowing the size of the pool would correspond to knowing something significant about the genome and its history. On the other hand, despite its elegance, with a purely coalescence-based model for genetic patterns, it is hard to perform statistical inference of ancestral features and many other interesting genetic variables (for a large population, the number of hidden variables in a coalescence tree is prohibitively large) (Stephens et al. 2001). In most practical population genetic problems, usually the detailed genealogical structure of a population (as provided by the coalescent trees) is of less importance than the population-level features such as the pattern of major common ancestor alleles (i.e., founders) in a population bottleneck [5], the age of such alleles, etc. In this case, the DP mixture offers a principled approach to generalize the finite mixture model for haplotypes to an infinite mixture model that models uncertainty regarding the size of the ancestor haplotype pool; at the same time, it provides a reasonable approximation to the coalescence model by utilizing the partition structure resulting therefrom (but allows further mutations within each partite to introduce further diversity among descents of the same founder, which correspond to the balls with the same color in the Pólya urn metaphor). Without further digression, below we summarize the Dirichlet process mixture model we proposed in Xing et al. (2004) for haplytope inheritance without recombination.

Write $H_i = [H_{i,1}, \ldots, H_{i,T}]$ for a haplotype over T SNPs from chromosome i [6]; let $A_k = [A_{k,1}, \ldots, A_{k,T}]$ denote an ancestor haplotype (indexed by k) and θ_k denote the *mutation rate* of ancestor k; and let C_i denote an *inheritance variable* that specifies the ancestor of haplotype H_i. Under a DP mixture, we have the following Pólya urn scheme for sampling modern haplotypes:

- Draw first haplotype:

  $a_1 \mid \mathrm{DP}(\tau, Q_0) \sim Q_0(\cdot)$, sample the 1st founder;

  $h_1 \sim P_h(\cdot|a_1, \theta_1)$, sample the 1st haplotype from an inheritance model defined on the 1st founder;

- for subsequent haplotypes:

  – sample the founder indicator for the ith haplotype:

  $$c_i|\mathrm{DP}(\tau, Q_0) \sim \begin{cases} p(c_i = c_j \text{ for some } j < i|c_1, \ldots, c_{i-1}) = \frac{n_{c_j}}{i-1+\alpha_0} \\ p(c_i \neq c_j \text{ for all } j < i|c_1, \ldots, c_{i-1}) = \frac{\alpha_0}{i-1+\alpha_0} \end{cases}$$

  where n_{c_i} is the *occupancy number* of class c_i—the number of previous samples belonging to class c_i.

[5] A stage in coalescence when there are only a very small number of founding haplotype patterns surviving and giving rise to all the haplotypes in the modern population.

[6] We ignore the parental origin index of haplotypes as used in Xing et al. (2004), and assume that the paternal and maternal haplotypes of each individual are given unambiguously (i.e., *phased*, as known in genetics), as is the case in many LD and haplotype-block analyses (Daly et al. 2001; Anderson and Novembre 2003). But it is noteworthy that our model can generalize straightforwardly to unphased genotype data by incorporating a simple genotype model as in Xing et al. (2004).

– sample the founder of haplotype i (indexed by c_i):

$$\phi_{c_i}|\mathrm{DP}(\tau, Q_0) \begin{cases} = \{a_{c_j}, \theta_{c_j}\} & \text{if } c_i = \{a_{c_j}, \theta_{c_j}\} \text{ for some } j < i \text{ (i.e., } c_i \text{ refers to an inherited founder)} \\ \sim Q_0(a, \theta) & \text{if } c_i \neq c_j \text{ for all } j < i \text{ (i.e., } c_i \text{ refers to a new founder)} \end{cases}$$

– sample the haplotype according to its founder:

$h_i \mid c_i \sim P_h(\cdot|a_{c_i}, \theta_{c_i})$.

The usefulness of the DP mixture framework for the haplotype problem should be clear—using a Dirichlet process prior we in essence maintain a pool of haplotype founders that grows as observed individual haplotypes are processed. But notice that the above generative process assumes each modern haplotype originates from a single ancestor, which is only true for haplotypes spanning a short region on a chromosomal. Now we consider long haplotypes possibly bearing multiple ancestors due to recombinations between an unknown number of founders.

2.2 Hidden Markov Dirichlet Process (HMDP)

In a standard HMM, state-transitions across a discrete time- or space-interval take place in a fixed-dimensional state space, thus it can be fully parameterized by, say, a K-dimensional initial-state probability vector π_0 and a $K \times K$ state-transition probability matrix $\Pi_{K\times K}$. As first proposed in Beal et al. (2002), and later discussed in Teh et al. (2006), one can "open" the state space of an HMM by treating the now infinite number of discrete states of the HMM as the support of a DP, and the transition probabilities to these states from some source as the masses associated with these states. In particular, for each source state (say, state j), the possible transitions to the target states need to be modeled by a unique DP Q_j. Since all possible source states and target states are taken from the same infinite state space, overall we need an open set of DPs with different mass distributions on the SAME support (to capture the fact that different source states can have different transition probabilities to any target state). In the sequel, we describe such a nonparametric Bayesian HMM using an intuitive hierarchical Pólya urn construction. We call this model a **hidden Markov Dirichlet process**.

In an HMDP, both the columns and rows of the transition matrix Π are infinite dimensional. To construct such an stochastic matrix, we will exploit the fact that in practice only a finite number of states (although we don't know what they are) will be visited by each source state, and we only need to keep track of these states. The following sampling scheme based on a hierarchical Pólya urn scheme captures this spirit and yields a constructive definition of HMDP.

We set up a single "stock" urn at the top level, which contains balls of colors that are represented by at least one ball in one or multiple urns at the bottom level. At the bottom level, we have a set of *distinct* urns which are used to define the initial and transition probabilities of the HMDP model (and are therefore referred as HMM-urns).

Specifically, one of the HMM urns, Q_0, is set aside to hold colored balls to be drawn at the onset of the HMM state-transition sequence [7]. Each of the remaining HMM urns is painted with a color represented by at least one ball in the stock urn, and is used to hold balls to be drawn during the execution of a Markov chain of state-transitions. Now let's suppose that at time t the stock urn contains n balls of K distinct colors indexed by an integer set $\mathcal{C} = \{1, 2, \ldots, K\}$; the number of balls of color k in this urn is denoted by $n_k, k \in \mathcal{C}$. For urn Q_0 and urns $Q_1, \ldots, Q_K$, let $m_{j,k}$ denote the number of balls of color k in urn Q_j, and $m_j = \sum_{k \in \mathcal{C}} m_{j,k}$ denote the total number of balls in urn Q_j. Suppose that at time $t-1$, we had drawn a ball with color k'. Then at time t, we either draw a ball randomly from urn $Q_{k'}$, and place back two balls both of that color; or with probability $\frac{\tau}{m_j+\tau}$ we turn to the top level. From the stock urn, we can either draw a ball randomly and put back two balls of that color to the stock urn and one to $Q_{k'}$, or obtain a ball of a new color $K+1$ with probability $\frac{\gamma}{n+\gamma}$ and put back a ball of this color to both the stock urn and urn $Q_{k'}$ of the lower level. Essentially, we have a master DP Q_0 (the stock urn) that serves as a source of atoms for infinite number of child DPs $\{Q_j\}$ (the HMM-urns). As pointed out in Teh et al. (2006), this model can be viewed as an instance of the hierarchical Dirichlet process mixture model, with an infinite number of DP mixtures as components. Specifically, we have:

$$
\begin{aligned}
Q_0|\alpha, F &\sim \mathrm{DP}(\alpha, F), && \text{The master DP over target states common for all sources;} \\
Q_j|\tau, Q_0 &\sim \mathrm{DP}(\tau, Q_0), && \text{The HMM DP over target states of source } j.
\end{aligned}
$$

From the above equation we see that the base measure of the DP mixture associated each of the source states in the HMM is itself drawn from a Dirichlet process $\mathrm{DP}(\alpha, F)$. Since a draw from a DP is a discrete measure with probability 1, atoms drawn from this measure—atoms which are used as targets for each of the (unbounded number of) source states—are not generally distinct. Indeed, the transition probabilities from each of the source states have the same support—the atoms in Q_0.

