A New Nonparametric Bayesian Model for Genetic Inference in Open Ancestral Space

Eric P. Xing Kyung-Ah Sohn

August 1, 2006
CMU-ML-06-111

Carnegie Mellon

A New Nonparametric Bayesian Model for Genetic Inference in Open Ancestral Space

Eric P. Xing **Kyung-Ah Sohn**

August 1, 2006
CMU-ML-06-111

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Keywords: Dirichlet Process, hidden Markov Model, recombination, clustering, Ancestral inference, population genetics

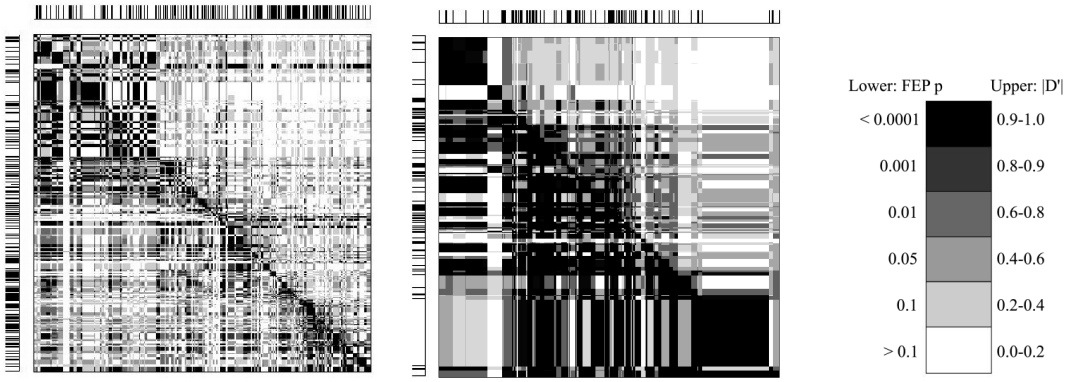
Abstract

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 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.

1 Introduction

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 this two loci can be characterized by a number of so-called *LD measures*. For example, the L1 distance between $p(X, Y)$ and $p(X)p(Y)$, where $p(\cdot, \cdot)$ and $p(\cdot)$ denotes the empirical joint and marginal distribution, respectively, of marker states in a population, can be used as a general LD measure for arbitrary markers. For bi-allelic markers (i.e., markers that have only two possible states), the most popular LD measures in the genetics community include the *gametic disequilibrium*, D'; and the p-value for Fisher's exact test. For a population that is typed at a sequence of polymorphic loci, the LD pattern over all loci-pairs is expected to offer some empirical picture of the aforementioned block structures on chromosomes (Fig 1). However, this kind of descriptive, population-level analysis offers limited insight of the underlying genetic processes (e.g., recombination) that generate these patterns, and 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.



(a) HapMap DB (b) Daly et al's data

Figure 1: The LD measurements, $|D'|$ (upper right), and the p-values for Fisher's exact test (lower left), of two real data sets. (a) The two-population haplotype dataset from the HapMap project [Thorisson *et al.*, 2005]. Note that as the two populations are mixed, the LD-block structures in the LD map are rather inobvious (compared to the LD patterns in Fig 9, where the populations are shown separately). (b) The single-population haplotype dataset from Daly *et al.* [2001]. In each of the LD maps, starting from the upper-left corner, all the markers are listed in top-down and left-right directions, and each marker is at a spatial position corresponding to its actual genetic distance with respect to the first marker (at the upper-left corner). The bars left-to and on-top-of the LD maps denote the genetic location of each marker with respect to the first marker.

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 (SNP) 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 [2004] 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 capturing 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 memembership 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 fix-dimensional multinomial-distributions of marker alleles. However, the admixture models developed so far do not 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 funding alleles [Excoffier and Hamilton, 2003].

These 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 on the Dirichlet process 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 process 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 (Fig 2).

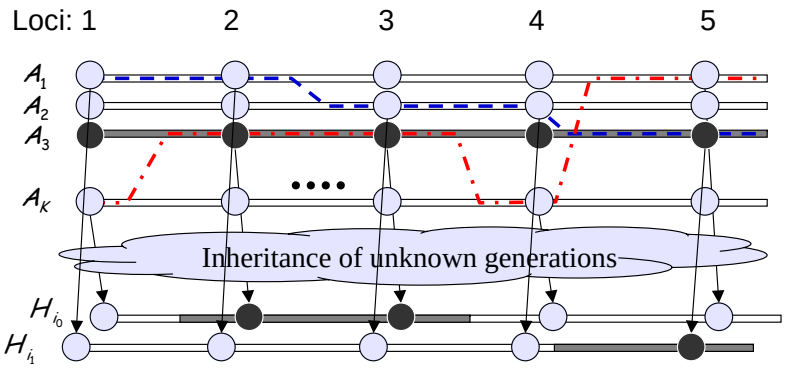


Figure 2: An illustration of a hidden Markov Dirichlet process for haplotype recombination and inheritance. Note that the total number of ancestors is unknown.

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.* [2001]), 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, 2003], 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.

2 Preliminary: The Data Sets and the Inference Problems

Before presenting our statistical model, we begin with a brief description of the data sets to be analyzed in this paper, and the specific inference problems we concern.

The data of interest are population samples of what is known as *haplotypes* of *single nucleotide polymorphisms*, or SNPs. Single nucleotide polymorphism represents the largest class of individual differences in DNA. A SNP refers to the existence of two possible kinds of nucleotides from $\{A, C, G, T\}$ at a single chromosomal locus in a population; each variant is called an *allele*. 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 (Fig 2).

We will analyze two 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 [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 [1].

[1] In general, the haplotypes of each individual is ambiguous given the genotype data; and inferring haplotypes

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 (so that here we don't have to devolve into large-scale computation and system implementation issue, a point we would like to address elsewhere); and we use data from two of the four populations, YRI and CEPH, because only these two populations contain trios of genotypes, which allows unambiguous determination of most of the true haplotypes. 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, assuming no mutation occurs in a single-generation inheritance). We concern ourselves with 254 SNPs (i.e., SNPs at 254 genomic loci), which are located in the region of $ENm010.7p15.2$ spanning 497.5 kilo-basepair (kb). The LD patterns of these SNPs are displayed in Fig 1a.

The Daly set [Daly *et al.*, 2001] 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, and the LD patterns based on traditional approaches (i.e., marker versus marker) are shown in Fig 1b. 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].

Given the haplotypes of a sequence of SNPs from a sample population, we are interested in the following questions: 1) Recovering a possible set of founding haplotypes that may give rise to all the haplotypes in the study population. We refer to this problem as *ancestral inference*. Note that *a priori* we have no knowledge about the exact number of possible founders, and typical solutions to problems of this nature employ a *model selection* procedure according to, for example, Bayes factor [Kass and Raftery, 1995]. We proposed a full Bayesian treatment of this problem that leverages the Dirichlet process models. 2) Inferring where recombinations take place in each individual and where are the recombination hotspots at the population level. We refer this problem as *recombination analysis*, and we propose a model-based posterior inference approach conditioning on the entire haplotype data rather than the pairwise LD measure as in conventional analysis. 3) Inferring the ancestral origin of each SNP in each individual haplotype, and thereby estimate the *ancestral composition* of each modern individual. We refer to this problem as *ancestral mapping*, and again we tackle it via Bayesian inference in an open, recombining, ancestral space. In the sequel we present a statistical model that addresses these inference problems jointly.

from genotypes from arbitrary population(s), a problem known as *haplotype phasing*, is itself a challenging statistical inference problem and has attracted significant amount of research in the statistics community (e.g., [Excoffier and Slatkin, 1995], [Niu *et al.*, 2002], [Stephens *et al.*, 2001]). To keep this paper focused, we omit an elaborated discussion on this problem, but see [Xing *et al.*, 2004; Xing *et al.*, 2006] for our resent works on this subject based on the nonparametric Bayesian formalism leveraged in this paper.

