Bayesian Haplotype Inference via the Dirichlet Process*

Eric P. Xing[†‡] Michael I. Jordan[§] Roded Sharan[¶]

Abstract

The problem of inferring haplotypes from genotypes of single nucleotide polymorphisms (SNPs) is essential for the understanding of genetic variation within and among populations, with important applications to the genetic analysis of disease propensities and other complex traits. The problem can be formulated as a mixture model, where the mixture components correspond to the pool of haplotypes in the population. 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. Thus methods for fitting the genotype mixture must crucially address the problem of estimating a mixture with an unknown number of mixture components. In this paper we present a Bayesian approach to this problem based on a nonparametric prior known as the Dirichlet process. The model also incorporates a likelihood that captures statistical errors in the haplotype/genotype relationship trading off these errors against the size of the pool of haplotypes. We describe an algorithm based on Markov chain Monte Carlo for posterior inference in our model. The overall result is a flexible Bayesian method, referred to as *DP-Haplotyper*, that is reminiscent of parsimony methods in its preference for small haplotype pools. We further generalize the model to treat pedigree relationships (e.g., trios) between the population's genotypes. We apply DP-Haplotyper to the analysis of both simulated and real genotype data, and compare to extant methods.

1 Introduction

The availability of a nearly complete human genome sequence makes it possible to begin to explore individual differences between DNA sequences on a genome-wide scale, and to search for associations of such genotypic variation with disease and other phenotypes [17]. The largest class of individual differences in DNA are the *single nucleotide polymorphisms (SNPs)*. Millions of SNPs have been detected thus far out of an estimated total of ten million common SNPs [18].

A SNP commonly has two variants, or *alleles*, in the population, corresponding to two specific nucleotides chosen from $\{A, C, G, T\}$. 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. Recall that for diploid organisms (such as humans) the chromosomes come in pairs. Thus two haplotypes go together to make up a *genotype*, which is the list of *unordered* pairs of alleles in a region. That is, a genotype is obtained from a pair of haplotypes by omitting the specification of the association of each allele with one of the two chromosomes—its *phase*. Common biological methods for assaying genotypes typically do not provide phase information; phase can be obtained at a considerably higher cost [16]. It is desirable to develop automatic methods for inferring haplotypes from genotypes and possibly other data sources (e.g., pedigrees). With a set of inferred haplotypes in hand, associations to disease can be explored.

From the point of view of population genetics, the basic model underlying the haplotype inference problem is a finite mixture model. That is, letting $\mathcal{H}$ denote the set of all possible haplotypes associated with a given

*A preliminary version of this paper appeared in [20].

†To whom correspondence should be addressed.

‡School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213. epxing@cs.cmu.edu.

§Computer Science Division, Department of Statistics, University of California at Berkeley, Berkeley, CA 94720-1776. jordan@cs.berkeley.edu.

¶School of Computer Science, Tel-Aviv University, Tel-Aviv 69978, Israel. roded@tau.ac.il.

region (a set of cardinality 2^k in the case of binary polymorphisms, where k is the number of heterozygous SNPs), the probability of a genotype is given by:

$$p(g) = \sum_{h_1, h_2 \in \mathcal{H}} p(h_1, h_2) \mathbb{I}(h_1 \oplus h_2 = g) \tag{1}$$

where $\mathbb{I}(h_1 \oplus h_2 = g)$ is the indicator function of the event that haplotypes h_1 and h_2 are consistent with g. Under the assumption of Hardy-Weinberg equilibrium (HWE), an assumption that is standard in the literature and will also be made here, the mixing proportion $p(h_1, h_2)$ is assumed to factor as $p(h_1)p(h_2)$.

Given this basic statistical structure, the simplest methodology for haplotype inference is maximum likelihood via the EM algorithm, treating the haplotype identities as latent variables and estimating the parameters $p(h)$ [5]. This methodology has rather severe computational requirements, in that a probability distribution must be maintained on the (large) set of possible haplotypes, but even more fundamentally it fails to capture the notion that small sets of haplotypes should be preferred. This notion derives from an underlying assumption that for relatively short regions of the chromosome there is limited diversity due to population bottlenecks and relatively low rates of recombination and mutation.

One approach to dealing with this issue is to formulate a notion of "parsimony," and to develop algorithms that directly attempt to maximize parsimony. Several important papers have taken this approach [1, 10, 4] and have yielded new insights and algorithms. Another approach is to elaborate the probabilistic model, in particular by incorporating priors on the parameters. Different priors have been discussed by different authors, ranging from simple Dirichlet priors [15] to priors based on the coalescent process [19] to priors that capture aspects of recombination [9]. These models provide implicit notions of parsimony, via the implicit "Ockham factor" of the Bayesian formalism.

We also take a Bayesian statistical approach in the current paper, but we attempt to provide more explicit control over the number of inferred haplotypes than has been provided by the statistical methods proposed thus far, and the resulting inference algorithm has commonalities with the parsimony-based schemes.

Our approach is based on a nonparametric prior known as the *Dirichlet process* [6]. In the setting of finite mixture models, the Dirichlet process—not to be confused with the Dirichlet distribution—is able to capture uncertainty about the number of mixture components [3]. The basic setup can be explained in terms of an urn model, and a process that proceeds through data sequentially. Consider an urn which 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, with a parameter defining the probabilities of these two possibilities. The association of data points to colors defines a "clustering" of the data.

To make the link with Bayesian mixture models, we associate with each color a draw from the distribution defining the parameters of the mixture components. This process defines a *prior distribution* for a mixture model with a random number of components. Multiplying this prior by a likelihood yields a *posterior distribution*. Markov chain Monte Carlo algorithms have been developed to sample from the posterior distributions associated with Dirichlet process priors [3, 14].

The usefulness of this framework for the haplotype problem should be clear—using a Dirichlet process prior we in essence maintain a pool of haplotype candidates that grows as observed genotypes are processed. The growth is controlled via a parameter in the prior distribution that corresponds to the choice of a new color in the urn model, and via the likelihood, which assesses the match of the new genotype to the available haplotypes.

To expand on this latter point, an advantage of the probabilistic formalism is its ability to elaborate the observation model for the genotypes to include the possibility of errors. In particular, the indicator function $\mathbb{I}(h_1 \oplus h_2 = g)$ in Eq. (1) is suspect—there are many reasons why an individual genotype may not match with a current pool of haplotypes, such as the possibility of mutation or recombination in the meiosis for that individual, and errors in the genotyping or data recording process. Such sources of small differences should not lead to the inference procedure spawning new haplotypes.

In the current paper we present, *DP-Haplotyper*, a statistical model for haplotype inference based on a Dirichlet process prior and a likelihood that includes error models for genotypes. We describe a Markov

chain Monte Carlo procedure, in particular a procedure that makes use of both Gibbs and Metropolis-Hasting updates, for posterior inference. We present results of applying our method to the analysis of both simulated and real genotype data, comparing to the state-of-the-art PHASE algorithm [19]. On the simulated data our predictions are comparable to those obtained by PHASE, and superior to those obtained by the EM algorithm. On a real dataset of [2] our results are again comparable to those of PHASE, and we outperform two other algorithms: HAP [11, 4] and HAPLOTYPER [15]. On data from [8], which is a difficult test case due to the small number of individuals in the sample, we outperform PHASE by a significant margin.

2 Haplotype Inference via the Dirichlet Process

The input to a phasing algorithm can be represented as a *genotype matrix* G with columns corresponding to SNPs in their order along the chromosome and rows corresponding to genotyped individuals. $G_{i,j}$ represents the information on the two alleles of the i-th individual for SNP j. We denote the two alleles of a SNP by 0 and 1, and $G_{i,j}$ can take on one of four values: 0 or 1, indicating a homozygous site; 2, indicating a heterozygous site; and '?', indicating missing data.[1]

We will describe our model in terms of a pool of ancestral haplotypes, or *templates*, from which each population haplotype originates [9]. The haplotype itself may undergo point mutation with respect to its template. The size of the pool and its composition are both unknown, and are treated as random variables under a Dirichlet process prior. We begin by providing a brief description of the Dirichlet process and subsequently show how this process can be incorporated into a model for haplotype inference.

2.1 Dirichlet process mixtures

Rather than present the Dirichlet process in full generality, we focus on the specific setting of mixture models, and make use of an urn model to present the essential features of the process. For a fuller presentation, see, e.g., [12]. We assume that data x arise from a mixture distribution with mixture components $p(x|\phi)$. We assume the existence of a *base measure* $G_0(\phi)$, which is one of the two parameters of the Dirichlet process. (The other is the parameter τ, which we present below). The parameter $G_0(\phi)$ is not the prior for ϕ, but is used to generate a prior for ϕ, in the manner that we now discuss.