The Pólya urn scheme described above is similar in spirit to the "Chinese restaurant franchise" scheme discussed in Teh et al. (2006), but it differs in that it avoids having separate occupancy counters in each lower-level DP for repeated draws of the same atom from a top-level DP, and it also motivates a simpler sampling scheme for inference as discussed in Section 3.

Associating each color k with an ancestor configuration $\phi_k = \{a_k, \theta_k\}$ whose values are drawn from the base measure F, and recalling our discussion in the previous section, we know that draws from the stock urn can be viewed as marginals from a random measure distributed as a Dirichlet Process Q_0 with parameter (α, F). Specifically, for n random draws $\boldsymbol{\phi} = \{\phi_1, \ldots, \phi_n\}$ from Q_0, the conditional prior for $(\phi_n|\boldsymbol{\phi}_{-n})$, where the subscript "$-n$" denotes the index set of all but the n-th ball, is

$$
\phi_n|\boldsymbol{\phi}_{-n} \sim \sum_{k=1}^{K} \frac{n_k}{n-1+\alpha} \delta_{\phi_k^*}(\phi_n) + \frac{\alpha}{n-1+\alpha} F(\phi_n), \tag{2}
$$

[7] Purposely, we overload the symbol Q_j to let it denote both the urns in the hierarchical Pólya urn scheme, and the Dirichlet processes distributions represented by each of these urns.

where $\phi_k^*, k = 1, \ldots, K$ denote the K distinct values (i.e., colors) of $\boldsymbol{\phi}$ (i.e., all the balls in the stock urn), n_k denote the number of balls of color k in the top urn, and $\delta_a(\phi_i)$ denotes a unit point mass at $\phi_i = a$.

Conditioning on the Dirichlet process underlying the stock urn, the samples in the jth bottom-level urn are also distributed as marginals under a Dirichlet measure:

$$\begin{aligned}\phi_{m_j}|\boldsymbol{\phi}_{-m_j} &\sim \sum_{k=1}^{K} \frac{m_{j,k} + \tau\frac{n_k}{n-1+\alpha}}{m_j - 1 + \tau}\delta_{\phi_k^*}(\phi_{m_j}) + \frac{\tau}{m_j - 1 + \tau}\frac{\alpha}{n-1+\alpha}F(\phi_{m_j}) \\ &= \sum_{k=1}^{K} \pi_{j,k}\delta_{\phi_k^*}(\phi_{m_j}) + \pi_{j,K+1}Q_0(\phi_{m_j}), \end{aligned} \tag{3}$$

where $\pi_{j,k} \equiv \frac{m_{j,k}+\tau\frac{n_k}{n-1+\alpha}}{m_j-1+\tau}$, $\pi_{j,K+1} \equiv \frac{\tau}{m_j-1+\tau}\frac{\alpha}{n-1+\alpha}$. Let $\boldsymbol{\pi}_j \equiv [\pi_{j,1}, \pi_{j,2}, \ldots]$. Now we have an infinite-dimensional Bayesian HMM that, given F, α, τ, and all initial states and transitions sampled so far, follows an initial states distribution parameterized by $\boldsymbol{\pi}_0$, and transition matrix Π whose rows are defined by $\{\boldsymbol{\pi}_j : j > 0\}$.

Finally, as in, e.g., Escobar and West (2002) and Rasmussen (2000), we can also introduce vague priors such as a Gamma or an inverse Gamma for the scaling parameters α and τ.

2.3 HMDP Model for Recombination and Inheritance

Now we describe a stochastic model, based on an HMDP, for generating individual haplotypes in a modern population from a hypothetical pool of ancestral haplotypes via recombination and mutations (i.e., random mating with neutral selection). See Figure 1 for an illustration.

First recall that a base measure F at the top of our hierarchical Pólya urn scheme is defined as a distribution from which ancestor haplotype templates ϕ_k are drawn. We define the base measure F as a joint measure on both ancestor A and mutation rate θ, and let $F(A, \theta) = p(A)p(\theta)$, where $p(A)$ is uniform over all possible haplotypes and $p(\theta)$ is a beta distribution, $Beta(\alpha_h, \beta_h)$, with a small value for $\beta_h/(\alpha_h + \beta_h)$ corresponding to a prior expectation of a low mutation rate. For simplicity, we assume each $A_{k,t}$ (and also each $H_{i,t}$) takes its value from an allele set B.

Now for each modern chromosome i, let $C_i = [C_{i,1}, \ldots, C_{i,T}]$ denote the sequence of inheritance variables specifying the index of the ancestral chromosome at each SNP locus. When no recombination takes place during the inheritance process that produces haplotype H_i (say, from ancestor k), then $C_{i,t} = k, \forall t$. When a recombination occurs, say, between loci t and $t+1$, we have $C_{i,t} \neq C_{i,t+1}$. We can introduce a Poisson point process to control the duration of non-recombinant inheritance. That is, given that $C_{i,t} = k$, then with probability $e^{-dr} + (1 - e^{-dr})\pi_{kk}$, where d is the physical distance between two loci, r reflects the rate of recombination per unit distance, and π_{kk} is the self-transition probability of ancestor k defined by HMDP, we have $C_{i,t+1} = C_{i,t}$; otherwise, the source state (i.e., ancestor chromosome k) pairs with a target state (e.g.,

ancestor chromosome k') between loci t and $t+1$, with probability $(1-e^{-dr})\pi_{kk'}$. Hence, each haplotype H_i is a mosaic of segments of multiple ancestral chromosomes from the ancestral pool $\{A_k\}_{k=1}^{\infty}$. Essentially, the model we described so far is a time-inhomogeneous infinite HMM. When the physical distance information between loci is not available, we can simply set r to be infinity (hence $e^{-dr} \approx 0$) so that we are back to a standard stationary HMDP model with infinite dimensional transition probability matrix $\Pi_{\infty\times\infty}$ described earlier.

The emission process of the HMDP corresponds to an inheritance model from an ancestor to the matching descendent. For simplicity, we adopt the *single-locus mutation model* in Xing et al. (2004):

$$p(h_t|a_t,\theta) = \theta^{\mathbb{I}(h_t=a_t)}\Big(\frac{1-\theta}{|B|-1}\Big)^{\mathbb{I}(h_t\neq a_t)}, \tag{4}$$

where h_t and a_t denote the alleles at locus t of an individual haplotype and its corresponding ancestor, respectively; θ indicates the ancestor-specific mutation rate; and $|B|$ denotes the number of possible alleles. As discussed in Liu et al. (2001), this model corresponds to a star genealogy resulting from infrequent mutations over a shared ancestor, and is widely used in statistical genetics as an approximation to a full coalescent genealogy starting from the shared ancestor.