3 Hidden Markov Dirichlet Process for Recombination

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, 2003], 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.*, 2001; 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.

3.1 Dirichlet Process mixtures

For self-containedness, we begin with a quick overview of the fundamentals of Dirichlet process and its connection to the coalescent process in population genetics, followed by a brief recap 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-mutation* (but without recombination), one can treat a haplotype from a modern individual in a study population as a descendent of their most resent common ancestor (MRCA), which is 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 [2] and each color to a possible haplotype, this partition is equivalent to the one resulted 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 [3], Blackwell and MacQueen [1973] showed that this Pólya urn model yields samples whose

[2] 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.

[3] Here we deviate from the conventional notations in the statistics literature (e.g., [Neal, 2000; Escobar and West, 1995; Ishwaran and James, 2001]) and use Q and Q_0, instead of G and G_0 (or H), to denote the random probability measure under DP and the base measure of DP, respectively, because in the genetic context, G and H have been

distributions are those of 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 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 [1994] 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 its 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, 1995] (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, 1995; 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 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

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].

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 pattern of major common ancestor alleles (i.e., founders) in a population bottleneck [4], 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, and at the same time it provides a reasonable approximation to the coalescence model by utilizing the partition structure resulted thereof (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, bellow we summarize the Dirichlet process mixture model we proposed in [Xing *et al.*, 2004] for haplytope inheritance without recombination.

Following Xing *et al.* [2004; 2006], let $H_i = [H_{i,1}, \ldots, H_{i,T}]$ denote a haplotype over T SNPs from chromosome i [5]; 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.

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

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

[5] We ignore the parental origin index of haplotype 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].

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

– 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 to be originated 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 ancestor due to recombinations between an unknown number of founders.

3.2 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.* [2001], 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 HMM urns, Q_0, is set aside to hold colored balls to be drawn at the onset of the HMM state-transition sequence [6]. Each of the remaining HMM urns is painted with a color represented by at least one ball in the stock urn,

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

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 state 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 $\phi = \{\phi_1, \ldots, \phi_n\}$ from Q_0, the conditional prior for $(\phi_n|\phi_{-n})$, where the subscript "$-n$" denotes the index set of all but the n-th ball, is

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

where $\phi_k^*, k = 1, \ldots, K$ denote the K distinct values (i.e., colors) of ϕ (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:

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

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 $\pi_j \equiv [\pi_{j,1}, \pi_{j,2}, \ldots]$, 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 π_0, and transition matrix Π whose rows are defined by $\{\pi_j : j > 0\}$.

Finally, as in, e.g., Escobar and West [1995] and Rasmussen [2000], we can also introduce vague priors such as a Gamma or an inverse Gamma for the scaling parameters α and τ.

3.3 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 Fig 1 for an illustration.

First recall that a base measure F at the root of HDP 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 HMDM, 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 HMDM 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 resulted 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 admit a Beta prior with hyperparameter (α_h, β_h) [7], 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 $\mathrm{Beta}(\alpha_h,\beta_h)$ (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 descried above points naturally to an algorithm for population genetic inference. Unlike the classical coalescence models for recombination [Hudson, 1983], which has been primarily used for theoretical analysis and simulation, but hardly feasible for reverse ancestral inference based on observed genetic data, the HMDP model described above for recombination and inheritance provides a semi-parametric Bayesian formalism that are well suited for data-driven posterior inference on the latent variables that can yield rich information of 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. If 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 HMDM.

[7] For simplicity, we assume that the mutation rates pertaining to different ancestors follow the same prior $\mathrm{Beta}(\alpha_h,\beta_h)$.

4 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 variables (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.

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 DPs (i.e., the urns corresponding to the recombination choices by each ancestor) — $\{m_k \ : \ \forall k\} \cup \{m_{k,k'} \ : \ \forall k, k'\}$. And let $\mathbf{l}_k$ denote the sufficient statistics associated with all haplotype instances originated from ancestor k. The predictive distribution of a δ-block of inheritance variables can be written as:

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

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 $p(c_{t+1:t+\delta} \,|\mathbf{c}^-, \mathbf{h}, \mathbf{a})$ playing the role of the prior. 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:

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

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:

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

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 boarder 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:

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

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 and 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 [8]:

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

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

[8] 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 $\mathrm{Beta}(\alpha_h, \beta_h)$. This is arguably a more accurate reflection of reality.

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.

5 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) Recombination Analysis: inferring the recombination sites in each individual haplotype and uncover population-level recombination hotspots on the chromosome region; 3) Ancestral Mapping: inferring the genetic origins of all loci of each individual haplotype in a population.

5.1 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 3, 5 (the true number of ancestors for the simulated) and 10 hidden states, respectively.

Following a *collapsed* Gibbs sampling scheme [Liu, 1994], we integrated out the mutation rate θ, and sampled variables $\{a_{k,t}\}$ and $\{c_{i,t}\}$ iteratively. We monitored convergence based on the occupancy counts of the top factors in the master DP. Typically, convergence was achieved after around 3000 samples (Fig 3), and the samples obtained after convergence (with proper de-

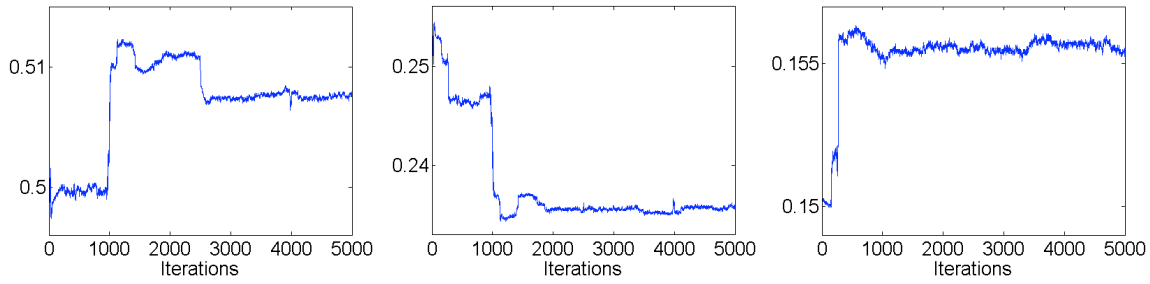


Figure 3: Sampling trace of the top three most occupied factors (ancestor chromosomes). The x-axis represents the sampling iteration, and the y-axis represent the fraction the occupancy (i.e., be chosen as recombination target) of each factor over total occupancy.

