Thesis proposal

Learning Ancestral Genetic Processes using Nonparametric Bayesian Models

Kyung-Ah Sohn

Computer Science Department
Carnegie Mellon University
ksohn@cs.cmu.edu

August 2009

Thesis Committee:
Eric P. Xing, Chair
Zoubin Ghahramani
Russell Schwartz
Kathryn Roeder
Matthew Stephens, University of Chicago

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy.

Abstract

Recent explosion of genomic data have fueled the long-standing interest of analyzing genetic variations to reconstruct the evolutionary history and ancestral structures of human populations that can provide essential clues for various medical applications. Although genetic properties such as linkage disequilibrium (LD) and population structure are closely related under a common inheritance process involving many different ancestral processes in the genetic history, the statistical methodologies developed so far mostly deal with those structural inferences separately using specialized models that do not capture their statistical and genetic relationships. Also, most of these approaches ignore the inherent uncertainty in the genetic complexity of the data and rely on inflexible models resulting from restrictive assumptions. These limitations may make it difficult to infer detailed and consistent structural information from rich genotypic data.

The goal of this proposal is to develop new nonparametric Bayesian models for learning the ancestral genetic processes under a unified inheritance framework. Our preliminary results include efficient models based on Dirichlet process for haplotype inference on multi-population data, and for the joint analysis of population structure and recombination events. We plan to generalize these models further to solve related problems in ancestral inference such as to recover local ancestries along chromosomes, or to detect the signatures of selective sweeps on the chromosomes. Rigorous statistical analysis will be performed as well as extensive validation on real large scaled data. The models would serve as valuable tools for downstream analysis such as to find genetic basis of phenotypic traits.

Contents

1 Introduction

Recent advances in biological assaying have led to an explosion of genomic data. These data pose challenging inference problems whose solutions could shed light on the evolutionary history of human population and the genetic basis of various phenotypic traits such as disease propensities [7, 8]. While lots of statistical models have been developed so far to uncover the mechanisms and properties of genetic processes (e.g. [9, 1, 22, 40] for recombination block structure analysis, or [10, 21, 10, 18] for haplotype inference, as will be described later), these approaches often rely on parametric models with restrictive assumptions, and ignore inherent uncertainty in the genetic complexity of the data. Considering that lots of parameters are needed to model the complex genetic system, if the underlying assumptions fail, these inflexible models can lead to seriously unreasonable inference results. Moreover, those models mostly deal with the structural inferences of correlated genetic properties separately using specialized models that do not capture their relationships.

Nonparametric Bayesian approaches can serve as a more flexible framework to address these issues. The models under nonparametric Bayesian framework are data-oriented, so the adequate model complexity is determined directly from data. Therefore, it generally provides very flexible models, without the need for model selection on parameters, which are especially well-suited for this type of genomic data. Computational complexity might be an issue in using nonparametric Bayesian models for large scaled data as they typically require higher computational costs than parametric models or frequentist approaches. However, existing simple parametric models working relatively fast have not been enough for sufficient understanding of genetic properties of interest, and the need for estimates with higher accuracy from rather complicated models than fast solutions obtainable from simple models is ever increasing. Also successful approaches have been appearing which present efficient algorithms for the newly developed nonparametric models, for example, approximate variational inferences [4], or analogies of parametric inferences for nonparametric models [36].

In this report, we propose to develop new methods for learning important ancestral processes in population genetics from genotypic data under a unified haplotype inheritance assumption. For realistic and reasonable biological result, our model would approximate the well-known coalescence which can be made possible under a nonparametric Bayesian framework. The models we develop would provide valuable information for downstream analysis such as phenotype association study. In the following sections, we first review the basic statistical and biological backgrounds, and introduce the related problems and works as well. We then describe preliminary works completed so far on applications in ancestral inference using Dirichlet process based models. Finally, the future work will be explained which includes extensions of current models to infer more generalized genetic properties.

2 Survey

This section consists of two major parts. We first introduce the necessary biological backgrounds, so ancestral genetic processes to be exploited in this work will be described. In the second part, we describe a nonparametric Bayesian model called Dirichlet process (DP) and its extensions. This would constitute the basic methodological component of the models which will be developed under this proposal.

2.1 Ancestral processes in population genetics

Modern individual chromosomes are typically believed to be the result of different but highly correlated genetic processes, originated from a pool of ancestral individuals. Some of these genetic processes with higher interest and importance will be introduced along with the necessary background concepts of population genetics in the following. The related problems and previous approaches will be briefly reviewed as well.

2.1.1 SNPs, haplotypes and genotypes

One of the most important kinds of genetic variations among individuals is a single nucleotide polymorphism (SNP), which refers to the existence of two (or more) possible nucleotide bases from $\{A, C, G, T\}$ at a chromosomal locus in a population. SNPs form the largest class of individual differences in DNA and have long been targeted for many biological and medical applications such as disease association study.

Contiguous sequences of multiple SNPs on a chromosome are often looked at together and these are called *haplotypes*. The haplotypes have recently started to gain popularity as alternative basis for the association study and other applications because of the richer information they contain about genetic history and processes than that of just the set of independent single SNPs. Interestingly, diploids like humans have two copies of each chromosome, one maternal copy and one paternal copy. These two haplotypes form a *genotype* that represents unordered pairs of alleles from the haplotypes. That is, it does not carry information about which allele is from which chromosome copy – its *phase*. Common biological methods for assaying genotypes typically do not provide phase information for individuals with heterozygous genotypes at multiple loci. Although phase can be obtained at a considerably higher cost via molecular haplotyping [22], or sometimes from analysis of trios [14], it is desirable to develop automatic and robust computational methods for inferring haplotypes from the inexpensive genotype data.

A lot of effort has been devoted to the problem of haplotype inference for reconstructing the most feasible haplotypes from genotypes of a study population. The PHASE [18, 32] program is one of the most widely used softwares so far. It is based on *Product of Approximate Conditionals* (PAC) that approximates the marginal probabilities of the current haplotypes in a population by assuming each individual haplotype as the progeny of a randomly-chosen existing haplotype. Although this leads to relatively accurate estimate of haplotype phases and has set the state-of-the-art benchmark in haplotype inference, it is not fast enough to be applied to large scale data commonly available these days. Another software called fastPHASE [28] greatly improves the speed, but at the expense of loss of accuracy. Other approaches have been proposed to imporve the accuracy and the speed, e.g. [6, 19], but the problem still remains to be open.

2.1.2 Genetic diversities: mutation, recombination, migration, and natural selection

The genetic diversities contained in chromosomes of modern populations come from many different sources: mutations, recombination, population migration, and so on. Mutation, the changes to the nucleotide when genetic materials are inherited from one's parents, is generally believed to be the major mechanism the natural selection acts on so that advantageous heritable traits to an organism in survival and reproduction become more and more common in a population over generations.

Recombination is the genetic process by which a strand of genetic material is broken and then joined to a different strand during meiosis. It plays a key role in shaping the patterns of linkage disequilibrium (LD)—the non-random association of alleles at different loci—which is one of the most important structural forms contained in genome. When a recombination occurs between two loci, it tends to decouple the alleles

carried at those loci in its descendants and thus reduce LD; uneven occurrence of recombination events along chromosomal regions during genetic history can lead to "block structures" in chromosomes such that within each block only low level of diversities are present in a population. Several combinatorial and statistical approaches have been developed for uncovering optimum block boundaries from SNP haplotypes [9, 1, 22, 40], and these advances have important applications in genetic analysis of disease propensities and other complex traits. Also the problem of inferring chromosomal recombination rates and hotspots is essential for understanding the origin and characteristics of genome variations, where different approaches have been tried to solve the problems [33, 12].

Population migration is another important source of diversities and structures stored in genomic sequences. Due to migration and admixing, the individual chromosome is typically decomposed of segments from different ancestral populations. The related structures about how many ancestral populations have formed the current one or which segments of chromosomes are from which ancestral populations can be very useful in many applications such as to correct the confounding effect and to improve the power in association study. A number of variants of statistical *admixture* models for genetic polymorphisms have been proposed for the analysis of population structure [23, 27, 11]. While these models aim to provide ancestry information for each individual and each locus, there is no explicit representation of "ancestors" as a real chromosome haplotype and the inferred population structural map emphasizes revealing the contributions of *abstract* population-specific ancestral proportion profiles, which does not directly reflect individual diversity. Also the related genetic processes to form the current population structure are not reflected explicitly in these models. Finer-scaled analysis under a more flexible framework would be still desired.

2.1.3 Coalescence

Under common genetic arguments, the ancestral relationships among a sample of individuals can be described by a genealogical tree known as the *coalescent* [17]. It traces the sample sequences of a population backward in time until a single ancestral sequence is met, known as the most recent common ancestor (MRCA). Different assumptions can lead to different statistical properties in coalescent theory. The simplest case can start from just assuming mutation as a single genetic process; consider two distinct sample sequences who differ at a single nucleotide by mutation. By tracing the ancestry of these two individuals backwards there will be time when the MRCA is encountered and then the two lineages will have coalesced forming a tree. Extensions for more complex processes such as recombination, selection, and population migration have been widely studied and their mathematical properties have been investigated rigorously.