Consider the following process for generating samples $\{x_1, x_2, \ldots, x_n\}$ from a mixture model consisting of an unspecified number of mixture components, or *equivalence classes*:

- The first sample x_1 is sampled from a distribution $p(x|\phi_1)$, where the parameter ϕ_1 is sampled from the base measure $G_0(\phi)$.
- The ith sample, x_i, is sampled from the distribution $p(x|\phi_{c_i})$, where:
  - The equivalence class of sample i, c_i, is drawn from the following distribution:

$$p(c_i = c_j \text{ for some } j < i | c_1, \ldots, c_{i-1}) = \frac{n_{c_j}}{i - 1 + \tau} \tag{2}$$

$$p(c_i \neq c_j \text{ for all } j < i | c_1, \ldots, c_{i-1}) = \frac{\tau}{i - 1 + \tau}, \tag{3}$$

    where n_{c_i} is the *occupancy number* of class c_i—the number of previous samples belonging to class c_i.
  - The parameter ϕ_{c_i} associated with the mixture component c_i is obtained as follows:

$$\begin{array}{lcll} \phi_{c_i} & = & \phi_{c_j} & \text{if } c_i = c_j \text{ for some } j < i \text{ (i.e., } c_i \text{ is a populated equivalence class),} \\ \phi_{c_i} & \sim & G_0(\phi) & \text{if } c_i \neq c_j \text{ for all } j < i \text{ (i.e., } c_i \text{ is a new equivalence class).} \end{array}$$

[1] Although we focus on binary data here, it is worth noting that our methods generalize immediately to non-binary data, and accommodate missing data.

Eqs. (2) and (3) define a conditional prior for the equivalence class indicator c_i of each sample during a sequential sampling process. They imply a self-reinforcing property for the choice of equivalence class of each new sample—previously populated classes are more likely to be chosen. The parameter ϕ_k for the k-th mixture component, $p(\cdot|\phi_k)$, has an interpretation which is problem-specific. In the case of Gaussian mixtures, this parameter defines the mean and covariance matrix of each mixture component. In the haplotype inference problem, ϕ_k defines underlying genetic parameters for a population. In particular, in the model we describe below, we let $\phi_k := \{A^{(k)}, \theta^{(k)}\}$, where $A^{(k)} := [A_1^{(k)}, ..., A_J^{(k)}]$ is a founding *haplotype configuration*, or *ancestral template*, for genetic loci $t = [1, ..., J]$, and where $\theta^{(k)}$ is the *mutation rate* of this founder.

It is important to emphasize that the process that we have discussed generates a *prior distribution*. We now embed this prior in a full model that includes a likelihood for the observed data. In Section 3 we develop Markov chain Monte Carlo inference procedures for this model.

2.2 DP-Haplotyper: a Dirichlet Process Mixture Model for Haplotypes

We now present a probabilistic model, *DP-haplotyper*, for the generation of haplotypes in a population and for the generation of genotypes from these haplotypes. We assume that each individual's genotype is formed by drawing two random *templates* from an ancestral pool, and that these templates are subject to random perturbation. To model such perturbations we assume that each locus is mutated independently from its ancestral state with the same error rate. Finally, we assume that we are given noisy observations of the resulting genotypes. The model is displayed as a graphical model in Figure 1.

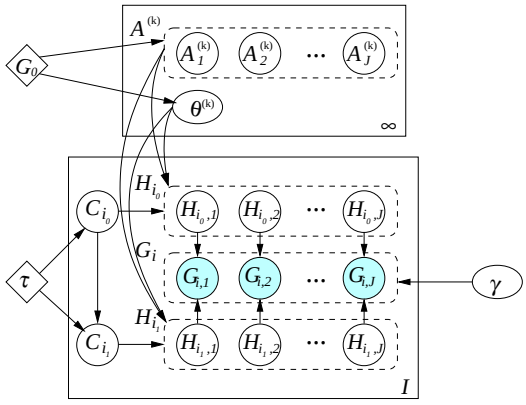

Figure 1: The graphical model representation of the haplotype model with a Dirichlet process prior. Circles represent the state variables, ovals represent the parameter variables, and diamonds represent fixed parameters. The dashed boxes denote sets of variables corresponding to the same ancestral template, haplotype, and genotype, respectively. The solid boxes correspond to i.i.d. replicates of sets of variables, each associated with a particular individual, or ancestral template, respectively.

Let J be an ordered list of loci of interest. For each individual i, we denote his/her paternal haplotype by $H_{i_0} := [H_{i_0,1}, \ldots, H_{i_0,J}]$ and maternal haplotype by $H_{i_1} := [H_{i_1,1}, \ldots, H_{i_1,J}]$. We denote a set of ancestral templates as $\mathbf{A} = \{A^{(1)}, A^{(2)}, \ldots\}$, where $A^{(k)} := [A_1^{(k)}, \ldots, A_J^{(k)}]$ is a particular member of this set.

In our framework, the probability distribution of the haplotype variable H_{i_t}, where the sub-subscript $t \in \{0, 1\}$ indexes paternal or maternal origin, is modeled by a mixture model with an unspecified number of mixture components, each corresponding to an equivalence class defined by the choice of a particular ancestor. For each individual i, we define the equivalence class variables C_{i_0} and C_{i_1} for the paternal and maternal haplotypes, respectively, to specify the ancestral origin of the corresponding haplotype. The C_{i_t} are the random variables corresponding to the equivalence classes of the Dirichlet process. The base measure G_0 of the Dirichlet process is a joint distribution on ancestral haplotypes A and mutation parameters θ, where the latter captures the probability that an allele at a locus is identical to the ancestor at this locus. We let $G_0(A, \theta) \equiv p(A)p(\theta)$, and we assume that $p(A)$ is a uniform distribution over all possible haplotypes. We let $p(\theta)$ be a beta distribution, Beta(α_h, β_h), and we choose a small value for $\beta_h/(\alpha_h + \beta_h)$, corresponding

to a prior expectation of a low mutation rate.

Given C_{i_t} and a set of ancestors, we define the conditional probability of the corresponding haplotype instance $h := [h_1, \ldots, h_J]$ to be:

$$\begin{aligned} p(H_{i_t} = h | C_{i_t} = k, \mathbf{A} = \mathbf{a}, \boldsymbol{\theta}) \\ = \; p(H_{i_t} = h | A^{(k)} = a, \theta^{(k)} = \theta) \\ = \; \prod_j p(h_j | a_j, \theta), \end{aligned} \tag{4}$$

where $p(h_j|a_j, \theta)$ is the probability of having allele h_j at locus j given its ancestor. Eq. (4) assumes that each locus is mutated independently with the same error rate. For haplotypes, $H_{i_t,j}$ takes values from a set B of alleles. We use the following *single-locus mutation model*:

$$p(h_j|a_j, \theta) = \theta^{\mathbb{I}(h_j = a_j)} \Big(\frac{1-\theta}{|B|-1} \Big)^{\mathbb{I}(h_j \neq a_j)} \tag{5}$$

where $\mathbb{I}(\cdot)$ is the indicator function.

The joint conditional distribution of haplotype instances $\mathbf{h} = \{h_{i_t} \; : \; t \in \{0,1\}, i \in \{1, 2, \ldots, I\}\}$ and parameter instances $\boldsymbol{\theta} = \{\theta^{(1)}, \ldots, \theta^{(K)}\}$, given the ancestor indicator $\mathbf{c}$ of haplotype instances and the set of ancestors $\mathbf{a} = \{a^{(1)}, \ldots, a^{(K)}\}$, can be written explicitly as:

$$p(\mathbf{h}, \boldsymbol{\theta} | \mathbf{c}, \mathbf{a}) \propto \prod_k \big[\theta^{(k)}\big]^{m_k + \alpha_h - 1} \Big(\frac{1-\theta^{(k)}}{|B|-1} \Big)^{m'_k} \big[1 - \theta^{(k)}\big]^{\beta_h - 1} \tag{6}$$

where $m_k = \sum_j \sum_i \sum_t \mathbb{I}(h_{i_t,j} = a_{k,j}) \mathbb{I}(c_{i_t} = k)$ is the number of alleles that were not mutated with respect to the ancestral allele, and $m'_k = \sum_j \sum_i \sum_t \mathbb{I}(h_{i_t,j} \neq a_{k,j}) \mathbb{I}(c_{i_t} = k)$ is the number of mutated alleles. The count $\mathbf{m}_k = \{m_k, m'_k\}$ is a sufficient statistic for the parameter θ_k and the count $\mathbf{m} = \{\mathbf{m}_k, \mathbf{m}'_k\}$ is a sufficient statistic for the parameter $\boldsymbol{\theta}$. The marginal conditional distribution of haplotype instances can be obtained by integrating out θ in Eq. (6):

$$p(\mathbf{h}|\mathbf{c}, \mathbf{a}) = \prod_k R(\alpha_h, \beta_h) \frac{\Gamma(\alpha_h + m_k)\Gamma(\beta_h + m'_k)}{\Gamma(\alpha_h + \beta_h + m_k + m'_k)} \Big(\frac{1}{|B|-1} \Big)^{m'_k} \tag{7}$$

where $\Gamma(\cdot)$ is the gamma function, and $R(\alpha_h, \beta_h) = \frac{\Gamma(\alpha_h + \beta_h)}{\Gamma(\alpha_h)\Gamma(\beta_h)}$ is the normalization constant associated with Beta(α_h, β_h). (For simplicity, we use the abbreviation R_h for $R(\alpha_h, \beta_h)$ in the sequel).