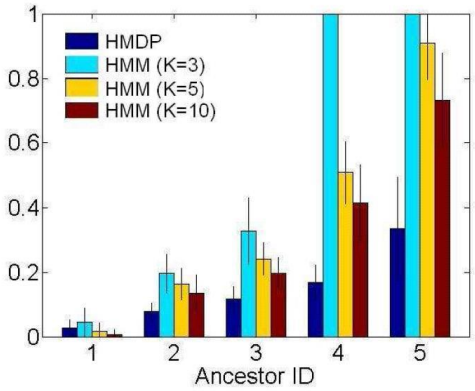
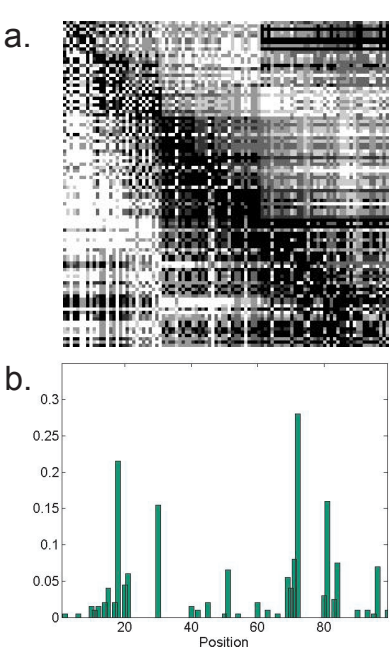


Figure 4: A comparison of ancestor reconstruction errors for the five ancestors (indexed along x-axis) in the simulated haplotype populations. The vertical lines show ± 1 standard deviation over 30 populations.

autocorrelation, i.e., by using samples from every 10 iterations) were used for computing relevant sufficient statistics. To increase the chance of proper mixing, 10 independent runs of sampling, with different random seeds, were simultaneously performed. Convergence were monitored at runtime using an on-line minimal pairwise Gelman-Rubin (GR) statistics [Gelman, 1998] of scalar summaries of the model parameters (e.g., average occupancy of top factors) obtained in each Markov chain.

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 Fig 4. 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



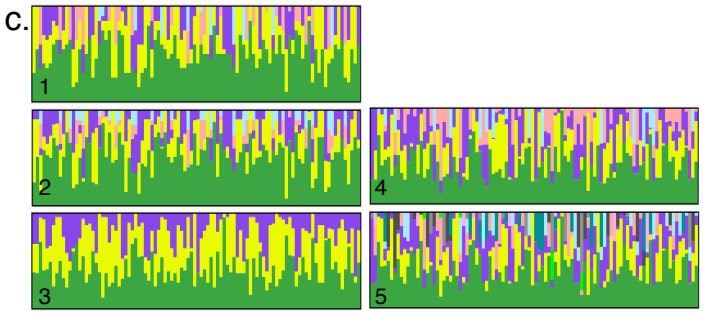


Figure 5: Analysis of simulated haplotype populations. (a) LD measurement $|D'|$ (upper right), and the p-value for Fisher's exact test (lower left) of 100 SNP loci in one of the 30 populations. (b) A plot of the empirical recombination rates along the 100 SNP loci in the same population. (c) The true (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].

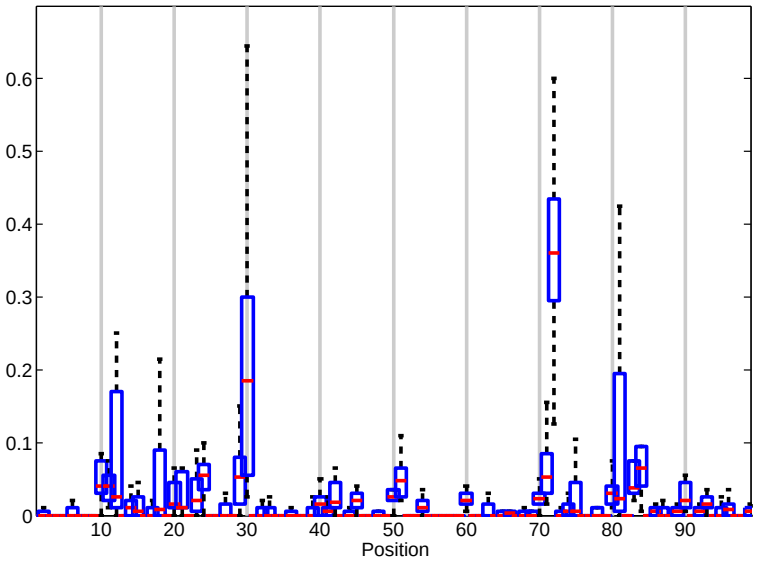


Figure 6: Boxplot of the empirical recombination rates at the 100 SNP loci over 30 different simulated population samples. The gray vertical lines show the pre-specified recombination hotspots used for simulating the data.

threshold	0.01			0.03		
tolerance window	0	$\pm$ 1	$\pm$ 2	0	$\pm$ 1	$\pm$ 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 Fig 6.

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 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 (Fig 4), 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 (comparing to a parametric prior on, e.g., the transition rates) on the model parameters, which allow them to be more accurately estimated. Whereas under a parametric setting, parameter estimation can easily go sub-optimum due to lack of appropriate smoothing or prior constraints, or deficiency of the learning algorithm (e.g., local-optimality of EM).

Recombination Analysis As discussed earlier, traditional LD-measures only capture pairwise couplings between every pair of SNPs, which may be inefficient to unravel the recombination structures in a population sample. In Fig 5a, we demonstrate such inability in an LD-map of one of the 30 simulated sample populations. The upper right triangle of the map shows the gametic disequilibrium, $|D'|$, and the lower left triangle shows the p-value for Fisher's exact test for every pair of loci. As one can see, although some sort of block structures are faintly visible in this LD-map, identifying the recombination hotspots directly from this map is apparently nontrivial, if possible. Alternatively, under the HMDP model, from samples of the inheritance variables $\{c_{i,t}\}$ obtained via Gibbs sampling, 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 are determined to have recombinations at that locus over the total number of haplotypes in the population. Fig 5b shows a plot of the λ_e in the same simulated population used for plotting Fig 5a. (For comparison we scale the width of the λ_e-plot to be the same as that of the LD-map, so that the horizontal positions in the λ_e-plot are aligned with those in the LD-map.) From the λ_e-plot, we can identify the recombination *hotspots* directly based on an empirical threshold λ_t. As we can see, some of the estimated recombination spots, as represented by the high peaks in Fig 5b, are not apparent in the LD-map shown in Fig 5a. This suggests that our model-based approach are more sensitive to the statistical dependencies that would have been resulted from natural recombinations, which are beyond pairwise couplings captured by the classical LD-analysis. Recall that the true recombination hotspots chosen in the ancestors for simulating the recombinant population are located at every 10th loci in the 100-loci long ancestral haplotypes; the inferred hotspots (i.e., the λ_e peaks) in the this example population sample show reasonable agreement with the reference.

More rigorously, Fig 6 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.

Ancestral Mapping 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. Fig 5c 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 Fig 5c. To assess the accuracy of 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.

5.2 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 and YRI) HapMap data [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.