However, the marginalization over all possible coalescent trees given sample sequences is widely known to be intractable. Therefore, it is nontrivial to use the full coalescence in general ancestral inference. Approximation approach such as *Product of Approximate Conditionals* (PAC) [18] was proposed for the task of haplotype inference and recombination rate estimation, but more general and principled schemes are sill in need.

2.2 Nonparametric Bayesian models based on Dirichlet process

2.2.1 Dirichlet process and its mixture models

Dirichlet process is a non-parametric Bayesian model which defines a distribution over distributions. Roughly, it can be viewed as an extension of the finite dimensional Dirichlet distribution to an infinite case. The formal definition of Dirichlet process is as follows: a random probability measure $\mathcal{Q}$ on a measurable space $(\Phi, \mathcal{B})$ is generated by a Dirichlet process DP(τ, Q_0) if for every measurable partition $(B_1, \ldots, B_k)$ of the

sample space Φ, the vector of random probabilities $\mathcal{Q}(B_i)$ follows a finite dimensional Dirichlet distribution: $(\mathcal{Q}(B_1), \ldots, \mathcal{Q}(B_k)) \sim \text{Dir}(\tau Q_0(B_1), \ldots, \tau Q_0(B_k))$ where $\tau > 0$ denotes a *scaling parameter* and Q_0 denotes a *base measure* defined on $(\Phi, \mathcal{B})$ [13].

Such a random probability measure is discrete with probability one and admits the following representation:

$$\mathcal{Q}(\cdot) = \sum_{k=1}^{\infty} \beta_k \delta_{\phi_k}(\cdot), \tag{1}$$

where $\delta_\phi(\cdot)$ denotes a point mass at ϕ; the distinct-valued atoms $\phi_k, k = 1, 2, \ldots$, are independent and identically distributed as Q_0; and their probabilities (i.e., weights) $\beta_k, k = 1, 2, \ldots$, are defined by a "stick-breaking" construction through the relationship: $\beta_k = \beta'_k \prod_{l=1}^{k-1}(1 - \beta'_l)$, where $\beta'_k \mid \tau \sim \text{Beta}(1, \tau)$ [29].

Samples from a DP tend to cluster around the distinct-valued atoms and this property is well reflected in the constructive definition of DP based on the following Pólya urn scheme [3]. Having observed n samples with values $(\phi_1, \ldots, \phi_n)$ from $DP(\tau, \mathcal{Q}_0)$, the conditional distribution of the value of the $(n+1)$th sample is given by:

$$\phi_{n+1} | \phi_1, \ldots, \phi_n, \tau, \mathcal{Q}_0 \ \sim \ \sum_{k=1}^{K} \frac{n_k}{n+\tau} \delta_{\phi_k^*}(\cdot) + \frac{\tau}{n+\tau} \mathcal{Q}_0(\cdot), \tag{2}$$

where ϕ_k^* denotes unique values in the n samples drawn so far, K denotes the number of such unique values, and n_k denotes the number of samples with value ϕ_k^*. This expression means that each new sample has positive probability of being equal to an existing value in the drawn samples, and moreover, the probability is proportional to n_k, the number of samples already having the value, hence creating a clustering effect.

Dirichlet process is especially useful in mixture scenarios, where DP acts as a nonparametric prior on parameters of the mixture model. Specifically, observations x_i follow some random distribution $F(\phi_i)$ where $\phi_i \mid \mathcal{Q} \sim \mathcal{Q}$ and $\mathcal{Q}$ is distributed according to a DP. This model is called a *Dirichlet process mixture model*. Note that the number of mixture components under DP prior is random and determined directly from data rather than being pre-specified. This allows the mixture model setting of unknown cardinality and gives more flexibility to the model and the inference.

2.2.2 Hierarchical Dirichlet process

A hierarchical Dirichlet process (HDP) [34] is another nonparametric Bayesian model which serves as a useful prior for data from multiple groups, especially when each group has unique characteristics which can be captured by Dirichlet process, but multiple groups need to be tied together. Suppose each group is associated with a probability measure $\mathcal{Q}_j$ distributed as a Dirichlet process with a base measure Q_0 and a scale parameter τ, that is, $\mathcal{Q}_j \mid \tau, Q_0 \sim \text{DP}(\tau, Q_0)$. Under HDP prior, the shared base measure Q_0 across groups is random and follows another Dirichlet process DP(γ, H). This hierarchical model allows the atoms of random measures $\mathcal{Q}_j$ to be shared across groups and induces a very useful mixture model where groups can share mixture components while admitting each of those to have its own components. The following conditional probabilities summarizes the HDP mixture model:

$$\begin{aligned} \mathcal{Q}_0 \mid \gamma, H &\sim DP(\gamma, H) \\ \mathcal{Q}_j \mid \tau, \mathcal{Q}_0 &\sim DP(\tau, \mathcal{Q}_0) \\ \phi_{ji} \mid \mathcal{Q}_j &\sim \mathcal{Q}_j \\ x_{ji} \mid \phi_{ji} &\sim F(\phi_{ji}) \end{aligned} \tag{3}$$

where x_{ji} denotes the i-th observation in group j, and ϕ_{ji} is the mixture component associated with the observation x_{ji}.

This HDP model can be extended to multiple levels, that is, a tree can be constructed such that each node is associated with a DP generating a base measure for its children and the atoms are shared across descendants, which enables the sharing of clusters at multiple levels of resolution [34].

2.2.3 Infinite Hidden Markov model

A hidden Markov model (HMM) is a very well known statistical model which is especially useful in modeling temporal patterns such as speech, handwriting, biological sequence data, and so on. It assumes the observation x_t for $t = 1, ..., T$ depends on its hidden state q_t where q_t has Markov property. This means that q_t is conditionally independent of $\{q_{t-2}, ..., q_2, q_1\}$ given q_{t-1}.

The HMM can be specified by a tuple $\lambda =< K, M, \boldsymbol{\pi}_0, \boldsymbol{\pi}, \mathbf{b} >$ [25]. Here, N is the number of possible hidden states and M is the number of observations; $\boldsymbol{\pi}_0$ denotes the initial probabilities ($\pi_{0i} = P(q_0 = i)$ for $i = 1, ..., K$); $\boldsymbol{\pi} = \{\pi_{ij}\}$ represents the transition probabilities between hidden states ($\pi_{ij} = P(q_t = j \mid q_{t-1} = i)$); and finally $\mathbf{b}$ defines the emission probabilities for a hidden state to emit each the observation variable ($b_i(k) = P(x_t = k \mid q_t = i)$).

Typical questions arising from HMM can be solved using standardized methods. For example, the probability of a given observation sequence $x_1x_2...x_T$ can be efficiently computed using the so-called forward-backward algorithm. Or the most probable path of hidden states given an observation sequence can be obtained using a dynamic programming scheme called the Viterbi algorithm. More details of these algorithms can be found in [25].

One caveat of this traditional hidden Markov model is that one needs to specify the number of hidden states which is not trivial to determine in many cases. To overcome these limitations, non-parametric extension of HMM to infinite state space was first introduced in [2], where the description of the model was based on two-level hierarchy of urns generating an infinite transition matrix. More recently, infinite hidden Markov model could have been defined more formally with the aid of hierarchical Dirichlet process. Note that both the columns and rows of the transition matrix are infinite dimensional under an infinite HMM. For each source state, the possible transitions to the target states can be modeled by a unique DP. Since all possible source states and target states are taken from the same infinite state space, overall an open set of DPs with different mass distributions on the same support is needed to capture the fact that different source states can have different transition probabilities to any target state. Therefore, the row-specific DPs are linked by a common base measure of another Dirichlet process as in the case of HDP.

Beam sampling algorithm which combines slice sampling and dynamic programming scheme is one good example of recent effort toward efficient inference algorithms for infinite HMM, which shows to be more robust and to outperform the traditional Gibbs sampling [36]. Still much work needs to be done to make this non-parametric model more practical for general use.

3 Proposed Work

Under this proposal, we will develop nonparametric Bayesian models for learning ancestral genetic processes which can provide essential clues to downstream applications. We will have a unified inheritance model which would serve as an approximate coalescence scheme on the assumption that the modern chromosomes are inherited from an unknown number of ancestral chromosomes through certain ancestral events

that can be modeled as probabilistic processes. Ancestral events and concepts such as recombination, mutation, natural selection, admixture and the resulting population structure will be exploited.

We have completed to build the framework of the inheritance model using Dirichlet process and its extensions, which is described in Section 4.1. On top of that, two main application models have been developed: a new model for haplotype inference from multi-population data which employs a hierarchical Dirichlet process mixture (Section 4.2), and another new model using an infinite HMM for the joint analysis of recombination event and population structure (Section 4.3).