Assume that the mutation rate θ admits a Beta prior with hyperparameter (α_h,β_h) [8], the marginal conditional likelihood of all the haplotype instances $\mathbf{h} = \{h_{i,t} \ : \ i \in \{1,2,\ldots,I\},\ t \in \{1,2,\ldots,T\}\}$ given the set of ancestors $\mathbf{a} = \{a_1,\ldots,a_K\}$ and the ancestor indicators $\mathbf{c} = \{c_{i,t} \ : \ i \in \{1,2,\ldots,I\},\ t \in \{1,2,\ldots,T\}\}$ can be obtained by integrating out θ from the joint conditional probability starting from Equation (4) as follows:

$$\begin{aligned} p(\mathbf{h}|\mathbf{c},\mathbf{a}) &= \prod_k \Big(\int \prod_{i,t|c_{i,t}=k} p(h_{i,t},\theta_k|a_{k,t}) R(\alpha_h,\beta_h)\theta_k^{\alpha_h-1}(1-\theta_k)^{\beta_h-1} d\theta_k \Big) \\ &= \prod_k R(\alpha_h,\beta_h)\frac{\Gamma(\alpha_h+l_k)\Gamma(\beta_h+l'_k)}{\Gamma(\alpha_h+\beta_h+l_k+l'_k)}\Big(\frac{1}{|B|-1}\Big)^{l'_k} \end{aligned} \tag{5}$$

where $\Gamma(\cdot)$ is the gamma function, $R(\alpha_h,\beta_h) = \frac{\Gamma(\alpha_h+\beta_h)}{\Gamma(\alpha_h)\Gamma(\beta_h)}$ is the normalization constant associated with Beta(α_h,β_h) (which is a prior distribution for θ), $l_k = \sum_t\sum_i \mathbb{I}(h_{i,t} = a_{k,t})\mathbb{I}(c_{i,t} = k)$ is the number of alleles that were not mutated with respect to the ancestral allele, and $l'_k = \sum_t\sum_i \mathbb{I}(h_{i,j} \neq a_{k,j})\mathbb{I}(c_{i,t} = k)$ is the number of mutated alleles. The counting record $\mathbf{l}_k = \{l_k, l'_k\}$ is a sufficient statistic for the parameter θ_k.

The generative process and likelihood functions described above point naturally to an algorithm for population genetic inference. Unlike the classical coalescence models for recombination (Hudson 1983), which have been primarily used for theoretical analysis and simulation, but are hardly feasible for reverse ancestral inference based on

[8] For simplicity, we assume that the mutation rates pertaining to different ancestors follow the same prior Beta(α_h,β_h).

observed genetic data, the HMDP model described above for recombination and inheritance provides a semi-parametric Bayesian formalism that is well suited for data-driven posterior inference on the latent variables that can yield rich information on the population ancestry and genetic structure of the study population. For example, under a HMDP, given the haplotype data, one can infer the ancestral pattern, LD structure and recombination hotspot of a population using the posterior distribution of inheritance variable $\mathbf{c}$ and ancestral state $\mathbf{a}$, as we will elaborate in the sequel. It is also possible to infer the age of the haplotype alleles and/or the time of recombination events by exploring the posterior estimates of the mutation and recombination rates under HMDP.

3 Posterior Inference

In this section, we describe a Gibbs sampling algorithm for posterior inference under HMDP. Recall that a Gibbs sampler draws samples of each random variable (or subset of random variables) in the model from the conditional distribution of the variable(s) given (previously sampled) values of all the remaining variables. The variables of interest in our model include $\{C_{i,t}\}$, the inheritance variables specifying the origins of SNP alleles of all loci on each haplotype, and $\{A_{k,t}\}$, the founding alleles at all loci of each ancestral haplotype. All other variables in the model, e.g., the mutation rate θ, are integrated out.

We assume that the individual haplotypes $\{H_{i_e,t}\}$ are given unambiguously for the study population, as is the case in many LD and haplotype-block analyses (Daly et al. 2001; Anderson and Novembre 2003); but it is noteworthy that our model can generalize straightforwardly to unphased genotype data by incorporating a simple genotype model as in Xing et al. (2004). Given that haplotypes are unambiguous, we can now treat the paternal and maternal haplotypes of N individual as $2N$ *iid* samples from the HMDP process and omit the parental index e.

The Gibbs sampler alternates between two sampling stages. First it samples the inheritance variables $\{C_{i,t}\}$, conditioning on all given individual haplotypes $\mathbf{h} = \{h_1, \ldots, h_{2N}\}$, and the most recently sampled configuration of the ancestor pool $\mathbf{a} = \{a_1, \ldots, a_K\}$; then given $\mathbf{h}$ and current values of the $C_{i,t}$'s, it samples every ancestor a_k.

To improve the mixing rate, we sample the inheritance variables one block at a time. That is, every time we sample δ consecutive states $c_{t+1}, \ldots, c_{t+\delta}$ starting at a randomly chosen locus $t+1$ along a haplotype. (For simplicity we omit the haplotype index i here and in the forthcoming expositions when it is clear from context that the statements or formulas apply to all individual haplotypes). Let $\mathbf{c}^-$ denote the set of previously sampled inheritance variables. Let $\mathbf{n}$ denote the totality of occupancy records of the top-level DP (i.e. the "stock urn") — $\{n\} \cup \{n_k \ : \ \forall k\}$, and $\mathbf{m}$ denote the totality of the occupancy records of each lower-level DP (i.e., the urns corresponding to the recombination choices by each ancestor) — $\{m_k \ : \ \forall k\} \cup \{m_{k,k'} \ : \ \forall k, k'\}$. Let $\mathbf{l}_k$ denote the sufficient statistics associated with all haplotype instances originating from ancestor k. The predictive distribution of a δ-block of inheritance variables can be

written as:

$$\begin{aligned} p(c_{t+1:t+\delta} \mid \mathbf{c}^{-}, \mathbf{h}, \mathbf{a}) &\propto p(c_{t+1:t+\delta} \mid c_t, c_{t+\delta+1}, \mathbf{m}, \mathbf{n}) p(h_{t+1:t+\delta} | a_{c_{t+1},t+1}, \ldots, a_{c_{t+\delta},t+\delta}) \\ &\propto \prod_{j=t}^{t+\delta} p(c_{j+1}|c_j, \mathbf{m}, \mathbf{n}) \prod_{j=t+1}^{t+\delta} p(h_j | a_{c_j,j}, \mathbf{l}_{c_j}). \end{aligned} \tag{6}$$

This expression is simply Bayes' theorem with $p(h_{t+1:t+\delta}|a_{c_{t+1},t+1}, \ldots, a_{c_{t+\delta},t+\delta})$ playing the role of the likelihood and $p(c_{t+1:t+\delta} \mid \mathbf{c}^{-}, \mathbf{h}, \mathbf{a})$ playing the role of the posterior. One should be careful that the sufficient statistics $\mathbf{n}$, $\mathbf{m}$ and $\mathbf{l}$ employed here should exclude the contributions by samples associated with the δ-block to be sampled. Note that naively, the sampling space of an inheritance block of length δ is $|A|^{\delta}$ where $|A|$ represents the cardinality of the ancestor pool. However, if we assume that the recombination rate is low and block length is not too big, then the probability of having two or more recombination events within a δ-block is very small and thus can be ignored. This approximation reduces the sampling space of the δ-block to $O(|A|\delta)$, i.e., $|A|$ possible recombination targets times δ possible recombination locations. Accordingly, Eq. (6) reduces to:

$$p(c_{t+1:t+\delta} \mid \mathbf{c}^{-}, \mathbf{h}, \mathbf{a}) \propto p(c_{t'} \mid c_{t'-1} = c_t, \mathbf{m}, \mathbf{n}) p(c_{t+\delta+1} \mid c_{t+\delta} = c_{t'}, \mathbf{m}, \mathbf{n}) \prod_{j=t'}^{t+\delta} p(h_j | a_{c_{t'},j}, \mathbf{l}_{c_{t'}}), \tag{7}$$

for some $t' \in [t+1, t+\delta]$. Recall that in an HMDP model for recombination, given that the total recombination probability between two loci d-units apart is $\lambda \equiv 1 - e^{-dr} \approx dr$ (assuming d and r are both very small), the transition probability from state k to state k' is:

$$\begin{aligned} & p(c_{t'} = k' \mid c_{t'-1} = k, \mathbf{m}, \mathbf{n}, r, d) \\ = & \begin{cases} \lambda \pi_{k,k'} + (1-\lambda)\delta(k,k') & \text{for } k' \in \{1, ..., K\}, \text{ i.e., transition to an existing ancestor,} \\ \lambda \pi_{k,K+1} & \text{for } k' = K+1, \text{ i.e., transition to a new ancestor,} \end{cases} \end{aligned} \tag{8}$$

where π_k represents the transition probability vector for ancestor k under HMDP, as defined in Eq. (3). Note that when a new ancestor a_{K+1} is instantiated, we need to immediately instantiate a new DP under F to model the transition probabilities from this ancestor to all instantiated ancestors (including itself). Since the occupancy record of this DP, $\mathbf{m}_{K+1} := \{m_{K+1}\} \cup \{m_{K+1,k} \ : \ k = 1, \ldots, K+1\}$, is not yet defined at the onset, with probability 1 we turn to the top-level DP when departing from state $K+1$ for the first time. Specifically, we define $p(\cdot|c_{t'} = K+1)$ according to the occupancy record of ancestors in the stock urn. For example, at the distal border of the δ-block, since $c_{t+\delta+1}$ always indexes a previously inherited ancestor (and therefore must be present in the stock-urn), we have:

$$p(c_{t+\delta+1} \mid c_{t+\delta} = K+1, \mathbf{m}, \mathbf{n}) \quad = \quad \lambda \times \frac{n_{c_{t+\delta+1}}}{n - 1 + \alpha}. \tag{9}$$