We now introduce a *noisy observation model* for the genotypes. We let $G_i = [G_{i,1}, \ldots, G_{i,J}]$ denote the *joint genotype* of individual i at loci $[1, \ldots, J]$, where each $G_{i,j}$ denotes the genotype at locus j. We assume that the observed genotype at a locus is determined by the paternal and maternal alleles of this locus as follows:

$$p(g_{i,j}|h_{i_0,j}, h_{i_1,j}, \gamma) = \gamma^{\mathbb{I}(h_{i,j} = g_{i,j})} [\mu_1(1-\gamma)]^{\mathbb{I}(h_{i,j} \overset{1}{\neq} g_{i,j})} [\mu_2(1-\gamma)]^{\mathbb{I}(h_{i,j} \overset{2}{\neq} g_{i,j})}$$

where $h_{i,j} \triangleq h_{i_0,j} \oplus h_{i_1,j}$ denotes the unordered pair of two actual SNP allele instances at locus j; "$\overset{1}{\neq}$" denotes set difference by exactly one element (i.e., the observed genotype is heterozygous, while the true one is homozygous, or vice versa); "$\overset{2}{\neq}$" denotes set difference of both elements (i.e., the observed and true genotypes are different and both are homozygous); and μ_1 and μ_2 are appropriately defined normalizing constants [2]. We place a beta prior Beta(α_g, β_g) on γ. Assuming independent and identical error models for each locus, the joint conditional probability of the entire genotype observation $\mathbf{g} = \{g_i \; : \; i \in \{1, 2, \ldots, I\}\}$

[2] For simplicity, we may let $\mu_1 = \mu_2 = 1/V$, where V is the total number of ways a single SNP haplotype $h_{i,j}$ and a single SNP genotype $g_{i,j}$ can differ (i.e., 2 for binary SNPs). When different μ_1 and μ_2 are desired to penalize single- and double-disagreement differently, one must be careful to treat the case of homozygous $h_{i,j}$ and heterozygous $h_{i,j}$ differently, because they are related to noisy genotype observations in different manners. For example, a heterozygous $h_{i,j}$ (e.g., 01) can not be related to any genotype with a double-disagreement, whereas a homozygous $h_{i,j}$ (e.g., 00) can (e.g., w.r.t. $g_{i,j} = 11$).

and parameter γ, given all haplotype instances is:

$$\begin{aligned} p(\mathbf{g}, \gamma | \mathbf{h}) &= \prod_i p(g_i, \gamma | h_{i_0}, h_{i_1}) \\ &= \gamma^{\alpha_g + u - 1} \big[1 - \gamma\big]^{\beta_g + u' + u'' - 1} \mu_1^{u'} \mu_2^{u''}, \end{aligned} \tag{8}$$

where the sufficient statistics $\mathbf{u} = \{u, u', u''\}$ are computed as $u = \sum_{i,j} \mathbb{I}(h_{i,j} = g_{i,j})$, $u' = \sum_{i,j} \mathbb{I}(h_{i,j} \overset{1}{\neq} g_{i,j})$, and $u'' = \sum_{i,j} \mathbb{I}(h_{j,i} \overset{2}{\neq} g_{j,i})$, respectively. Note that $u + u' + u'' = IJ$. To reflect an assumption that the observational error rate is low we set $\beta_g/(\alpha_g + \beta_g)$ to a small constant (0.001). Again, the marginal conditional distribution of $\mathbf{g}$ is computed by integrating out γ.

Having described the Bayesian haplotype model, the problem of phasing individual haplotypes and estimating the size and configuration of the latent ancestral pool can be solved via posterior inference given the genotype data. In Section 3 we describe Markov chain Monte Carlo (MCMC) algorithms for this purpose.

2.3 Haplotype Modeling Given Partial Pedigrees

A diploid individual carries two chromosomes, or haplotypes, one of paternal origin and one of maternal origin. When a parent-offspring triplet (or even other close biological relatives) are (geno)typed, the ambiguity of haplotypes of an individual can sometimes be resolved by exploiting the dependencies among the haplotypes of family members induced by genetic inheritance and segregation. For example, if both parents are homozygous, i.e., $g_1 = a \oplus a$, $g_0 = b \oplus b$, and the offspring is heterogeneous, i.e., $g_{\lambda_{10}} = a \oplus b$, where λ_{10} denotes the offspring of subjects "1" and "0," then we can infer that the haplotypes of the offspring are $h_{\lambda_{10}} = (a, b)$. However, inheritance of haplotypes may be more than mere faithful copying. In particular, chromosomal inheritance could be accompanied by single-generation mutations, which alter single or multiple SNPs on the chromosomes, and recombinations, which disrupt and recombine some chromosome pairs in gamete donors to generate novel (i.e., mosaic) haplotypes. Although genotypes of this nature do not directly lead to full resolution of each individual's haplotypes, undoubtedly the strong dependencies that exist among the genotype data (in contrast to the *iid* genotypes we studied in the last section) could be exploited to reduce the ambiguity of the phasing. In order to exploit pedigree information, we need to introduce a few new ingredients into the basic DP-haplotyper model described in the last section and in particular to model the distribution of individual haplotypes in a population consisting of now partially coupled (rather than conditionally independent) individuals (Fig. 2). We refer to this expanded model as the *Pedi-haplotyper* model. Formally, we introduce a segregation random variable, $S_{i_t,j}$, for each one of the two SNP alleles of each locus of an individual, to indicate its meiotic origin (i.e., from which one of the two SNP alleles of a parent it is inherited). For example, $S_{i_t,j} = 1$ indicates that allele $H_{i_t,j}$ is inherited from the maternal allele of individual i's t-parent (where $t = 0$ means father and $t = 1$ means mother). We denote the t-parent of individual i by $\pi(i_t)$, and his/her paternal (resp. maternal) allele by $\pi_0(i_t)$ (resp. $\pi_1(i_t)$). We use the following conditional distribution to model possible mutation during single generation inheritance:

$$p(h_{i_t,j} | s_{i_t,j} = r, h_{\pi_0(i_t),j}, h_{\pi_1(i_t),j}, \epsilon_t) = \big[\epsilon_t\big]^{\mathbb{I}(h_{i_t,j} = h_{\pi_r(i_t),j})} \big[\frac{1 - \epsilon_t}{|B| - 1}\big]^{\mathbb{I}(h_{i_t,j} \neq h_{\pi_r(i_t),j})}, \tag{9}$$

where $1 - \epsilon_t$ is the mutation rate during inheritance, and $r \in \{0, 1\}$ represents the choice of the paternal or maternal alleles of a parent subject by an offspring. Note that this *single generation inheritance model* allows different mutational rates for the parental and maternal alleles if desired (e.g., to reflect the difference in gamete environment in a male or a female body), by letting ϵ_0 and ϵ_1 take different values, or giving them different beta prior distributions. To model possible recombination events during single generation inheritance, we assume that the list of segregation random variables, $[S_{i_t,1}, \ldots, S_{i_t,J}]$, associated with individual haplotype H_{i_t} forms a first-order Markov chain, with transition matrix $\boldsymbol{\xi}$:

$$\begin{aligned} p(S_{i_t,j+1} = r' | S_{i_t,j} = r) &= \xi_{rr'} \\ &= \big[\xi\big]^{\mathbb{I}(r = r')} \big[1 - \xi\big]^{\mathbb{I}(r \neq r')}, \end{aligned} \tag{10}$$