In the future, we will expand our models to infer more generalized genetic processes related to natural selection and admixture. So, we aim to estimate local ancestries along the chromosome in an admixed population, and also to detect signatures of selective sweeps on chromosomes in a principled way, by the new models (Section 5).

4 Completed Work

4.1 Dirichlet process based inheritance model

Having a realistic and efficient inheritance model that describes how ancestral materials are passed into modern individuals is a crucial starting step in learning ancestral processes. We have presented a new statistical haplotype inheritance model based on Dirichlet process which was originally introduced in [38]. The model is exchangeable, unbounded, and also has interesting connection with the well-known coalescence as described below.

Our model starts from the assumption that a haplotype population H is originated from an unknown number of founders (ancestral chromosomes), which has gone through mutation. Then H can be naturally modeled as a mixture model by considering modern chromosomes as mixtures of founder chromosomes. Dirichlet process mixture model is especially well suited for this purpose as it allows the number and the configuration of founder chromosomes to be unknown a priori and inferred from data. We associate a mixture component with a founder haplotype (with its mutation rate), that is, $\phi = (a, \theta)$, and each sample with an individual haplotype h. The founder can be mapped to an individual i by an indicator variable c_i such that h_i is inherited as a unit from an ancestor a_{c_i}. Then the following generative scheme is defined as an inheritance model by a DP mixture with a scale parameter τ and a base measure Q_0 [31]:

- Draw first haplotype:

  $a_1, \theta_1 \mid \mathrm{DP}(\tau, Q_0) \sim Q_0(\cdot)$, sample the 1st founder (and its mutation rate);

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

- for subsequent haplotypes:

  – sample the founder indicator for the ith haplotype:

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

  where n_{c_i} is the *occupancy number* of founder a_{c_i}.

  – sample the founder of haplotype i:

  $$a_{c_i}, \theta_{c_i}|\mathrm{DP}(\tau, Q_0) \begin{cases} = \{a_{c_j}, \theta_{c_j}\} \text{ if } c_i = c_j \text{ for some } j < i \\ \sim Q_0(a, \theta) \text{ if } c_i \neq c_j \text{ for all } j < i \end{cases}$$

  – sample the haplotype according to its founder:

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

- sample all genotypes (according to a mapping between haplotype index i and allele index i_e):

  $g_i \mid h_{i_0}, h_{i_1} \sim P_g(\cdot|h_{i_0}, h_{i_1})$.

Here, $P_h(\cdot \mid a, \theta)$ defines the inheritance model to generate an individual haplotype h from a founder a with a mutation rate θ. Note that the index i_e for $e = 0, 1$ represents the maternal and paternal copy of the haplotype pair in each individual. We define our inheritance model to be a *single-locus mutation model* as follows:

$$P_h(h_t|a_t, \theta) = (1-\theta)^{\mathbb{I}(h_t = a_t)} \left(\frac{\theta}{|\mathcal{A}| - 1}\right)^{\mathbb{I}(h_t \neq a_t)} \tag{4}$$

where $\mathbb{I}(\cdot)$ is the indicator function and $|\mathcal{A}|$ is the size of the allele space. This model corresponds to a star genealogy resulting from infrequent mutations over a shared ancestor, and is widely used as an approximation to a full coalescent genealogy starting from the shared ancestor (e.g., [20]).

Given this inheritance model, and under a beta prior $Beta(\alpha_h, \beta_h)$ for the mutation rate θ, it can be shown that the marginal conditional distribution of a haplotype sample $\mathbf{h} = \{h_{i_e} : e \in \{0, 1\}, i \in \{1, 2, ..., I\}\}$ takes the following form resulted from an integration of θ in the joint conditional:

$$p(\mathbf{h}|\mathbf{a}, \mathbf{c}) = \prod_{k=1}^{K} R(\alpha_h, \beta_h) \frac{\Gamma(\alpha_h + l_k)\Gamma(\beta_h + l'_k)}{\Gamma(\alpha_h + \beta_h + l_k + l'_k)} \left(\frac{1}{|\mathcal{A}| - 1}\right)^{l'_k}, \tag{5}$$

where $R(\alpha_h, \beta_h) = \frac{\Gamma(\alpha_h + \beta_h)}{\Gamma(\alpha_h)\Gamma(\beta_h)}$, $l_k = \sum_{i,e,t} \mathbb{I}(h_{i_e,t} = a_{k,t})\mathbb{I}(c_{i_e} = k)$ is the number of alleles which are identical to the ancestral alleles, and $l'_k = \sum_{i,e,t} \mathbb{I}(h_{i_e} \neq a_{k,t})\mathbb{I}(c_{i_e} = k)$ is the total number of mutated alleles.

$P_g(\cdot \mid h_0, h_1)$ defines the genotype observation model under which the genotype allele is stochastically determined by the paternal and maternal alleles with some random noise:

$$P_g(g|h_{i_0,t}, h_{i_1,t}) = \xi^{\mathbb{I}(h=g)}[\mu_1(1-\xi)]^{\mathbb{I}(h \neq^1 g)}[\mu_2(1-\xi)]^{\mathbb{I}(h \neq^2 g)} \tag{6}$$

where $h \triangleq h_{i_0,t} \oplus h_{i_0,t}$ denotes the unordered pair of two actual SNP allele instances at locus t; " $\neq^1$ " denotes set difference by exactly one element; " $\neq^2$ " denotes set difference of both elements, and μ_1 and μ_2 are appropriately defined normalizing constants. Again we place a beta prior $Beta(\alpha_g, \beta_g)$ on ξ for smoothing.

To capture uncertainty over the scaling parameter τ, we use a vague inverse Gamma prior:

$$p(\tau^{-1}) \sim \mathcal{G}(1,1) \Rightarrow p(\tau) \propto \tau^{-2} \exp(-1/\tau)). \tag{7}$$

In general, the probability density function of inverse Gamma distribution with shape parameter ι and scale parameter κ is given as follows:

$$p(x; \iota, \kappa) = \frac{\kappa^{\iota}}{\Gamma(\iota)} x^{-\iota-1} \exp\left(\frac{-\kappa}{x}\right).$$

Under this prior, the posterior distribution of τ depends only on the number of instances n, and the number of components K, but not on how the samples are distributed among the components:

$$p(\tau|k,n) \quad \propto \quad \frac{\tau^{k-2} \exp(1/\tau)\Gamma(\tau)}{\Gamma(n+\tau)}. \tag{8}$$

The distribution $p(\log(\tau)|k,n)$ is log-concave, so we may efficiently generate independent samples from this distribution using adaptive rejection sampling [26].

Under the above model specifications, it is standard to derive the posterior distribution of each haplotype h_{i_e} given all other haplotypes and all genotypes, and the posterior of any missing genotypes, by integrating out parameters θ or ξ and resorting to the Bayes theorem, which enables collapsed Gibbs sampling step where necessary.

As mentioned earlier, there is an interesting connection between our DP based haplotype model and the coalescence. On a coalescent tree with n lineages under an *infinitely-many-alleles* (IMA) model with rate $\tau/2$, a new haplotype is created with probability $\tau/(n-1+\tau)$, and an existing haplotype is replicated with probability $(n-1)/(n-1+\tau)$ [15]. This is identical to the Pólya urn scheme described in Section 2.2.1 with a scaling parameter τ and a uniform base distribution.

4.2 Application in haplotype inference from multi-population data

We have presented a new haplotype inference program, *Haploi*, by extending the DP-based haplotype model described in Section 4.1 to more general cases of multi-populations [31]. Although the problem of haplotype inference has long been studied for its importance in many biological and medical applications, and there has been many previous approaches [32, 18, 6, 19] which has shown to work well for a genetically homogeneous population (e.g. a single ethnic group), few existing programs have explicitly leveraged the individual population labels in haplotype inference while a lot of existing data come from genetically diverse populations. Our model makes use of such information explicitly, and shows comparable and often superior performance compared to the state-of-the-art programs.

4.2.1 Statistical Model

Haploi is developed from a new haplotype distribution model that is based on the hierarchical Dirichlet process mixture. When there exist multi-population data, each sub-population can be modeled as a DP mixture as described in Section 4.1. Instead of treating the sub-populations independently using unrelated

DP mixtures, these can be tied together by incorporating a hierarchical Dirichlet process as a prior. Then the founders can be shared across different sub-populations as well as they are defined population-specifically.