Now we can substitute the relevant terms in Eq. (6) with Eqs. (8) and (9). The marginal likelihood term in Eq. (6) can be readily computed based on Eq. (4), by

integrating out the mutation rate θ under a Beta prior (and also the ancestor a under a uniform prior if $c_{t'}$ refers to an ancestor to be newly instantiated) (Xing et al. 2004). Putting everything together, we have the proposal distribution for a block of inheritance variables. Upon sampling every c_t, we update the sufficient statistics $\mathbf{n}$, $\mathbf{m}$ and $\{\mathbf{l}_k\}$ as follows. First, before drawing the sample, we erase the contribution of c_t to these sufficient statistics. In particular, if an ancestor gets no occupancy in either the stock or the HMM urns afterwards, we remove it from our repository. Then, after drawing a new c_t, we increment the relevant counts accordingly. In particular, if $c_t = K+1$ (i.e., a new ancestor is to be drawn), we update $n = n+1$, set $n_{K+1} = 1$, $m_{c_t} = m_{c_t}+1, m_{c_t,K+1} = 1$, and set up a new (empty) HMM urn with color $K+1$ (i.e. instantiating $\mathbf{m}_{K+1}$ with all elements equal to zero).

Now we move on to sample the founders $\{a_{k,t}\}$. From the mutation model in Equation (4), we can derive the following posterior distribution to sample the founder a_k [9]:

$$p(a_{k,t}|\mathbf{c},\mathbf{h}) \propto \int \Big(\prod_{i|c_{i,t}=k} p(h_{i,t}|a_{k,t},\theta) \Big) Beta(\theta|\alpha_h,\beta_h) d\theta$$
$$= \frac{\Gamma(\alpha_h + l_{k,t})\Gamma(\beta_h + l'_{k,t})}{\Gamma(\alpha_h+\beta_h+l_{k,t}+l'_{k,t})(|B|-1)^{l'_{k,t}}} R(\alpha_h,\beta_h), \quad (10)$$

where $l_{k,t}$ is the number of allelic instances originating from ancestor k at locus t that are identical to the ancestor, when the ancestor has the pattern $a_{k,t}$; and $l'_{k,t} = \sum_i \mathbb{I}(c_{i,t} = k|a_{k,t}) - l_{k,t}$ represents the complement. The normalization constant of this proposal distribution can be computed by summing the R.H.S. of Eq. (10) over all possible allele states of an ancestor at the locus being sampled. If k is not represented previously, we can just set $l_{k,t}$ and $l'_{k,t}$ both to zero. Note that when sampling a new ancestor, we can only condition on a small segment of an individual haplotype. To instantiate a complete ancestor, after sampling the alleles in the ancestor corresponding to the segment according to Eq. (10), we first fill in the rest of the loci with random alleles. When another segment of an individual haplotype needs a new ancestor, we do not naively create a new full-length ancestor; rather, we use the *empty* slots (those with random alleles) of one of the previously instantiated ancestors, if any, so that the number of ancestors does not grow unnecessarily.

4 Experiments

We applied the HMDP model to both simulated and real haplotype data. Our analyses focus on the following three popular problems in statistical genetics: 1. Ancestral Inference: estimating the number of founders in a population and reconstructing the ancestor haplotypes; 2) LD-block Analysis: inferring the recombination sites in each individual haplotype and uncover population-level recombination hotspots on the chro-

[9] In deriving Equation (10), instead of assuming a common mutation rate θ_k for all loci of ancestor a_k, we endow each locus with its own mutation parameter $\theta_{k,t}$, with all parameters admitting the same prior Beta(α_h, β_h). This is arguably a more accurate reflection of reality.

mosome region; 3) Population Structural Analysis: mapping the genetic origins of all loci of each individual haplotype in a population.

4.1 Analyzing simulated haplotype population

To simulate a population of individual haplotypes, we started with a fixed number, K_s (unknown to the HMDP model), of randomly generated ancestor haplotypes, on each of which a set of recombination hotspots were pre-specified. Then we applied a hand-specified recombination process, which is defined by a K_s-dimensional HMM, to the ancestor haplotypes to generate N_s individual haplotypes, via sequentially recombining segments of different ancestors according to the simulated HMM states at each locus, and mutating certain ancestor SNP alleles according to the emission model. All the ancestor haplotypes were set to be 100 SNPs long. At the hotspots (pre-specified at every 10-th loci in the ancestor haplotypes), we defined the recombination rate to be 0.05, otherwise it is 0.00001. We simulated the recombination process for each progeny haplotype; but to force every progeny haplotype to have at least one recombination, in the rare cases where no recombination event was simulated for an progeny haplotype, we sampled one of the hotspots randomly and forced it to recombine with another ancestor chosen at random at that loci. (Thus our simulated samples were not exactly distributed according to the generative model we used, but such samples were arguably more close to the real data.) Overall, 30 datasets each containing 100 individuals (i.e., 200 haplotypes) with 100 SNPs were generated from $K_s = 5$ ancestor haplotypes.

As baseline models, we also implemented 3 standard fixed-dimensional HMM, with K' equal to 3, 5 (the true number of ancestors for the simulated) and 10 hidden states, respectively, which correspond to the number of ancestors available for recombination. For these baseline HMMs, we follow the same mutation model for emission as that of the HMDP (i.e., Eq. (4)), and we also subject the mutation rate to a Beta prior. In these HMMs, the SNP-types of the ancestors at every locus, e.g., $a_{t,k}$, are treated as the mean parameters of the observed SNPs samples at the corresponding locus; the inheritance variables $\{C_{i,t}\}$ correspond to the latent states following a 1-st order Markov process; and the transition models governing recombinations amongst the ancestors as indicated by the values $c_{i,t}$'s are parameterized by a K'-dimensional stochastic matrix. We estimate these parameters via a maximal likelihood principle using the Balm-Welch algorithm. Note that since K' is chosen *a priori*, we cannot estimate the number of ancestors using these HMMs.

Following a *collapsed* Gibbs sampling scheme (Liu 1994), we integrated out the mutation rate θ, and sample variables $\{A_{k,t}\}$ and $\{C_{i,t}\}$ iteratively. We monitor convergence based on the occupancy counts of the top factors in the master DP. Typically, convergence was achieved after around 3000 samples (Figure 2), and the samples obtained after convergence (with proper de-autocorrelation, i.e., by using samples from every 10 iterations over $5000 \sim 10000$ samples) are used for computing relevant sufficient statistics. To increase the chance of proper mixing, 10 independent runs of sampling, with different random seeds, are simultaneously performed. Convergence is monitored at runtime using an on-line minimal pairwise Gelman-Rubin (GR) statistic (Gelman

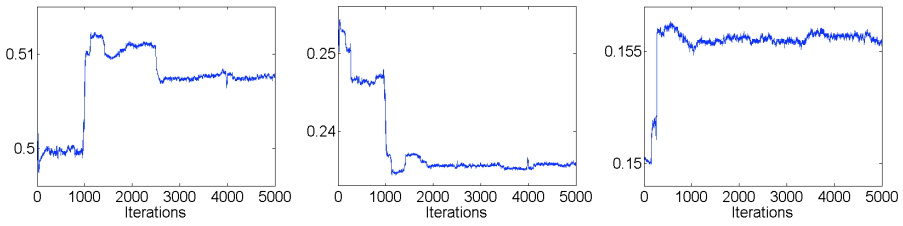

Figure 2: Sampling trace of the top three most occupied factors (ancestor chromosomes). The x-axis represents the sampling iteration, and the y-axis represent the fraction of the occupancy (i.e., be chosen as recombination target) of each factor over total occupancy.