The single-population Daly dataset We first analyzed the 256 individuals (i.e., 512 haplotypes) from the Daly data set (after excluding one person due to severe missing data). As in the **recombination analysis** for simulated data (§5.1), we computed the empirical recombination rate λ_e for

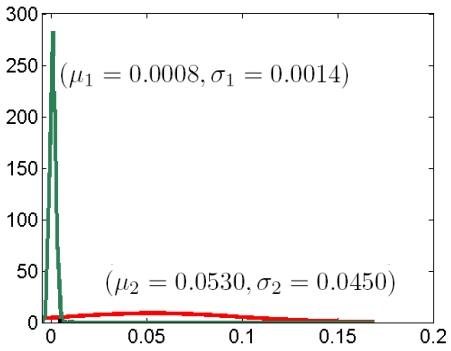


Figure 7: Analysis of the Daly data: a mixture of Gaussian fitting of the estimated λ_e.

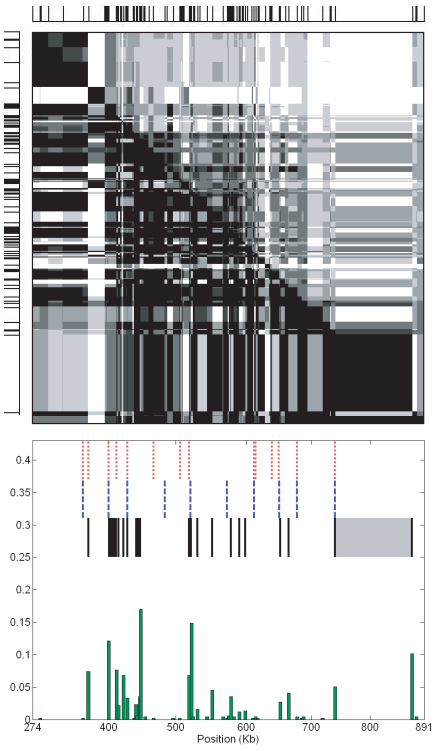


Figure 8: Analysis of the Daly data. Upper panel: the LD-map of the data. Lower panel: a plot of λ_e estimated via HMDP; and the haplotype block boundaries according to HMDP (black solid line), HMM [Daly *et al.*, 2001] (red dotted line), and MDL [Anderson and Novembre, 2003] (blue dashed line). Note that the width of the λ_e-plot is scaled to be the same as that of the LD-map, and the spacing of the loci are scaled accordingly, so that the horizontal positions in the λ_e-plot are aligned with those in the LD-map. Also note that the thickness of the black solid lines delineating the haplotype blocks is proportional to the width of the hotspot regions between adjacent blocks. (But the long span of the last hotspot region between 738.46-877.57 with only 3 SNPs is depicted with gray shade.

each locus, based on which the locations of possible recombination hotspots can be estimated, and the haplotype blocks can be identified. Rather than picking an empirical threshold as in §5.1, here we determined the recombination hotspot as follows. We fitted the estimated λ_e's of all loci with a one-dimensional mixture of Gaussians (Fig 7). Then we used the intersection point of the two Gaussian components as the threshold for determining hotspot-loci. This threshold is essentially the point where the posterior probabilities of λ_e being a baseline recombination rate or a hotspot recombination rate are equal. The mass in the area where the two Gaussians overlap represents the Bayes-error of loci classification under this model. One can also employ more rigorous model-based methods for hotspot classification, we will return to this point in Section 6. Figure 8 shows the plot of the empirical recombination rates over all loci, and the estimated hotspots. Note that according to the HMDP model, certain estimated recombination hotspots are very close to each other, for example, at loci 398kb, two hotspots are right next to each other. This finding suggests that the actual LD patterns in a population sample may not simply fall into blocks with sharp boundaries universal to all individuals, as assumed in Daly's HMM model. It is more appropriate to define "hotspot regions" (i.e., stretch of consecutive hotspot loci) rather than point "hotspot loci", where necessary, to delineate haplotype blocks, as discussed in [Li and Stephens, 2003]. For example, according to the estimated λ_e's shown in Fig 8, 15 hotspot loci/regions (represented as thick solid vertical bars in Fig 8) were identified, and they divide the entire study region into 16 haplotype blocks of low diversity. Note that in Fig 8, the x-axis represents the actual genetic locations of the SNP loci (starting from 274kb at the leftmost with respect to a genetic reference), not the integer indices of the SNPs. Since the SNPs of interest are not located uniformly in this region, the spatial-intervals as seen from Fig 8 between hotspots may not reflect the "length" of the haplotype blocks. For example, the block between 445-518kb contains 15 SNPs. Whereas the seemingly longest interval between 738-877kb contains only 3 SNPs, two of which have high recombination rates, which render this interval to be a hotspot region as explained below. Biologically, this is not surprising because the probability of recombination between adjacent SNPs increases with their physical distance, in addition to depending on the intrinsic recombination rate. This "hotspot region" between 738-877kb is more likely to be merely a consequence of sparse location-sampling of SNPs in this region, rather than a biologically meaningful hotspot region.

Table 2 summarizes the summary-statistics that characterize each haplotype block (and hotspot regions). We used the threshold of 0.005 determined by the mixture of Gaussians as described above to identify recombination hotspots. The blocks were determined accordingly, with the constrain that the lengths of the identified blocks are at least three-SNPs long, to avoid over-fragmenting the haplotypes. (This means that within our "hotspot regions" there may be 1 or 2 "cold-spots" separating the hotspots). In column 1 of Table 2, the blocks with blockID starting with an "r" represent the hotspots regions which contain more than 2 SNPs, others represent the haplotype blocks. The numbers of SNPs within the blocks varied from 3 to 15 (the second column of Table 2). The actual genomic region and length of each block are shown in the third and the fourth columns, respectively. The lengths of the smallest and the biggest blocks were 1.3kb and 93kb, respectively, while the average was 22kb. We also report the total number of distinct haplotypes as a reflection of diversity for each block, of which the most diverse is, not suprisingly, one of the largest blocks (which spans 71kb), which contains 17 different haplotypes. This is sig-

blockID	#SNPs	region (Kb)	length (Kb)	#hap.	Anc.freq						#hap. (frq>3)	coverage (%)	#hap. (95%)	#hap. (90%)
1	9	(274.04-366.81)	92.8	12	0.805	0.190	0.001	0.002	0.002	0.000	3	0.98	3	2
2	5	(395.08-398.35)	3.3	7	0.816	0.176	0.004	0.002	0.002	0.000	2	0.98	2	2
(r1)	3	(398.35-411.87)	13.5											
3	3	(411.87-413.23)	1.4	7	0.633	0.164	0.199	0.002	0.002	0.000	6	0.99	4	3
4	3	(415.58-419.85)	4.3	5	0.613	0.162	0.219	0.002	0.002	0.002	4	1.00	2	2
5	3	(424.28-425.55)	1.3	4	0.548	0.162	0.278	0.002	0.008	0.002	2	0.99	2	2
6	3	(433.47-437.68)	4.2	5	0.534	0.161	0.262	0.014	0.027	0.002	3	1.00	3	2
(r2)	5	(437.68-445.34)	7.7											
7	15	(445.34-518.48)	73.1	17	0.636	0.157	0.164	0.010	0.029	0.004	9	0.95	9	6
(r3)	5	(518.48-522.60)	4.1											
8	3	(522.60-529.56)	7.0	5	0.585	0.282	0.076	0.010	0.043	0.004	4	1.00	4	3
9	3	(532.36-553.19)	20.8	6	0.594	0.275	0.081	0.005	0.041	0.004	3	0.99	3	2
10	9	(570.98-579.82)	8.8	6	0.583	0.286	0.065	0.014	0.049	0.004	3	0.99	3	2
11	6	(582.65-590.59)	7.9	8	0.614	0.286	0.033	0.014	0.049	0.004	5	0.99	3	2
12	3	(594.12-598.80)	4.7	5	0.621	0.287	0.031	0.008	0.049	0.004	4	1.00	3	2
13	15	(601.29-649.90)	48.6	17	0.627	0.291	0.020	0.009	0.049	0.004	10	0.95	11	9
14	3	(657.23-662.82)	5.6	4	0.605	0.289	0.043	0.010	0.049	0.004	4	1.00	3	2
15	8	(676.69-738.46)	61.8	13	0.563	0.297	0.076	0.009	0.051	0.004	9	0.97	8	5
(r4)	3	(738.46-877.57)	139.1											
16	4	(877.57-890.71)	13.1	6	0.489	0.384	0.066	0.006	0.045	0.010	3	0.99	3	3