Using the notation in Equation (3) and the formulation in Equation (2), a hierarchical Pólya urn scheme for generating samples under HDP induces the following conditional probability for $(\phi_{jm_j}|\phi_{-jm_j})$ for m_j random draws $\phi_{j1}, \ldots, \phi_{jm_j}$ from $\mathcal{Q}_j$, where the subscript $-jm_j$ denotes the index set of all but the m_j-th sample in j-th population [37]:

$$\begin{aligned}\phi_{jm_j}|\boldsymbol{\phi}_{-jm_j} &\sim \sum_{k=1}^{K} \frac{m_{jk} + \tau\frac{n_k}{n-1+\gamma}}{m_j - 1 + \tau}\delta_{\phi_k^*}(\phi_{jm_j}) + \frac{\tau}{m_j - 1 + \tau}\frac{\gamma}{n-1+\gamma}H(\phi_{jm_j}) \\ &= \sum_{k=1}^{K} \pi'_{jk}\delta_{\phi_k^*}(\phi_{jm_j}) + \pi'_{j,K+1}H(\phi_{jm_j}) \qquad (9)\end{aligned}$$

where n_k denotes the number of samples under $\mathcal{Q}_0$ drawn from the global measure F and equal to ϕ_k^*; m_{jk} denotes the number of samples in the j-th group which are equal to ϕ_k^*; and $\pi'_{jk} := \frac{m_{jk}+\tau\frac{n_k}{n-1+\gamma}}{m_j-1+\tau}$, $\pi'_{j,K+1} = \frac{\tau}{m_j-1+\tau}\frac{\gamma}{n-1+\gamma}$. The vector $\vec{\pi}'_j = (\pi'_{j1}, \pi'_{j2}, \ldots)$ gives the *a priori* conditional probability of a new sample in group j. As shown later, this formula will be useful for implementing a Gibbs sampler for posterior inference under HDP mixtures.

The following summarizes the generative scheme for genotypes in multiple populations under HDP mixture model:

$\mathcal{Q}_0(\phi_1, \phi_2, \ldots)\|\gamma, H \sim \mathrm{DP}(\gamma, H)$,	sample a DP of founders for all populations;
$\mathcal{Q}_j(\phi_{j1}, \phi_{j2}, \ldots)\|\tau, Q_0 \sim \mathrm{DP}(\tau, Q_0)$,	sample the DP of founders for each population;
$\phi_{ji_e}\|\mathcal{Q}_j \sim \mathcal{Q}_j$,	sample the founder of haplotype i_e in population j
$h_{i_e}^{(j)}\|\phi_{ji_e} \sim P_h(\cdot\|\phi_{ji_e})$,	sample haplotype i_e in population j;
$g_i^{(j)}\|h_{i_0}^{(j)}, h_{i_1}^{(j)} \sim P_g(\cdot\|h_{i_0}^{(j)}, h_{i_1}^{(j)})$,	sample genotype i in population j,

where the first three steps describe the HDP scheme for sampling founder haplotypes, the fourth step corresponds to the mixture formulism for the inheritance model, and the last step describes the noisy genotyping model.

4.2.2 Inference

Given genotype data, the individual haplotypes can be inferred using collapsed Gibbs sampling along with the founder haplotypes and other parameters of interest.

An efficient MCMC algorithm can be derived to sample from the posterior associated with HDP mixtures. Specifically, the variables of interest in our model include $\{c_{i_e}^{(j)}\}$: the inheritance variables specifying the origin of each haplotype, $\{a_{k,t}\}$: the founding alleles at all loci of each ancestral haplotype, $\{h_{i_e,t}^{(j)}\}$: the alleles at all loci of individual haplotypes, γ, and τ. All other variables in the model, e.g., the mutation rate θ, are integrated out. The sampler alternates between three coupled stages. First, it samples the scaling parameters γ and τ of the DPs, following the predictive distribution given by Equation (8). Then, it samples the $c_{i_e}^{(j)}$'s and $a_{k,t}$'s given the current values of the hidden haplotypes and the scaling parameters according to the following Equations (10) and (11), respectively.

$$p(c_{i_e}^{(j)} = k|\mathbf{c}^{[-j,i_e]}, \mathbf{h}, \mathbf{a}) \propto p(c_{i_e}^{(j)} = k|\mathbf{c}^{[-j,i_e]}, \mathbf{m}, \mathbf{n}) p(h_{i_e}^{(j)}|a_k, \mathbf{c}, \mathbf{h}^{[-j,i_e]})$$
$$\propto (m_{jk}^{[-j,i_e]} + \tau\beta_k) p(h_{i_e}^{(j)}|a_k, \mathbf{l}_k^{[-j,i_e]}), \text{ for } k = 1, ..., K+1 \tag{10}$$

$$p(a_{k,t}|\mathbf{c}, \mathbf{h}) \propto \tag{11}$$
$$\prod_{j,i_e|c_{i_e,t}^{(j)}=k} p(h_{i_e,t}^{(j)}|a_{k,t}, l_{k,t}^{(j)}) = \frac{\Gamma(\alpha_h + l_{k,t})\Gamma(\beta_h + l'_{k,t})}{\Gamma(\alpha_h + \beta_h + m_k)(|\mathcal{A}| - 1)^{l'_{k,t}}} R(\alpha_h, \beta_h)$$

where $m_{jk}^{[-j,i_e]}$ represents the number of $c_{i'_{e'}}^{(j)}$ that are equal to k, except $c_{i_e}^{(j)}$ in group j, and $m_{j,K+1} = 0$; $\mathbf{l}_k^{[-j,i_e]}$ denotes the sufficient statistics associated with all haplotype instances originating from ancestor k, except $h_{i_e}^{(j)}$; $l_{k,t}$ is the number of allelic instances originating from ancestor k at locus t across the groups that are identical to the ancestor, when the ancestor has the pattern $a_{k,t}$. If k was not represented previously, we can just use zero values of $l_{k,t}$ which is equivalent to using the probability $p(a|h_{i_e}^{(j)})$.

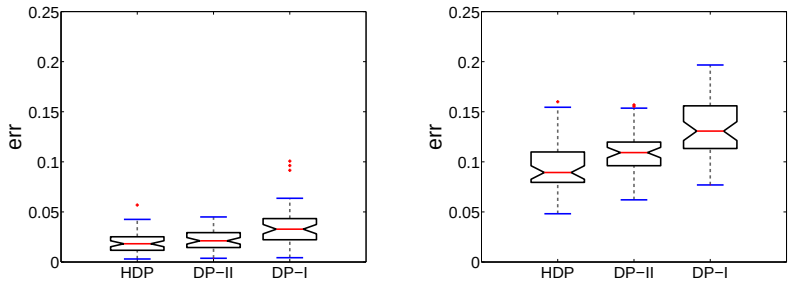
Finally, given the current state of the ancestral pool, the ancestor assignment for each individual and the observed genotypes, it samples the $h_{i_e,t}^{(j)}$ variables according to the following conditional distribution:

$$p(h_{i_e,t}^{(j)}|\mathbf{h}_{[-i_e,t]}^{(j)}, \mathbf{c}, \mathbf{a}, \mathbf{g}) \propto p(g_{i,t}^{(j)}|h_{i_e,t}^{(j)}, h_{i_{\overline{e}},t}^{(j)}, \mathbf{u}_{[-i_e,t]}^{(j)}) p(h_{i_e,t}^{(j)}|a_{k',t}, \mathbf{l}_{k',[-i_e,t]}^{(j)}) = \tag{12}$$
$$R_g \frac{\Gamma(\alpha_g + u)\Gamma(\beta_g + (u' + u''))}{\Gamma(\alpha_g + \beta_g + IJ)} [\mu_1]^{u'} [\mu_2]^{u''} \times R_h \frac{\Gamma(\alpha_h + l_{k',i_e,t}^{(j)})\Gamma(\beta_h + l'^{(j)}_{k',i_e,t})}{\Gamma(\alpha_h + \beta_h + n_k)(|\mathcal{A}| - 1)^{l'^{(j)}_{k',i_e,t}}}$$

where $k' \equiv c_{i_e}^{(j)}$, $l_{k,i_e,t}^{(j)} = l_{[-i_e,t]}^{(j)} + \mathbb{I}(h_{i_e,t}^{(j)} = a_{k,t})$, and $\mathbf{u}_{[-i_e,t]}^{(j)}$ are the set of sufficient statistics recording the inconsistencies between the haplotypes and genotypes in population j.

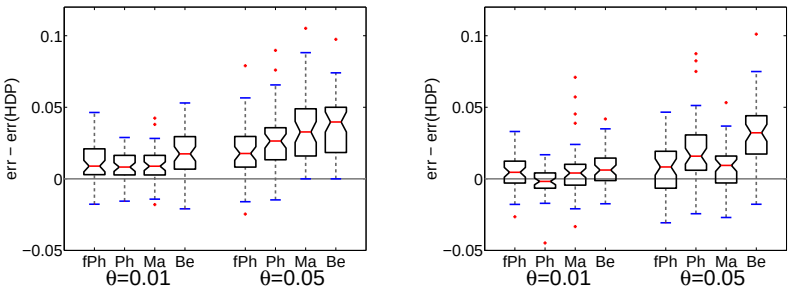
4.2.3 Experimental result

We compare *Haploi* (i.e. HDP) and benchmark algorithms of PHASE 2.1.1 [32, 33], fastPHASE [28], MACH1.0 [19], and Beagle 2.1.3 [6] applied in two modes on synthetic data. Two kinds of multi-population data have been generated, the conserved data using mutation rate of $\theta = 0.01$, and the diverse data using mutation rate of $\theta = 0.05$. Each dataset contains 100 individuals from five populations, and the simulation was repeated 50 times. Given multi-population genotype data, to use DP or other extant methods, one can either adopt mode-I: pool all populations together and jointly solve a single haplotype inference problem that ignore the population label of each individual; or follow mode-II: apply the algorithm to each population and solve multiple haplotype inference problems separately. *Haploi* takes a different approach, by making explicit use of the population labels and jointly solving multiple coupled haplotype inference problems. Note that when only a single population is concerned, or no population label is available, *Haploi* is still applicable and is equivalent to a baseline DP with one more layer of DP hyper-prior over the base measure. We compare the overall performance of *Haploi* on the whole data with other algorithms run in mode-I; and also the accuracy of *Haploi* within each population with those of other methods run in mode-II. Since fastPHASE can also take account of populations labels, we supplied the labels to fastPHASE in mode-I experiments.