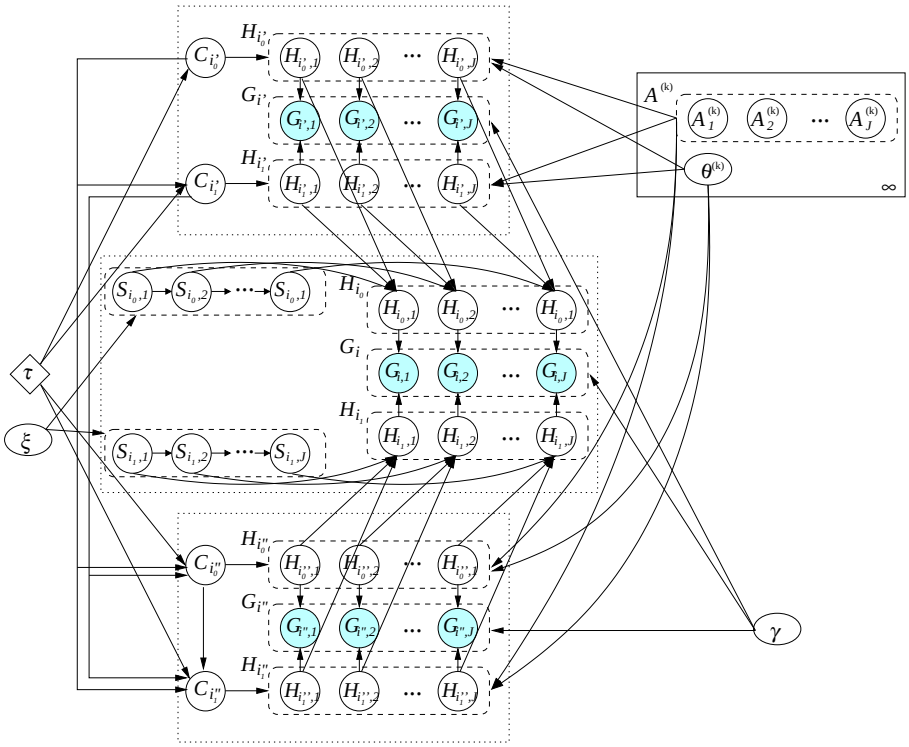

Figure 2: The graphical model representation of the Pedi-haplotyper model.

where $1 - \xi$ is the probability of a recombination event (i.e., a swap of parental origin) at position j. This model is equivalent to assuming that the recombination events follow a Poisson point process of rate ξ along the chromosome. If desired, a beta prior Beta(α_s, β_s) can be introduced for ξ. Again, the recombination rates in males and females can be different if desired. Considering the overall graphical topology of the Pedi-haplotyper model, as illustrated in Figure 2, for founding members in the pedigree (i.e., those without parental information), or half-founding members (i.e., those with information from only one of the two parents), we assume that their un-progenitored haplotype(s) are inherited from missing ancestors, thus following the basic haplotype model. For the haplotypes of the offspring in the pedigree, we couple them to their parents using the single generation mutation and recombination model described in the previous paragraphs. Thus, the Pedi-haplotyper model proposed in this section is fully generalizable to any pedigree structure. We note that this model has some commonalities with the probabilistic model for linkage analysis developed by Fishelson and Geiger [7].

3 Markov chain Monte Carlo for Haplotype Inference

In this section, we describe a Gibbs sampling algorithm for exploring the posterior distribution under our DP-haplotyper model, including the latent ancestral pool. We also present a Metropolis-Hastings variant of this algorithm that appears to mix better in practice.

3.1 A Gibbs Sampling Algorithm

Recall that the Gibbs sampler draws samples of each random variable from a conditional distribution of that variable given (previously sampled) values of all the remaining variables. The variables needed in our algorithm are: C_{i_t}, the index of the ancestral template of a haplotype instance t of individual i; $A_j^{(k)}$, the

allele pattern at the jth locus of the kth ancestral template; $H_{i_t,j}$, the tth allele of the SNP at the jth locus of individual i; and $G_{i,j}$, the genotype at locus j of individual i (the only observed variables in the model). All other variables in the model—θ and γ—are integrated out. The Gibbs sampler thus samples the values of C_{i_t}, $A_j^{(k)}$ and $H_{i_t,j}$. Conceptually, the Gibbs sampler alternates between two coupled stages. First, given the current values of the hidden haplotypes, it samples the c_{i_t} and subsequently $a_j^{(k)}$, which are associated with the Dirichlet process prior. Second, given the current state of the ancestral pool and the ancestral template assignment for each individual, it samples the $h_{i_t,j}$ variables in the basic haplotype model. In the first stage, the conditional distribution of c_{i_t} is:

$$\begin{aligned} & p(c_{i_t} = k \mid \mathbf{c}_{[-i_t]}, \mathbf{h}, \mathbf{a}) \\ \propto\ & p(c_{i_t} = k \mid \mathbf{c}_{[-i_t]}) \int p(h_{i_t} \mid c_{i_t} = k, \theta_k, a^{(k)}) p(\theta^{(k)} \mid \{h_{i'_{t'}} \ :\ i'_{t'} \neq i_t, c_{i'_{t'}} = k\}, a^{(k)}) d\theta^{(k)} \\ =\ & p(c_{i_t} = k \mid \mathbf{c}_{[-i_t]}) p(h_{i_t} \mid a^{(k)}, \mathbf{c}, \mathbf{h}_{[-i_t]}) \\ =\ & \begin{cases} \frac{n_{[-i_t],k}}{n-1+\tau} p(h_{i_t} \mid a^{(k)}, \mathbf{m}_{[-i_t],k}) & \text{if } k = c_{i'_{t'}} \text{ for some } i'_{t'} \neq i_t \\ \frac{\tau}{n-1+\tau} \sum_{a'} p(h_{i_t} \mid a') p(a') & \text{if } k \neq c_{i'_{t'}} \text{ for all } i'_{t'} \neq i_t \end{cases} \end{aligned} \tag{11}$$

where $[-i_t]$ denotes the set of indices excluding i_t; $n_{[-i_t],k}$ represents the number of $c_{i'_{t'}}$ for $i'_{t'} \neq i_t$ that are equal to k; n represents the total number of instances sampled so far; and $\mathbf{m}_{[-i_t],k}$ denotes the sufficient statistics m associated with all haplotype instances originating from ancestor k, except h_{i_t}. This expression is simply Bayes theorem with $p(h_{i_t}|a^{(k)}, \mathbf{c}, \mathbf{h}_{[-i_t],})$ playing the role of the likelihood and $p(c_{i_t} = k \mid \mathbf{c}_{[-i_t]})$ playing the role of the prior. The likelihood $p(h_{i_t}|a^{(k)}, \mathbf{m}_{[-i_t],k})$ is obtained by integrating over the parameter $\theta^{(k)}$, as in Eq. (7), up to a normalization constant:

$$p(h_{i_t}|a^{(k)}, \mathbf{m}_{[-i_t],k}) \quad \propto \quad R(\alpha_h, \beta_h) \frac{\Gamma(\alpha_h + m_{i_t,k})\Gamma(\beta_h + m'_{i_t,k})}{\Gamma(\alpha_h + \beta_h + m_{i_t,k} + m'_{i_t,k})} \Big(\frac{1}{|B|-1}\Big)^{m'_{i_t,k}}, \tag{12}$$

where $m_{i_t,k} = m_{[-i_t],k} + \sum_j \mathbb{I}(h_{i_t,j} = a_j^{(k)})$ and $m'_{i_t,k} = m'_{[-i_t],k} + \sum_j \mathbb{I}(h_{i_t,j} \neq a_j^{(k)})$, both functions of h_{i_t} (note that $m_{i_t,k} + m'_{i_t,k} = nJ$) [3]. It is easy to see that the normalization constant is the marginal likelihood $p(\mathbf{m}_{[-i_t],k} \mid a^{(k)})$, which leads to:

$$p(h_{i_t}|a^{(k)}, \mathbf{m}_{[-i_t],k}) = \frac{\Gamma(\alpha_h + m_{i_t,k})\Gamma(\beta_h + m'_{i_t,k})}{\Gamma(\alpha_h + m_{[-i_t],k})\Gamma(\beta_h + m'_{[-i_t],k})} \frac{\Gamma(\alpha_h + \beta_h + (n_k - 1)J)}{\Gamma(\alpha_h + \beta_h + n_k J)} \Big(\frac{1}{|B|-1}\Big)^{J}. \tag{13}$$

For $p(h_{i_t}|a)$, the computation is similar, except that the sufficient statistics $\mathbf{m}_{[-i_t],k}$ are now null (i.e., no previous matches with a newly instantiated ancestor):

$$p(h_{i_t}|a) \quad = \quad R(\alpha_h, \beta_h) \frac{\Gamma(\alpha_h + m_{i_t})\Gamma(\beta_h + m'_{i_t})}{\Gamma(\alpha_h + \beta_h + J)} \Big(\frac{1}{|B|-1}\Big)^{m'_{i_t}}, \tag{14}$$

where $m_{i_t} = \sum_j \mathbb{I}(h_{j,i_t} = a_j)$ and $m'_{i_t} = J - m_{i_t,k}$ are the relevant sufficient statistics associated only with haplotype instance h_{i_t}. The conditional probability for a newly proposed equivalence class k that is not populated by any previous samples requires a summation over all possible ancestors: $p(h_{i_t}) = \sum_{a'} p(h_{i_t}|a')p(a')$. Since the gamma function does not factorize over loci, computing this summation takes time that is exponential in the number of loci. To skirt this problem we endow each locus with its own mutation parameter $\theta_j^{(k)}$, with all parameters admitting the same prior Beta(α_h, β_h) [4]. This gives rise to a closed-form formula

[3] Recall that in Section 2.2 we use the symbol m_k to denote the count of matching SNP alleles in those individual haplotypes associated with ancestor $a^{(k)}$ (and m'_k for those inconsistent with the ancestor $a^{(k)}$). Here, we use a variant of these symbols to denote the pair of random counts (as indicated by the additional subscript i_t) resulting from the original m_k (or m'_k) for individual haplotypes known to associate with $a^{(k)}$ plus a randomly assigned haplotype h_{i_t} (whose actual associated ancestor is unknown).

