Learning Ancestral Genetic Processes using Nonparametric Bayesian Models

Kyung-Ah Sohn

CMU-CS-11-136

November 2011

Computer Science Department
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

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.

This research was sponsored by the National Science Foundation under grant number DBI-0546594. The views and conclusions contained in this document are those of the author and should not be interpreted as representing the official policies, either expressed or implied, of any sponsoring institution, the U.S. government or any other entity.

Keywords: ancestral inference, Dirichlet process, hierarchical Dirichlet process, infinite hidden Markov model, haplotypes, recombination, admixture, local ancestry

To Celine, KiHyun, and my parents

Abstract

Recent explosion of genomic data have enabled in-depth investigation of complex genetic mechanisms for various applications such as the inference on the human evolutionary history or the search for the genetic basis of phenotypic traits. Although great advances have been made in the analysis of genetic processes underlying such data, most statistical methods developed so far deal with the closely related genetic objects separately using specialized methods, and do not capture the intrinsic relatedness among multiple properties that have resulted from a common inheritance process. Moreover, these approaches often ignore the inherent uncertainty about the genetic complexity of the data and rely on inflexible models resulting from restrictive assumptions.

In this thesis, we develop nonparametric Bayesian models for learning ancestral genetic processes, which provide more flexible control over the complexity of the genetic data, and at the same time, utilize the structured data in a more principled way. Under a unified inheritance framework built on the assumption of hypothetical founder haplotypes that generate modern individual chromosomes, hierarchical Bayesian models based on Dirichlet process are developed for the following related applications in population genetics: the problem of haplotype inference from multi-population genotype data, joint inference of population structure and the recombination events, and the local ancestry estimation in admixed populations. This new approach allows one to explicitly exploit the shared structural information in the data from multiple populations. The resulting methods have shown to significantly outperform other existing methods that do not utilize such relatedness properly.

Acknowledgments

First and foremost, I would like to express my deepest gratitude to my advisor Prof. Eric P. Xing for his guidance during my study in Carnegie Mellon University. This thesis would not have been possible without his continuous encouragement and excellent advising. I am also very grateful to Prof. Zoubin Ghahramani who has given me lots of helpful advices and kind suggestions on the thesis work. I wish to thank the rest of my committee members as well: Prof. Russell Schwartz, Prof. Kathryn Roeder, and Prof. Matthew Stephens, for their insightful comments and supports at my thesis defense and throughout my thesis writing.

I would like to thank all the former and current members of the sailing group. My special thanks goes to Fan Guo and Wenjie Fu, my dear friends and sources for lots of fun and help over the many years in CMU. I also thank Prof. Seyoung Kim, who as an excellent colleague and a good friend, has helped me to develop various skills as a researcher and has also given heart-full advices on everything. I also want to thank Jing and Prof. Wei Wu for their positive energies and fun experience we shared through the recent project. All the other people including Amr, Hetu, Steve, Pradipta, Judie, Suyash, Kriti, Mladen, Seunghak, Gunhee, Ross, Leon, Ankur, Qirong, Le, Zun, Andrew, and Jacob have also inspired me in various aspects. I would also like to thank Yanxin and Xi for their warm friendships.

I would like to show my gratitude to Prof. Martin Kreitman and Prof. Joy Bergelson for their kind welcome and insightful discussion during my visit to University of Chicago. I am also grateful to Prof. Yee Whye Teh and Prof. Michael I. Jordan for their help and comments as co-authors. Many thanks go to other professors in School of Computer Science as well, especially Prof. Christos Faloutsos, Prof. Ziv Bar-Joseph, Prof. Mor Harchol-Balter, and Prof. Tai-Sing Lee for giving me invaluable advice on research projects, teaching, writing, and life. I would like to thank Sharon, Deb, Michelle, Sophie, Martha, and Catherine for their professional support and help during my graduate years.

I am also deeply thankful to Prof. Myung-Soo Kim for suggesting and advising me to work on computer science, and to Prof. Hyeonbae Kang for guiding me to study applied fields in mathematics when I was a mathematics major. I wish to thank my friends Juhi and Sungwon for helping me get through the difficult times and for sharing important moments with me.

Most importantly, I wish to thank my family. I am indebted to my parents who have always encouraged me with their best wishes, and cared for my daughter with endless love in a foreign country they have never visited before. Without them, this dissertation is just impossible. My little daughter Celine has given me a life-time project with a lot of happiness and challenges, thanks for coming to me. Lastly, Kihyun, he is a perfect husband to me. Thanks for being with me in every moment.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Overview

Recent advances in biotechnology have led to an explosion of genomic data. Understanding of hidden mechanisms underlying such data is crucial for many applications such as the inference on the evolutionary history of human population or the search for the genetic basis of various phenotypic traits (Chakravarti, 2001; Clark, 2003; Li et al., 2009; Price et al., 2006; Wang et al., 2010; Xu et al., 2008; Xu and Jin, 2008). A lot of statistical methods have been developed to uncover the genetic mechanisms and ancestral processes from the genetic data, for example, for the analysis of recombination rates and hotspots (Anderson and Novembre, 2003; Daly et al., 2001; Patil et al., 2001; Zhang et al., 2002), for the reconstruction of haplotypes given genotype sequences (Browning and Browning, 2009; Excoffier and Slatkin, 1995; Li et al., 2010; Qin et al., 2002; Scheet and Stephens, 2006; Stephens and Scheet, 2005), or for the population structure and ancestry estimation in admixed populations (Falush et al., 2003; Pasaniuc et al., 2009; Patterson et al., 2004; Price et al., 2009; Sundquist et al., 2008). Although great advances have been made in these studies through efficient utilization of the increasing amount of data, conventional approaches developed so far often rely on the restrictive parametric models that do not capture the intrinsic relatedness among multiple genetic objects, and deal with the closely related genetic properties separately using specialized methods. The overall goal of this thesis is to propose a more flexible statistical framework that addresses these issues in a principled way. For the inference of ancestral genetic processes that can enhance our understanding about the genetic mechanisms, we develop non-parametric Bayesian models that provide more flexible control over the complexity of the genetic data and at the same time utilize the structured data in a more principled way.

We especially focus on the haplotype data constructed from genetic polymorphisms called single nucleotide polymorphisms (SNPs). On the assumption of hypothetical founders that generate haplotypes in modern populations, we employ a new haplotype inheritance model in Xing et al. (2007) that allows one to incorporate various genetic processes in a unified framework. Under this framework, the distribution of haplotypes in a population is modeled as a Dirichlet process (DP) mixture model. It offers a principled approach to take into account the inherent uncertainty regarding the size of the hypothetical founder pool, so the number of the founder haplotypes does not need to be pre-specified and can be naturally inferred from the given population data. Furthermore, it provides a reasonable approximation to the well-known theory called the coalescence in population genetics by utilizing the partition structure resulting from the Dirichlet process.

Using the DP-based haplotype inheritance model as a building block, we develop flexible non-parametric Bayesian models for ancestral genetic processes in the following three major applications. First, we consider the problem of inferring haplotypes using genotypes from multiple populations. Most previous approaches for haplotype inference either ignore the sub-population structure, or handle each of the sub-population separately and therefore also ignore the close relationship between different populations. We adopt a hierarchical Dirichlet process that enables one to overcome this limitation systematically. The resulting haplotype model explicitly exploits the population labels and shows significantly enhanced performance over previous methods.

We further generalize this model to incorporate the recombination process as well as the mutation process from the hypothetical founders to the modern individuals. The haplotype inheritance under these two processes is modeled by an infinite hidden Markov process in which the hidden state corresponds to a founder haplotype and the observation corresponds to the individual haplotype. It enables one to infer the population structure and the recombination events jointly in a single framework by tracing the association between the founders and the individuals along the chromosome. Moreover, this extended model offers an alternative way of characterizing a population in terms of the association pattern between the founders and the modern individuals, which can be reflected in the estimated infinite hidden Markov model parameters. This alternative population representation can provide richer information about the genome than the traditional representations such as the allele frequency profiles.

Finally, this generalized inheritance model is applied to the problem of local ancestry estimation in an admixed population. When multiple ancestral populations have contributed to a modern admixed population over generations, the information about which allele in an modern admixed individual is inherited from which ancestral population can reveal essential clues in disease association studies. We associate each of the ancestral populations with an infinite hidden Markov model that captures the population-specific characteristics, and hierarchically link these infinite HMMs together to model an admixture event among these populations. This hierarchical model is able to utilize the genetic relatedness among the ancestral populations effectively, and hence the resulting model leads to a robust estimation of local ancestry in an admixed population, which significantly outperforms the existing methods that mostly ignore such relationship.

1.2 Summary of contributions

The main contribution of this thesis is two-fold. Statistically, it provides well-defined applications of the Dirichlet process and its extensions. Unlike typical applications of the Dirichlet process such as document modeling or image analysis, in which the accuracy of the application or the advantage of the non-parametric models is hard to measure directly, the applications we show allow direct evaluation of such models in terms of the quantitative accuracy measure and highlight the effectiveness of the flexible non-parametric Bayesian models. Biologically, it produces accurate and robust tools for various kinds of ancestral inference using genetic polymorphism data.

Specifically, the contributions of this thesis work can be detailed as follows.

- We efficiently exploit the shared structural information contained in the genetic data from multiple populations by using hierarchical statistical models that describe grouped data in an effective way. A hierarchical Dirichlet process or hierarchically linked infinite hidden Markov models applied to multi-population data utilize the population labels or shared genetic characteristics systematically and enhances the performance of the resulting methods substantially.
- The genetic inheritance models based on the Dirichlet process allow one to model the inherent uncertainty about the size of the genetic components in the data. The number of founder haplotypes that correspond to the mixture components in the DP mixture model can be inferred from the given data, which also offers valuable information about the complexity of the given population data.
- The proposed models are built on a unified inheritance framework on the assumption of hypothetical founders. This serves as a very flexible framework that can be generalized into various scenarios, for example, to model multiple population data or to incorporate admixture events to the original model designed for the homogeneous population. It makes it easy to further incorporate other important genetic processes such as natural selection or to consider more complex demographic scenarios.
- Important genetic parameters can be jointly inferred from the model in a single unified framework, for example, the mutation rate that reflects the relative age of the study population, the recombination rate, or population diversity and sub-structure. These parameters can play critical roles in elucidating the genetic history of study populations.
- These applications highlight the effectiveness of the non-parametric Bayesian models in real applications where the accuracy can be explicitly assessed. The developed models can serve as valuable resources that can extract important information from the genetic data essential for various kinds of downstream analyses.

The remainder of this thesis is organized as follows. We first introduce the basic terms and biological background in Chapter 2, and explain the theoretical background of non-parametric models based on the Dirichlet process in Chapter 3. Chapter 4 describes a haplotype inheritance framework modeled as a Dirichlet process mixture, which would be used as a building block for the models developed in this thesis. Then we include three major applications under this inheritance framework using non-parametric Bayesian models: the haplotype inference from multi-population data using a hierarchical Dirichlet process (Sohn and Xing, 2009; Xing et al., 2006)(Chapter 5), joint inference of population structure and recombination events by an infinite Hidden Markov model (Sohn and Xing, 2007a,b; Xing and Sohn, 2007) (Chapter 6), and the local ancestry estimation in admixed populations using hierarchically linked infinite hidden Markov models (Sohn et al., 2011) (Chapter 7). We summarize and conclude the thesis in Chapter 8.

Chapter 2

Background

The genetic diversities observed in DNA sequences of modern individuals come from many different sources: inheritance processes such as mutation and recombination, or population migration and the resulting admixture between different populations. By putting the main focus on the genetic data we analyze, in this chapter, we introduce the basic biological terms used in population genetics and explain the common genetic processes that affect the characteristics of the genetic data. This is explored in different perspectives depending on at which level the genetic diversity is created. We first explain the basic *inheritance* mechanism that passes genetic materials from the parental chromosomes to the chromosomes of offsprings within a population. We then consider more global scale of effect, *admixture*, that involves interaction between different populations. The well-known genealogical tree model called the coalescent is also briefly introduced.

2.1 Genetic inheritance process: mutation and recombination

Diploids like humans have two copies of each chromosome, one maternal copy and one paternal copy. When the two parental chromosomes join and create new offspring chromosomes during meiosis, the genetic information in the parental chromosomes is not identically copied to the offspring, and instead, certain genetic processes can change the chromosomal composition during the inheritance. The mutation and recombination processes are the most commonly considered genetic processes. A simple example about the effect of the mutation is that when a parental chromosome has a nucleotide ‘A’ at a certain locus on the chromosome, the genetic mutation can change the nucleotide to ‘C’ during meiosis, and as a result, the chromosome of its offspring has ‘C’ instead of ‘A’ at the locus. Therefore, this creates new alleles in individual chromosomes and thus adds a new genetic sequence to a population. The increased genotypic diversity in turn increases the phenotypic diversity as well, and it is generally believed that natural selection works by this genetic mutation as a major source. That is, among the various heritable traits generated by the genetic mutations, those traits that are advantageous 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 into a different strand. When a pair of parental chromosomes are copied and inherited to the offspring, parts of their genetic materials can be exchanged by the recombination and produce offspring chromosomes that can be decomposed into segments from both of the parents. When a recombination occurs between two loci, it tends to decouple the alleles carried at those loci in its descendants. Since the probability of recombination at different loci is different, this plays a key role in producing a block-like pattern on the chromosome called Linkage Disequilibrium (LD) 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 on the chromosome (Anderson and Novembre, 2003; Daly et al., 2001; Patil et al., 2001; Zhang et al., 2002), 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 (Fearnhead and Donnelly, 2001; Stephens and Scheet, 2005).

2.2 SNPs, genotypes and haplotypes

Genetic polymorphisms refer to the differences in DNA sequences between individuals or populations. One of the most important kinds of such genetic variations is a *single nucleotide polymorphism* (SNP), which is a single-nucleotide-based polymorphism. It 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 as these genetic variations underlie differences in our susceptibility to various types of heritable diseases.

Contiguous sequences of multiple SNPs on a chromosome are often looked at together and these are called *haplotypes*. The haplotypes have recently gained great popularity as an alternative basis for the association study and other applications because of the richer information they convey than just the set of independent single SNPs. In diploids, a pair of haplotypes, one from each of one's parents, 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 (Patil et al., 2001), or sometimes from analysis of trios (Hodge et al., 1999), the automatic and robust computational methods for inferring haplotypes from the inexpensive genotype data are still desired.

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 (Li, 2003; Stephens et al., 2001) program is one of the most widely used softwares with its notable accuracy. 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. This inheritance model has been successfully used in wide range of applications dealing with ancestral inheritance processes such as recombination analysis (Li, 2003), gene conversion rate estimation (Gay et al., 2007), and local ancestry estimation (Price et al., 2009). However, it does not scale up to the recent large scale datasets due to the high computational cost. More recent approaches such as fastPHASE (Scheet and

Stephens, 2006), MACH (Li et al., 2010), or BEAGLE (Browning and Browning, 2007) have improved the speed considerably, but at the expense of accuracy.