(a) $\theta = 0.01$ (b) $\theta = 0.05$

Figure 1: A comparison of HDP with the baseline DP on the synthetic multi-population data. DP-II: DP run on each separate population (mode-II). DP-I: DP run on a merged population (mode-I). The errors measured by site-discrepancies over 50 random samples are presented for (a) conserved datasets ($\theta = 0.01$) and (b) diverse datasets ($\theta = 0.05$).



(a) Mode-I (b) Mode-II

Figure 2: A comparison of HDP with other methods (fPh:fastPHASE, Ph:Phase, Ma:Mach, Be:Beagle) running in (a) mode-I, and (b) mode-II, on synthetic multi-population data. Boxplots for the differences between the error rate of each algorithm and that of HDP (i.e, $err_{\{alg\}} - err_{HDP}$) are presented.

We first test how much HDP can gain by the hierarchical structure on multiple populations compared to the baseline DP. Figure 1 compares the result of HDP with the baseline-DP in mode-I (denoted by DP-I) and that in mode-II (denoted by DP-II) on synthetic multiple populations. On both the conserved samples, which are presumably easier to phase, and the diverse samples, which are more challenging, HDP significantly outperformed DP in both modes (with $p = 0.0336$ against DP-II on the conserved samples, and $p \leq 1.83 \times 10^{-6}$ in all other comparisons, according to a paired t-test). In addition, as a baseline case, we applied HDP to each single-population separately as DP in mode-II, assuming the scenario of a single population or individuals without population labels. Again, HDP applied to all populations jointly outperformed this *baseline HDP* significantly as the latter is deprived of the gain by information sharing. Moreover, this baseline HDP also dominates DP in mode-II significantly, especially on diverse datasets ($p \leq 0.0017$). It appears that the hierarchical structure of HDP which introduces a non-parametric hyper-prior over the base measure of a DPM allows more flexibility in the model and gives better performance than a plain DPM with fixed base measure.

Figure 2 shows boxplots for the differences between the error rate of each benchmark algorithm and that of HDP (i.e., $err_{\{alg\}} - err_{HDP}$). Note that the regions above the horizontal line $y = 0$ correspond to the cases where HDP outperforms others. When other algorithms are run in mode-I (Fig 2 (a)), *Haploi*

outperforms all of them significantly on both the conserved and diverse samples ($p \leq 8.9 \times 10^{-5}$). *Haploi* remains competitive in comparison with other methods when the latter are run in mode-II, i.e., on each population separately (Fig 2 (b)). On the conserved data, PHASE shows the best result, but the differences between algorithms are not significant ($p \leq 0.11$). Whereas on the diverse data, *Haploi* outperforms other algorithms significantly ($p \leq 0.0043$). Again, all significant scores were computed according to a paired t-test.

A thorough sensitivity analysis with respect to the hyper-parameters in our model is detailed in Table 1. The proposed HDP model has two scale parameters, γ and τ, for the upper and lower level DP, which are under inverse Gamma priors. To see the sensitivity of the K and θ estimations under different priors, we applied various values of hyper parameters ι and κ (the same for both γ and τ) on a synthetic dataset. Columns 4 – 9 in Table 1 show the number of recovered founders within each sub-population (the correct number is 5 for each), and the total number of distinct founders over all the populations. Overall, over a wide range of values for the hyper-parameters, *Haploi* gives low-bias and low-variance estimation of the number of founders of each sub-population as well as the total number of distinct founders. In columns 10-11, we show the inferred mutation rate and the haplotyping error. Even when incorrect numbers of founders are recovered, the actual haplotyping errors are not significantly affected, which shows the robustness of the proposed approach for haplotype recovering application.

Next, we compare the haplotype error of *Haploi* and benchmark algorithms on real multi-population data with varying data size. Four population data from real HapMap project DB have been used, and the test was done on randomly selected 100 sets of 6-SNPs segment from chromosome 21. The evaluation is done only on two populations of CEPH and Yoruba due to the limited availability of ground truth. For inference, three population-composition scenarios are considered; FourPops: all the four populations are merged together and used for inference (with or without population labels depending on the software's adaptability). TwoPops: only the two populations of CEPH and Yoruba are used together for inference. OnePop: inference is done for CEPH and Yoruba separately. For each of the three population-composition scenarios, we applied all methods to different population sizes, i.e., 60, 30, 20, and 10 individuals per population, to examine the effect of population size on phasing accuracy. Figure 3 summarizes the result which shows that *Haploi* improves significantly as more populations are added in the inference while other benchmark algorithms do not show such tendency. Also the performance gain through information sharing enabled by HDP tends to be greater when the population sizes decrease, suggesting that HDP is especially advantageous for the data scarcity situation. Comparing the results from the most preferred scenarios of each algorithm, *Haploi* and PHASE work equally well when all the available data were used (i.e. #individuals per pop=60), and *Haploi* starts to surpass others more substantially when the population size decreases.

Partition-Ligation scheme for long sequences As for most haplotype inference models proposed in the literature, the state space of the proposed HDP mixture model scales exponentially with the length of the genotype sequence, and therefore it cannot be directly applied to genotype data containing hundreds or thousands of SNPs. To deal with haplotypes with a large number of linked SNPs, [21] proposed a divide-and-conquer heuristic known as Partition-Ligation (PL), which was adopted by a number of haplotype inference algorithms including PL-EM [24], PHASE [32, 18], and CHB [41]. We equipped *Haploi* with a variant of the PL heuristic for haplotype inference of multiple population genotype data over long SNPs sequences. We omit the technical details here, and more description can be found in [31]. We tested *Haploi* with PL scheme on 10 ENCODE regions from the HapMap DB, each spanning roughly 500 Kb and containing from 254 to 972 common SNPs across all four populations. We performed haplotype inference under three different population-composition scenarios as before, but due to the extremely high cost in computational time in

Table 1: A sensitivity analysis to the hyper-parameters of HDP on a synthetic dataset. Result with different hyper-parameters ι and κ for inverse Gamma prior is shown. The number of founders for each population (K_i) and the total number of ancestors across all the populations are shown in columns 4–9. The estimated mutation rate θ and the haplotyping errors (err_s) are also shown through columns 10 – 11. The sensitivity of θ estimate to the hyper prior is examined over a wide range of both different magnitudes (0.1 to 1000) and ratios (0.0001 to 10000) of ι and κ.

κ	ι	κ/ι	K_1	K_2	K_3	K_4	K_5	total K (17)	θ (0.005)	err_s
0.1	0.1	1	5.0	5.0	5.0	5.0	5.0	17.8	0.005	0.0058
	0.5	0.2	5.0	5.0	5.0	5.0	5.0	17.5	0.004	0.0116
	1	0.1	5.0	5.0	5.0	5.0	5.0	18.0	0.004	0.0000
	10	0.01	5.0	5.0	5.0	5.0	5.0	18.0	0.004	0.0087
	100	0.001	5.0	4.0	5.0	5.0	4.0	16.0	0.007	0.0029
	1000	0.0001	5.0	5.0	5.0	5.0	4.0	17.0	0.004	0.0029
0.5	0.1	5	5.0	5.1	5.0	5.0	5.0	18.1	0.004	0.0087
	0.5	1	5.0	4.1	5.0	5.0	5.0	17.1	0.007	0.0029
	1	0.5	5.0	5.0	5.0	5.0	5.0	18.0	0.004	0.0029
	10	0.05	5.0	5.0	5.0	5.0	5.0	18.0	0.004	0.0145
	100	0.005	5.0	5.0	5.0	5.0	4.0	17.0	0.004	0.0029
	1000	0.0005	5.0	5.0	5.0	5.0	4.0	17.0	0.005	0.0087
1	0.1	10	5.0	5.0	5.0	6.0	5.0	18.0	0.006	0.0116
	0.5	2	5.0	5.0	5.0	5.0	5.0	18.0	0.004	0.0058
	1	1	5.0	5.0	5.0	5.0	5.0	18.0	0.004	0.0087
	10	0.1	5.0	5.0	5.0	5.0	5.0	18.0	0.004	0.0029
	100	0.01	5.0	4.0	5.0	5.0	4.0	16.0	0.007	0.0087
	1000	0.001	5.0	4.9	5.0	5.0	4.0	16.9	0.005	0.0087
10	0.1	100	5.0	5.0	5.0	5.3	5.0	17.1	0.004	0.0000
	0.5	20	5.0	5.0	5.0	5.0	5.0	18.0	0.004	0.0087
	1	10	5.0	5.0	5.0	5.0	5.0	18.1	0.004	0.0029
	10	1	5.0	5.0	5.0	5.0	5.0	18.0	0.004	0.0000
	100	0.1	5.0	4.0	5.0	5.0	5.0	17.0	0.007	0.0058
	1000	0.01	5.0	5.0	5.0	5.0	4.0	17.0	0.004	0.0087
100	0.1	1000	5.8	5.5	5.6	6.1	6.0	18.2	0.010	0.0116
	0.5	200	5.2	5.2	5.2	5.8	5.5	18.4	0.008	0.0116
	1	100	5.1	6.2	5.4	5.5	5.2	17.3	0.006	0.0087
	10	10	5.0	5.0	5.1	5.0	5.1	18.1	0.005	0.0029
	100	1	5.0	5.0	5.0	5.0	5.0	18.0	0.004	0.0000
	1000	0.1	5.0	5.0	5.0	5.0	4.0	17.0	0.004	0.0000
1000	0.1	10000	6.8	6.3	8.5	6.0	10.3	25.6	0.003	0.0087
	0.5	2000	7.1	7.0	7.4	6.6	8.5	24.5	0.006	0.0116
	1	1000	6.4	6.5	7.7	6.4	8.4	22.8	0.005	0.0145
	10	100	5.3	6.5	6.3	5.8	7.0	17.8	0.010	0.0260
	100	10	5.1	5.1	5.0	5.0	5.1	18.1	0.005	0.0087
	1000	1	5.0	5.0	5.0	5.0	5.0	18.0	0.004	0.0029