[4] Note that now we also need to split counts $m_{[-i_t],k}$, $m_{i_t,k}$ and m_{i_t} into site-specific counts, $m_{[-i_t],k,j}$, $m_{i_t,k,j}$ and $m_{i_t,j}$, respectively, where j denotes a single SNP site.

for the summation and also for the normalization constant in Eq. (11). It is also, arguably, a more accurate reflection of reality. Specifically,

$$\begin{aligned} p(h_{i_t}|a) &= \prod_j R(\alpha_h,\beta_h)\frac{\Gamma(\alpha_h+m_{i_t,j})\Gamma(\beta_h+m'_{i_t,j})}{\Gamma(\alpha_h+\beta_h+1)}\left(\frac{1}{|B|-1}\right)^{m'_{i_t,j}} \\ &= \prod_j \left(\frac{\alpha_h}{\alpha_h+\beta_h}\right)^{\mathbb{I}(h_{i_t,j}=a_j)}\left(\frac{\beta_h}{(|B|-1)(\alpha_h+\beta_h)}\right)^{\mathbb{I}(h_{i_t,j}\neq a_j)}. \end{aligned} \tag{15}$$

Assuming that loci are also independent in the base measure $p(a)$ of the ancestors and that the base measure is uniform, we have:

$$\begin{aligned} \sum_a p(h_{i_t}|a)p(a) &= \prod_j \Big(\sum_{l\in B} p(a_j=l)p(h_{i_t,j}|a_j=l)\Big) \\ &= \prod_j \left(\sum_{l\in B}\frac{1}{|B|}\left(\frac{\alpha_h}{\alpha_h+\beta_h}\right)^{\mathbb{I}(h_{i_t,j}=l)}\left(\frac{\beta_h}{(|B|-1)(\alpha_h+\beta_h)}\right)^{\mathbb{I}(h_{i_t,j}\neq l)}\right) \\ &= \left(\frac{1}{|B|}\right)^J. \end{aligned} \tag{16}$$

In this case (that each locus has its own mutation parameter), the conditional likelihood computed in Eq. (13) is:

$$\begin{aligned} & p(h_{i_t,j}|a_j^{(k)},\mathbf{m}_{[-i_t],k,j}) \\ = & \prod_j \left(\frac{\alpha_h+m_{[-i_t],k,j}}{\alpha_h+\beta_h+n_k-1}\right)^{\mathbb{I}(h_{i_t,j}=a_j^{(k)})}\left(\frac{\beta_h+m'_{[-i_t],k,j}}{(|B|-1)(\alpha_h+\beta_h+n_k-1)}\right)^{\mathbb{I}(h_{i_t,j}\neq a_j^{(k)})}. \end{aligned} \tag{17}$$

Note that during the sampling of c_{i_t}, the numerical values of c_{i_t} are arbitrary, as long as they index distinct equivalence classes.

Now we need to sample the ancestor template $a^{(k)}$, where k is the newly sampled ancestor index for c_{i_t}. When k is not equal to any other existing index $c_{i'_{t'}}$, a value for a_k needs to be chosen from $p(a|h_{i_t})$, the posterior distribution of A based on the prior $p(a)$ and the single dependent haplotype h_{i_t}. On the other hand, if k is an equivalence class populated by previous samples of $c_{i'_{t'}}$, we draw a new value of $a^{(k)}$ from $p(a|\{h_{i_t},\ :\ c_{i_t}=k\})$. If, after a new sample of c_{i_t}, a template is no longer associated with any haplotype instance, we remove this template from the pool. The conditional distribution for this Gibbs step is therefore:

$$\begin{aligned} p(a^{(k)}|\mathbf{a}^{(-k)},\mathbf{h},\mathbf{c}) &= p(a^{(k)}|\{h_{i_t},\ :\ c_{i_t}=k\}) \\ &= \frac{p(\{h_{i_t},\ :\ c_{i_t}=k\}|a^{(k)})}{\sum_a p(\{h_{i_t},\ :\ c_{i_t}=k\}|a^{(k)}=a)} \\ &= \prod_j \frac{p(m_{k,j}|a_j^{(k)})}{\sum_{l\in B} p(m_{k,j}|a_j^{(k)}=l)}. \end{aligned} \tag{18}$$

We can sample $a_1^{(k)}, a_2^{(k)}, \ldots,$ sequentially:

$$p(a_j^{(k)} \mid \{h_{i_t,j} \ : \ c_{i_t} = k\}) =$$

$$\begin{cases} \frac{1}{Z} p(h_{i_t,j} | a_j^{(k)}) & \\ = \left(\frac{\alpha_h}{\alpha_h+\beta_h}\right)^{\mathbb{I}(h_{i_t,j}=a_j^{(k)})} \left(\frac{\beta_h}{(|B|-1)(\alpha_h+\beta_h)}\right)^{\mathbb{I}(h_{i_t,j}\neq a_j^{(k)})} & \text{if } k \text{ is not previously instantiated} \\ \frac{1}{Z} p(\{h_{i_t,j} \ : \ c_{i_t} = k\} | a_j^{(k)}) & \\ = \frac{1}{Z} \frac{\Gamma(\alpha_h+m_{k,j})\Gamma(\beta_h+m'_{k,j})}{\Gamma(\alpha_h+\beta_h+n_k)\cdot(|B|-1)^{m'_{k,j}}} & \\ = \frac{\Gamma(\alpha_h+m_{k,j})\Gamma(\beta_h+m'_{k,j})/(|B|-1)^{m'_{k,j}}}{\sum_{l\in B}\Gamma(\alpha_h+m_{k,j}(l))\Gamma(\beta_h+m'_{k,j}(l))/(|B|-1)^{m'_{k,j}(l)}} & \text{if } k \text{ is previously instantiated,} \end{cases} \tag{19}$$

where $m_{k,j}$ (respectively, $m'_{k,j}$) is the number of allelic instances originating from ancestor k at locus j that are identical to (respectively, different from) the ancestor, when the ancestor has the pattern $a_j^{(k)}$; and $m_{k,j}(l)$ (respectively, $m'_{k,j}(l)$) is the value of $m_{k,j}$ (respectively, $m'_{k,j}$) when $a_j^{(k)} = l$. [5]

We now proceed to the second sampling stage, in which we sample the haplotypes h_{i_t}. We sample each $h_{i_t,j}$, for all j, i and t, sequentially according to the following conditional distribution:

$$\begin{aligned} & p(h_{i_t,j} | \mathbf{h}_{[-(i,j)]}, h_{i_{\bar{t}},j}, \mathbf{c}, \mathbf{a}, \mathbf{g}) \\ \propto \ & p(g_i | h_{i_t,j}, h_{i_{\bar{t}},j}, \mathbf{u}_{[-(i,j)]}) p(h_{i_t,j} | a_j^{(k)}, \mathbf{m}_{[-(i_t,j)],k}) \\ = \ & R_g \frac{\Gamma(\alpha_g+u)\Gamma(\beta_g+(u'+u''))}{\Gamma(\alpha_g+\beta_g+IJ)} [\mu_1]^{u'} [\mu_2]^{u''} \times \\ & R_h \frac{\Gamma(\alpha_h+m_{i_t,k,j})\Gamma(\beta_h+m'_{i_t,k,j})}{\Gamma(\alpha_h+\beta_h+n_k)\cdot(|B|-1)^{m'_{i_t,k,j}}}, \end{aligned} \tag{20}$$

where $[-(i_t,j)]$ denotes the set of indices excluding (i_t,j) and $m_{i_t,k,j} = m_{[-(i_t,j)],k,j} + \mathbb{I}(h_{i_t,j} = a_j^{(k)})$ (and similarly for the other sufficient statistics). Note that during each sampling step, we do not have to recompute the $\Gamma(\cdot)$, because the sufficient statistics are either not going to change (e.g., when the newly sampled $h_{i_t,j}$ is the same as the old sample), or only going to change by one (e.g., when the newly sampled $h_{i_t,j}$ results in a change of the allele). In such cases the new gamma function can be easily updated from the old one.