2.3 Admixture and genetic ancestry

Population migration is another important source of variation in genomic sequences. When populations that are genetically different meet through migration and the individuals mate to produce descendants over generations, the chromosomes in the *admixed* population contain the genetic materials from both of the ancestral populations. The investigation of the genetic ancestry in such an admixed population allows us to track the migration history of the populations and also provides important clues about the disease related genes especially when the ancestral populations have significantly different allele frequencies or disease susceptibility.

A number of statistical admixture models for genetic polymorphisms have been proposed for the analysis of population structure. In a *global* ancestry estimation as in Alexander et al. (2009); Falush et al. (2003); Patterson et al. (2006); Pritchard et al. (2000); Rosenberg et al. (2002), the information about the ancient populations is typically assumed to be unknown and the ancestry of a modern individual is represented as the average proportion of each contributing population across the genome. Therefore, this can be considered as an unsupervised problem. The admixture models identify each ancestral population mostly by focusing on the specific allele frequency profile for each ancestral population.

On the other hand, the *local* ancestry estimation problem is more concerned with a locus-by-locus ancestry given reference population data that are close to the real ancestral population data (Pasaniuc et al., 2009; Price et al., 2009; Sundquist et al., 2008; Tang et al., 2006). As mentioned earlier, genetic recombination tends to break the LD and generates block-structure on the chromosomes. Therefore, the chromosomes of the admixed individual can be partitioned into blocks of distinct ancestry. A common example is to decompose the chromosomes of modern African Americans into blocks with either African or European ancestry given the population data close to ancient African and European populations. The locus-specific ancestries are typically traced along the chromosome using statistical models such as hidden Markov models. These approaches for the local ancestry are either based on the allele frequency profiles as reference information (Pasaniuc et al., 2009; Tang et al., 2006), or utilize the haplotypes from the reference population data directly as in Price et al. (2009); Sundquist et al. (2008). Although significant progress has been made by these previous approaches, they share the limitation of ignoring the possible sub-structures among the ancestral populations. In addition, the restrictive modeling assumptions such as two-way admixture involving only two ancestral populations can also limit the general applicability of these models to the analysis of detailed ancestral structure in an admixed population under complex migration histories.

2.4 Coalescence

We include the brief description of the genealogical model called *coalescent* (Kingman, 1982) that has been widely studied in population genetics. It describes the theoretical inheritance model for a group of individuals in a population. The ancestral relationships among a sample of modern individuals can be described by a tree model known as the coalescent. By associating the modern individuals with the leaf nodes in the tree, it traces the parental individuals of the sample sequences backward in time until a single ancestral sequence is met, known as the most recent common ancestor (MRCA). Different assumptions regarding the genetic processes involved and the demographic scenarios under consideration can lead to different statistical properties in the coalescent theory. The simplest case can start from just assuming the mutation as a single genetic process. Consider two distinct sample sequences who differ at a single nucleotide by mutation. At each step backward in time, either these two samples find their distinct parents, or *coalesce* into a single parent, implying the occurrence of mutation at the corresponding time span and forming a tree. The common parent encountered by this later case corresponds to the MRCA of these sample individuals. Extensions for more complex processes such as recombination, selection, and population migration have also been studied and their mathematical properties have been investigated rigorously.

Despite its mathematical elegance, however, the marginalization over all the possible coalescent trees given sample sequences is widely known as intractable. Therefore, the full coalescence model is not easily applicable to the general ancestral inference problems. Alternatively, an approximation scheme such as *Product of Approximate Conditionals* (PAC) (Li, 2003) has been employed for different applications. However, the PAC model makes the implicit assumption that there exists an ordering of the given individual samples, and therefore the resulting likelihood is not exchangeable. Moreover, the latent demographic information such as founding chromosomes and their mutation rates are not directly captured in the PAC model as it involves no explicit ancestral genealogy over existing individual chromosomes.

Chapter 3

Dirichlet process and its extensions

A non-parametric Bayesian model called a Dirichlet process has gained great popularity in recent years especially for its usefulness in mixture scenarios. In this chapter, we introduce the non-parametric Bayesian models based on Dirichlet process, which include a hierarchical Dirichlet process and an infinite Hidden Markov Model.

3.1 Dirichlet process and its mixture models

The Dirichlet process describes a distribution over distributions and is formally defined as follows: a random probability measure Q on a measurable space $(\Phi, \mathcal{B})$ is generated by a Dirichlet process DP(γ, Q_0) if for every measurable partition $(B_1, \ldots, B_k)$ of the sample space Φ, the vector of random probabilities $Q(B_i)$ follows a finite dimensional Dirichlet distribution: $(Q(B_1), \ldots, Q(B_k)) \sim$ Dir$(\gamma Q_0(B_1), \ldots, \gamma Q_0(B_k))$ where $\gamma > 0$ denotes a *scaling parameter* and Q_0 denotes a *base measure* defined on $(\Phi, \mathcal{B})$ (Ferguson, 1973). Therefore, the draw Q from the Dirichlet process is itself a random measure and we write $Q \sim$ DP(γ, Q_0).

A useful representation of DP(γ, Q_0) is the stick-breaking construction by Sethuraman (1994). This representation is based on sequences of independent random samples $\{\pi'_i\}_{i=1}^{\infty}$ and $\{\phi_i\}_{i=1}^{\infty}$ generated in the following way:

$$\begin{aligned} \pi'_i &\sim \mathrm{Beta}(1, \gamma) \\ \phi_i &\sim Q_0 \end{aligned} \tag{3.1}$$

where $Beta(a, b)$ is the Beta distribution with parameters a and b. Analogous to a process of repetitively breaking a stick at fraction π'_l, the following sequence of π_i can be constructed from the sequence of π'_i:

$$\pi_i = \pi'_i \prod_{l=1}^{k-1} (1 - \pi'_l). \tag{3.2}$$

Sethuraman (1994) showed that the random measure Q arising from DP(γ, Q_0) admits the representation

$$Q = \sum_{i=1}^{\infty} \pi_i \delta_{\phi_i}. \tag{3.3}$$

The discrete atoms ϕ_i's can be thought of as the *locations* of samples in their space, and the π_i's are the *weights* of these samples. Note that $\sum_{i=1}^{\infty} \pi_i = 1$ with probability one. Therefore, we may think the sequence $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots)$ as a distribution on the positive integers. Following the notation in Teh et al. (2010), we write $\boldsymbol{\pi} \sim$ GEM(γ) if $\boldsymbol{\pi}$ is defined by Equations (3.1) and (3.2).

The discrete nature of the DP, as obviated from the stick-breaking construction, is well suited for the problem of placing priors on the parameters of the mixture model. This property can also be easily explained by another constructive definition of DP called Pólya urn scheme (Blackwell and MacQueen, 1973). Consider an urn that 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 with a probability proportional to γ, we are given a ball of a new color which we place in the urn. Such a scheme leads to a partition of the balls according to their colors. By mapping each ball to a sample and each color to its mixture component, this naturally defines the clustering of samples. Blackwell and MacQueen (1973) showed that this Pólya urn model yields samples whose distributions are those of the marginal probabilities under the Dirichlet process.

Suppose we have observed n samples with values $(\phi_1, \ldots, \phi_n)$ from DP(γ, Q_0). Considering this urn model, the conditional distribution of the value of the $(n+1)$th sample is given by :

$$\begin{aligned} \phi_{n+1}|\phi_1, \ldots, \phi_n, \tau, Q_0 &\sim \sum_{i=1}^{n} \frac{1}{n+\gamma} \delta_{\phi_i}(\cdot) + \frac{\gamma}{n+\gamma} Q_0(\cdot) \\ &= \sum_{k=1}^{K} \frac{n_k}{n+\gamma} \delta_{\phi_k^*}(\cdot) + \frac{\gamma}{n+\gamma} Q_0(\cdot), \end{aligned} \tag{3.4}$$

where K denotes the number of unique values in the n samples drawn so far, ϕ_k^* denotes the distinct values of ϕ_is, and n_k denotes the number of samples with value ϕ_k^*. This expression implies that each new sample has positive probability of being equal to an existing unique value in the drawn samples, and moreover, the probability is proportional to n_k. This creates a clustering effect on the samples and the *popular* components that have larger values of n_k tend to become more popular as more samples are considered.

In a DP mixture model, these samples ϕ_i from the Dirichlet process serve as the mixture components to which each observation x_i is assigned. This DP mixture model can be defined by using the following conditional probabilities:

$$\begin{aligned} Q \mid \gamma, Q_0 &\sim DP(\gamma, Q_0) \\ \phi_i \mid Q &\sim Q \\ x_i \mid \phi_i &\sim F(\phi_i) \end{aligned} \tag{3.5}$$

where x_i denotes the i-th observation, and ϕ_i is the mixture component associated with the observation x_i, and F denotes the likelihood function that generates the observation x_i given its mixture component.

Equivalently, we can incorporate an indicator variable $c_i \in \{1, 2, \ldots\}$ that selects the mixture component ϕ_i for each observation x_i such that $\phi_i = \phi_{c_i}^*$ for the distinct values ϕ_k^* of ϕ_is. Then the DP mixture model can also be expressed as follows:

$$\begin{aligned} \boldsymbol{\pi} \mid \gamma &\sim \text{GEM}(\gamma) \\ c_i \mid \boldsymbol{\pi} &\sim \boldsymbol{\pi} \\ \phi_k^* \mid Q_0 &\sim Q_0 \\ x_i \mid c_i, (\phi_k^*)_{k=1}^{\infty} &\sim F(\phi_{c_i}^*) \end{aligned} \tag{3.6}$$

Note that a DP mixture requires no prior specification of the number of components, which is typically unknown in general data clustering problems. This allows the mixture model setting of unknown cardinality and gives more flexibility to the model and the inference. It is important to emphasize that the Dirichlet process is used as a *prior distribution* of mixture components. Multiplying this prior by a likelihood that relates the mixture components to the actual data yields a *posterior distribution* of the mixture components, and the design of the likelihood function is completely up to the modeler based on specific problems. MCMC algorithms have been developed to sample from the posterior associated with DP priors (Escobar and West, 1995; Ishwaran and James, 2001; Neal, 2000). This nonparametric Bayesian formalism forms the technical foundation of the ancestral inference algorithms developed in this thesis.

3.2 Hierarchical Dirichlet process

A hierarchical Dirichlet process (HDP) (Teh et al., 2010) is a non-parametric Bayesian model that is very useful for describing data from multiple related groups, especially when each group has unique characteristics that can be captured by Dirichlet process, but multiple groups still need to be coupled together. For example, in document modeling, the distribution of words in a document is typically modeled as a mixture model in which the observation corresponds to the number of appearances of each word in the document and the mixture component corresponds to the *topic* that is assumed to generate the word. The DP mixture model described in the previous section allows to model this scenario without pre-specifying how many topics we should consider. Now, suppose we have a collection of such documents, each of which is modeled as a DP mixture model. While each document may have been written under a different theme, it is often more desirable to assume a common set of possible topics across the multiple documents, rather than to use a separate set of topics for each of the documents. More generally, given data that can be partitioned into a set of groups, we may want to cluster the data within each group, while still allowing the clusters to be shared across the groups.

A hierarchical Dirichlet process provides a model-based approach for clustering such grouped data. Suppose we have data from J groups, and each group j for $j = 1, \ldots, J$ is associated with a probability measure Q_j distributed as a Dirichlet process for generating mixture components in group j. Let the scale parameter τ and the base measure Q_0 shared by all the groups:

$$\mathcal{Q}_j \sim \text{DP}(\tau, Q_0)$$

Nonetheless, the use of a common base measure Q_0 does not necessarily ensure the mixture components to be shared across the multiple groups. If Q_0 is a continuous distribution, for instance, then the random draws from this distribution would be distinct with probability one, so different groups would have disjoint sets of mixture components with probability one. To allow the clusters to be shared across groups, an additional mechanism is necessary.

A hierarchical Dirichlet process handles this by assuming that the shared base measure Q_0 follows another Dirichlet process with a scale parameter γ and the base measure H:

$$Q_0 \mid \gamma, H \sim DP(\gamma, H)$$

Since the distribution Q_0 drawn from a Dirichlet process is discrete as seen in the stick-breaking construction in Equation (3.3), the individual values drawn from the distribution Q_0 can be repeated even if the base measure H is continuous. Therefore, this hierarchical model enables the atoms of random measures $\mathcal{Q}_j$ to be shared across groups and induces a very useful mixture model where multiple groups share mixture components while admitting each of those to have its own components.

The stick-breaking construction makes it clear how the atoms of Q_j under HDP are shared and how the weights of atoms are related to the global weight $\boldsymbol{\pi}$. Since Q_0 is distributed as DP(γ, H), it can be written as follows:

$$Q_0 = \sum_{k=1}^{\infty} \pi_k \delta_{\phi_k^*}$$

where $\phi_k^* \sim H$ and the sequence of π_k is constructed from the stick-breaking process in Equations (3.1) and (3.3). Since Q_j has the same support as its base measure Q_0, it also allows the following representation:

$$Q_j = \sum_{k=1}^{\infty} \pi_{jk} \delta_{\phi_k^*} \tag{3.7}$$

The weights π_{jk} have the following correspondence to the global weights as derived in Teh et al. (2010):

$$\begin{aligned} \pi'_{jk} &\sim \text{Beta}\big(\tau\pi_k, \tau(1 - \sum_{l=1}^{k} \pi_l)\big) \\ \pi_{jk} &= \pi'_{jk} \prod_{l=1}^{k-1} (1 - \pi'_{jl}) \end{aligned}$$

A modified Pólya urn scheme gives an intuitive explanation about how samples are generated under a hierarchical Dirichlet process prior. At the bottom level, we set up J urns which are used to define the DP mixture for data in each group j. Additionally, we also set up a single urn at the top level that contains balls of colors that are represented by at least one ball in the urns at the bottom level. To draw a sample for a group j, we either draw a ball randomly and put back two balls of the same color to the urn j, or we go to the top level urn with probability proportional to τ, instead of getting a ball of a new color immediately as in the plain Dirichlet process. At the top level urn, we can either draw a ball from the urn and put back two balls of the same color to the top level urn and also to the urn for group j, or, with probability proportional to γ, we now get a ball of a new color and put back a ball of this color to both the top-level urn and the urn

j. Essentially, the top-level urn defines the master DP that generates atoms for the bottom level DPs. While each urn has its own color distribution of the balls in it, the colors can be shared across groups, which demonstrates how the mixture components are shared across groups in a hierarchical Dirichlet process mixture model.