Table 2: Haplotype block structures and the summary statistics of the blocks for the Daly data. The block boundaries correspond to the x-coordinates of the λ_e peaks in Fig. 9a.

nificantly lower than the 2^{17} possible different haplotypes one could observe had there existed no co-inheritance (i.e., due to fully random recombination) among loci in this block. Note that the 17 haplotypes reported here are the actual total observed diversity in this region among the study population, not the number of prototypes underlying these haplotypes that parsimoniously account for the majority of the observed diversity when small amount mutation are allowed, as reported in [Daly *et al.*, 2001]. The actual demographic diversity of these blocks is actually much lower than that is reflected by the total number of haplotypes, as obviated from the results in column 6-15. In column 6-11 of Table 2, we report the ancestor association frequencies of haplotypes within each block, where the associations were directly estimated from the inheritance variable $c_{i,t}$'s sampled by our algorithm. We can see that overall 6 founders sufficed to fully account for our data, and indeed within each block, only 3-4 of them were significantly used. It is worth pointing out that our HMDM model explicitly allows mutations during inheritance, therefore there can be many different individual haplotypes even if they originated from the same ancestor. For practical applications such as association studies, we can focus on the dominant haplotypes which cover the majorities of the population such that the remaining few can be derived by mutating only one or two sites from the dominant ones. We present the number of necessary haplotypes to cover over 95% and 90% of the entire population, which were mostly around 3 with a few blocks with higher diversity around 10.

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). Again in Fig 8 we show the plot of the empirical recombination rates estimated under HMDP, side-by-side with the reported recombination hotspots. For such real biological data, there is no ground truth to judge which one is correct because we do not know where are the true recombination hotspots; 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 9 shows a comparison of these scores for haplotype blocks

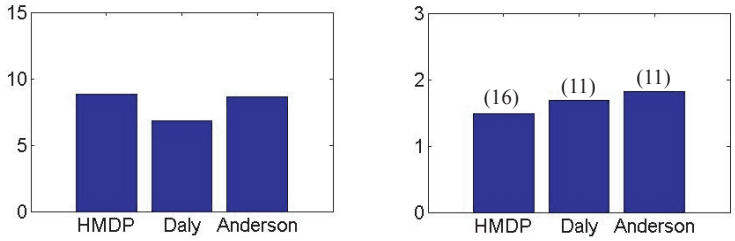


Figure 9: The IT scores of the haplotype blocks inferred by different methods for the Daly data. The left panel shows cross-block MI and the right shows the average within-block entropy. The total number of blocks inferred by each method are given on top of the bars.

Figure 10: The estimated population map of the Daly dataset. The ordering of all individuals in the sample population was determined by a K-means clustering with $K = 6$, followed by a within-cluster ordering of samples based on their distances to the cluster centroid. The black vertical bars show the K-means cluster boundaries.

obtained from HMDP and the other two sources. The left panel of Fig 9 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 was similar to that of the MDL structure, but that of Daly was the smallest. Not that due to the presence of hotspot regions in our segmentation of the population haplotypes, the transition probabilities between blocks are sometimes not directly defined if between the two blocks there exists a hotspot region containing multiple highly heterogeneous SNPs. In our IT-score calculation, we took into account the mutual information pertaining to all adjacent blocks and hotspots, and between all adjacent hotspots. Thus our cross-block MI is a pessimistic reflection of the compactness of the block structures due to inclusion of the hotspot regions. The within-block entropy is arguably a more faithful reflection of this property. As shown in the right panel of Fig 9, the average entropy of the HMDP structure was indeed lower than either of the other methods.

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 reflected in the population map shown in Fig 10, 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.

For a more informative revelation of the underlying population structure captured by the empirical ancestor composition vector η_e, in Fig 10 we clustered the individuals based on their η_e's, and

then ordered all individuals accordingly. Specifically, all individuals were clustered into $K = 6$ clusters (which is an empirical choice just for illustration) using the K-means algorithm; then within each group, individual ordering were determined by their distances to the cluster centroid. Interestingly, we can see that although the Daly data are reported to be from a European-derived population that is expected to be genetically less diverse, our ancestral map clearly suggests that in this population there exists multiple genetically well-defined and distinct sub-populations each with a unique ancestral composition. As of now, we do not have access to the ethnic-label of each individual, thus we defer a more qualitative validation of this finding.

The two-population HapMap dataset The HapMap data contains 60 individuals from CEPH and 60 from YRI. We applied HMDP to the union of the populations, with a random individual order. Interestingly, although no population label is given to the HMDP model, the two-population structure of the HapMap data was clearly retrieved from the population map constructed from the population composition vectors η_e for every individual. As seen in Fig 11a, 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 YRI population appeared to be more difficult than for CEPH [Marchini *et al.*, 2006].

The inferred two-population structures prove to be very useful for more informative recombination analysis of the population sample (especially when the true population labels are not known or not discriminative enough as in the Daly data) because such information can be used to separate sub-populations, and uncover population-specific recombination patterns originally confounded by the mixing of the populations. To show this, we invite the readers to inspect again the LD-map of the whole HapMap data shown in Fig 1a, which reveals no significant LD patterns. Now, knowing the population devision of the data, we can plot the LD-maps separately for each population, i.e., CEPH and YRI. As shown in the top panels of Fig 11b, the now population-specific LD-maps start to exhibit more obvious, and distinctive LD patterns. Echoing the emergence of new patterns in the LD maps, the recombination maps of the two different populations also show noticeably different spatial patterns of recombination hotspots (Fig. 11b), which may reflect different recombination histories of the founders of the two populations. The LD plots here as well as the recombination plots have been scaled according to the genomic distance between adjacent loci, where the x-ticks (and the symmetric y-ticks) in the LD plots show the relative positions of each SNP .