3.2 A Metropolis-Hasting Sampling Algorithm

Note that for a long list of loci, a prior $p(a)$ that is uniform over all possible ancestral template patterns will render the probability of sampling a new ancestor infinitesimal, due to the small value of the smoothed marginal likelihood of any haplotype pattern h_{i_t}, as computed from Eq. (11). This could result in slow mixing. An alternative sampling strategy is to use a partial Gibbs sampling strategy with the following Metropolis-Hasting updates, which could allow more complex $p(a)$ (e.g., non-factorizable and non-uniform) to be readily handled. To sample the equivalence class of h_{i_t} from the target distribution $\pi(c_{i_t}) = p(c_{i_t} | \mathbf{c}_{[-i_t]}, \mathbf{h}, \mathbf{a})$ described in Eq. 11, consider the following proposal distribution:

$$q(c^*_{i_t} = k | c_{[-i_t]}) = \begin{cases} \frac{n_{[-i_t],k}}{n-1+\tau} \ : \ \text{if} \ \ k = c_{i'_{t'}} \ \text{for some} \ i'_{t'} \neq i_t \\ \frac{\tau}{n-1+\tau} \ : \ \text{if} \ \ k \neq c_{i'_{t'}} \ \text{for all} \ i'_{t'} \neq i_t \end{cases} \tag{21}$$

[5] Note that here the counts m_k (and m'_k) vary with different possible configurations of the ancestor $a^{(k)}$ given $\mathbf{h}$, unlike previously in Eqs. (12)-(17), in which they vary with different possible configurations of h_{i_t} given $a^{(k)}$.

Then we sample $a^{(c^*_{i_t})}$ from the prior $p(a)$. For the target distribution $p(c_{i_t} = k|\mathbf{c}_{[-i_t]}, \mathbf{h}, \mathbf{a})$, the proposal factor cancels when computing the acceptance probability ξ, leaving[6]:

$$\xi(c^*_{i_t}, c_{i_t}) = \min\Big[1, \frac{p(h_{i_t}|a^{c^*_{i_t}}, \mathbf{c}, \mathbf{h}_{[-i_t]})}{p(h_{i_t}|a^{c_{i_t}}, \mathbf{c}, \mathbf{h}_{[-i_t]})}\Big]. \tag{23}$$

In practice, we found that the above modification to the Gibbs sampling algorithm leads to substantial improvement in efficiency for long haplotype lists (even with a uniform base measure for A), whereas for short lists, the Gibbs sampler remains better due to the high (100%) acceptance rate.

3.3 A Sketch of MCMC Strategies for the Pedi-Haplotyper Model

The MCMC sampling strategy for the Pedi-haplotype model is similar to that of the basic DP-haplotyper described above, except that we need to sample a few more variables on top of the DP-haplotyper model, which requires collecting a few more sufficient statistics for updating the predictive distributions of these variables. In addition to the sufficient statistics $\mathbf{m}$ (for the consistency between the ancestral and individual haplotypes (i.e., the number of cases of which the ancestral and individual haplotypes agree in a single sweep during sampling), and $\mathbf{u}$ (for the consistency between the individual haplotypes and genotype (i.e., the number of cases of which the genotype and its corresponding haplotype pair agree in a single sweep during sampling), needed in the DP-haplotyper model, we need to update the following sufficient statistics during each sampling step that sweeps all the random variables:

- $\mathbf{w}$: the sufficient statistics of the transition probability ζ,

$$w_{rr'} = \sum_t \sum_i \sum_j \mathbb{I}(s_{i_t,j} = r)\mathbb{I}(s_{i_t,j+1} = r').$$

  If we prefer to model the recombination rates in males and females differently, then we compute $\mathbf{w}_t$ separately for $t = 0$ and $t = 1$.

- $\mathbf{v}$: the sufficient statistics of the single generation inheritance (i.e., non-mutation) rate ϵ,

$$v = \sum_t \sum_r \sum_i \sum_j \mathbb{I}(h_{i_t,j} = h_{\pi_r(i_t),j})\mathbb{I}(s_{i_t,j} = r).$$

The ancestral template indicators associated with the founding subjects and the ancestor pool can be sampled as in the basic DP-haplotyper model. Now we derive the additional predictive distributions needed for collapsed Gibbs sampling for the Pedi-haplotyper model. For each predictive distribution of the hidden variables, we integrate out the model parameters given their (conjugate) priors.

[6]The cancellation of the proposal in ξ can be seen from the following derivation:

$$\begin{aligned}
\frac{q(c_{i_t}|\mathbf{c}_{[-i_t]})}{q(c^*_{i_t}|\mathbf{c}_{[-i_t]})}\frac{\pi(c^*_{i_t})}{\pi(c_{i_t})} &= \frac{q(c_{i_t}|\mathbf{c}_{[-i_t]})}{q(c^*_{i_t}|\mathbf{c}_{[-i_t]})}\frac{p(c^*_{i_t}|\mathbf{c}_{[-i_t]}, \mathbf{h}, \mathbf{a})}{p(c_{i_t}|\mathbf{c}_{[-i_t]}, \mathbf{h}, \mathbf{a})} \\
&= \frac{q(c_{i_t}|\mathbf{c}_{[-i_t]})}{q(c^*_{i_t}|\mathbf{c}_{[-i_t]})}\frac{p(c^*_{i_t}|\mathbf{c}_{[-i_t]})p(h_{i_t}|a^{(c^*_{i_t})}, \mathbf{c}, \mathbf{h}_{[-i_t]})}{p(c_{i_t}|\mathbf{c}_{[-i_t]})p(h_{i_t}|a^{(c_{i_t})}, \mathbf{c}, \mathbf{h}_{[-i_t]})} \\
&= \frac{p(h_{i_t}|a^{(c^*_{i_t})}, \mathbf{c}, \mathbf{h}_{[-i_t]})}{p(h_{i_t}|a^{(c_{i_t})}, \mathbf{c}, \mathbf{h}_{[-i_t]})},
\end{aligned} \tag{22}$$

- Sample a founding haplotype:

$$
\begin{aligned}
& p(h_{i_t,j}|\mathbf{h}_{[-(i,j)]}, h_{i_{\bar{t}},j}, \mathbf{s}, \mathbf{c}, \mathbf{a}, \mathbf{g}) \\
= \;& p(h_{i_t,j}|h_{i_{\bar{t}},j}, h_{\lambda(i),j}, s_{\lambda(i),j}, a_{c_{i_t},j}, g_i, \mathbf{v}_{[-(i,j)]}, \mathbf{u}_{[-(i,j)]}, \mathbf{m}_{[-(i,j)]}) \\
\propto \;& p(h_{i_t,j}, h_{\lambda(i),j}, g_i|h_{i_{\bar{t}},j}, s_{\lambda(i),j}, a_{c_{i_t},j}, \mathbf{v}_{[-(i,j)]}, \mathbf{u}_{[-(i,j)]}, \mathbf{m}_{[-(i,j)]}) \\
= \;& p(h_{\lambda(i),j}|h_{i_t,j}, h_{i_{\bar{t}},j}, s_{\lambda(i),j}, \mathbf{v}_{[-(i,j)]}) p(g_i|h_{i_t,j}, h_{i_{\bar{t}},j}, \mathbf{u}_{[-(i,j)]}) p(h_{i_t,j}|a_{c_{i_t},j}, \mathbf{m}_{[-(i,j)]}) \\
= \;& R_m \frac{\Gamma(\alpha_m + v(h_{i_t,j}))\Gamma(\beta_m + v'(h_{i_t,j}))}{\Gamma(\alpha_m + \beta_m + v(h_{i_t,j}) + v'(h_{i_t,j}))} \times \\
& R_g \frac{\Gamma(\alpha_g + u(h_{i_t,j}))\Gamma(\beta_g + u'(h_{i_t,j}) + u''(h_{i_t,j}))}{\Gamma(\alpha_m + \beta_m + IJ)} \mu_1^{u'} \mu_2^{u''} \times \\
& R_h \frac{\Gamma(\alpha_h + m(h_{i_t,j}))\Gamma(\beta_h + m'(h_{i_t,j}))}{\Gamma(\alpha_h + \beta_h + m(h_{i_t,j}) + m'(h_{i_t,j})) \cdot (|B|-1)^{m'(h_{i_t,j})}},
\end{aligned} \tag{24}
$$

  where $h_{\lambda(i),j}$ refers to the allele in the child of i that is inherited from i. For simplicity, we suppose only one child. For the case of multiple children, the first term of Eq. (24) becomes a product of such terms, each corresponding to one child.