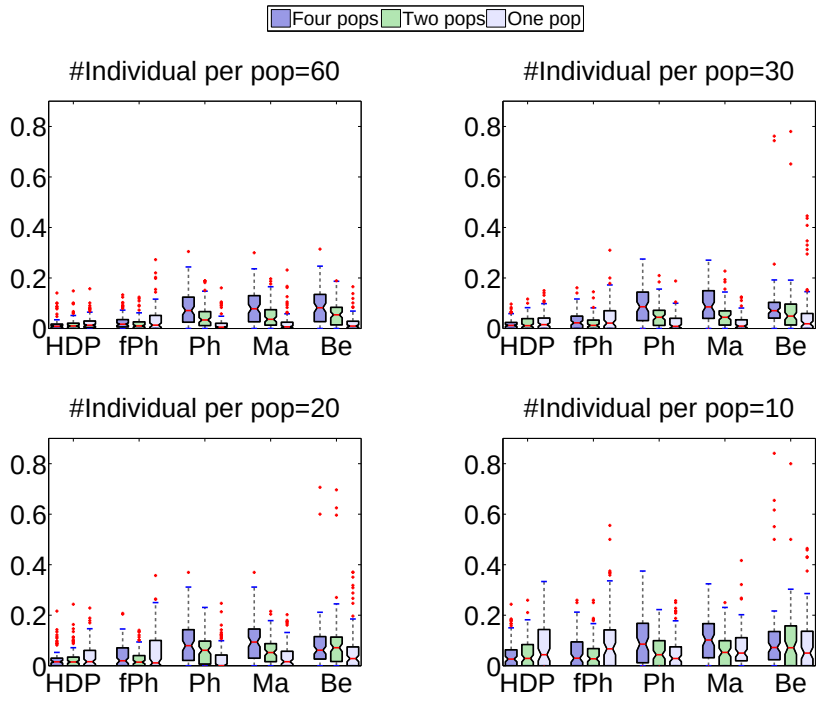
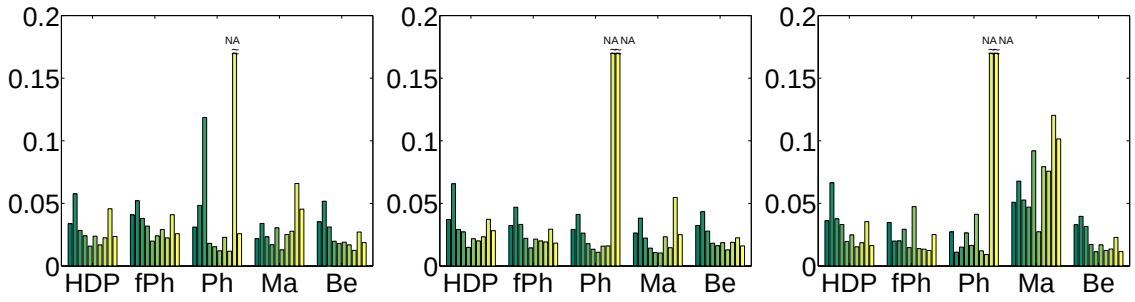


Figure 3: A comparison of haplotyping error on CEPH+Yoruba population over randomly chosen 100 sets of 6-SNP segments from Chromosome 21 [31] . The results were obtained under three population-composition scenarios: (i) FourPops: when data from all the four populations were used (blue) for inference; (ii) TwoPops: when data from CEPH and Yoruba populations were used together (green); (iii) OnePop: when each of CEPH and Yoruba population was used separately (gray). Different sample sizes, with 60, 30, 20, and 10 individuals per each population, were used.

these experiments, we only worked on the full-size data sets. Figure 4 shows a comparison of haplotype reconstruction quality. Out of the 30 experiments we performed (10 regions and three scenarios), the PHASE program failed to yield results in 5 experiments after a 31-day runtime, so we omit the corresponding results in our summary figure.

The conclusion from Figure 4 is less clear than the ones from experiments on short SNP sequences. Overall, Beagle dominates all the algorithms with a small margin, PHASE also shows comparable result to Beagle when converged, but all the other algorithms work comparably in most cases across different datasets and different scenarios. In terms of computational cost, Beagle was the fastest, it took less than a minute for each task; fastPHASE and MACH mostly took less than 1 hour for each task, *Haploi* took from 1-10 hours, depending on the length of the sequence; whereas PHASE took one to two orders of magnitude longer, and was indeed impractical for phasing very long sequence.

In summary, our result shows that *Haploi* is competent and robust for phasing long SNP sequences from diverse genetic origins at reasonable time cost, even though it has not yet employed any sophisticated way for processing long sequences, such as the recombination process. Since *Haploi* appeared to outperform other methods over short SNPs, we believe that the competence of *Haploi* on long SNPs is due to a better inference power endowed by the HDP model for multi-population haplotypes; and we expect that an upgrade that incorporates explicit recombination models in conjunction with HDP for long SNPs are likely to lead to more accurate haplotype reconstructions.



(a) (b) (c)

Figure 4: Performance on the full sequences of the selected ten ENCODE regions. (a) Error rates under four population scenario (b) Under the two-population scenario. (c) Under the one population scenario. For cases of which the program does not converge (NC) within a tolerable duration (i.e., 800 hours), we cap the bar with a "≈" to indicate that the results are not available (NA).

4.3 Application in population structure and recombination analysis

The linkage disequilibrium (LD) pattern and population structure are closely related under a common inheritance process, but the statistical methods developed so far mostly deal with these genetic properties separately. We presented a new model-based approach to address these issues through joint inference of population structure and recombination events under a nonparametric Bayesian framework [30, 39]

4.3.1 Statistical Model

The model is based on the same basic assumption described in Section 4.1, but recombination is now added in the inheritance process in addition to mutation. Thus, modern chromosomes are assumed to be formed from ancestral chromosomes via biased random recombination and mutation. If the number of founders is known and can be fixed, sequential selection of recombination targets from a set of founder chromosomes can be modeled as a hidden Markov process, where the hidden states correspond to the founders, 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 in the ancestors to the descendants. As first proposed in [2] and later discussed in [34], 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. We call this a *hidden Markov Dirichlet process* model for inheritance under recombination and mutation. Hence, this model extends the traditional hidden Markov model to an infinite ancestral space. In addition, it is an extension of the DP-based haplotype model such that the association of an individual haplotype with the founder can now change along the chromosome positions due to the possible ancestral recombination, which serves as more realistic inheritance model for dealing with long chromosome sequences.