1998) of scalar summaries of the model parameters (e.g., average occupancy of top factors) obtained in each Markov chain. The total running time for posterior inference on a simulated data set described below was around 3.5 hours using a matlab implementation on a Dell PowerEdge 1850 workstation with an Intel Xeon 3.6 GHz processor. (This computation includes a huge disk-writing overhead for recording the running trace. The actual CPU time for computing is less than 10% of that. We intend to soon release a C++ implementation which is expected to further reduce computation cost.)

Ancestral Inference Using HMDP, we successfully recovered the correct number (i.e., $K = 5$) of ancestors in 21 out of 30 simulated populations; for the remaining 9 populations, we inferred 6 ancestors. From samples of ancestor states $\{a_{k,t}\}$, we reconstructed the ancestral haplotypes under the HMDP model. For comparison, we also inferred the ancestors under the 3 standard HMM using an EM algorithm. We define the *ancestor reconstruction error* ϵ_a for each ancestor to be the ratio of incorrectly recovered loci over all the chromosomal sites. The average ϵ_a over 30 simulated populations under 4 different models are shown in Figure 3a. In particular, the average reconstruction errors of HMDP for each of the five ancestors are 0.026, 0.078, 0.116, 0.168, and 0.335, respectively. There is a good correlation between the reconstruction quality and the population frequency of each ancestor. Specifically, the average (over all simulated populations) fraction of SNP loci originated from each ancestor among all loci in the population is 0.472, 0.258, 0.167, 0.068 and 0.034, respectively. As one would expect, the higher the population frequency of an ancestor is, the better its reconstruction accuracy. Interestingly, under the fixed-dimensional HMM, even when we use the correct number of ancestor states, i.e., $K = 5$, the reconstruction error is still very high (Figure 3), typically 2.5 times or higher than the error of HMDP. We conjecture that this is because the non-parametric Bayesian treatment of the transition rates and ancestor configurations under the HMDP model leads to a desirable adaptive smoothing effect and also less constraints on the model parameters, which allow them to be more accurately estimated. Whereas under a parametric setting, parameter estimation can easily be sub-optimal due to lack of appropriate smoothing or prior constraints, or deficiency of the learning algorithm (e.g., local-optimality of EM).

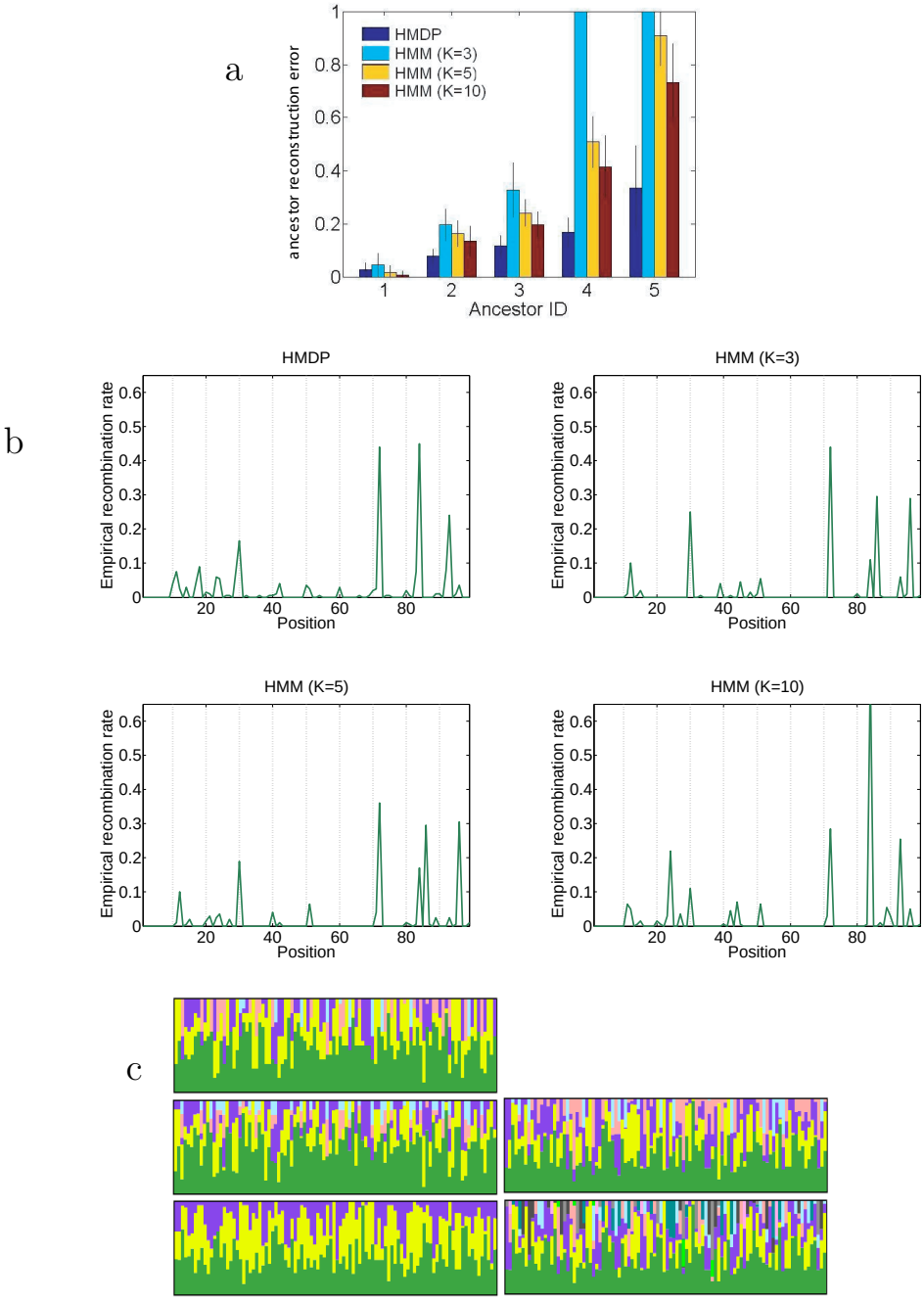

Figure 3: Analysis of simulated haplotype populations. (a) A comparison of ancestor reconstruction errors for the five ancestors (indexed along x-axis). The vertical lines show ± 1 standard deviation over 30 populations. (b) Plots of the empirical recombination rates along 100 SNP loci in one of the 30 populations for HMDP and 3 HMMs. The dotted lines show the pre-specified recombination hotspots. (c) The true (panel 1) and estimated (panel 2 for HMDP, and panel 3-5 for 3 HMMs) population maps of ancestral compositions in a simulated population. Figures were generated using the software *distruct* from Rosenberg *et al* [2002].

threshold	0.01			0.03		
tolerance window	0	± 1	± 2	0	± 1	± 2
False positive rate	0.16	0.12	0.067	0.08	0.04	0.03
False negative rate	0	0	0	0.77	0.55	0.55

Table 1: False positive and false negative rates for recombination hotspot detection using medians of the empirical recombination rates over 30 population samples as shown in Figure 4.