- To sample a non-founding haplotype:

$$
\begin{aligned}
& p(h_{i_t,j}|\mathbf{h}_{[-(i,j)]}, h_{i_{\bar{t}},j}, \mathbf{s}, \mathbf{c}, \mathbf{a}, \mathbf{g}) \\
= \;& p(h_{i_t,j}|\mathbf{h}_{[-(i,j)]}, h_{i_{\bar{t}},j}, h_{\lambda(i),j}, h_{\pi(i_t)_0,j}, h_{\pi_t(i_t),j}, s_{i_t,j}, s_{\lambda(i),j}, g_i, \mathbf{v}_{[-(i,j)]}, \mathbf{u}_{[-(i,j)]}) \\
\propto \;& p(h_{i_t,j}, h_{\lambda(i),j}, g_i|\mathbf{h}_{[-(i,j)]}, h_{i_{\bar{t}},j}, h_{\pi(i_t)_0,j}, h_{\pi_t(i_t),j}, s_{i_t,j}, s_{\lambda(i),j}, \mathbf{v}_{[-(i,j)]}, \mathbf{u}_{[-(i,j)]}) \\
= \;& p(h_{i_t,j}|h_{\pi(i_t)_0,j}, h_{\pi(i_t)_1,j}, s_{i_t,j}, \mathbf{v}_{[-(i,j)]}) p(h_{\lambda(i),j}|h_{i_t,j}, h_{i_{\bar{t}},j}, s_{\lambda(i),j}, \mathbf{v}_{[-(i,j)]}) \\
& p(g_i|h_{i_t,j}, h_{i_{\bar{t}},j}, \mathbf{u}_{[-(i,j)]}) \\
= \;& R_m \frac{\Gamma(\alpha_m + v(h_{i_t,j}))\Gamma(\beta_m + v'(h_{i_t,j}))}{\Gamma(\alpha_m + \beta_m + v(h_{i_t,j}) + v'(h_{i_t,j}))} \times \\
& R_g \frac{\Gamma(\alpha_g + u(h_{i_t,j}))\Gamma(\beta_g + u'(h_{i_t,j}) + u''(h_{i_t,j}))}{\Gamma(\alpha_m + \beta_m + IJ)} \mu_1^{u'} \mu_2^{u''}.
\end{aligned} \tag{25}
$$

- Sample the segregation variable:

$$
\begin{aligned}
& p(s_{i_t,j}|\mathbf{h}, \mathbf{s}_{[-(i,j)]}, s_{i_{\bar{t}},j}, \mathbf{c}, \mathbf{a}, \mathbf{g}) \\
= \;& p(s_{i_t,j}|h_{i_t,j}, h_{\pi_0(i_t),j}, h_{\pi_1(i_t),j}, s_{i_t,j-1}, s_{i_t,j+1}, \mathbf{v}_{[-(i,j)]}, \mathbf{w}_{[-(i_t,j)]}) \\
\propto \;& p(h_{i_t,j}|h_{\pi_0(i_t),j}, h_{\pi_1(i_t),j}, s_{i_t,j}, \mathbf{v}_{[-(i,j)]}) p(s_{i_t,j-1}|s_{i_t,j}, \mathbf{w}_{[-(i_t,j)]}) \\
= \;& p(s_{i_t,j}|s_{i_t,j+1}, \mathbf{w}_{[-(i_t,j)]}) \\
= \;& R_m \frac{\Gamma(\alpha_m + v(s_{i_t,j}))\Gamma(\beta_m + v'(s_{i_t,j}))}{\Gamma(\alpha_m + \beta_m + v(s_{i_t,j}) + v'(h_{i_t,j}))} \times \\
& R_s \frac{\Gamma(\alpha_s + w_{00}(s_{i_t,j}) + w_{11}(s_{i_t,j}))\Gamma(\beta_s + + w_{01}(s_{i_t,j}) + w_{10}(s_{i_t,j}))}{\Gamma(\alpha_s + \beta_s + |\mathbf{w}|)},
\end{aligned} \tag{26}
$$

  where $|\mathbf{w}| = \sum_{r,r'} w_{r,r'}$.

4 Experimental Results

We validated our algorithm by applying it to simulated and real data and compared its performance to that of the state-of-the-art PHASE algorithm [19] and other current algorithms. We report on the results of both

variants of our algorithm: The Gibbs sampler, denoted DP(Gibbs), and the Metropolis-Hasting sampler, denoted DP(MH). Throughout the experiments, we set the hyperparameter τ in the Dirichlet process to be roughly 1% of the population size; i.e., for a data set of 100 individuals, $\tau = 1$. We used a burn-in of 2000 iterations (or 4000 for datasets with more than 50 individuals), and used the next 6000 iterations for estimation.

4.1 Simulated data

In our first set of experiments we applied our method to simulated data ("short sequence data") from [19]. This data contains sets of $2n$ haplotypes, randomly paired to form n genotypes, under an infinite-sites model with parameters $\eta = 4$ and $R = 4$ determining the mutation and recombination rates, respectively (see [19] for additional details). We used the first 40 datasets for each combination of individuals and sites, where the number of individuals ranged between 10 and 50, and the number of sites ranged between 5 and 30.

	DP(MH)			PHASE			EM
#individuals	err_s	err_i	d_s	err_s	err_i	d_s	err_i
10	0.060	0.216	0.051	0.046	0.182	0.054	0.424
20	0.039	0.152	0.039	0.029	0.136	0.046	0.296
30	0.036	0.121	0.038	0.024	0.101	0.027	0.231
40	0.030	0.094	0.029	0.019	0.071	0.026	0.195
50	0.028	0.082	0.024	0.019	0.072	0.025	0.167
Average	**0.039**	**0.133**	**0.036**	**0.027**	**0.112**	**0.036**	**0.263**

Table 1: Performance on data from [19]. The results for the EM algorithm are adapted from [19].

To evaluate the performance of the algorithms we used the following error measures: err_s, the ratio of incorrectly phased SNP sites over all non-trivial heterozygous SNPs (excluding individuals with a single heterozygous SNP); err_i, the the ratio of incorrectly phased individuals over all non-trivial heterogeneous individuals; and d_s, the *switch distance*, which is the number of phase flips required to correct the predicted haplotypes over all non-trivial heterogeneous SNPs. The results are summarized in Table 1. Overall, we perform slightly worse than PHASE on the first two measures, and similar to PHASE on the switch distance measure (which uses 100,000 sampling steps). Both algorithms provide a substantial improvement over EM.

		DP(Gibbs)			DP(MH)			PHASE			HAP	HAPLOTYPER
block id.	length	err_s	err_i	d_s	err_s	err_i	d_s	err_s	err_i	d_s	err_s	err_s
1	14	0.223	0.485	0.229	0	0	0	0.003	0.030	0.003	0.007	0.039
2	5	0	0	0	0.007	0.026	0.007	0.007	0.026	0.007	0.036	0.065
3	5	0	0	0	0	0	0	0	0	0	0	0.008
4	11	0.143	0.262	0.128	0	0	0	0	0	0	0.015	-
5	9	0.020	0.066	0.020	0.011	0.033	0.011	0.011	0.033	0.011	0.027	0.151
6	27	0.071	0.191	0.074	0.005	0.043	0.005	0	0	0	0.018	0.041
7	7	0.005	0.018	0.005	0.005	0.018	0.005	0.005	0.018	0.005	0.068	0.214
8	4	0	0	0	0	0	0	0	0	0	0	0.252
9	5	0.029	0.097	0.029	0.012	0.032	0.012	0.012	0.032	0.012	0.057	0.152
10	4	0.007	0.025	0.007	0.007	0.025	0.007	0.008	0.025	0.008	0.042	0.056
11	7	0.010	0.034	0.005	0.005	0.017	0.005	0.011	0.034	0.011	0.033	0.093
12	5	0.010	0.037	0.020	0	0	0	0	0	0	0	0.077
Average	**8.58**	**0.043**	**0.101**	**0.043**	**0.004**	**0.016**	**0.004**	**0.005**	**0.017**	**0.005**	**0.025**	**0.104**

Table 2: Performance on the data of [2], using the block structure provided by [11]. The results of HAP and HAPLOTYPER are adapted from [11]. Since the error rate in [11] uses the number of both heterozygous and missing sites as the denominator, whereas we used only the non-trivial heterozygous ones, we rescaled the error rates of the two latter methods to be comparable to ours.

4.2 Real data

We applied our algorithm to two real datasets and compared its performance to that of PHASE [19] and other algorithms.

		DP(MH)			PHASE		
region	length	err_s	err_i	d_s	err_s	err_i	d_s
16a	13	0.185	0.480	0.141	0.174	0.440	0.130
1b	16	0.100	0.250	0.160	0.200	0.450	0.180
25a	14	0.135	0.353	0.115	0.212	0.588	0.212
7b	13	0.105	0.278	0.066	0.145	0.444	0.092
Average	**14**	**0.131**	**0.340**	**0.121**	**0.183**	**0.481**	**0.154**

Table 3: Performance on the data of [8].