In summary, the following conditional probabilities define the HDP mixture model:

$$\begin{aligned} \mathcal{Q}_0 \mid \gamma, H &\sim DP(\gamma, H) \\ \mathcal{Q}_j \mid \tau, \mathcal{Q}_0 &\sim DP(\tau, \mathcal{Q}_0) \\ \phi_{ji} \mid \mathcal{Q}_j &\sim \mathcal{Q}_j \\ x_{ji} \mid \phi_{ji} &\sim F(\phi_{ji}) \end{aligned} \tag{3.8}$$

where x_{ji} denotes the i-th observation in group j, ϕ_{ji} is the mixture component associated with the observation x_{ji}, and F is the likelihood function that is specific to the mixture problem to be considered.

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 (Teh et al., 2010).

3.3 Infinite Hidden Markov model

A hidden Markov model (HMM) is a widely used statistical model for describing sequential data such as speech signals or DNA sequences that can be written as $(x_1, x_2, \ldots .x_T)$. Under a hidden Markov model, the observation sequence x_t depends on its hidden state q_t such that given the state q_t, the observation x_t is independent of other observations x_t' and states q_t' for $t' \neq t$. Moreover, q_t is assumed to have Markov property which means q_t is conditionally independent of $\{q_{t-2}, ..., q_2, q_1\}$ given q_{t-1}, that is, $p(q_t \mid q_{t-1}, q_{t-2}, \ldots, q_1) = p(q_t \mid q_{t-1})$. Therefore, the HMM can be defined by the following three components:

- the initial probabilities $\pi_{i0} = P(q_0 = i)$ for generating the initial hidden state q_0
- the transition probabilities $\pi_{ij} = P(q_t = j \mid q_{t-1} = i)$ that define the probability of each transition from hidden state i to state j
- the emission probabilities $b_i(x_t) = P(x_t \mid q_t = i)$ for a hidden state to emit each of the observation variables .

A traditional HMM assumes K possible hidden states and thus $q_t \in \{1, \ldots, K\}$. Then the transition probabilities are represented as a K by K matrix where each row of the matrix sums to one. The initial probabilities are written as a K-dimensional vector which also sums to one. In many practical applications, however, it is not straightforward to determine the number of hidden states and we may often want to infer the number as well as the hidden state sequence or other HMM parameters.

A non-parametric extension of the traditional HMM to an infinite state space was first introduced in Beal et al. (2002) and formally defined later in a context of a hierarchical Dirichlet process in Teh et al. (2010). Since each row i of the transition matrix defines the probability of transition from the source state i to all the states, the transition probabilities in an infinite Hidden Markov model are represented by an infinite matrix in which both the columns and rows are infinite dimensional. Formally, the followings summarize the infinite Hidden Markov Model:

$$\begin{aligned} \beta \mid \gamma &\sim \text{GEM}(\gamma) \\ \boldsymbol{\pi_i} \mid \tau, \beta &\sim \text{DP}(\tau, \beta) \\ \phi_i \mid H &\sim H \\ q_t \mid q_{t-1}, (\boldsymbol{\pi_i})_{i=1}^{\infty} &\sim \boldsymbol{\pi_{q_{t-1}}} \\ x_t \mid q_t, (\phi_i)_{i=1}^{\infty} &\sim F(\phi_{q_t}) \end{aligned}$$

where F defines the emission probability. Here, the DP representation using the indicator variables as in Equation (3.6) has been adopted because the hidden state variable q_t actually corresponds to the indicator variable to select the atom from the Dirichlet process. We can see that each row of the infinite-dimensional transition matrix is described by $\boldsymbol{\pi}$ and these are coupled by the common base measure β under the Dirichlet process. Since a draw from a DP is a discrete measure with probability 1, atoms drawn from this measure—atoms which are used as targets for each of the (unbounded number of) source states—are not generally distinct. Indeed, the transition probabilities from each of the source states have the same support.

To construct such a stochastic matrix of infinite dimensionality, we can exploit the fact that in practice only a finite number of states will be visited by each source state, and we only need to keep track of those states. The following sampling scheme based on a hierarchical Pólya urn model captures this spirit and shows how to generate a transition matrix in an infinite HMM. As in the urn model for a hierarchical Dirichlet process, a two-level hierarchy of the urn model is considered. The "stock" urn at the top level contains balls of colors that are represented by at least one ball in the urns at the bottom level. At the bottom level, we have a set of urns which are used to define the initial and the transition probabilities from each source state. Recall that in a mixture model scenario, the color of the ball represents the mixture component that the ball (or the observation) is associated with. In an infinite HMM, the color corresponds to the hidden state the observation is generated from, and each urn at the bottom level defines the probabilities of state-transition from each source state observed so far. Therefore, each bottom-level urn is used to describe the Dirichlet process mixture for each row of the transition matrix. Specifically, the transition probability from a source state i to a target j at the current step is proportional to the number of times the same transition occurs so far, which is equal to the number of the balls of the color j in the urn i. But with the probability proportional to the scale parameter τ, we refer to the top-level urn to select the target state. At this top level, the transition probability to the source state j is either proportional to the number of previous visits to j by this top level urn that corresponds to the number of balls of color j at the stock urn. Or with probability proportional to γ, a ball of a new color is created, which means a new state has been initiated. In this case, we set up a new urn to define the DP mixture at the newly initiated state. As pointed out in Teh et al. (2010), this model can be viewed as an instance of the hierarchical Dirichlet process mixture model, with row-specific DP mixtures that are coupled by the top level DP.

The inference under an infinite hidden Markov model becomes more tricky because the traditional method for the standard HMM such as the forward-backward algorithm or Viterbi decoding is not directly applicable due to the dimensionality. We can apply a traditional MCMC sampling, although this involves book-keeping about the number of previous transitions between each pair of states. In Van Gael et al. (2008), a more efficient inference algorithm called the Beam sampling algorithm has also been introduced. This extends the traditional forward-backward algorithm to an infinite state space by combining a slice sampling and dynamic programming scheme, which is shown to be more robust and to outperform the traditional Gibbs sampling. In the following chapters, we show the application of these non-parametric Bayesian models using both the traditional MCMC sampling schemes and the beam sampling algorithm.

Chapter 4

Haplotype inheritance model based on Dirichlet process

Before describing the specific applications considered in this thesis, we first describe the general haplotype inheritance model adopted in this thesis. The distribution of haplotypes in a population can be formulated as a mixture model, where the set of mixture components corresponds to the pool of ancestral haplotypes, or *founders*, of the population (Excoffier and Slatkin, 1995; Kimmel and Shamir, 2004; Qin et al., 2002). However, the size of this pool is unknown. Indeed, knowing the size of the pool would correspond to knowing something significant about the genome and its history. On the other hand, while pure coalescence-based models can provide elegant mathematical properties for the genetic patterns in the populations, it is hard to perform statistical inference of ancestral features and many other interesting genetic variables because for a large population, the number of hidden variables in a coalescence tree is prohibitively large. (Stephens et al., 2001). In most practical population genetic problems, usually the detailed genealogical structure of a population as provided by the coalescent trees is of less importance than the population-level features such as the pattern of major common ancestor alleles or founders in a population bottleneck, or the age of such alleles. In this case, the Dirichlet process mixture offers a principled approach to generalize the finite mixture model for haplotypes to an infinite mixture that models uncertainty regarding the size of the ancestor haplotype pool. At the same time, it provides a reasonable approximation to the coalescence model by utilizing the partition structure resulting from it but still allowing further mutations within each partite to introduce further diversity among descents of the same founder.

The Dirichlet process mixture model for describing haplotypes was first proposed in Xing et al. (2004) although with no consideration about the recombination process. As this model will be used as a basic building block for the applications developed in this thesis, we include the description of each component of the statistical model in this chapter. In more recent work in Sohn and Xing (2009), we notice that there is an interesting connection of the DPM-based methods to the Wright-Fisher model and Kingman's coalescent with an infinitely-many-alleles (IMA) mutation process for allele evolution. 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 $(n-1)/(n-1+\tau)$ (Hoppe, 1984). This is identical to the Pólya urn scheme described in Section 3.1 with a scaling parameter τ and a uniform base distribution. We include brief discussion about this connection as well.

4.1 DP mixture model for haplotype modeling

The model starts from the assumption that a haplotype population H is originated from an unknown number of founder 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. The 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. As a brief recap of the Dirichlet proces, the distinct atoms ϕ_k^* from a Dirichlet process in Equation (3.3) act as the mixture components in a Dirichlet process mixture model. Under the haplotype inheritance model as a DP mixture, each unique value ϕ_k^* from a DP is associated with a possible founder and its mutation probability, i.e., $\{a_k, \theta_k\}$. Specifically, let $h_i = [h_{i1}, \ldots, h_{iT}]$ denote the haplotype of individual i over T contiguous SNPs. Let $a_k = [a_{k1}, \ldots, a_{kT}]$ denote an ancestor haplotype and θ_k denote the *mutation rate* of ancestor k; and let c_i denote the indicator variable that specifies the ancestor of haplotype h_i. $P_h(h|a,\theta)$ represents the *inheritance model* according to which individual haplotypes are derived from a founder. Let γ be the scale parameter of the Dirichlet process, and F be the base measure that generates the founder haplotype a_k and its mutation rate θ_k jointly. The complete DP mixture model for haplotype inheritance can be summarized as follows:

$$\begin{aligned} \boldsymbol{\pi} \mid \gamma &\sim \mathrm{GEM}(\gamma) \\ c_i \mid \boldsymbol{\pi} &\sim \boldsymbol{\pi} \\ (a_k, \theta_k) \mid F &\sim F \\ h_i \mid c_i, (a_k, \theta_k)_{k=1}^{\infty} &\sim P_h(\cdot \mid a_{c_i}, \theta_{c_i}) \end{aligned}$$

We let $F(A, \theta) = p(A)p(\theta)$, where $p(A)$ is uniform over all possible haplotypes and $p(\theta)$ is a beta distribution introducing a prior belief of a low mutation rate.

The haplotype inheritance model P_h is defined as a *single-locus mutation model* where $P_h(h \mid a, \theta)$ is decomposed into the product of the likelihood at each locus represented as:

$$P_h(h_{it}|c_i, (a_{kt}, \theta_k)_{k=1}^{\infty}) = (1-\theta)^{\mathbb{I}(h_{it}=a_{c_i t})} \left(\frac{\theta}{|\mathcal{A}|-1}\right)^{\mathbb{I}(h_{it}\neq a_{c_i t})} \tag{4.1}$$

where $\mathbb{I}(\cdot)$ is the indicator function and $|\mathcal{A}|$ is the size of the allele space. It defines the model to generate an individual haplotype h from a founder a with a mutation rate θ. 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 such as in the BLADE model for mapping (Liu et al., 2001), and numerous models for haplotype inference (Zhang et al., 2006).

To allow the inference of haplotypes given genotypes under this inheritance model, Xing et al. (2007) has adopted the following additional components to the basic model above. Since diploids like human have two copies of each haplotype, one can write the individual haplotypes using the notation h_{i_e} for $e \in \{0, 1\}$, where e denotes the index to indicate either the maternal or the paternal copy of individual i. Then the genotype at a locus is determined by the paternal and maternal alleles of this site with some random noise via the following genotyping model:

$$P_g(g_{it}|h_{i_0t}, h_{i_1t}; \xi) = \xi^{\mathbb{I}(h_{it}=g_{it})}[\mu_1(1-\xi)]^{\mathbb{I}(h_{it}\neq^1 g_{it})}[\mu_2(1-\xi)]^{\mathbb{I}(h_{it}\neq^2 g_{it})} \tag{4.2}$$

where $h_{it} \triangleq h_{i_0t} \oplus h_{i_1t}$ 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. A beta prior $Beta(\alpha_g, \beta_g)$ is placed on ξ for smoothing.

The complete process that generates individual haplotypes and genotypes from the founder haplotypes under the DP mixture model are summarized as the following generative scheme.

- Draw first haplotype:

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

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

- for subsequent haplotypes:

– sample the founder indicator for the ith haplotype:

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

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

– sample the founder of haplotype i:

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

– 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:

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

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 $\mathbf{h} = \{h_i : i \in \{1, 2, ..., I\}\}$ takes the following form resulted from an integration of θ in the joint conditional:

$$p(\mathbf{h}|\mathbf{a}, \mathbf{c}) = \prod_{k=1}^{K} R(\alpha_h, \beta_h) \frac{\Gamma(\alpha_h + l_k)\Gamma(\beta_h + l'_k)}{\Gamma(\alpha_h + \beta_h + l_k + l'_k)} \left(\frac{1}{|\mathcal{A}| - 1}\right)^{l'_k}, \tag{4.3}$$

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

The only observed variable in this problem is the individual genotypes g_i and all the other variables of h_i, c_i, a_k will be inferred from the posterior inference. 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.

By using a Dirichlet process prior we essentially maintain a pool of haplotype founders that grows as observed individual haplotypes are processed. But notice that the above generative model assumes each modern haplotype originates from a single ancestor where no recombination is involved, which is only true for haplotypes spanning a short region on a chromosomal.

4.2 Population genetic implication of DP haplotype model

The Dirichlet-process-based models relate to the fundamental stochastic models from population biology in a very interesting way, somewhat justifying their application to haplotype modeling from a statistical genetic point of view. Given a sample of n chromosomes, under neutrality and random-mating assumptions as in Wright (1931), Fisher (1930),and Kingman (1982), the distribution of the *genealogy trees* of the sample can be approximated by that of a random tree known as the n-coalescent (Kingman, 1982). Additionally, on each lineage there can be a point process of mutation events. In an *infinitely-many-alleles* (IMA) model, each mutation in the lineage produces a novel mutant that is independent of the parental allele; thus IMA can be understood as an independent Poisson process with rate, say, $\tau/2$ (note the intentional use of the same symbol as the scaling parameter in the Dirichlet process, which implies a close relationship between IMA on n-coalescent with DP, which we shall reveal shortly), which is determined by the size of the evolving population N (usually $N >> n$) and the per-generation mutation rate μ (i.e., $\tau = 4N\mu$) (Kingman, 1982). The IMA model refers to such a situation where each mutation produces a novel haplotype a. (Without loss of generality, here we assume that the haplotype-generating mutation does not have to be a point mutation that changes one SNP locus only, but can be a “macroscopic” event that produces an entirely new T-loci haplotype.) Hoppe (1984) observed that the IMA model with rate $\tau/2$ on an n-coalescent extends haplotype lineages on the tree according to the following law: with probability $\tau/(n - 1 + \tau)$ it instantiates a new haplotype, and with probability $(n - 1)/(n - 1 + \tau)$ it replicates an existing haplotype lineage.

This is exactly the Pólya urn scheme described in Eq (3.4) with scaling parameters τ and uniform base distribution over $\mathcal{A}$, a Dirichlet process $DP(\tau, \mathrm{Uniform})$.

There is a mapping between the distinct founders $\phi_k^* \equiv \{a_k, \theta_k\}, \forall k$ arising from a DP, to the novel haplotypes generated according to IMA on a coalescent tree at the birth of every new lineage. Samples from the DP that share a common haplotype corresponds to the descendant (i.e., non-mutating) lineages rooted from the founder; but the genealogical relationships between distinct haplotypes are not preserved under an IMA model (once a new haplotype is instantiated from a mutation, it "forgets" its "progenitor" because the mutation is independent of the parental haplotype). Thus a basic DP cannot capture relationships between different haplotypes in a population.