LD-block Analysis From samples of the inheritance variables $\{c_{i,t}\}$ under HMDP, we can infer the recombination status of each locus of each haplotype. We define the empirical recombination rates λ_e at each locus to be the ratio of individuals who had recombinations at that locus over the total number of haploids in the population. Figure 3b shows plots of the λ_e from HMDP and the 3 HMMs in one of the 30 simulated populations. We can identify the recombination *hotspots* directly from such a plot based on an empirical threshold λ_t (i.e., $\lambda_t = 0.05$). For comparison, we also give the true recombination hotspots (depicted as dotted vertical lines) chosen in the ancestors for simulating the recombinant population. The inferred hotspots (i.e., the λ_e peaks) show reasonable agreement with the reference in both HMDP and HMMs, but it appears that in the HMMs the hotspots around position 20 and 60 are less obvious. Figure 4 shows a boxplot of the empirical recombination rates at the 100 SNP loci estimated from the the 30 different population samples simulated from these ancestors. The gray vertical lines along the x-axis correspond to the locations of pre-specified recombination hotspots. A simple thresholding at 0.01 would identify 24 hotspots which include all the 9 true hotspots and 15 false positive sites. This leads to the false negative rate to be 0 and the false positive rate to be 0.16. To give credit to the false positive sites which are close to the true hotspots, we may allow small discrepancy between the true hotspots and the detected ones. By allowing ±2 sites discrepancy and eliminating possibly redundant ones in the detection, (e.g., the two detected sites 70 and 71 would be just counted as 1 site of 70), the number of false positive sites decreased to 6, which resulted in the false positive rate of 0.067 and the false negative rate unchanged. Using a threshold of 0.03, 10 hotspots would be detected, among which two sites agree with the true ones. After allowing ±2 sites discrepancy 4 true hotspots could be identified with 3 remaining false positive sites. The false positive and negative rates using these two thresholds are summarized in Table 1.

Population Structural Analysis Finally, from samples of the inheritance variables $\{c_{i,t}\}$, we can also uncover the genetic origins of all loci of each individual haplotype in a population. For each individual, we define an empirical *ancestor composition vector* η_e, which records the fractions of every ancestor in all the $c_{i,t}$'s of that individuals. Figure 3c displays a *population map* constructed from the η_e's of all individual. In the population map, each individual is represented by a thin vertical line which is partitioned into colored segments in proportion to the ancestral fraction recorded by η_e. Five population maps, corresponding to (1) true ancestor compositions, (2) ancestor compositions inferred by HMDP, and (3-5) ancestor compositions inferred by HMMs with 3, 5, 10 states, respectively, are shown in Figure 3c. To assess the accuracy of

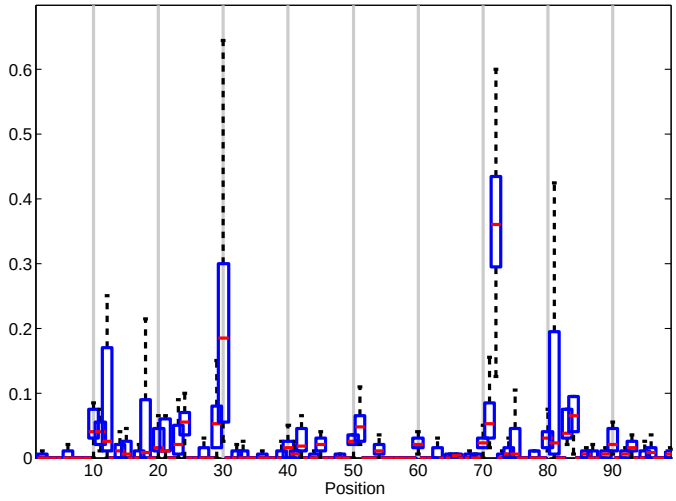

Figure 4: Boxplot of the empirical recombination rates at the 100 SNP loci over 30 different simulated population samples. The gray vertical lines show the pre-specified recombination hotspots used for simulating the data.

our estimation, we calculated the distance between the true ancestor compositions and the estimated ones as the mean squared distance between true and the estimated η_e over all individuals in a population, and then over all 30 simulated populations. We found that the distance between the HMDP-derived population map and the true map is 0.190 ± 0.0748, whereas the distance between HMM-map and true map is 0.319 ± 0.0676, significantly worse than that of HMDP even though the HMM is set to have the true number of ancestral states (i.e., $K = 5$). Because of dimensionality incompatibility and apparent dissimilarity to the true map for other HMMs (i.e., $K = 3$ and 10), we forgo the above quantitative comparison for these two cases.

To summarize our analyses on the simulated data, although the fixed dimensional HMMs are fast and easy to implement, they appear to offer much less accurate results than that of the HMDP model on ancestor reconstruction, and population-map estimation, even when the number of HMM states is set to the true number of haplotype ancestors (which is in practice unknown). When the number of HMM states is chosen incorrectly, the inference results degrade significantly. For hotspot prediction, qualitatively we have not seen significant differences in the accuracy, although the HMDP model appeared to be slightly better. We will look into this issue via a more quantitative analysis in our later study.

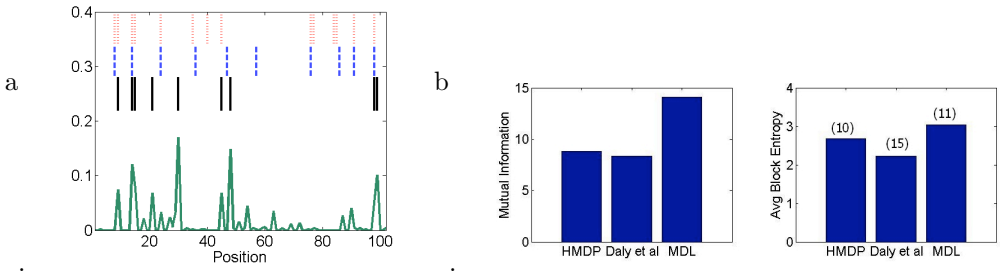

Figure 5: Analysis of the Daly data. (a) A plot of λ_e estimated via HMDP; and the haplotype block boundaries according to HMDP (black solid line), HMM (Daly et al. 2001) (red dotted line), and MDL (Anderson and Novembre 2003) (blue dashed line). (b) IT scores for haplotype blocks from each method. The left panel shows cross-block MI and the right shows the average within-block entropy. The total number of blocks inferred by each method are given on top of the bars.

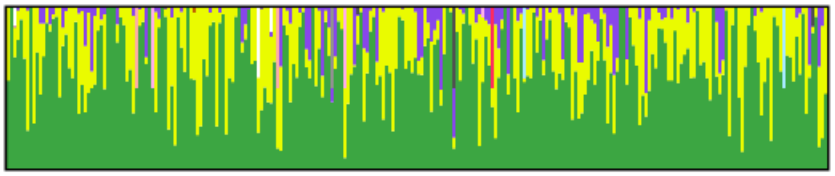

Figure 6: The estimated population map of the Daly dataset.

4.2 Analyzing two real haplotype datasets

We applied HMDP to two real haplotype datasets, the single-population Daly data (Daly et al. 2001), and the two-population (CEPH: Utah residents with northern/western European ancestry; and YRI: Yoruba in Ibadan and Nigeria) HapMap data (Consortium" 2005; Thorisson et al. 2005). These data consist of trios of genotypes, so most of the true haplotypes can be directly inferred from the genotype data. Note that for these real biological data, there is no ground truth regarding the ancestral history, hotspot location, and population composition, based on which we can validate our results, or compare to other methods. We present our analysis as a demonstration of the utilities of our model, which, to our knowledge, are not offered jointly under a unified model by extant methods in statistical genetics. (As we discuss in the sequel, some extant methods can perform some of the inference tasks that HMDP does, and in these cases we show a comparison.)

The single-population Daly dataset We first analyzed the 256 individuals from Daly data. This data set consists of the haplotypes 103 SNPs across a 616.7-kb region on chromosome 5q31 of 129 trios from a European-derived population. Earlier studies indicate that this region contains a genetic risk factor for Crohn disease. Earlier analysis of this data set using a hidden Markov model revealed the existence of discrete haplotype

blocks, each with low diversity, in this region (Daly et al. 2001).