Table 3 and 4 summarize the within-block properties of the two populations of CEPH and YRI as described for the Daly's dataset. For CEPH population, we found 21 blocks, where the block lengths range from 1kb to 80.1kb, with an average of 18.0kb. Except for the 11th and 16th blocks, at most 7 different haplotypes were enough to describe over 90 $\%$ of the population. For YRI population, 24 blocks were found, where the block lengths range from 1.5 kb (3 SNPs) to 95.6 kb (42 SNPs) with an average of 17.4 kb. In 21 out of the total 24 blocks, 90$\%$ of the individuals of the YRI population could be covered by 9 haplotypes. It appears that for the YRI

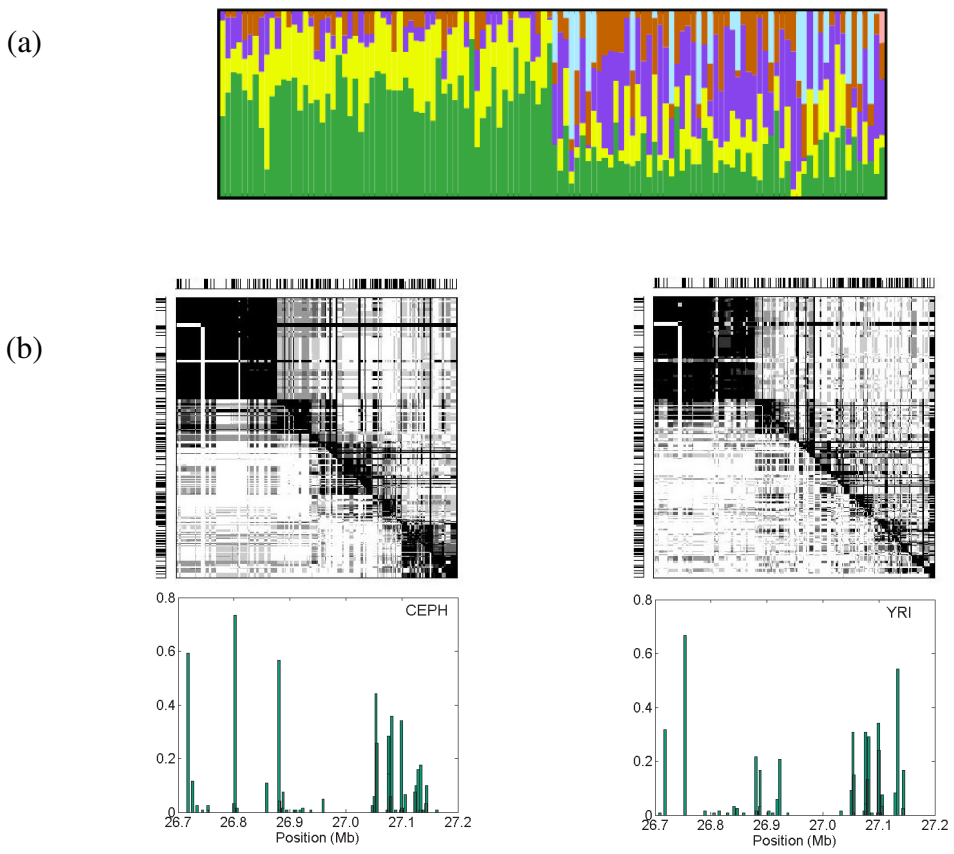


Figure 11: Result on the two-population (CEPH and YRI) HapMap data. (a) The estimated population map of the whole dateset with two populations. (b) The LD measure with the estimated recombination rates along the chromosomal position in the population of CEPH (left) and YRI (right).

population, the genomic regions between 26.94-27.04Mb and 27.14-27.20Mb harbor much more diverse haplotypes than in similar genomic regions in the CEPH population, which suggests that the amount of genetic drift due to mutations are different for the two populations, and the identified regions are of interest to probe for underlying reasons, such as selections or purifying effects.

6 Conclusion

We have proposed a new Bayesian approach for jointly modeling 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 well-defined transition process between infinite ancestor space, 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.

blockID	#SNPs	region (Mb)	length (Kb)	#hap.	Anc.freq						#hap. (frq>3)	coverage (%)	#hap. (95%)	#hap. (90%)
1	9	(26.70-26.71)	10.0	7	0.279	0.408	0.079	0.143	0.091	0.000	4	0.96	4	3
(r1)	5	(26.71-26.74)	24.9											
2	3	(26.74-26.75)	9.6	3	0.408	0.183	0.175	0.146	0.087	0.000	3	1.00	3	3
(r2)	5	(26.75-26.76)	10.8											
3	11	(26.76-26.80)	46.7	6	0.412	0.183	0.168	0.157	0.079	0.000	4	0.97	4	3
(r3)	3	(26.80-26.81)	3.7											
4	3	(26.81-26.81)	2.4	3	0.200	0.392	0.175	0.154	0.079	0.000	2	0.99	2	2
5	18	(26.81-26.86)	52.3	9	0.201	0.404	0.161	0.154	0.079	0.000	4	0.91	6	4
6	6	(26.86-26.89)	22.8	7	0.146	0.417	0.208	0.154	0.075	0.000	5	0.97	5	4
(r4)	10	(26.89-26.90)	12.8											
7	7	(26.90-26.91)	13.8	5	0.445	0.146	0.200	0.129	0.079	0.000	3	0.97	3	3
8	3	(26.91-26.91)	1.0	4	0.454	0.138	0.200	0.129	0.079	0.000	4	1.00	4	3
(r5)	5	(26.91-26.93)	14.0											
9	9	(26.93-26.94)	12.3	7	0.554	0.108	0.204	0.054	0.079	0.000	3	0.95	3	3
10	12	(26.94-26.96)	19.3	14	0.550	0.108	0.208	0.054	0.079	0.000	8	0.94	9	7
11	39	(26.97-27.05)	80.1	34	0.525	0.106	0.210	0.079	0.079	0.000	6	0.68	28	22
(r6)	8	(27.05-27.06)	16.7											
12	15	(27.06-27.08)	16.4	11	0.108	0.496	0.146	0.192	0.058	0.000	6	0.93	7	6
(r7)	5	(27.08-27.08)	2.9											
13	3	(27.08-27.08)	1.5	2	0.271	0.325	0.192	0.171	0.042	0.000	2	1.00	2	2
14	3	(27.08-27.09)	2.1	4	0.350	0.308	0.208	0.092	0.042	0.000	4	1.00	3	3
15	6	(27.09-27.09)	5.1	7	0.370	0.117	0.404	0.063	0.042	0.004	5	0.97	5	4
16	11	(27.09-27.10)	4.7	36	0.371	0.113	0.404	0.067	0.042	0.004	9	0.63	30	24
(r8)	6	(27.10-27.10)	5.2											
17	9	(27.10-27.11)	3.7	7	0.409	0.276	0.217	0.062	0.033	0.004	4	0.97	4	4
18	11	(27.11-27.12)	15.4	11	0.346	0.333	0.221	0.067	0.029	0.004	6	0.95	6	6
(r9)	12	(27.12-27.13)	10.2											
19	8	(27.13-27.14)	3.7	8	0.408	0.287	0.200	0.075	0.025	0.004	5	0.94	6	5
(r10)	6	(27.14-27.14)	5.7											
20	12	(27.14-27.16)	16.6	13	0.471	0.250	0.158	0.092	0.025	0.004	5	0.93	7	4
21	11	(27.16-27.20)	39.5	13	0.471	0.254	0.158	0.087	0.025	0.004	7	0.93	8	7