The *parental-dependent-mutation* model posits that, in a sequential generation process of haplotypes, if the next haplotype does not match exactly with an existing haplotype, it will tend to differ by a small number of mutations from an existing one, rather than be completely different. Under a DP mixture, modern individual haplotypes h_i are marginally dependent, because similar but nonidentical haplotypes can be grouped around possible founders according to an inheritance model $P_h(H|A, \theta)$ that permits further changes on top on founders. As discussed later, this leads to an exchangeable $P(H)$ that captures the effect of parent-dependent mutations.

Chapter 5

Haplotype inference from multi-population data

5.1 Introduction

We now consider the specific applications of the non-parametric Bayesian models described in Chapters 3 and 4. SNPs represent the largest class of individual differences in DNA sequences. Recall that a SNP refers to the existence of two possible nucleotide bases from $\{A, C, G, T\}$ at a chromosomal locus in a population. Each variant, denoted as 0 or 1, is called an *allele*. A haplotype refers to the joint allelic identities of a contiguous list of polymorphic loci within a study region on a given chromosome. As introduced in Chapter 2, diploid organisms such as human beings have two haplotypes in each individual, one from each of the parents. When the parental chromosomes come in pairs, two haplotypes go together and make up a genotype which consists of the list of allele-pairs at every locus. A genotype is resulted from a pair of haplotypes by omitting the phase information regarding the specific association of each allele with one of the two chromosomes at every locus. Common biological methods for assaying genotypes typically do not provide phase information for individuals with heterozygous genotypes at multiple loci. The problem of haplotype inference concerns determining which phase reconstruction among many alternatives is more plausible.

Key to the inference of individual haplotypes based on a given genotype sample is the formulation and tractability of the marginal distribution of the haplotypes of the study population. Consider the set of haplotypes, denoted as $H = \{h_1, h_2, \ldots, h_{2n}\}$, of a random sample of $2n$ chromosomes of n individuals. Under common genetic arguments, the ancestral relationships among the sample back to its most recent common ancestor (MRCA) can be described by a genealogical tree known as the coalescent. Computing $P(H)$ involves a marginalization over all possible coalescent trees compatible with the sample, which is widely known to be intractable. In Li (2003), it was suggested to approximate $P(H)$ by a *Product of Approximate Conditionals* (PAC). The PAC model tries to incorporate a desirable evolution assumption known as the *parental-dependent-mutation* (PDM) by modeling each h_i as the progeny of a randomly-chosen existing haplotype, and it forms the basis of the PHASE program, which has set the state-of-the-art benchmark in haplotype inference. However, the PAC model implicitly assumes existence of an ordering in the haplotype sample, therefore the resulting likelihood is not *exchangeable* as one would expect for the true $P(H)$. Moreover, since PAC involves no explicit ancestral genealogy over existing haplotypes, certain latent demographic information such as founding haplotypes and their mutation rates are not directly captured in the model.

The finite mixture models represent another class of haplotype models that rely very little on demographic and genetic assumptions of the sample (Excoffier and Slatkin, 1995; Kimmel and Shamir, 2004; Qin et al., 2002; Zhang et al., 2006). Under such a model, haplotypes are treated as latent variables associated with specific frequencies, and the haplotype inference problem can be viewed as a *missing value inference* and *parameter estimation* problem, for which numerous statistical inference approaches have been developed, such as the maximum likelihood approaches via the EM algorithm (Excoffier and Slatkin, 1995; Fallin and Schork, 2000; Hawley and Kidd, 1995; Long and Williams, 1995), and a number of parametric Bayesian inference methods based on Markov Chain Monte Carlo (MCMC) sampling (Qin et al., 2002; Zhang et al., 2006). However, this class of methods has rather severe computational requirements in that a probability distribution must be maintained on a large set of possible haplotypes. Indeed, the size of the haplotype pool, K, which reflects the diversity of the genome, is unknown for any given population data and needs to be inferred. There is a plethora of combinatorial algorithms based on various hypothesis such as the "parsimony" principles that offer control over the complexity of the inference problem (see Gusfield (2004) for an excellent survey).

Recently, substantial efforts have also been made to speed up haplotype inference on large scale data. Notable programs include Beagle (Browning and Browning, 2007) which uses a localized haplotype model based on variable-length Markov chains, and MACH (Li and Abecasis, 2006), which significantly improves PHASE in terms of computation time.

It is noteworthy that current progresses on approximating $P(H)$, K, and on scalability to long SNP sequences, are made while ignoring potentially useful information on population structures in a genetic sample. In particular, statistical models developed so far are inadequate for addressing the multi-population haplotype sharing problems concerned in this chapter. Consider for example a genetic demography study, in which one seeks to uncover ethnic- or geographic-specific genetic patterns based on a sparse census of multiple populations. In particular, suppose that we are given a sample that can be divided into a set of subpopulations; e.g., African, Asian and European. We may not only want to discover the sets of haplotypes within each subpopulation, but also which haplotypes are shared between subpopulations, and what their frequencies are. Empirical and theoretical evidence suggests that an early split of an ancestral population following a populational bottleneck may lead to ethnic-group-specific population diversity, which features both ancient haplotypes shared among different ethnic groups, and modern haplotypes uniquely present in different ethnic groups (Pritchard, 2001). This structure is analogous to a co-clustering in which different groups comprising multiple clusters may share clusters with common centroids, and its implication on haplotype reconstruction has not been thoroughly investigated.

We have developed a new haplotype model for multi-population data based on a *hierarchical Dirichlet process* (HDP) (Teh et al., 2010; Xing et al., 2006). Recall that a hierarchical Dirichlet process over a measurable space $(\Phi, \mathcal{B})$ specifies a set of coupled random distributions $\{\mathcal{Q}_1, \mathcal{Q}_2, \ldots, \mathcal{Q}_J\}$ on Φ for data from J groups. For modeling haplotypes in multiple populations, we let $\Phi \equiv \mathcal{A} \times \mathcal{E}$ where $\mathcal{E} = [0, 1]$ and $\mathcal{A} = \{0, 1\}^T$ denote the space of the mutation rates and joint allele configurations, respectively, of the *ancestral haplotypes* of T SNP loci. Each $\mathcal{Q}_j$ is a population-specific Dirichlet process (DP) (Blackwell and MacQueen, 1973; Ferguson, 1973) which defines a nonparametric prior over the ancestral haplotypes and their frequencies of being inherited within the population, and thereby induces a Dirichlet process mixture (DPM) model for all the individual haplotypes in that population. To allow every ancestral haplotype in a particular population to also have non-zero probability of being inherited in a different population (albeit with different frequencies), a hyper-prior $\mathcal{Q}_0$, which is also a Dirichlet process and therefore discrete on Φ, is used to define the base measures of each population-specific $\mathcal{Q}_j$, ensuring that they are all realized on a common set of supports (i.e.,founders) in Φ. Our model differs from other methods reviewed earlier in the following ways: 1) Instead of resorting to empirical assumptions or model selection over the number of population haplotypes, we introduce a nonparametric prior over haplotype ancestors, which facilitates posterior inference of the haplotypes in an "open" state space accommodating arbitrary sample size. 2) Our model explicitly exploits the population labels of individuals and potentially latent sub-population structures to improve haplotyping accuracy. 3) Our model captures similar genetic properties as those emphasized in Stephens et al. (2001), including the parent-dependent-mutations, but with an exchangeable likelihood function.

We have developed an efficient MCMC-based software program *Haploi*, based on our proposed model, and using a variant of the Partition-Ligation scheme by Niu et al. (2002) to handle complexity explosion due to long input sequences. It can be readily applicable to multi-population genotype sequences, at a time-cost often at least two-orders of magnitude less than that of the state-of-the-art PHASE program, with competitive performance. We also show that *Haploi* can significantly outperform other popular haplotype inference algorithms on both simulated and real short SNPs data.

5.2 The Statistical Model

Our proposed model for multi-population haplotype inference is based on a basic Dirichlet process mixture model described in Chapter 3 developed for a simple demographic scenario where individual ethnic labels are ignored and no recombination is assumed in the sample. In the model, the DP is used as the prior over the components in an unbounded ancestral space. This prior requires no specification of the size of the ancestor pool.

In this section, we describe the hierarchical Dirichlet process mixture for haplotypes from multiple population in detail.

5.2.1 Hierarchical DP mixture for multi-population haplotypes

We consider the case where there exist multiple ethnic or geographic populations. Instead of modeling these populations independently by unrelated Dirichlet process mixtures, we place all the population-specific Dirichlet process mixtures under a common prior such that the ancestors in any of the population-specific mixtures can be shared across all the mixtures, but the *weight* of an ancestral haplotype in each mixture is unique.

To tie population-specific DP mixtures together in this way, we employ a hierarchical DP (HDP) mixture model (Teh et al., 2010) described in Section 3.2, in which the base measures of the all population-specific DPMs admit a common discrete prior defined by another Dirichlet process $\mathrm{DP}(\gamma, F)$. An HDP defines a distribution over a set of dependent random probability measures, $\{\mathcal{Q}_j \;\; j = 1, \ldots, J\}$, and another master random probability measure $\mathcal{Q}_0$ that controls all the $\mathcal{Q}_j$'s. Each $\mathcal{Q}_j$ is a population specific DP with common (or population-specific) scaling parameter τ, and a shared base measure defined by $\mathcal{Q}_0$. Moreover, $\mathcal{Q}_0$ itself follows a Dirichlet process DP(γ, F). Following a hierarchical Pólya urn scheme, for m_j random draws $\phi_j = \phi_{j,1}, \ldots, \phi_{j,m_j}$ from $\mathcal{Q}_j$ described in Section 3.2, we can derive the following conditional probability for $(\phi_{m_j}|\phi_{-m_j})$ (Xing et al., 2006), where the subscript $-m_j$ denotes the index set of all but the m_j-th sample:

$$\begin{aligned}\phi_{m_j}|\boldsymbol{\phi}_{-m_j} &\sim \sum_{k=1}^{K} \frac{m_{j,k} + \tau \frac{n_k}{n-1+\gamma}}{m_j - 1 + \tau} \delta_{\phi_k^*}(\phi_{m_j}) + \frac{\tau}{m_j - 1 + \tau} \frac{\gamma}{n - 1 + \gamma} F(\phi_{m_j}) \\ &= \sum_{k=1}^{K} \pi'_{j,k} \delta_{\phi_k^*}(\phi_{m_j}) + \pi'_{j,K+1} F(\phi_{m_j}) \qquad (5.1)\end{aligned}$$

where n_k denotes the number of samples under $\mathcal{Q}_0$ drawn from the global measure F and equal to ϕ_k^*; $m_{j,k}$ denotes the number of samples in the j-th group which are equal to ϕ_k^*; and

$$\pi'_{j,k} := \frac{m_{j,k} + \tau \frac{n_k}{n-1+\gamma}}{m_j - 1 + \tau}$$

$$\pi'_{j,K+1} = \frac{\tau}{m_j - 1 + \tau} \frac{\gamma}{n - 1 + \gamma}$$

The vector $\vec{\pi}'_j = (\pi'_{j,1}, \pi'_{j,2}, \ldots)$ gives the *a priori* conditional probability of a new sample in group j. As shown later, this formula in Equation (5.1) will be useful for implementing a Gibbs sampler for posterior inference under HDP mixtures.

Based on the HDP described above, we now define an HDP mixture (HDPM) model for the genotypes in J populations. Elaborating on the notational scheme used earlier in Section 4.1, let $G_i^{(j)} = [G_{i1}^{(j)}, \ldots, G_{iT}^{(j)}]$ denote the *genotype* of T contiguous SNPs of individual i from ethnic group j; and let $H_{i_e}^{(j)} = [H_{i_e 1}^{(j)}, \ldots, H_{i_e T}^{(j)}]$ denote a haplotype of individual i from ethnic group j. The atoms $\phi_{i_e}^{(j)}$ under a hierarchical Dirichlet process correspond to the founder haplotype and its mutation rate associated with the individual haplotype h_{i_e} in population j. The basic generative structure of multi-population genotypes under an HDPM is defined as follows.

$$\begin{aligned} Q_0 \mid \gamma, F &\sim \mathrm{DP}(\gamma, F) \\ Q_j \mid \tau, Q_0 &\sim \mathrm{DP}(\tau, Q_0) \\ \phi_{i_e}^{(j)} | \mathcal{Q}_j &\sim Q_j \\ h_{i_e}^{(j)} | \phi_{i_e}^{(j)} &\sim P_h(\cdot | \phi_{i_e}^{(j)}) \\ g_i^{(j)} | h_{i_0}^{(j)}, h_{i_1}^{(j)} &\sim P_g(\cdot | h_{i_0}^{(j)}, h_{i_1}^{(j)}) \end{aligned}$$

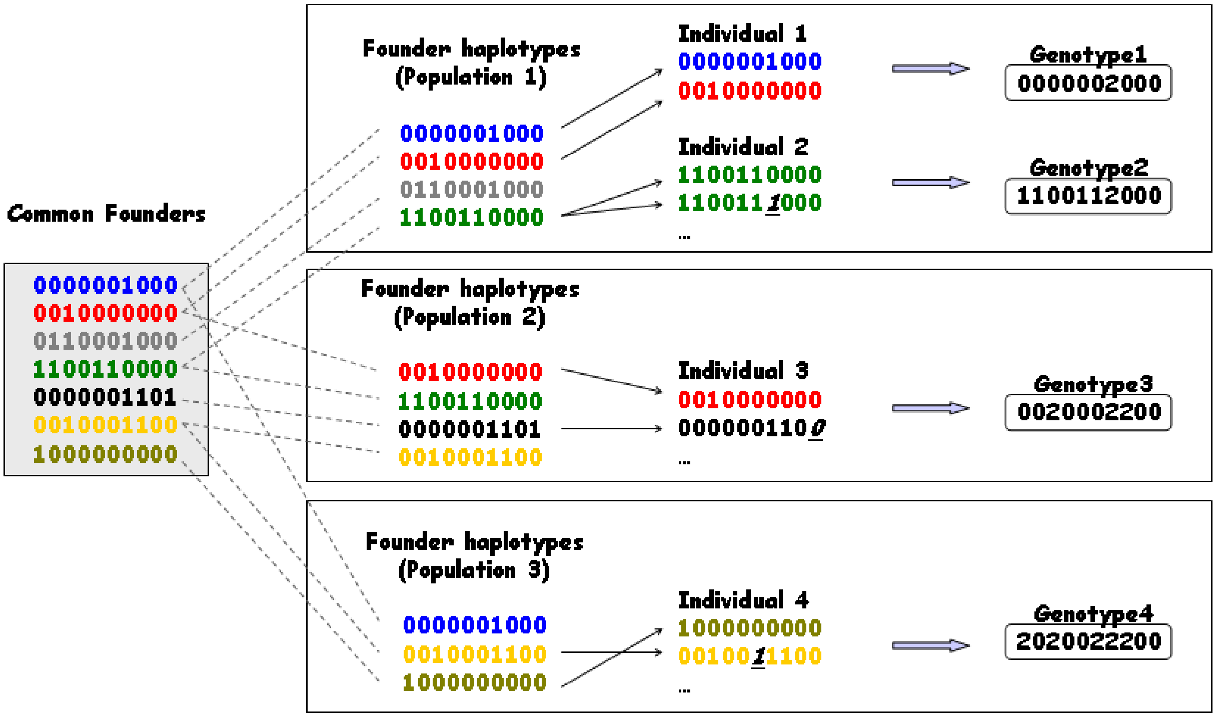


Figure 5.1: The haplotype-genotype generative process under HDPM, illustrated by an example concerning three populations. At the first level, all haplotype founders from different populations are drawn from a common pool via a Pólya urn scheme, which leads to the following effects: 1. The same founder can be drawn by either multiple populations (e.g., the red founder in population 1 and 2, and the blue one in population 1 and 3), or only a single population (e.g., the grey founder in population 1); 2. Shared founders can have different frequencies of being inherited. Then at the second level, individual haplotypes were drawn from a population-specific founder pool also via a Pólya urn scheme, but this time through an inheritance models P_h that allows mutations with respect to the founders, as indicated by the underscores at the mutated loci in the individual haplotypes. Finally, genotypes are related to the haplotype pairs of every individual via a noisy channel P_g.