We compared the recovered recombination hotspots with those reported in Daly et al. (2001) (which is based on an HMM employing different number of states at different chromosome segments) and in Anderson and Novembre (2003) (which is based on a minimal description length (MDL) principle applied to Daly's HMM). Note that the HMM used by Daly et al. (2001) and Anderson and Novembre (2003) is different from the ones we used in our simulation study in section 4.1. Their HMM models a stochastic process that selects haplotype-segments from pools of "ancestors" without mutation for a concatenating list of haplotype-block regions constituting the study SNP sequences. Each region has their own ancestor pool of possibly unequal sizes; thus between each pair of adjacent blocks, the HMM needs a unique (possibly rectangular) stochastic matrix for ancestor transitions. The block boundaries are fixed under this HMM (and the only stochasticity lies in the choice of local "ancestors" for each block), and determining the block boundaries is treated as a model-selection problem based on a maximal-likehood (Daly et al. 2001) or MDL (Anderson and Novembre 2003) principle. Strictly speaking, Daly's HMM model itself offers little means to infer recombination events and the ancestor association map, because the "ancestors" thereof are defined independently for each block rather than as whole founding chromosomes; different blocks have different number of ancestors; and the determination of these "local ancestors" employs an initial heuristic scan for regions of low haplotype diversity, whose formal connection to the HMM model is not clear.

Figure 5a shows the plot of empirical recombination rates estimated under HMDP, side-by-side with the reported recombination hotspots. There is no ground truth to judge which one is correct; hence we computed information-theoretic (IT) scores based on the estimated within-block haplotype frequencies and the between-block transition probabilities under each model for a comparison. Figure 5b shows a comparison of these scores for haplotype blocks obtained from HMDP and the other two sources. The left panel of Figure 5b shows the total pairwise mutual information between adjacent haplotype blocks segmented by the recombination hotspots uncovered by the three methods. The right panel shows the average entropies of haplotypes within each block. The number above each bar denotes the total number of blocks. The pairwise mutual information score of the HMDP block structure is similar to that of the Daly structure, but smaller than that of MDL. Similar tendencies are observed for average entropies. Note that the Daly and the MDL methods allow the number of haplotype founders to vary across blocks to get the most compact local ancestor constructions. Thus their reported scores are an underestimate of the true global score because certain segments of an ancestor haplotype that are not or rarely inherited are not counted in the score. Thus the low IT scores achieved by HMDP suggest that HMDP can effectively avoid inferring spurious global and local ancestor patterns. This is confirmed by the population map shown in Figure 6, which shows that HMDP recovered 6 ancestors and among them the 3 dominant ancestors account for 98% of all the modern haplotypes in the population.

We did not compare our results with that of Daly et al. (2001) and Anderson and Novembre (2003) exhaustively, e.g., on ancestor reconstruction and population map estimation, because their methods cannot perform these inferential tasks. Indeed, to

our knowledge there is no single model that does all the inferential tasks HMDP is capable of. Thus we can only compare HMDP with specialized models on certain tasks, as described above. Since implementations of the methods in Daly et al. (2001) and Anderson and Novembre (2003) are not available, we can only compare with their results reported on the original papers, which are obtained on the Daly data. But we cannot apply their methods to our simulated data or the HapMap data for more informative comparisons. The total running time of our algorithm on the Daly data set (with the 3000 burn-in steps, 3000 samples, and 1 per 5 sample deceleration sampling interval) is about 14hr, which includes the disk-writing overhead for trace-recording.

The two-population HapMap dataset The HapMap data was generated by the International HapMap Project that attempts to identify and catalog genetic similarities and differences in human beings of different ethnic origins (Consortium" 2005; Thorisson et al. 2005). The current release of the whole HapMap data contains over 1 million SNPs, from 269 individuals belonging to four populations. In this study, we only focus on a small subset of SNPs common to all populations; we use data from two of the four populations, YRI and CEPH. Specifically, we have 30 trios of YRI and 30 trios of CEPH (i.e., 180 individuals in total), of which the 120 unrelated phase-known individuals corresponding to the parents in the trios were used in the experiment (the children's haplotypes are inherited from the parents and are redundant in the population). We concern ourselves with 254 SNPs, which are located in the region of $ENm010.7p15.2$ spanning 497.5 kilo-basepair (kb). The computation time for analyzing this data set is comparable to that of the Daly data set.

We applied HMDP to the union of the populations, with a random individual order. Delightfully, the two-population structure is clearly retrieved from the population map constructed from the population composition vectors η_e for every individual. As seen in Figure 7a, the left half of the map clearly represents the CEPH population and the right half the YRI population. We found that the two dominant haplotypes covered over 85% of the CEPH population (and the overall breakup among all four ancestors is 0.5618, 0.3036, 0.0827, 0.0518). On the other hand, the frequencies of each ancestor in YRI population are 0.2141, 0.1784, 0.3209, 0.1622, 0.1215 and 0.0029, showing that the YRI population is much more diverse than CEPH. This might explain an earlier observation that genetic inference on the YRI population appeared to be more difficult than for CEPH (Marchini et al. 2006). The recombination maps of the two different populations also show noticeably different spatial patterns of recombination hotspots (Figure 7b), which may reflect different recombination histories of the founders of the two populations.

Note that the population partition result reported in Figure 7b is trivial because it is inferred purely based on SNPs haplotypes without knowledge of ethnic labels of the samples. In most genetic samples, ethnic labels are either not available or ambiguous (e.g., the Daly data has no subpopulation details). By discovering the right population separation, one can perform hotspot estimation for each population and capture population-specific LD (as in Figure 7a); whereas in a mixed population, one may not be able to correctly estimate such patterns.

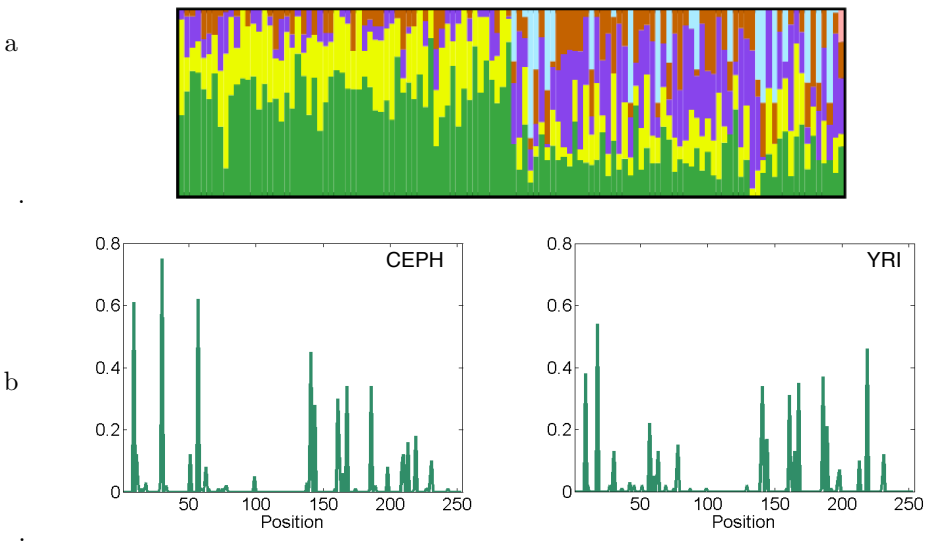

Figure 7: Result on the two-population (CEPH and YRI) HapMap data. (a) The estimated population map of the whole dateset with two populations. (b) The estimated recombination rates along the chromosomal position in the two populations.

5 Conclusion

We have proposed a new Bayesian approach for joint modeling of genetic recombinations among possibly infinite founding alleles and coalescence-with-mutation events in the resulting genealogies. By incorporating a hierarchical DP prior to the stochastic matrix underlying an HMM, which facilitates a well-defined transition process between infinitely many ancestors, our proposed method can efficiently infer a number of important genetic variables, such as recombination hotspot, mutation rates, haplotype origin, and ancestor patterns, jointly underly a unified statistical framework.