As discussed in [38], associating each hidden state k with an ancestor configuration $\phi_k = \{a_k, \theta_k\}$ whose values are drawn from the top level base measure $H \equiv Beta(\theta)p(a)$, conditioning on the Dirichlet process DP(γ, H) which stochastically determines a common base measure for row-specific DPs, the

samples from the j-th DP (i.e. transitions from source state j) are distributed as follows:

$$\phi_{m_j}|\boldsymbol{\phi}_{-m_j} \sim \sum_{k=1}^{K} \frac{m_{j,k} + \tau\frac{n_k}{n-1+\gamma}}{m_j - 1 + \tau}\delta_{\phi_k^*}(\phi_{m_j}) + \frac{\tau}{m_j - 1 + \tau}\frac{\gamma}{n-1+\gamma}H(\phi_{m_j})$$
$$= \sum_{k=1}^{K} \pi_{j,k}\delta_{\phi_k^*}(\phi_{m_j}) + \pi_{j,K+1}H(\phi_{m_j}), \tag{13}$$

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

Based on this HMDP model, for each modern chromosome i, let $c_i = [c_{i,1}, \ldots, c_{i,T}]$ denote the sequence of inheritance variables specifying the index of the ancestral chromosome at each SNP locus. When no recombination takes place during the inheritance process that produces haplotype h_i (say, from ancestor k), then $c_{i,t} = k, \forall t$. When a recombination occurs, say, between loci t and $t+1$, we have $c_{i,t} \neq c_{i,t+1}$. We can introduce a Poisson point process to control the duration of non-recombinant inheritance. That is, given that $c_{i,t} = k$, then with probability $e^{-dr} + (1 - e^{-dr})\pi_{kk}$, where d is the physical distance between two loci, r reflects the rate of recombination per unit distance, and π_{kk} is the self-transition probability of ancestor k defined by HMDP, we have $c_{i,t+1} = c_{i,t}$; otherwise, the source state (i.e., ancestor chromosome k) pairs with a target state (e.g., ancestor chromosome k') between loci t and $t+1$, with probability $(1 - e^{-dr})\pi_{kk'}$. Hence, each haplotype h_i is a mosaic of segments of multiple ancestral chromosomes from the ancestral pool $\{a_{k,\cdot}\}_{k=1}^{\infty}$. 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 so that we are back to a standard stationary HMDP model.

The emission process of the HMDP corresponds to an inheritance model from an ancestor to the matching descendent. We adopt the *single-locus mutation model* introduced in Section 4.1.

4.3.2 Inference

We briefly describe a Gibbs sampling algorithm for posterior inference under HMDP. 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. Here, we assume that individual haplotypes as well as genotypes are already known for simplicity, but the model can be easily extended to unknown haplotypes as in the application for haplotype inference.

The Gibbs sampler alternates between two 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 $\mathbf{n}$ and $\mathbf{m}$ denote the sufficient statistics for the transitions between ancestors in HMDP Pólya urn scheme. 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:

$$P(c_{t+1:t+\delta} \,|\mathbf{c}^-, \mathbf{h}, \mathbf{a}) \propto \prod_{j=t}^{t+\delta} P(c_{j+1}|c_j, \mathbf{m}, \mathbf{n}) \prod_{j=t+1}^{t+\delta} P(h_j|a_{c_j,j}, \mathbf{l}_{c_j}) \tag{14}$$

This expression is simply Bayes' theorem with $\prod_{j=t+1}^{t+\delta} p(h_j|a_{c_j,j}, \mathbf{l}_{c_j})$ playing the role of the likelihood and $p(c_{t+1:t+\delta} \,|\mathbf{c}^-, \mathbf{h}, \mathbf{a})$ playing the role of the posterior. 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, Equation (14) reduces to:

$$\begin{aligned}
&p(c_{t+1:t+\delta} \mid \mathbf{c}^-, \mathbf{h}, \mathbf{a}) \\
&\sim p(\text{at most one recombination in} [t, t+\delta] \,|\mathbf{c}^-, \mathbf{h}, \mathbf{a}) \\
&\propto p(c_{t'} \,|c_{t'-1} = c_t, \mathbf{m}, \mathbf{n}) p(c_{t+\delta+1} \,|c_{t+\delta} = c_{t'}, \mathbf{m}, \mathbf{n}) \times \\
&\quad \prod_{j=t'}^{t+\delta} p(h_j|a_{c_{t'},j}, \mathbf{l}_{c_{t'}})
\end{aligned}$$

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:

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

where $\pi_{k,\cdot}$ represents the transition probability vector for ancestor k under HMDP. Putting everything together, we have the proposal distribution for a block of inheritance variables.

To sample the ancestors $\{a_{k,t}\}$, we can derive the posterior distribution similar to Equation (11).

4.3.3 Experimental results

Spectrum, an efficient implementation of our new model has been validated on simulated data and applied also to real SNP datasets of ENm010 region on chromosome 7 in HapMap DB. While the algorithm was run with all the populations together, according to the implications about the distinct genetic structure reflected in the ancestral map (Figure 6), we estimated the empirical recombination rates separately for each population (i.e., CEPH, YRI and HCB+JPT) by using the posterior samples belonging to each population only.

Figure 5 shows the recombination rate estimates and the detected recombination hotspots, together with the corresponding LD-measurement. While each recombination pattern largely agrees with the given LD patterns, noticeably different patterns of recombination hotspots of the three groups are observed, which may reflect different recombination histories of the ancestors of these populations and the need for the

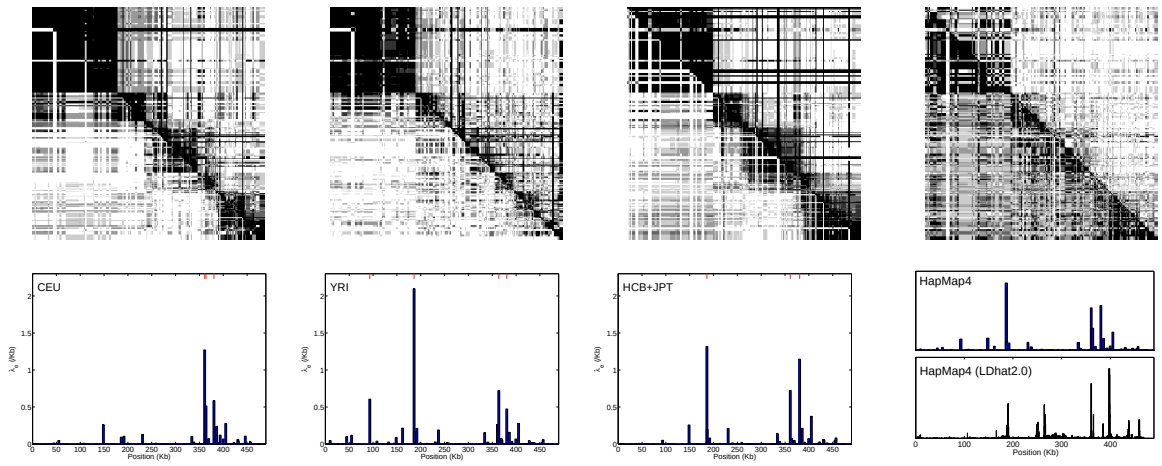


Figure 5: For each population of HapMap data, the LD measure with the estimated recombination rates along the chromosomal position are shown together with the detected recombination hotspots [30]. The last column shows the result on the mixed four populations from both *Spectrum* and *LDhat 2.0*.

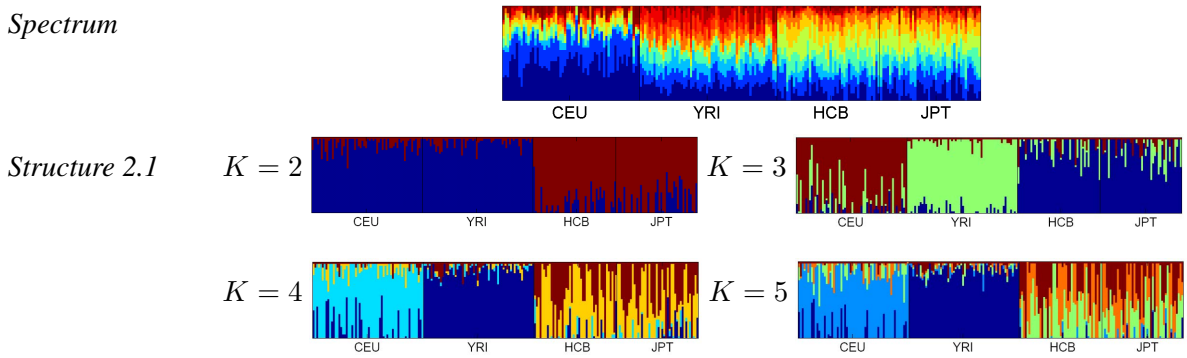


Figure 6: Inferred population structure of HapMap four population data from *Spectrum*, and *Structure 2.1* with different pre-specified numbers of population K [30].

population-based recombination analysis. For comparison, the result on the mixed populations are also shown together for *Spectrum* and LDhat 2.0 [12] in the last column of Figure 5. Overall, it performs well relative to LDhat 2.0 in estimating the recombination rates and hotspots.

More interestingly, *Spectrum* generates an ancestral spectrum for representing population structures which not only displays sub-population structure but also reveals the genetic diversity of each individual. Note that *Spectrum* uncovers the genetic origins of all loci of each individual haplotype in a population from Gibbs samples of the inheritance variables $\{c_{i,t}\}$. We define an empirical *ancestor composition vector* η_e for each individual, which records the fractions of every ancestor in all the $c_{i,t}$'s of that individual. Figure 6 displays an *ancestral spectrum* constructed from the η_e's of all individuals. In this spectrum, each individual is represented by a vertical line which is partitioned into colored segments in proportion to the ancestral fraction recorded by η_e. It offers an alternative view of the population structures to that offered by Structure 2.1 [23, 11], which ignores chromosome-level mutation and recombination with respect to founders.