Here, the first three steps follow the general HDP scheme to generate the atoms of founder haplotypes and their mutation rates. The fourth step describes the mixture formalism in which the individual haplotype is generated given its founder from the haplotype inheritance model P_h. The last step corresponds to the noisy genotyping model to generate the genotype given a pair of haplotypes in each individaul. Recall that in an HDP, the base measure $\mathcal{Q}_0$ is a random distribution of the pool of haplotype founders and their associated mutation rates. It ensures that all the population-specific child DPs can be defined on a common unbounded pool of candidate founder patterns. The child DPs place different mass distributions, i.e., *a priori* frequencies of haplotype founders, on this common support, in a population-specific fashion. This generative procedure is also illustrated graphically in Figure 5.1:

The base measure F in the above generative process is defined as a distribution from which haplotype founders $\phi_k \equiv \{A_k, \theta_k\}$ are drawn. Thus it is a joint measure on both A and θ. As defined in Section 4.1, we let $F(A, \theta) = p(A)p(\theta)$, where $p(A)$ is uniform over all possible haplotypes and $p(\theta)$ is a beta distribution introducing a prior belief of low mutation rate. For other building blocks of the haplotype inheritance model P_h and the noisy genotype observation model P_g, we adopt the model described in Equations (4.1) and (4.2).

5.2.2 Hyperprior for scaling parameters

To capture uncertainty over the scaling parameters of Dirichlet process, e.g., γ, we use a vague inverse Gamma prior:

$$p(\gamma^{-1}) \sim \mathcal{G}(1,1) \Rightarrow p(\gamma) \propto \gamma^{-2}\exp(-1/\gamma)). \tag{5.2}$$

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 γ 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:

$$p(\gamma|k,n) \quad \propto \quad \frac{\gamma^{k-2}\exp(1/\gamma)\Gamma(\gamma)}{\Gamma(n+\gamma)}. \tag{5.3}$$

The distribution $p(\log(\gamma)|k,n)$ is log-concave, so we may efficiently generate independent samples from this distribution using adaptive rejection sampling (Rasmussen, 2000). It is noteworthy that in an HDPM we need to define vague inverse Gamma priors also for the scaling parameters τ of population-specific DPs at the bottom level. We use a single concentration parameter τ for these DPs; it is also possible to allow separate concentration parameters for each of the lower-level DPs, possibly tied distributionally via a common hyperparameter.

5.2.3 Posterior inference via Gibbs sampling

Based on the two-level Pólya urn implementation of the HDP mixture model described in Section 3.2, an efficient MCMC algorithm can be derived to sample from the posterior associated with the HDP mixture model. Recall that the mixture components correspond to the ancestral haplotypes a_k with their mutation rates θ_k, and the samples correspond to individual haplotypes h. Therefore, after integrating out θ_k according to Equation (4.3), the variables of interest are a_{kt}, $h_{i_e t}^{(j)}$, $c_{i_e t}^{(j)}$, γ and τ , and $g_{it}^{(j)}$ (the only observed variables). We may assume that the represented mixture components are indexed by $1, ..., K$, the weights of the founders at the top level DP is

$$\hat{\beta} = \left(\frac{n_1}{n-1+\gamma}, ..., \frac{n_K}{n-1+\gamma}, \frac{\gamma}{n-1+\gamma}\right)$$

where $\frac{\gamma}{n-1+\gamma}$ is the total weight corresponding to some unrepresented founder $K+1$; and the weights of founders at the bottom-level DP for, say, the jth population, are

$$\hat{\pi}^j = \left(\frac{m_{j,1}}{m_j-1+\tau}, ..., \frac{m_{j,K}}{m_j-1+\tau}, \frac{\tau}{m_j-1+\tau}\right)$$

where $\frac{\tau}{m_j-1+\tau}$ corresponds to the probability of consulting the top-level DP. The Gibbs sampler alternates between three coupled stages. First, we sample the scaling parameters γ and τ of the DPs according to Equation (5.3).

Then, we sample the $c_{i_e}^{(j)}$ and a_{kt} given the current values of the hidden haplotypes and the scaling parameters. Before sampling $c_{i_e}^{(j)}$, we first erase its contribution to the sufficient statistics of the model. If the old $c_{i_e}^{(j)}$ was k', set $m_{jk'} = m_{jk'} - 1$. If it was sampled from the top level DP, we also set $n_{k'} = n_{k'} - 1$. Note that $c_{i_e}^{(j)} \leq K + 1$ (i.e., indicating existing founders, plus a new one to be instantiated). Now we can sample $c_{i_e}^{(j)}$ from the following conditional distribution:

$$
\begin{aligned}
p(c_{i_e}^{(j)} = k|\mathbf{c}^{[-j,i_e]}, \mathbf{h}, \mathbf{a}) \quad &\propto \quad p(c_{i_e}^{(j)} = k|\mathbf{c}^{[-j,i_e]}, \mathbf{m}, \mathbf{n})p(h_{i_e}^{(j)}|a_k, \mathbf{c}, \mathbf{h}^{[-j,i_e]}) \\
&\propto \quad (m_{jk}^{[-j,i_e]} + \tau\beta_k)p(h_{i_e}^{(j)}|a_k, \mathbf{l}_k^{[-j,i_e]}) \qquad (5.4)
\end{aligned}
$$

for $k = 1, ..., K + 1$, 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 $h_{i_e}^{(j)}$. If as a result of sampling $c_{i_e}^{(j)}$ a formerly represented founder is left with no haplotype associated with it, we remove it from the represented list of founders. If on the other hand the selected value k is not equal to any other existing index $c_{i_e}^{(j)}$, i.e, $c_{i_e}^{(j)} = K + 1$, we increment K by 1, set $n_{K+1} = 1$, update β accordingly, and sample a_{K+1} from its base measure F.

Now, from Equations (4.1) and (4.2), we can use the following posterior distribution to sample a_k:

$$
\begin{aligned}
&p(a_{k,t}|\mathbf{c}, \mathbf{h}) \propto \qquad (5.5) \\
&\prod_{j,i_e|c_{i_et}^{(j)}=k} p(h_{i_et}^{(j)}|a_{kt}, l_{kt}^{(j)}) = \frac{\Gamma(\alpha_h + l_{kt})\Gamma(\beta_h + l'_{kt})}{\Gamma(\alpha_h + \beta_h + m_k)(|\mathcal{A}| - 1)^{l'_{kt}}} R(\alpha_h, \beta_h)
\end{aligned}
$$

where l_{kt} is the number of allelic instances originating from the founder haplotype k at locus t across the groups that are identical to the founder, when the founder has the pattern a_{kt}. If k was not represented previously, we can just use zero values of l_{kt} which is equivalent to using the probability $p(a|h_{i_e}^{(j)})$.

We now proceed to the third sampling stage, in which we sample the haplotypes $h_{i_e}^{(j)}$, given the current state of the ancestral pool and the ancestral haplotype assignment for each individual, according to the following conditional distribution:

$$
\begin{aligned}
p(h_{i_et}^{(j)}|\mathbf{h}_{[-i_et]}^{(j)}, \mathbf{c}, \mathbf{a}, \mathbf{g}) &\propto p(g_{it}^{(j)}|h_{i_et}^{(j)}, h_{i_{\bar{e}},t}^{(j)}, \mathbf{u}_{[-i_et]}^{(j)})p(h_{i_et}^{(j)}|a_{k't}, \mathbf{l}_{k',[-i_et]}^{(j)}) \qquad (5.6) \\
&= R_g \frac{\Gamma(\alpha_g + u)\Gamma(\beta_g + (u' + u''))}{\Gamma(\alpha_g + \beta_g + IJ)} [\mu_1]^{u'} [\mu_2]^{u''} \\
&\quad \times \; R_h \frac{\Gamma(\alpha_h + l_{k',i_et}^{(j)})\Gamma(\beta_h + l_{k',i_et}'^{(j)})}{\Gamma(\alpha_h + \beta_h + n_k)(|\mathcal{A}| - 1)^{l_{k',i_et}'^{(j)}}}
\end{aligned}
$$

where $k' \equiv c_{i_e}^{(j)}$, $l_{k,i_et}^{(j)} = l_{[-i_et]}^{(j)} + \mathbb{I}(h_{i_et}^{(j)} = a_{kt})$, and $\mathbf{u}_{[-i_et]}^{(j)}$ are the set of sufficient statistics recording the inconsistencies between the haplotypes and genotypes in population j.

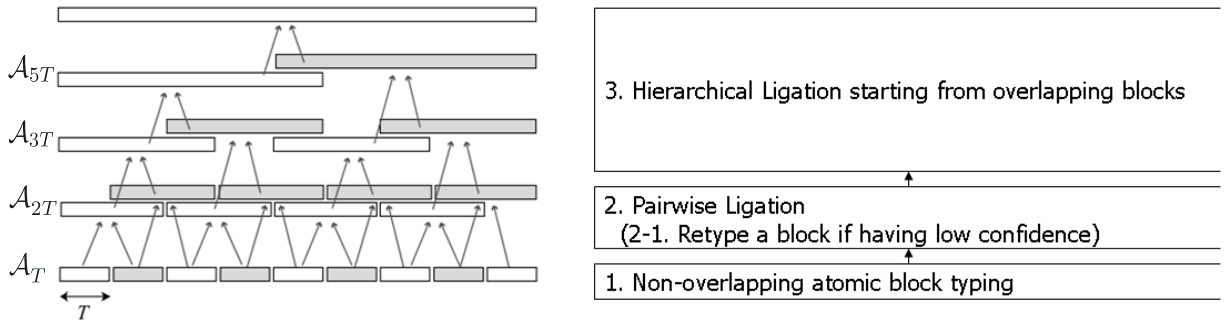

$\mathcal{A}_{5T}$
$\mathcal{A}_{3T}$
$\mathcal{A}_{2T}$
$\mathcal{A}_{T}$
T
3. Hierarchical Ligation starting from overlapping blocks
2. Pairwise Ligation
(2-1. Retype a block if having low confidence)
1. Non-overlapping atomic block typing

Figure 5.2: The partition-ligation scheme used in *Haploi*.

5.3 Partition-ligation and the *Haploi* program

As for most haplotype inference models proposed in the literature, the state space of the proposed HDPM 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, (Niu et al., 2002) proposed a divide-and-conquer heuristic known as Partition-Ligation (PL), which was adopted by a number of haplotype inference algorithms including PL-EM (Qin et al., 2002), PHASE (Li, 2003; Stephens et al., 2001), and CHB (Zhang et al., 2006). We equipped the HDPM model with a variant of the PL heuristic, and present a new tool, *Haploi* for *haplo*type *i*nference of multiple population genotype data over long SNPs sequences.

The original PL-scheme in Niu et al. (2002) first divides the entire sequence into disjoint short blocks and reconstructs haplotypes within each block. Then pairs of blocks are recursively ligated into larger (non-overlapping) haplotypes via Gibbs sampling under a fixed-dimensional Dirichlet prior over the frequencies of the ligated haplotype in the *product space* (or a subset) of all the "atomistic haplotypes" of every pair of blocks. This bottom-up approach can recover haplotypes of every individual either hierarchically or progressively. However, this PL scheme does not scale well to long sequences because the number of possible haplotypes in the product space can quickly become intractable as the size of the non-overlapping blocks to be ligated grows multiplicatively during the iteration. Unlike their approach, our PL-scheme generates partially overlapping intermediate blocks from smaller blocks phased at the lower level. The pairs of overlapping blocks are recursively merged into larger ones by leveraging the redundancy of information from overlapping regions, as well as overall parsimonious criteria. Empirically we found that this strategy can lead to a significant reduction of the size of the haplotype search space for long genotypes, and therefore facilitates a more efficient inference algorithm.

Figure 5.2 outlines the PL-procedure adopted by *Haploi*, which can be divided into three steps. In step 1, we begin by partitioning given genotype sequences into L short blocks of length T (e.g., $T \leq 10$ as suggested in Niu et al. (2002)). Then we phase each atomistic block using the proposed HDPM (Figure 5.2 step 1). By doing this, we obtain all the individual haplotypes and also the population haplotype pool (i.e., founders) for each block. In the next step, we ligate every pair of neighboring blocks. Naively the candidate population haplotype pool for the ligated segment can be a Cartesian product of the haplotype pools in neighboring blocks.

But such an unconstrained product is in fact unnecessary. Since each individual harbors only two possible haplotypes within each blocks, for each pair of adjacent blocks, we can impute at most four new stitched haplotypes from an individual, but in practice we get much fewer because an individual can be homozygous on one or both blocks and the stitched haplotypes may have been imputed already from earlier individuals; also, not all combinations of haplotypes in the two pools are necessary because some combinations may never exist in any individual. We pool such stitched haplotypes imputed from all individuals, which usually leads to only a small subset of the Cartesian product of the two haplotype pools. Then based on a finite dimensional Dirichlet prior over the candidate pool, we do Gibbs sampling as in Niu *et al.*'s PL scheme to obtain individual haplotypes for each overlapping $2T$ region. Essentially, our procedure produces a more parsimonious set of population haplotypes by using an individual-based population haplotype imputation scheme. In addition, comparing to the ligation in Niu *et al.*'s scheme, we stitch every neighboring pairs of blocks (ith with $(i+1)$th) whereas they ligate every odd numbered block with the next even numbered block (i.e., $(2i-1)$th with $(2i)$th). In step 3, we hierarchically ligate overlapping adjacent blocks from the previous iteration, until the full sequence is covered (Figure 5.2, step 3). The ligation strategy is again different from that of Niu *et al.*'s due to the haplotype consistency constraints imposed by overlapping SNPs, which helps to reduce the candidate haplotype space of the merged blocks. More details about the entire partition-ligation process can be found in Appendix A.1.