Table 3: Haplotype block structure of CEPH population in HapMap data.

blockID	#SNPs	region (Mb)	length (Kb)	#hap.	Anc.freq						#hap. (frq>3)	coverage (%)	#hap. (95%)	#hap. (90%)
1	7	(26.70-26.71)	7.9	4	0.279	0.408	0.079	0.142	0.092	0.000	4	1.00	4	4
(r1)	4	(26.71-26.72)	9.9											
2	9	(26.72-26.76)	37.1	8	0.411	0.168	0.188	0.146	0.087	0.000	4	0.96	4	4
3	9	(26.76-26.79)	37.7	7	0.412	0.183	0.167	0.158	0.079	0.000	4	0.97	4	4
4	6	(26.79-26.81)	14.8	6	0.309	0.285	0.175	0.152	0.079	0.000	3	0.97	3	3
5	3	(26.81-26.82)	5.3	8	0.200	0.396	0.179	0.146	0.079	0.000	6	0.97	6	5
6	3	(26.82-26.83)	13.6	6	0.200	0.396	0.171	0.154	0.079	0.000	4	0.98	4	4
7	3	(26.84-26.85)	9.8	7	0.204	0.396	0.167	0.154	0.079	0.000	4	0.96	4	3
8	3	(26.85-26.85)	1.8	4	0.204	0.404	0.150	0.163	0.079	0.000	4	1.00	3	3
9	6	(26.85-26.86)	9.6	6	0.200	0.417	0.150	0.154	0.079	0.000	5	0.99	5	4
10	6	(26.86-26.89)	22.8	11	0.146	0.417	0.208	0.154	0.075	0.000	8	0.97	8	7
(r2)	8	(26.89-26.89)	6.3											
11	6	(26.89-26.90)	11.3	9	0.442	0.149	0.201	0.129	0.079	0.000	5	0.95	5	5
12	6	(26.91-26.91)	5.2	8	0.452	0.140	0.200	0.129	0.079	0.000	5	0.94	6	5
(r3)	5	(26.91-26.93)	14.0											
13	9	(26.93-26.94)	12.3	13	0.554	0.108	0.204	0.054	0.079	0.000	5	0.88	9	6
14	42	(26.94-27.04)	95.6	73	0.532	0.108	0.208	0.072	0.079	0.000	4	0.23	67	61
15	11	(27.04-27.06)	17.0	17	0.527	0.098	0.217	0.079	0.079	0.000	8	0.88	11	9
(r4)	6	(27.06-27.06)	5.5											
16	15	(27.06-27.08)	16.4	11	0.108	0.496	0.146	0.192	0.058	0.000	10	0.99	9	8
(r5)	5	(27.08-27.08)	2.9											
17	3	(27.08-27.08)	1.5	2	0.271	0.325	0.192	0.171	0.042	0.000	2	1.00	2	2
18	3	(27.08-27.09)	2.1	5	0.350	0.308	0.208	0.092	0.042	0.000	4	0.99	3	3
19	5	(27.09-27.09)	5.0	9	0.371	0.117	0.404	0.063	0.042	0.004	6	0.95	6	5
20	12	(27.09-27.10)	5.2	46	0.370	0.113	0.404	0.067	0.042	0.004	10	0.57	40	34
(r6)	15	(27.10-27.11)	9.0											
21	15	(27.11-27.13)	19.7	16	0.335	0.330	0.221	0.081	0.029	0.004	10	0.93	11	9
22	6	(27.13-27.13)	3.1	9	0.329	0.297	0.212	0.128	0.029	0.004	8	0.99	7	6
23	11	(27.13-27.14)	5.0	16	0.408	0.286	0.200	0.077	0.025	0.004	9	0.92	11	9
(r7)	3	(27.14-27.14)	4.4											
24	23	(27.14-27.20)	57.2	39	0.471	0.252	0.158	0.090	0.025	0.004	11	0.65	33	27

Table 4: Haplotype block structure of YRI population in HapMap data.

Empririically, 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 state, i.e., the ancestors, is correctly specified) and a few other specialized methods, on ancestral inference, haplotype-block uncovering and population structural analysis.

From a statistical modeling point of view, the proposed HMDP model fo recombination conjoins the advantages of two classes of earlier approaches for analyzing LD across multiple genomic loci. The first class of approaches, as used in, e.g., [Daly *et al.*, 2001; Anderson and Novembre, 2003; Patil *et al.*, 2001], adopt a similar assumption as in HMDP, that each observed haplotypes is a "mosaic" of ancestral haplotypes and the formation of the mosaic is governed by a hidden Markov process over the ancestor states. Specifically, it is assumed that each modern chromosome is a concatenation of a sequence of "haplotype blocks"; for each block the haplotype is stochastically chosen from a finite pool of common haplotype patterns—referred to as "ancestors"—without mutation; and as a result, every individual chromosome in the population has identical haplotype-block boundaries. Strictly speaking, this HMM model itself offers little means to infer recombination events, because the block boundaries (which conceptually correspond to the recombination sites) of all individual chromosomes are determined *a priori*, and the only stochasticity lies in the choice of the "ancestors" at each block for each chromosome rather than the genomic locations of recombination events in each chromosome. Furthermore, the recombination templates, i.e., the physical ancestral chromosomes existed in the past that gave rise to modern individual chromosomes via mutation and recombination, are not well-defined in the model. Indeed, the "ancestors" as posited in this model has little connection to what one might expected as complete physical ancestral chromosomes — in this HMM model the "ancestors" are defined independently for each block rather than as whole chromosomes; different blocks have different number of ancestors (which is biologically inconsistent if we view each modern chromosome as mosaic of a common pool of multiple complete ancestral chromosomes via recombination and mutation); 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. It is possible to employ a model selection approach, as in [Greenspan and Geiger, 2004], to couple the inference of the boundary and haplotypes of the "ancestral-blocks" with the parameter estimation of the HMM model, but it is unclear to what extent this class of approaches might be helpful for applications that involves explicit ancestral inference (i.e., whole-genome based ancestral-composition estimation as in, e.g., [Rosenberg *et al.*, 2002]) and for interpreting LD patterns that do not have sharp block boundaries, as seen in Figure 1, which is likely due to the stochastic deviations of the recombination sites among individual chromosomes from the common recombination hotspots. It also does not facilitate statistical estimation of recombination rates over chromosomal region. In contrast, the HMDP proposed in this paper repents a well-defined generative model for the observed haplotypes based on spatial point process for stochastic recombination (in the sense of both choice of recombination participants and choice of recombination locations) and random mutations over an open pool of complete ancestral chromosomes. Although such a generative process is still a simplification of the real biological mechanism underlying LD, it enables joint statistical characterization of a number of genetic variables of interest via posterior inference based on well-founded statistical principles, and it strikes a reasonable tradeoff between being biologically meaningful and computationally

manageable.