Empirically, on both simulated and real data, our approach compares favorably to its parametric counterpart—a fixed-dimensional HMM (even when the number of its hidden states, i.e., the ancestors, is correctly specified) and a few other specialized methods, on ancestral inference, haplotype-block uncovering and population structural analysis. We are interested in further investigating the behavior of an alternative scheme based on reverse-jump MCMC over Bayesian HMMs with different latent states in comparison with HMDP; we also intend to apply our methods to genome-scale LD and demographic analysis using the full HapMap data. While our current model employs only phased haplotype data, it is straightforward to generalize it to unphased genotype data as provided by the HapMap project. HMDP can also be easily adapted to many engineering and information retrieval contexts such as object and theme tracking in open space.

References

Anderson, E. C. and Novembre, J. (2003). "Finding haplotype block boundaries by using the minimum-description-length principle." *Am J Hum Genet*, 73: 336–354.

Antoniak, C. E. (1973). "Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems." *Annals of Statistics*, 2: 1152–1174.

Beal, M. J., Ghahramani, Z., and Rasmussen, C. E. (2002). "The Infinite Hidden Markov Model." In Dietterich, T., Becker, S., and Ghahramani, Z. (eds.), *Advances in Neural Information Processing Systems 14*, 577–584. MIT Press.

Blackwell, D. and MacQueen, J. B. (1973). "Ferguson Distributions Via Pólya Urn Schemes." *Annals of Statistics*, 1: 353–355.

Consortium", I. H. (2005). "A haplotype map of the human genome." *Nature*, 437: 1299–1320.

Daly, M. J., Rioux, J. D., Schaffner, S. F., Hudson, T. J., and Lander, E. S. (2001). "High-resolution haplotype structure in the human genome." *Nature Genetics*, 29(2): 229–232.

Erosheva, E., Fienberg, S., and Lafferty, J. (2004). "Mixed-membership models of scientific publications." *Proc Natl Acad Sci U S A*, 101 (Suppl 1): 5220–5227.

Escobar, M. D. and West, M. (2002). "Bayesian density estimation and inference using mixtures." *Journal of the American Statistical Association*, 90: 577–588.

Excoffier, L. and Hamilton, G. (2003). "Comment on Genetic Structure of Human Populations." *Science*, 300(5627): 1877b–.

Excoffier, L. and Slatkin, M. (1995). "Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population." *Molecular Biology and Evolution*, 12(5): 921–7.

Falush, D., Stephens, M., and Pritchard, J. K. (2003). "Inference of population structure: Extensions to linked loci and correlated allele frequencies." *Genetics*, 164(4): 1567–1587.

Ferguson, T. S. (1973). "A Bayesian analysis of some nonparametric problems." *Annals of Statistics*, 1: 209–230.

Gelman, A. (1998). "Inference and monitoring convergence." In Gilks, W. E., Richardson, S., and Spiegelhalter, D. J. (eds.), *Markov Chain Monte Carlo in Practice*. Boca Raton, Florida: Chapman & Hall/CRC.

Greenspan, G. and Geiger, D. (2004a). "High density linkage disequilibrium mapping using models of haplotype block variation." *Bioinformatics*, 20 (Suppl.1): 137–144.

— (2004b). "Model-Based Inference of Haplotype Block Variation." *Journal of Computational Biology*, 11(2/3): 493–504.

Hoppe, F. M. (1984). "Pólya-like urns and the Ewens' sampling formula." *Journal of Math. Biol.*, 20(1): 91–94.

Hudson, R. R. (1983). "Properties of a neutral allele model with intragenic recombination." *Theor Popul Biol.*, 23(2): 183–201.

Ishwaran, H. and James, L. F. (2001). "Gibbs sampling methods for stick-breaking priors." *Journal of the American Statistical Association*, 90: 161–173.

Kimmel, G. and Shamir, R. (2004). "Maximum likelihood resolution of multi-block genotypes." In *proceedings of the Eighth Annual International Conference on Research in Computational Molecular Biology (RECOMB 2004)*, 2–9. The Association for Computing Machinery.

Kingman, J. (1982). "On the genealogy of large populations." *J. Appl. Prob.*, 19A: 27–43.

Li, N. and Stephens, M. (2003). "Modelling Linkage Disequilibrium, and identifying recombination hotspots using SNP data Genetics." *Genetics*, 165: 2213–2233.

Liu, J. S. (1994). "The collapsed Gibbs sampler with applications to a gene regulation problem." *J. Amer. Statist. Assoc*, 89: 958–966.

Liu, J. S., Sabatti, C., Teng, J., Keats, B., and Risch, N. (2001). "Bayesian analysis of Haplotypes for Linkage Disequilibrium Mapping." *Genome Res.*, 11: 1716–1724.

Marchini, J., Cutler, D., Patterson, N., Stephens, M., Eskin, E., Halperin, E., Lin, S., Qin, Z., Munro, H., Abecasis, G., Donnelly, P., and Consortium, I. H. (2006). "A Comparison of Phasing Algorithms for Trios and Unrelated Individuals." *The American Journal of Human Genetics*, 78: 437–450.

Neal, R. M. (2000). "Markov chain sampling methods for Dirichlet process mixture models." *J. Computational and Graphical Statistics*, 9(2): 249–256.

Niu, T., Qin, S., Xu, X., and Liu, J. (2002). "Bayesian haplotype inference for multiple linked single nucleotide polymorphisms." *American Journal of Human Genetics*, 70: 157–169.

Patil, N., Berno, A. J., et al. (2001). "Blocks of Limited Haplotype Diversity Revealed by High-Resolution Scanning of Human Chromosome 21." *Science*, 294: 1719–1723.

Pritchard, J. K., Stephens, M., Rosenberg, N., and Donnelly, P. (2000). "Association mapping in structured populations." *Am. J. Hum. Genet.*, 67: 170–181.

Rannala, B. and Reeve, J. P. (2001). "High-resolution multipoint linkage-disequilibrium mapping in the context of a human genome sequence." *Am J Hum Genet.*, 69(1): 159–78.

Rasmussen, C. E. (2000). "The Infinite Gaussian Mixture Model." In *Advances in Neural Information Processing Systems 12*, 554–560. Cambridge, MA: MIT Press.

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

Sethuraman, J. (1994). "A Constructive Definition of Dirichlet Priors." *Statistica Sinica*, 1(4): 639–50.

Sohn, K.-A. and Xing, E. (2006). "Hidden Markov Dirichlet Process: Modeling Genetic Recombination in Open Ancestral Space." In *Advances in Neural Information Processing Systems 19*. Cambridge, MA: MIT Press.

Stephens, M., Smith, N., and Donnelly, P. (2001). "A new statistical method for haplotype reconstruction from population data." *American Journal of Human Genetics*, 68: 978–989.

Tavare, S. and Ewens, W. (1998). "The Ewens Sampling Formula." *Encyclopedia of Statistical Sciences*, Update Volume 2.: 230–234.

Teh, Y., Jordan, M. I., Beal, M., and Blei, D. (2006). "Hierarchical Dirichlet processes." *Journal of the American Statistical Association* (to appear).

Thorisson, G. A., Smith, A. V., Krishnan, L., and Stein, L. D. (2005). "The International HapMap Project Web site." *Genome Research*, 15: 1591–1593.

Xing, E., Sharan, R., and Jordan, M. (2004). "Bayesian Haplotype Inference via the Dirichlet Process." In *Proceedings of the 21st International Conference on Machine Learning*, 879–886. New York: ACM Press.

Zhang, K., Deng, M., Chen, T., Waterman, M., and Sun, F. (2002). "A dynamic programming algorithm for haplotype block partitioning." *Proc. Natl. Acad. Sci. USA*, 99(11): 7335–39.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. 0523757, and and by the Pennsylvania Department of Health's Health Research Program under Grant No. 2001NF-Cancer Health Research Grant ME-01-739. E.P.X. is also supported by a NSF CAREER Award under Grant No. DBI-0546594.