As we reduce the search space based on feasible individual haplotype pairs, there may be possibility of missing some haplotypes in the haplotype space construction if the ligation is only based on disjoint blocks. However, our ligation process considers two blocks with an overlapping region and takes into account all the possible inconsistencies for the every heterozygous locus. Therefore, the actual number of haplotypes added to the space can be greater than four in general except for the first pairwise ligation stage in Step 2 (see Appendix for a detailed example of this). Moreover, even in the pairwise ligation from the non-overlapping atomic blocks, this risk can be reduced by considering every neighboring pairs, not every odd-numbered and even-numbered pairs as noted above, as the information in one block can be propagated into both side of neighbors and can be preserved better. Empirically, this new scheme leaded to more accurate result than the original PL scheme with greatly improved computational cost, as the original PL scheme cannot be applied to more than a few hundreds of SNPs.

The underlying intuition of our ligation procedure is to allow recombination-like transition on the overlapping regions for including only all the necessary new haplotype configurations, but also to maximally preserve the haplotypes obtained at previous steps. This heuristic typically results in a population haplotype space of the merged block that is much smaller than the naive product-space of non-overlapping lower-level blocks. Moreover, individuals whose atomistic haplotypes of the pre-merged blocks have no discrepancy in the overlapping region would not only contribute only very few but high-confidence population haplotypes to the pool, but also they need not to be phased again in that ligation step. This constitutes the main source of efficiency and effectiveness of our algorithm.

In summary, comparing to the PL scheme in Niu et al. (2002), our method attempts to build more parsimonious set of population haplotypes at each ligation iteration by using an individual-based population-haplotype imputation scheme that leverages haplotypic diversity constraints imposed by individual genotypes and overlapping blocks. However, these modifications only help to better trim the population haplotype space; statistically, our haplotype inference still follows a well-defined MCMC scheme.

5.4 Results

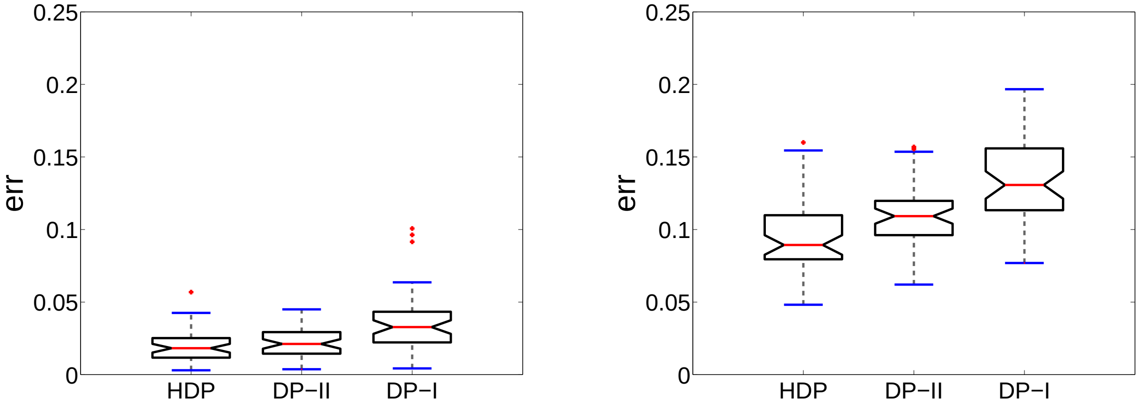
We evaluated the proposed HDPM model on both simulated genotype data and real genotype sequences from the International HapMap database. The haplotype inference accuracy under HDPM via the *Haploi* program is compared to that of the the baseline DP mixture model, and to PHASE 2.1.1 (Stephens and Scheet, 2005; Stephens et al., 2001), fastPHASE (Scheet and Stephens, 2006), MACH1.0 (Li and Abecasis, 2006), and Beagle 2.1.3 (Browning and Browning, 2007), in their default parameter settings unless otherwise specified. Two different error measures are used: err_s, the ratio of incorrectly phased SNP sites over all non-trivial heterozygous SNPs, and d_w, the switch distance, which is the number of phase flips required to correct the predicted haplotypes over all non-trivial cases. For short SNP sequences, we primarily use err_s; whereas for long sequences we compare d_w according to common practice. In addition to haplotype inference, we also estimated other metrics of interest, such as the haplotype frequencies, the mutation rates θ of each founding haplotypes, and the number of reconstructed haplotype founders K to assess the consistency of our model.

5.4.1 Simulated multi-population SNP data

To simulate multi-population genotypes, we used a pool of haplotypes taken from the coalescent-based synthetic dataset in Stephens et al. (2001), each containing 10 SNPs, as the hypothetical founders; and we drew each individual's haplotypes and genotype by randomly choosing two ancestors from these founders and applying the mutation and noisy genotyping models described in the methodology section. For each of our synthetic multi-population data set, we simulated five populations each with 20 individuals. Each population is derived from 5 founders, where two of them are shared across all the populations, and the other three are population-specific. Thus the total number of founders across the five populations is 17. We test our algorithm on two data sets with different degrees of sequence diversity. In the *conserved* data set, we set the mutation rate θ to be 0.01 for all populations and all loci in the simulation; in the *diverse* data set, θ is set to be 0.05. All populations and loci are assumed to have the same genotyping error rate. Fifty random samples were drawn from both the conserved and the diverse data sets.

5.4.2 Haplotype Accuracy

We compare *Haploi* using the HDP mixture and other methods applied in two modes on synthetic data. 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



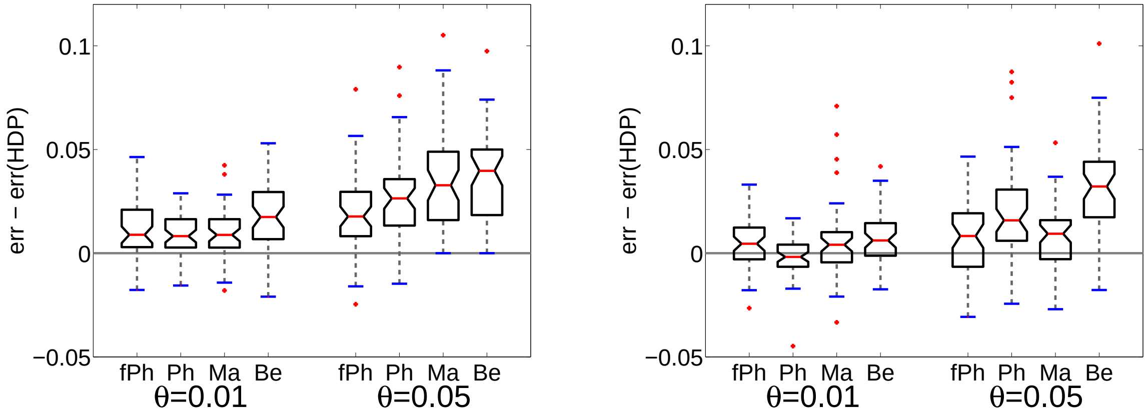
(a) $\theta = 0.01$ (b) $\theta = 0.05$

Figure 5.3: 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 ($\theta = 0.05$).

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 if specified, we supplied the labels to fastPHASE in mode-I experiments.

We first test how much HDP can gain by the hierarchical structure on multiple populations compared to the baseline DP. Figure 5.3 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 5.4 compares the performance of *Haploi* with those of the benchmark algorithms. When other algorithms are run in mode-I (Figure 5.4 (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



(a) Mode-I

(b) Mode-II

Figure 5.4: A comparison of HDP with other algorithms (fPh:fastPHASE, Ph:Phase, Ma:Mach, Be:Beagle) running in (a) mode-I, and (b) mode-II, on synthetic multi-population data.

separately (Figure 5.4 (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 ($p \leq 0.0043$).

5.4.3 Parameter estimation and sensitivity analysis

Typically, with random initialization, the Gibbs sampler for *Haploi* converges within 1000 iteration on the synthetic data. This contrasts sampling algorithms used in some of the other haplotype models, which typically need tens of thousands of iterations to reach convergence. The fast convergence is possibly due to *Haploi*'s ability to quickly infer the correct number of founding haplotypes underlying the genotypes samples, which leads to a model significantly more compact (i.e., parsimonious) than that derived from other methods.

Estimating K and θ

We compared the estimated K— the number of recovered ancestors via both HDP and DP mixtures. Recall that we expect K to be 17. Overall, the estimated K under both the DP and HDP models turns out to be very close to this number on the conserved datasets. From the diverse datasets, HDP can still offer a good estimate of the number of ancestors, whereas DP recovered more ancestors (around 25 on average) than the true number. This is not surprising since a haplotype which appears in more than one population can have different frequencies in different populations, the baseline DP cannot capture such sub-population structure, and the higher divergence due to both mutation and population diversification can make it generate more ancestors to describe the given dataset.

Our Gibbs sampler also provides reasonable estimates of the mutation rates of each haplotype founder. We observe that for the conserved data sets, HDP yields highly consistent and low

Table 5.1: A sensitivity analysis to the hyper-parameters of HDP on conserved 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.

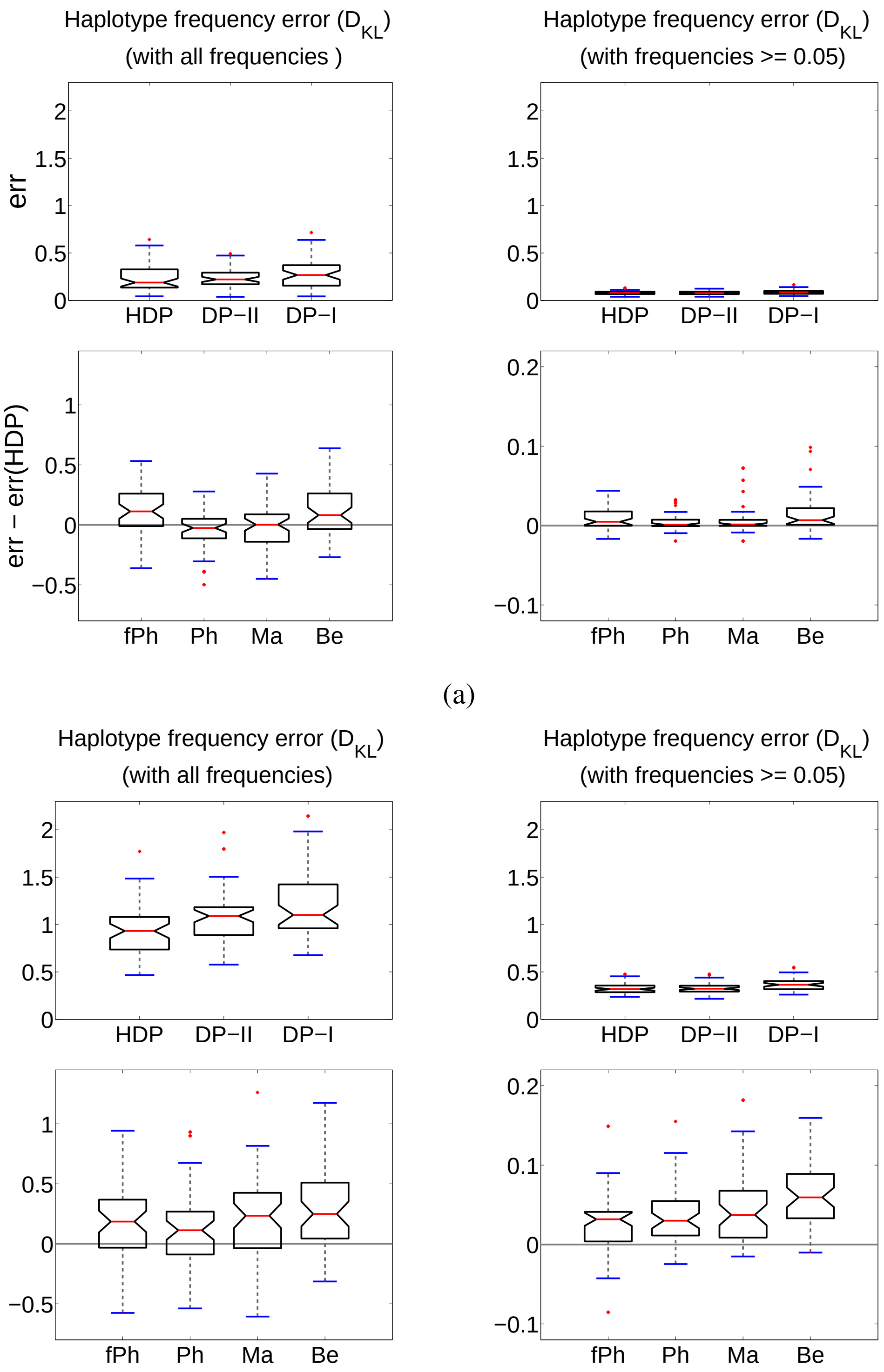
κ	ι	κ/ι	K_1	K_2	K_3	K_4	K_5	total K (17)	θ (0.005)	err_s
0.1	0.1	1	5.0	5.0	5.0	5.0	5.0	17.8	0.005	0.0058
	0.5	0.2	5.0	5.0	5.0	5.0	5.0	17.5	0.004	0.0116
	1	0.1	5.0	5.0	5.0	5.0	5.0	18.0	0.004	0.0000
	10	0.01	5.0	5.0	5.0	5.0	5.0	18.0	0.004	0.0087
	100	0.001	5.0	4.0	5.0	5.0	4.0	16.0	0.007	0.0029
	1000	0.0001	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

variance estimations of θ, and the quality of the estimates due to DP is slightly worse. For the diverse data both algorithms tend to slightly underestimate the mutation rates, and the variance is also higher. It is noteworthy that in principal, high haplotype diversity of a population can be explained by two competing sources: high mutation rate from ancestors to descendants, and large number of ancestors. Indeed K and θ cannot be independently determined, following a similar argument of the un-identifiability of the evolution time and population size under IAM model. But empirically, HDP appears to strike a reasonable balance between K and θ, and offers plausible estimates of both.

A more thorough sensitivity analysis with respect to the hyper-parameters in our model is detailed in Table 5.1. The proposed HDP model has two scale parameters, γ and τ, for the upper and lower level DP, which are under inverse Gamma priors as discussed in Section 5.2.2. 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 one of the 50 random conserved datasets. Columns 4 – 9 in Table 5.1 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. The test on the diverse dataset shows similar tendency while the result is slightly less stable (see Table A.1 in Appendix for more details).