The proposed HMDP model is also closely related to another class of LD models, as in [Rannala and Reeve, 2001; Li and Stephens, 2003], which place more emphasis on capturing the mechanistic underpinning of LD patterns, and resort to tractable approximations to the recombinational coalescence. As in HMDP, these models assume that the observed haplotypes are descendants of a randomly-mating population, with stochastic recombination and mutation. Although the data likelihood under this assumption can be in theory defined by a recombinational coalescent process [Hudson, 1983], no closed-form expression of is currently known. In Li and Stephens [2003]'s PAC model, the data likelihood is approximated by a "product of approximate conditionals" of each individual haplotype given some other haplotypes determined by an *ad hoc* ordering of the data. This approach offers a very flexible platform for modeling the recombination processes and facilitates estimation of site specific recombination rates with remarkable computational efficiency and accuracy. However, as a cost of such approximation, PAC likelihood essentially "marginalizes" out the coalescent history of the data, thus one can not infer the ancestral structure of the data as in the previous class of approaches. Moreover, although empirically appears to be non-consequential, the PAC likelihood is non-exchangeable (i.e., depends on the order of the data). As disused in §3.1 and §3.2, HMDP represents a new type of approximation to the coalescent likelihood, which is exchangeable and bears an explicit ancestral structure (which is roughly equivalent to marginalizing out the "partial" coalescent history starting from certain generations before present, and approximating the remaining branches of the coalescent tree with star genealogies). Thus our proposed method does not need the heuristic averaging over sample ordering adopted in [Li and Stephens, 2003], and allows flexible statistical query of the ancestral history of the observed genetic data, including the number of haplotypes of the founders, evolutionary time from founders to current population, time of recombination, etc.

As of now, there are a number of statistical and biological aspects of real data that we have not accounted for here. For example, so far the HMDP model does not intrinsically capture the heterogeneity of recombination rates over loci, and the recombination rates are determined by the posterior distribution of recombination events (i.e., as capture by the indicators $\{C_{i,t}\}$) under a universal recombination rate, rather than directly by an maximum likelihood estimation of site-specific recombination rates as in [Li and Stephens, 2003]. Also, we have not addressed the issue of threshold calculation and confidence measure of hotspot prediction as in [Li and Stephens, 2003]. These problems are of importance in various applications such as linkage-based quantitative train locus mapping and disease-gene mapping. One way of addressing these issues is to explicitly introduce more recombination states (e.g., for both base-line recombination and hotspot-recombination) into the infinite HMM we proposed, and/or to introduce priors for for site-specific recombination rates for Bayesian inference.

Another aspect we have not dealt with extensively is regarding estimating the time of existence of the hypothetical founders and recombination. These queries are of interest in genetic demography studies concerning human divergence, migration and mating history. It is possible to address this by replacing the simplistic recombination and mutation models in the current HMSP with richer, biologically more plausible alternatives that explicitly incorporate the time factor, i.e., based on continuous-time Markov processes, as used in many phylogenetic models for sequence

evolution.

We are also 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; and we 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.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. 0523757.

References

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

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

[Beal *et al.*, 2001] M. J. Beal, Z. Ghahramani, and C. E. Rasmussen. The infinite hidden Markov model. In *Advances in Neural Information Processing Systems 13*, 2001.

[Blackwell and MacQueen, 1973] D. Blackwell and J. B. MacQueen. Ferguson distributions via polya urn schemes. *Annals of Statistics*, 1:353–355, 1973.

[Consortium, 2005] International HapMap Consortium. A haplotype map of the human genome. *Nature*, pages 1229–1320, 2005.

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

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

[Escobar and West, 1995] M. D. Escobar and M. West. Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 90:577–588, 1995.

[Excoffier and Hamilton, 2003] L. Excoffier and G. Hamilton. Comment on 'genetic structure of human populations'. *Science*, 300:1877, 2003.

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

[Falush *et al.*, 2003] D. Falush, M. Stephens, and J. K. Pritchard. Inference of population structure using multilocus genotype data: Linked loci and correlated allele frequencies. *Genetics*, 164:1567–1587, 2003.

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

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

[Greenspan and Geiger, 2003] D. Greenspan and D. Geiger. Model-based inference of haplotype block variation. In *Proceedings of RECOMB 2003*, 2003.

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

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

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

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

[Kass and Raftery, 1995] R. Kass and A. Raftery. Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795, 1995.

[Kimmel and Shamir, 2004] G. Kimmel and R. Shamir. Maximum likelihood resolution of multi-block genotypes. In *Proceedings of RECOMB 2004*, pages 847–56, 2004.

[Kingman, 1982] J.F.C Kingman. On the genealogy of large populations. *J. Appl. Prob.*, 19A:27–43, 1982.

[Li and Stephens, 2003] N. Li and M. Stephens. Modelling linkage disequilibrium, and identifying recombination hotspots using snp data genetics. *Genetics*, 165:2213–2233, 2003.

[Liu *et al.*, 2001] J. S. Liu, C. Sabatti, J. Teng, B.J.B. Keats, and N. Risch. Bayesian analysis of haplotypes for linkage disequilibrium mapping. *Genome Res.*, 11:1716–1724, 2001.

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

[Marchini *et al.*, 2006] J. Marchini, D. Cutler, N. Patterson, M. Stephens, E. Eskin, E. Halperin, S. Lin, Z.S. Qin, H.M. Munro, G.R. Abecasis, P. Donnelly, and International HapMap Consortium. A comparison of phasing algorithms for trios and unrelated individuals. *American Journal of Human Genetics*, 78:437–450, 2006.

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

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

[Patil *et al.*, 2001] N. Patil, A. J. Berno, D. A. Hinds, et al. Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science*, 294:1719–1723, 2001.

[Pritchard *et al.*, 2000] J. K. Pritchard, M. Stephens, N. A. Rosenberg, and P. Donnelly. Association mapping in structured populations. *American Journal of Human Genetics*, 67:170–181, 2000.

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

[Rasmussen, 2000] C. E. Rasmussen. The infinite Gaussian mixture model. In *Advances in Neural Information Processing Systems 12*, 2000.

[Rosenberg *et al.*, 2002] N. A. Rosenberg, J. K. Pritchard, J. L. Weber, H. M. Cann, K. K. Kidd, L. A. Zhivotovsky, and M. W. Feldman. Genetic structure of human populations. *Science*, 298:2381–2385, 2002.

[Sethuraman, 1994] J. Sethuraman. A constructive definition of Dirichlet priors. *Statistica Sinica*, 1(4):639–50, 1994.

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

[Tavare and Ewens, 1998] S. Tavare and W.J. Ewens. The Ewens sampling formula. *Encyclopedia of Statistical Sciences*, Update Volume 2.:230–234, 1998.

[Teh *et al.*, 2006] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 2006. Forthcoming.

[Thorisson *et al.*, 2005] G.A. Thorisson, A.V. Smith, L. Krishnan, and L.D. Stein. The international hapmap project web site. *Genome Research*, 15:1591–1593, 2005.

[Xing *et al.*, 2004] E.P. Xing, R. Sharan, and M.I Jordan. Bayesian haplotype inference via the Dirichlet process. In *Proceedings of the 21st International Conference on Machine Learning*, 2004.

[Xing *et al.*, 2006] E.P. Xing, K.-A. Sohn, M.I Jordan, and Y. W. Teh. Bayesian multi-population haplotype inference via a hierarchical Dirichlet process mixture. In *Proceedings of the 23st International Conference on Machine Learning*, 2006.

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

Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213

Carnegie Mellon