5 Future work

5.1 Inference of local ancestry in admixed populations

We propose to develop a new statistical method for inferring local ancestry along the individual chromosomes in recently admixed populations. Through migration, human has been forming admixed populations, e.g. African American or Latino American which can be thought of as admixture of ancestral populations of Europeans, Africans, native Americans. Interestingly, many phenotypes such as disease susceptibility show great difference across populations depending on which ancestral populations have formed the admixed one on particular genetic loci, reflecting the influence of natural selective forces. Hence, local ancestry information if available would give essential clues for finding selective sweeps on chromosomes and also for which loci are associated with which phenotypes.

Most previous work has focused on models assuming low marker density so that each locus can be considered independently. Moreover, those methods have often relied on allele frequencies at each locus, not on the real haplotypes, and this makes it difficult for the resulting likelihood models to reflect the underlying ancestral genetic processes properly. This means that those approaches can only reveal limited view for local ancestry with low resolution coming from less realistic assumptions. We intend to develop more biologically realistic and statistically sound model. The model would assume high density SNP markers and hence LD between markers, and also incorporate the necessary genetic processes starting from founder haplotypes to the observed modern haplotypes as defined in our previous works. For instance, each ancestral population can be characterized by a unique HMDP model, while different ancestral populations can share their founders and the recombining patterns. Chromosomes in the admixed population then can be analyzed in reference to these ancestral populations.

Computational complexity is one of the major challenges, so efficient inference algorithms for nonparametric Bayesian models will be exploited rigorously, for example, beam sampling algorithm developed for an infinite HMM [36].

Validation on synthetic data will be necessary because ground truth data for local ancestry is not available in practice. For realistic simulation, real multi-population data from HapMap project will be used as ancestral populations, and the admixed population then can be generated from those ancestral populations under various demographic scenarios, especially under which existing methods have difficulty in accurate estimates. These would include the case of more than two populations mixing at different times or admixing of very close populations such as Japanese and Chinese. In addition to this simulation study, there are real admixed population data publicly available, so those will be also analyzed for biologically interesting findings.

5.2 A new model for detecting recent selective sweeps

Signatures of selective sweeps that provide valuable clues for association study have been mostly analyzed so far by using some summary statistics often defined heuristically [35]. For example, the typical approaches often search for certain patterns in the genomic sequences such as a low haplotype frequency or a skewed site frequency. However, it is not obvious to score which summary statistics is better, and moreover, structural information contained in the genome is often ignored in the search. A new approach using a composite-likelihood of the allele frequencies spectrum [5] has shown significant improvement over existing methods, but it is rather sensitive to demographic effects because of a simple demographic model assumption. More recently, model-based approaches have been appearing, for instance, [16] employed a hidden Markov model to detect the selective sweeps based on allele frequency spectrum as observations. While these methods

show superior and more reliable performance on the study case, those models are still preliminary in a sense as they are often based on a very simple demographic scenario with little consideration about underlying genetic history. More realistic models need to be developed further to provide reliable results for downstream analysis.

We propose to develop a new model-based approach for detecting selective sweeps and apply it to real data such as Arabidopsis data from University of Chicago group. The Bayesian and biological framework we have developed so far will be adapted for this purpose so that the inheritance processes are modeled by treating founder chromosomes as mixture components and by employing necessary genetic processes as probabilistic processes for generating observations. For example, we may assume the alleles are determined by some hidden states (e.g. *sweep* and *neutral*) and then a different haplotype inheritance model can be used to generate the individual allele at a specific site depending on the hidden state where *sweep* sites follow more skewed founder distribution. This new approach is expected to be more robust to different demographic history by utilizing correlation structure in chromosome more systematically and also easily adaptable to more complicated population scenarios.

5.3 Software Release

All the softwares developed and to be developed will be publicly available on a project website. All the related information will be maintained together including source codes, examples, references and related documents.

5.4 Timeline

The approximate timeline for the remaining thesis work is as follows.

Activity	Months	Start date
(Side-project on time-varying network)	2	September 2009
Model development for future work 1	1	
Implementation, experiments and paper submission for future work 1	3	
Model development for future work 2	1.5	March 2010
Implementation, experiments and paper submission for future work 2	3.5	
Future work 3	1	August 2010
Thesis writing	4	September 2010

References

[1] 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.

[2] Matthew J. Beal, Zoubin Ghahramani, and Carl Edward Rasmussen. The infinite hidden markov model. In *Advances in Neural Information Processing Systems 14*. MIT Press, 2002.

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

[4] David M. Blei and Michael I. Jordan. Variational inference for dirichlet process mixtures. *Journal of Bayesian Analysis*, 1[1]:121–144, 2006.

[5] Simon Boitard, Christian Schlotterer, and Andreas Futschik. Detecting selective sweeps: A new approach based on hidden markov models. *Genetics*, 181:1567–1578, 2009.

[6] Sharon R. Browning and Brian L. Browning. Rapid and accurate haplotype phasing and missing data inference for whole genome association studies using localized haplotype clustering. *Am J Hum Genet.*, 81:1084–1097, 2007.

[7] A. Chakravarti. Single nucleotide polymorphisms: . . .to a future of genetic medicine. *Nature*, 409:822–823, 2001.

[8] A. Clark. Finding genes underlying risk of complex disease by linkage disequilibrium mapping. *Curr Opin Genet Dev*, 13(3):296–302, 2003.

[9] 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.

[10] 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.

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

[12] P. Fearnhead and P. J. Donnelly. Estimating recombination rates from population genetic data. *Genetics*, 159:1299–1318, 2001.

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

[14] S. E. Hodge, M. Boehnke, and M. A. Spence. Loss of information due to ambiguous haplotyping of SNPs. *Nat Genet*, 21:360–361, 1999.

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

[16] Yuseob Kim and Wolfgang Stephan. Detecting a local signature of genetic hitchhiking along a recombining chromosome. *Genetics*, 160:765–777, 2002.

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

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

[19] Yun Li and GR Abecasis. Mach 1.0: Rapid haplotype reconstruction and missing genotype inference. *Am J Hum Genet.*, S79:2290, 2006.

[20] 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.

[21] 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.

[22] 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.

[23] J. K. Pritchard, M. Stephens, and P. Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155:945–959, 2000.

[24] Zhaohui S. Qin, Tianhua Niu, and Jun S. Liu. Partition-ligation expectation-maximization algorithm for haplotype inference with single-nucleotide polymorphisms. *Am J Hum Genet*, 71:1242–1247, 2002.

[25] Lawrence R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. In *Proceedings of the IEEE*, volume 77:2, pages 257–286, 1989.

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

[27] 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.

[28] Paul Scheet and Matthew Stephensu. A fast and flexible statistical model for large-scale population genotype data: Applications to inferring missing genotypes and haplotypic phase. *Am J Hum Genet*, 78:629–644, 2006.

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

[30] Kyung-Ah Sohn and Eric P. Xing. Spectrum: Joint bayesian inference of population structure and recombination event. In *Proc of The Fifteenth International Conference on Intelligence Systems for Molecular Biology*, 2007.

[31] Kyung-Ah Sohn and Eric P. Xing. A hierarchical dirichlet process mixture model for haplotype reconstruction from multi-population data. *Annals of Applied Statistics*, 2009.

[32] 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.

[33] Matthew Stephens and Paul Scheet. Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. *American Journal of Human Genetics*, 76:449–462, 2005.

[34] Yee Whye Teh, Michael I. Jordan, Matthew J. Beal, and David M. Blei. Hierarchical Dirichlet processes. *JASA*, 101(476):1566–1581, 2006.

[35] G. Coop Teshima, K. and M. Przeworski. How reliable are empirical genomic scans for selective sweeps. *Genome Res.*, 16:702–712, 2006.

[36] Jurgen Van Gael, Yunus Saatci, Yee Whye Teh, and Zoubin Ghahramani. Beam sampling for the infinite hidden markov model. In *ICML '08: Proceedings of the 25th international conference on Machine learning*, pages 1088–1095, New York, NY, USA, 2008. ACM.

[37] 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 *Proc 23th Int Conf on Machine Learning*, pages 1049–1056, New York, 2006. ACM Press.

[38] Eric P. Xing, Roded Sharan, and Michael I. Jordan. Bayesian haplotype inference via the Dirichlet process. In *Proceedings of the 21st International Conference on Machine Learning*, 2004.

[39] Eric P. Xing and Kyung-Ah Sohn. Hidden Markov Dirichlet process: Modeling genetic recombination in open ancestral space. *Bayesian Analysis*, 2(3):501–528, 2007.

[40] 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.

[41] Yu Zhang, Tianhua Niu, and Jun S. Liu. A coalescence-guided hierarchical bayesian method for haplotype inference. *Am J Hum Genet*, 79:313–322, 2006.