Estimating haplotype frequencies

Figure 5.5 summarizes the accuracy of population haplotype frequencies estimated by each algorithm. The discrepancy between the true frequencies and estimated ones is measured by the KL-Divergence $D_{KL}(p||q) = \sum_x p(x) \log \frac{p(x)}{q(x)}$. The top row shows the accuracy of HDP along with those of DP in mode-II and in mode-I, and the bottom row shows the comparison of HDP with other benchmark algorithms. The left column of Figure 5.5 (a) reports D_{KL} computed on ALL haplotypes frequencies estimated by different algorithms from the conserved data sets and the right column of Figure 5.5 (a) shows the result when measured only on the frequent haplotypes (i.e., with frequencies ≥ 0.05). Comparing to the baseline-DP, HDP is as accurate when only frequent haplotypes are considered. When all the frequencies are considered, however, the margin of HDP over DP becomes significant, especially on the diverse dataset (p=0.0009). Overall, $Haploi$, PHASE, and MACH work equally well without significant difference in performance on conserved datasets. For more difficult diverse data sets (Figure 5.5 (b)), HDP achieves the lowest discrepancy by a significant margin over all the other algorithms. The runner-up, PHASE beats fastPHASE and MACH with a small margin. When measured only on the frequent haplotypes (i.e., the right column of Figure 5.5 (b)), the discrepancies decrease significantly, but the relative ordering of all the compared algorithms remains similar, except that now fastPHASE outperforms PHASE ($p = 0.0036$).



(a)

(b)

Figure 5.5: A comparison of the accuracies of haplotype frequencies. Top: the result from HDP, DP in mode-II (DP-II), and DP in mode-I (DP-I). Bottom: the result from HDP and three benchmark algorithms. (a) Box-plots of D_{KL}'s estimated from the conserved data sets. Left column shows measurements on all haplotypes, right column shows measurements on only the frequent haplotypes. (b) Same measurements on the diverse datasets.

5.4.4 Result on HapMap Data

We also test *Haploi* on both short SNP segments (i.e., ~ 6 SNPs), and long SNP sequences (i.e., $\sim 10^2 - 10^3$ SNPs) available from the International HapMap Project. This data contains SNP genotypes from four populations: Utah residents with ancestry from northern and western Europe (CEU); Yoruba in Ibadan, Nigeria (YRI); Han Chinese in Beijing (CHB); and Japanese in Tokyo (JPT), with 60, 60, 45, and 44 unrelated individuals, respectively. Although haplotype inference can be, and in some test scenarios, was performed on all populations, evaluation of the outcome is on only the CEPHs and Yorubas since the true haplotypes can be almost unambiguously deduced from trios only in these two populations. The individual genotypes that cannot be unambiguously phased from the trios were ignored in the scoring. We consider three different population-composition scenarios in our experiments below: 1) using all the four populations together for haplotype inference (FourPop); 2) using only CEPH and Yoruba populations for inference (TwoPop); and 3) phasing CEPH and Yoruba separately (OnePop). Essentially, in the FourPop and TwoPop scenarios we solve a bigger haplotype inference problem on data that contain richer population information.

Short SNP sequences

Phasing short SNPs is the basic operation of large-scale haplotype inference problems that rely either on a partition-ligation heuristic, or on a model-based methods, such as recombination processes, to integrate short phased haplotype segments into long haplotypes. Figure 5.6 shows a comparison of the phasing accuracy on 6-SNP segments (following a recommendation in Niu et al. (2002) on the optimal size-range of basic units for subsequent ligation) by four algorithms. The test was done on randomly selected 100 sets of 6-SNPs segment from chromosome 21. 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.

Several aspects of *Haploi*'s performance on real data are revealed by Figure 5.6. First, comparing the performances of *Haploi* under the three different population-composition scenarios, we observe that *Haploi* improves steadily as more populations are included in haplotype inference, and the improvements are statistically significant. The p-values of the differences between FourPop and OnePop scenarios are 0.00024, 0.000038, 0.0016, and 0.000022 for data with 60, 30, 20, 10 individuals per population, respectively; and the p-values of the margins of TwoPop over OnePop are 0.0014, 0.0002, 0.0053, and 0.00047, respectively, in the same order. The improvement in FourPop over TwoPop is less significant, with p-values 0.35, 0.11, 0.16, and 0.023, respectively, suggesting that the possible gain in haplotype accuracy enabled by the HDP model via exploring shared information among populations can be capitalized the most when we change from single-population inference to joint-inference in multiple population; whereas the effect of having more populations in the multi-population scenario appears to be less obvious in this dataset.

Second, comparing the performances of *Haploi* under different population sizes, we observe that the performance-gain through information sharing among populations tends to be greater when the population sizes decrease. For example, the performance differences of *Haploi* in multi-population over single-population become most significant when the number of individuals per population is the smallest (#Individual per pop=10). This observation suggests that HDP is especially advantageous under data scarcity situation where information from each population becomes insufficient to warrant reliable inference within the population.

Third, other methods such as PHASE, MACH, and Beagle, appear not able to benefit from increased population diversity as indicated by the significant drop of their accuracies when more populations are involved. The performance of fastPHASE (with known population labels) improves substantially when two populations are used together, while the performance becomes slightly worse in the case of four populations. Comparing the results from the most preferred scenario of each algorithm, that is, *Haploi* under FourPop, fastPHASE under TwoPop, and all the others under OnePop, *Haploi* and PHASE worked similarly well when all the available data were used (i.e. #Individual per pop=60), with mean error rate of each algorithm at 0.0174, 0.0198, 0.0173, 0.0229, and 0.0222, respectively (with p =0.05,0.89,0.10,0.01 over differences of *Haploi* with other algorithms). When the population sizes decrease, *Haploi* starts to surpass others more substantially, and works more reliably than others. For example, on 10 individuals per population, the mean error rates of the five algorithms were 0.0424, 0.0460, 0.0512, 0.0777, and 0.0945, and the p-values of the margin of *Haploi* over others are 0.17, 0.02, 1.2×10^5, 6.7×10^6, respectively.

Long SNP sequences

Finally we test *Haploi* on very long genotype sequences with $10^2 \sim 10^3$ SNPs. We selected 10 ENCODE regions from the HapMap DB, each spanning roughly 500 Kb and containing from 254 to 972 common SNPs across all four populations (see Table A.2 in Appendix for more details). We performed haplotype inference under three different population-composition scenarios as before, but due to the extremely high cost in computational time in these experiments, we only worked on the full-size data sets. Figure 5.7 shows a comparison of haplotype reconstruction quality, using PHASE, fastPHASE, MACH, Beagle and *Haploi* equipped with the PL heuristic [1]. 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 5.7 is less clear than the ones from previous sections from experiments on short SNP sequences and on simulation data. 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

[1] We could not get output of PHASE for these long sequences within acceptable running time ($>$ 800hours). Instead we included fastPhase result, which is said to be much faster than PHASE with a slight performance degradation (Scheet and Stephens, 2006).

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.

5.5 Discussion

We have proposed a new Bayesian approach to haplotype inference for multiple populations using a hierarchical Dirichlet process mixture. By incorporating an HDP prior which couples multiple heterogeneous populations and facilitates sharing of mixture components (i.e., founder haplotypes) across multiple Dirichlet process mixtures, the proposed method can infer the true haplotypes in a multi-ethnic group with an accuracy superior to the state-of-the-art haplotype inference algorithms.

There emerged new models related to our HDP model, the closest being the nested Dirichlet process (NDP) by Rodriguez et al. (2006). In an NDP, instead of using a hyper-DP as a common base measure as in HDP to allow sharing of founders across populations, the population-specific DPs are directly drawn from a prior DP, so that not only the founders, but also their frequencies can be shared across populations. Although this model can be more expressive in many applications, it may be less appropriate than HDP for multi-population haplotype problems where excessive structural sharing across populations is not warranted, especially when different populations bear very distinct demography and genetic prototypes. Another strategy proposed by Muller et al. (2004) employs an explicit stochastic convex combination of a population-specific prior and a universal prior for each founder. Under such a model, once a founder is destined to be shared across populations, it will appear with equal frequency in all populations. HDP subsumes this scenario, but also allows more flexible sharing of the founders.

The proposed model achieves the desirable properties of PAC regarding mutation dynamics (Li, 2003), including the parental-dependent-mutation effect, albeit in a very different way. For example, to see the PDM property, note that when a next haplotype is to be sampled according to Equation (3.4), we pick an ancestor of some previously drawn haplotypes, and apply a mutation process to the ancestor, rather than to one of the previously drawn haplotypes as in PAC. This operation implicitly results in a PDM effect among haplotypes by relating them to their corresponding founder via a tractable star genealogy equipt with a common mutation process $P_h(|founder)$. A new haplotype generated from this process will bear mutations over its corresponding founders rather than been completely random. Above these founders, we model their genealogy and type history by a *coalescent-with-IMA* model, whose resulting marginal is equivalent to that of the Dirichlet process. Here a new founder can be sampled independent of the type-history in the coalescent from the base measure, rather than according to a PDM, with probability proportional to the IMA mutation rate. Putting everything together, the DP mixture model essentially implements a combination of IMA and PDM: it models the genealogy and type history of hypothetical ancestors presumably corresponding to a bottleneck with a coalescent-with-IMA model; below the bottleneck, it uses multiple (indeed, can be countably infinite many) star genealogies rooted at the ancestors present in the bottleneck and equipt with an ancestor-dependent Poisson mutation process, to approximate the coalescent-with-PDM model. The time of the bottleneck depends on the value of the scaling parameter α of the DP. One can introduce a prior to this parameter so that it can be estimated *a posteriori* from data.

It is well-known that under Kingman's n-coalescent, a dominant portion of the depth of the coalescent tree is spent waiting for the earliest few lineages to coalesce to the MRCA and the majority of lineages of even a very large population can actually coalesce very rapidly into a few ancestors, which means that the net mutation rates from each of these ancestors to their descendants in a modern haplotype sample do not vary dramatically among the descendants. Thus qualitatively a star genealogy provides a reasonable approximation to the actual (heavily time-compressed) genealogy of a modern haplotype sample up to these ancestors. As a reward of such approximation, a well-known property of DP mixture is that it defines an exchangeable distribution of the samples. Furthermore, the Pólya urn construction of DP enables simple and efficient Monte Carlo for posterior inference of haplotypes and other parameters of interest, and the DPM formalism offers a convenient path for extensions that capture more complex demographic and genetic scenarios of the sample, such as the multi-population haplotype distribution as we explored.

Unlike the models underlying PHASE and fastPhase, the PL heuristic used in the *Haploi* program does not explicitly model the recombination process that shapes the LD patterns of long SNP sequences. Since an HDP model without the aid of the PL-scheme dominates PHASE and fastPhase over short SNPs, we believe that an upgrade that incorporates an explicit recombination model in conjunction with HDP is likely to lead to more accurate reconstruction of long haplotypes. The hidden Markov Dirichlet process recently developed by us to model recombination in open ancestral space offers a promising path for such an upgrade (Xing and Sohn, 2007). Under the proposed statistical framework for modeling haplotype and genotype distribution, it is also straightforward to handle various missing value problems in a principled way. In another possible extension, although in the present study we have assumed that the population labels of individuals are known, it is straightforward to generalize our method to situations in which the ethnic group labels are unknown and to be inferred. This opens the door to applications of our method to large-scale genetic studies involving joint inference over markers and demography. The HDP model is also a natural formalism for applications outside of population genetics, such as in text modeling, where one can use an HDPM to model co-clustering of documents from different journals (analogous to different populations here) according to both shared and unique topics defined by, e.g, a latent Dirichlet allocation model (Blei et al., 2003); and also in network modeling, where the neighbor profiles of every node can be modeled by a low-level DPM whose likelihood function is defined by, e.g., a mixed membership stochastic block model (Airoldi et al., 2006), and the entire network corresponds to an HDP over all nodes.

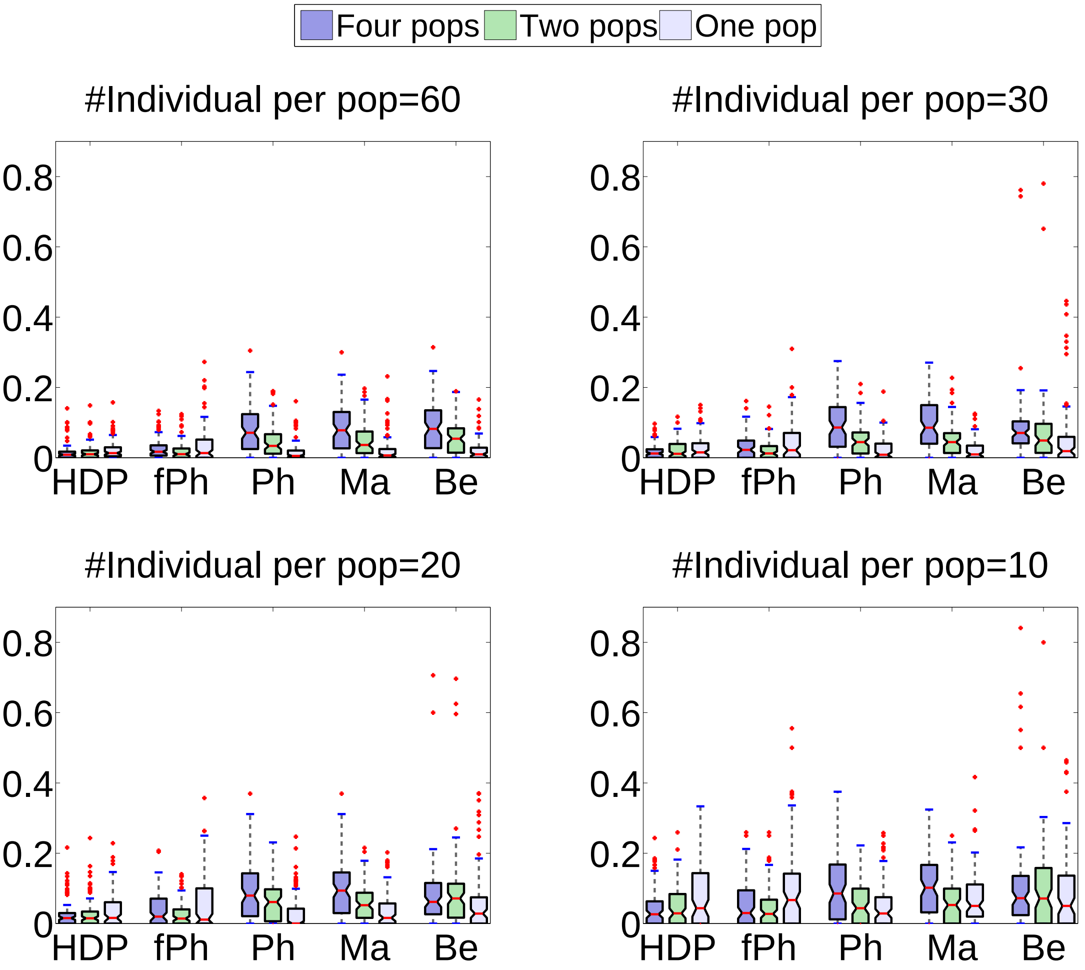
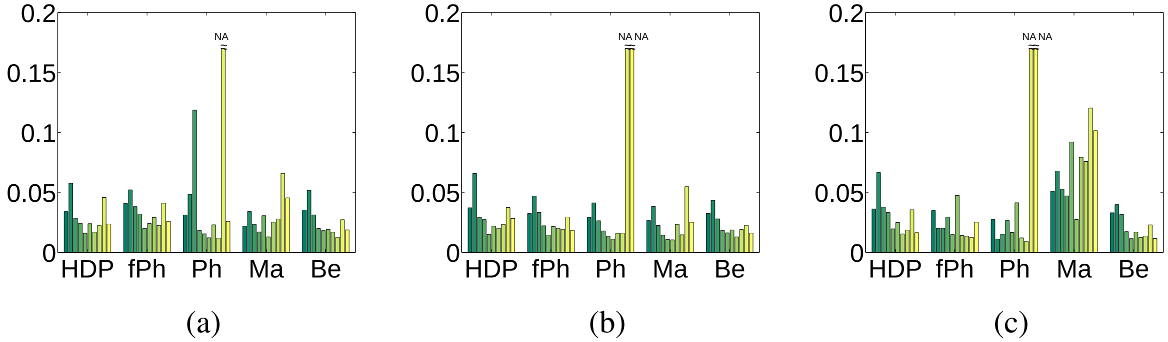


Figure 5.6: A comparison of haplotyping error on CEPH+Yoruba population over randomly chosen 100 sets of 6-SNP segments from Chromosome 21. The results were obtained under three population-composition scenarios: (i) FourPops: when data from all the four populations were used (blue) for inference; (ii) 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.