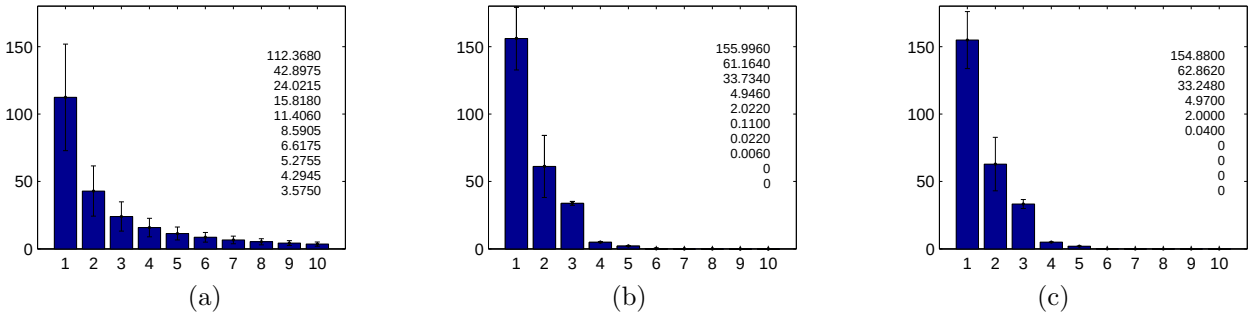

Figure 3: The top ten ancestral templates during Metropolis-Hasting sampling for block 1 of the data of [2]. (The numbers in the panels are the posterior means of the frequencies of each template). (a) Immediately after burn-in (first 2000 samples). (b) 3000 samples after burn-in. (c) 6000 samples after burn-in.

The first dataset contains the genotypes of 129 individuals over 103 polymorphic sites [2]. In addition it contains the genotypes of the parents of each individual, which allows the inference of a large portion of the haplotypes as in [4]. The results are summarized in Table 2. It is apparent that the Metropolis-Hasting sampling algorithm significantly outperforms the Gibbs sampler, and is to be preferred given the relatively limited number of sampling steps (~ 6000). The overall performance is comparable to that of PHASE and better than both HAP [11, 4] and HAPLOTYPER [15].

It is important to emphasize that our methods also provide a posteriori estimates of the ancestral pool of haplotype templates and their frequencies. We omit a listing of these haplotypes, but provide an illustrative summary of the evolution of these estimates during sampling (Figure 3).

The second dataset contains genotype data from four populations, 90 individuals each, across several genomic regions [8]. We focused on the Yoruban population (D), which contains 30 trios of genotypes (allowing us to infer most of the true haplotypes) and analyzed the genotypes of 28 individuals over four medium-sized regions (see below). The results are summarized in Table 3. All methods yield higher error rates on these data, compared to the analysis of the data of [2], presumably due to the low sample size. In this setting, over all but one of the four regions, our algorithm outperformed PHASE on all three types of error measures. A preliminary analysis suggests that our performance gain may be due to the bias toward parsimony induced by the Dirichlet process prior. We found that the number of template haplotypes in our algorithm is typically small, whereas in PHASE the haplotype pool can be very large (e.g., region 7b has 83 haplotypes, compared to 10 templates in our case and 28 individuals overall).

In terms of computational efficiency, we noticed that PHASE typically required 20,000 to 100,000 steps until convergence, while our DP-based method required around 2,000 to 6,000 steps to convergence (Fig. 4a). The posterior distribution of K, the number of ancestor haplotypes underlying the population, is sharply peaked at a single mode (Fig. 4b).

5 Conclusions

We have proposed a Bayesian approach to the modeling of genotypes based on a Dirichlet process prior. We have shown that the Dirichlet process provides a natural representation of uncertainty regarding the size and composition of the pool of haplotypes underlying a population. We have developed several Markov

(a) (b)

Figure 4: (a) Sampling trace of the number of population haplotypes derived from the genotypes. As can be seen, the Markov chain starts from a rather non-parsimonious estimation, and converges to a parsimonious solution after about two thousand samples. (b) The histogram representation of the posterior distribution of the number of ancestors obtained via Gibbs sampling.

chain Monte Carlo algorithms for haplotype inference under either a basic DP mixture haplotype model intended for an *iid* population, or, an extended graphical DP mixture model—the Pedi-haplotyper model—for a population containing both *iid* subjects and subjects coupled by partial pedigrees. The experiments on the basic DP mixture haplotype model show that this model leads to effective inference procedures for inferring the ancestral pool and for haplotype phasing based on a set of genotypes. The model accommodates growing data collections and noisy and/or incomplete observations. The approach also naturally imposes an implicit bias toward small ancestral pools, reminiscent of parsimony methods, doing so in a well-founded statistical framework that permits errors.

Our focus here has been on adapting the technology of the Dirichlet process to the setting of the standard haplotype phasing problem. But an important underlying motivation for our work, and a general motivation for pursuing probabilistic approaches to genomic inference problems, is the potential value of our model as a building block for more expressive models. In particular, as in [9] and [13], the graphical model formalism naturally accommodates various extensions, such as segmentation of chromosomes into haplotype blocks and the inclusion of pedigree relationships. In Section 2.3, we have outlined a preliminary extension of the basic Dirichlet process mixture model that incorporates pedigree relationships and briefly discussed how to model realistic biological processes that might influence haplotype formation and diversification, such as recombination and mutation during single generation inheritance. We recognize that many other important issues also deserve careful attention, for example, haplotype recombinations among the ancestral haplotype pools (so far, we assume that these ancestral haplotypes relate to modern individual haplotypes only via mutations), aspects of evolutionary dynamics (e.g., coalescence, selection, etc.), and linkage analysis under joint modeling of complex traits and haplotypes. We believe that the graphical model formalism we proposed can readily accommodate such extensions. In particular, it appears reasonable to employ an ancestral recombination hypothesis (rather than single generation recombination) to account for common individual haplotypes that are distant from any single ancestral haplotype template, but can be matched piecewise to multiple ancestral haplotypes. This may be an important aspect of chromosomal evolution and can provide valuable insight into the dynamics of populational genetics in addition to point-mutation-based coalescence theory, and can potentially improve the efficiency and quality of haplotype inference. The Dirichlet process parameterization also provides a natural upgrade path for the consideration of richer models; in particular, it is possible to incorporate more elaborate base measures G_0 into the Dirichlet process framework—the coalescence-based distribution of [19] would be an interesting choice. From an implementation point of view, our model, as many other basic haplotype inference programs, can be straightforwardly wrapped into a simple *Partition-Ligation* scheme (or more sophisticated HMM-based model) as in [15, 19], to phase long sequences of SNP genotype data.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. 0523757, and and by the Pennsylvania Department of Health's Health Research Program under Grant No. 2001NF-Cancer Health Research Grant ME-01-739. EPX. is also supported by a NSF CAREER Award under Grant No. DBI-0546594. MIJ was supported by grant R33 HG003070 from the National Institutes of Health and RS was supported by an Alon Fellowship.

References

[1] A. Clark et al. Haplotype structure and population genetic inferences from nucleotide-sequence variation in human lipoprotein lipase. *American Journal of Human Genetics*, 63:595–612, 1998.

[2] M. J. Daly et al. High-resolution haplotype structure in the human genome. *Nature Genetics*, 29(2):229–232, 2001.

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

[4] E. Eskin, E. Halperin, and R.M. Karp. Efficient reconstruction of haplotype structure via perfect phylogeny. *Journal of Bioinformatics and Computational Biology*, 1:1–20, 2003.

[5] 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.

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

[7] M. Fishelson and D. Geiger. Exact genetic linkage computations for general pedigrees. *Bioinformatics*, 18 (Suppl. 1):S189–S198, 2002.

[8] S. B. Gabriel et al. The structure of haplotype blocks in the human genome. *Science*, 296:2225–2229, 2002.

[9] G. Greenspan and D. Geiger. Model-based inference of haplotype block variation. *Journal of computational biology*, 11:493–504, 2004.

[10] D. Gusfield. Haplotyping as perfect phylogeny: Conceptual framework and efficient solutions. In *Proc. RECOMB*, pages 166–175, 2002.

[11] E. Halperin and E. Eskin. Haplotype reconstruction from genotype data using imperfect phylogeny. Technical Report, Columbia University, 2002.

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

[13] S. L. Lauritzen and N. A. Sheehan. Graphical models for genetic analysis. TR R-02-2020, Aalborg University, 2002.

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

[15] 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.

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

[17] N. J. Risch. Searching for genetic determinants in the new millennium. *Nature*, 405(6788):847–56, 2000.

[18] R. Sachidanandam et al. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature*, 291:1298–2302, 2001.

[19] 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.

[20] E.P. Xing, R. Sharan, and M.I. Jordan. Bayesian haplotype inference via the Dirichlet process. In *Proc. ICML*, pages 879–886, 2004.