(a) (b) (c)

Figure 5.7: 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).

Chapter 6

Joint inference of population structure and recombination events

6.1 Introduction

SNPs are remnants of ancient DNA alterations dated back to a time measured at a genealogical scale. They contain finer-grained information on molecular evolution than that revealed by orthologous genomic sequences from multiple species. In general, the higher the frequency of a SNP allele, the older the mutation that produced it, so high-frequency SNPs largely predate human population diversification whereas low-frequency ones appeared afterwords. Therefore, population-specific alleles may bear important information about human evolution such as specific migrations and genetic diversifications (Stoneking, 2001).

A number of variants of statistical admixture models for genetic polymorphisms have been proposed for the analysis of current population structure (Falush et al., 2003; Pritchard et al., 2000; Rosenberg et al., 2002). These models are instances of a more general class of hierarchical Bayesian models known as *mixed membership models* (Erosheva et al., 2004), which postulate that genetic markers of each individual are *iid* (Pritchard et al., 2000) or spatially coupled (Falush et al., 2003) samples from multiple population-specific fixed-dimensional multinomial distributions (known as *ancestry proportions* (Falush et al., 2003), or AP) of marker alleles. Under this assumption, the *admixture* model identifies each ancestral population by a specific AP (that defines a unique allele frequency profile for each ancestral population for each marker) and displays the fraction of contributions from each AP in a modern individual chromosome as a *structural map*. Fig. 6.1 shows an example of structural maps of four modern populations inferred from a portion of the HapMap multi-population dataset by *Structure 2.1* (Falush et al., 2003; Pritchard et al., 2000). In this *population structural map*, each individual is represented as a thin vertical line which shows the fraction of the individual's chromosome which originated from each ancestral population, as given by a unique AP.

However, since an AP merely represents the *frequency* of alleles in an ancestral population, rather than the actual allelic content or haplotypes of the alleles themselves, the admixture model does not model genetic drift due to mutations from the ancestral alleles. Moreover, in the extant admixture models, the correlations between loci along the chromosome are only captured by the

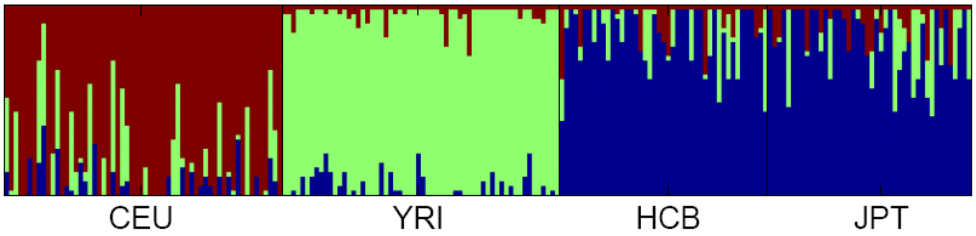
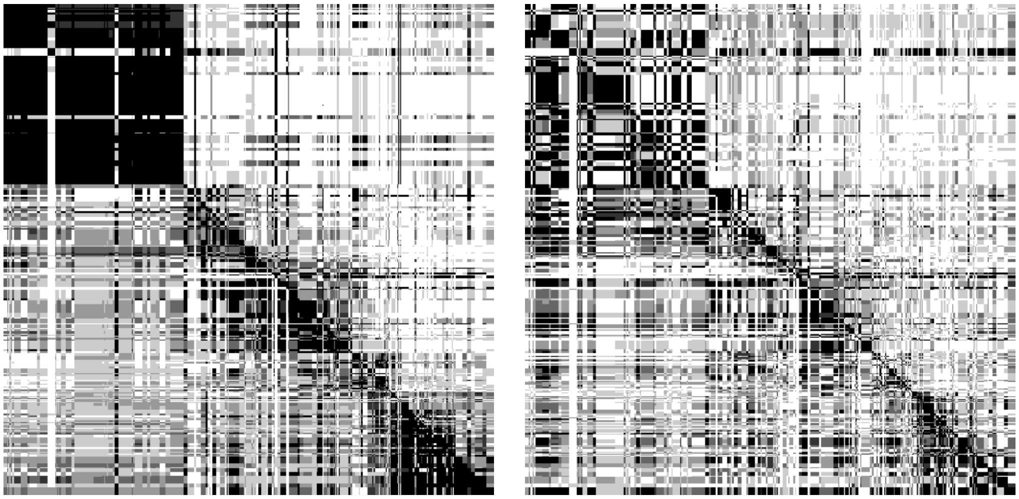


Figure 6.1: Population structural map inferred by *Structure 2.1* on HapMap multi-population data consisting of CEU, YRI, HCB and JPT populations.

linkage disequilibrium due to variation in the AP fractions over all markers among individuals, or due to a "recombination" process between APs , rather than ancestral chromosomes, for sampling markers along a modern chromosome. These two scenarios are known as "mixture LD" and "admixture LD" respectively (Falush et al., 2003). Neither one captures the actual recombination events at the ancestral chromosome level, so they do not enable inference of the founding genetic patterns, the recombination events, the age of the founding alleles, or the composition of individual chromosomes at founding chromosome level (Excoffier and Hamilton, 2003). Actually, while this model aims to provide ancestry information for each individual and each locus, there is no explicit representation of "ancestors" as a real chromosome haplotype. Therefore, the inferred population structural map emphasizes revealing the contributions of abstract population-specific ancestral proportion profiles, which does not directly reflect individual diversity. This representation may not be optimal, as seen in Figure 6.1: each modern population is represented by a very homogenous, but distinct population structural sub-map, which reflects little about the actual genetic diversity of each population and individual and little about the relative similarity between populations. For example, the YRI population from Africa is known to be genetically diverse, but in Figure 6.1 it appears to be the most homogeneous.

We have presented a new method, *Spectrum*, for inferring and representing population structures, using a unified statistical framework for modeling the genetic inheritance process that allows both recombination among an unspecified number of founding haplotypes and mutations from these founders. Based on this model, which represents a well-defined generative model for the observed chromosomes, we represent the population structure in terms of an *ancestral spectrum* which shows the ancestral composition of each modern individual chromosome in terms of its origin among the chromosomal ancestors. By considering the different ancestral association patterns among populations, this spectrum helps to separate the sub-populations, as well as reveal the diversity among individuals and populations. Moreover, our model allows us to recover the recombination events in each individual chromosome. In fact, the population structure can play an important role for the LD analysis. Figure 6.2 shows the LD measurements for all pairwise loci on the ENm010 region from HapMap DB. When we compute LD in three populations of CEU (European ancestry), HCB and JPT (Asian ancestry) together (Figure 6.2(a)), some degree of block-like patterns are visible, but when CEU (European ancestry) and YRI (African ancestry) populations are mixed (Figure 6.2(b)), the block structure is less obvious. This result implies the existence of different genetic processes in the evolutionary history of the two populations. Hence, if we perform LD or recombination analysis on a population which may



(a) CEU + HCB + JPT

(b) CEU + YRI

Figure 6.2: The LD measurements, $|D'|$ (upper right), and the p-values for Fisher's exact test (lower left), of HapMap DB (Thorisson et al., 2005). In each of the LD maps, starting from the upper-left corner, all the markers are listed in top-down and left-right directions, and each marker is at a spatial position corresponding to its actual genetic distance with respect to the first marker at the upper-left corner. Note the LD-block structures on the mixed populations of CEU and YRI in shown (b) are rather opaque compared to the LD patterns of CEU+HCB+JPT populations in (a).

have a concealed sub-population structure, it would be more informative to perform LD analysis on each sub-population separately, and our ancestral spectrum offers a way to classify such sub-populations on genetic basis. While the statistical methodologies developed so far mostly deal with ancestral inference and LD analysis separately using specialized models that do not capture the close statistical and genetic relationships of these two problems, we propose a unified framework which allows joint inference of the population structure and the recombination patterns.

We assume that individual chromosomes in a modern population originated from a number of ancestral chromosomes via biased random recombination and mutation. By associating each ancestor with a hidden state, the recombination between the ancestors can follow a state transition process, and the mutation can follow an emission process in the hidden Markov model. Hence each individual chromosome can be thought of as a "mosaic" of ancestral chromosomes under this model.

Several existing methods have employed similar ideas. For example, Daly et al. (2001) and Greenspan and Geiger (2004) have developed hidden Markov models for locating recombination hotspots in haplotypes; Anderson and Novembre (2003) proposed a minimum description length (MDL) method for optimal haplotype block finding. While these models are based on a similar assumption that each observed haplotype is a "mosaic" of ancestral haplotypes and the formation of the mosaic is governed by a hidden Markov process over the ancestor space, these HMMs cannot be used easily to infer individual recombination events because the block boundaries which conceptually correspond to the recombination sites of all individual chromosomes are decided outside the model via model selection, and the only intrinsic stochasticity lies in the choice of the "ancestors" at each block for each chromosome rather than the genomic locations of recombination events in each chromosome. It is also unclear to what extent this class of approaches might be helpful for applications involving explicit ancestral map inference as in Rosenberg et al. (2002) and for interpreting LD patterns that do not have sharp block boundaries as in Figure 6.2(b).

While most of the previous approaches ignore the inherent uncertainty in the genetic complexity (e,g., the number of genetic founders of a population) of the data, our new approach employs a non-parametric Bayesian model of infinite hidden Markov model which we call *Hidden Markov Dirichlet Process* to extend a *closed* genetic inheritance model based on a fixed number of founders to an *open* ancestral space, which allows more flexible control over the number of genetic founders than has been provided by the statistical methods proposed thus far. We report validation of *Spectrum* on both simulated data and on two real datasets of HapMap and Daly data, and compare with a number of established methods.

6.2 The statistical model

We describe a statistical model for generating individual haplotypes in a modern population from a hypothetical pool of ancestral haplotypes via recombination and mutations. We begin our exposition with a parametric Bayesian model of genetic inheritance involving recombination and mutation over a fixed number of ancestors; then we extend the model to open ancestral space which requires no *ad hoc* specification of the number of ancestors, via a nonparametric Bayesian approach.

6.2.1 Hidden Markov model for recombination and mutation in *closed* ancestral space

We begin with the assumption that modern chromosomes are derived from ancestral chromosomes via biased random recombination and mutation. This assumption corresponds to an idealized noninterference model for chromosomal crossover and a star genealogy over every inherited site. If the number of ancestors is known to be K, sequential selection of recombination targets from a set of ancestral 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 founders to the descendants.

Assuming that individual haplotypes over T SNPs $H_{i_e} = [H_{i_e,1}, \ldots, H_{i_e,T}]$ for $e = 1, 2$ are given unambiguously for the study population, as is the case in many LD and haplotype-block analyses (Anderson and Novembre, 2003; Daly et al., 2001), we can now treat the paternal and maternal haplotypes of N individual as $2N$ *iid* samples and omit the parental index e. Although this assumption may seem stringent, our model can easily generalize to unphased genotype data by incorporating a simple genotype model, as will be explained later in this section.

Now, let $A_k = [A_{k,1}, \ldots, A_{k,T}]$ for $k = 1, \ldots, K$ be the K ancestral haplotypes, and let $C_i = [C_{i,1}, \ldots, C_{i,T}]$ denote the sequence of inheritance variables that specify the index of the ancestral chromosome at each SNP locus for each chromosome i. Also suppose that the transition probabilities of the HMM are given as a $K \times K$ matrix π. When no recombination takes place during the inheritance process that produces the haplotype H_i from an ancestor k as assumed in the HDP model in Chapter 5, then $C_{i,t} = k$ for all $t = 1, \ldots, T$. When recombination occurs between a locus 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 $e^{-d_t r} + (1 - e^{-d_t r})\pi_{kk}$, where d_t 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 HMM, 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 $(1 - e^{-dr})\pi_{kk'}$. That is,

$$P(C_{i,t+1} = k' \mid C_{i,t} = k) = e^{-dr}\pi_{k,k'} + (1 - e^{-dr})\delta(k, k') \tag{6.1}$$

Hence, each haplotype H_i can be thought of as a mosaic of segments of multiple ancestral chromosomes from the ancestral pool $\{A_k\}_{k=1}^{K}$.

The emission process of this model corresponds to a mutation model from an ancestor to the matching descendent.We adopt the *single-locus mutation model* explained in Equation (4.1). As discussed in Liu et al. (2001), this model corresponds to a star genealogy resulting from infrequent mutations over a shared ancestor and is widely used in statistical genetics as an approximation to a full coalescent genealogy starting from the shared ancestor. Assuming that the mutation rate θ_k admits a Beta prior with hyperparameter (α_h, β_h), the marginal conditional likelihood of all the haplotype instances $\mathbf{h} = \{h_{i,t} : i \in \{1, 2, \ldots, I\},\ t \in \{1, 2, \ldots, T\}\}$ given the set of ancestors $\mathbf{a} = \{a_1, \ldots, a_K\}$ and the ancestor indicators $\mathbf{c} = \{c_{i,t} : i \in \{1, 2, \ldots, I\},\ t \in \{1, 2, \ldots, T\}\}$ can be obtained by integrating out θ from the joint conditional probability starting from Equation (4.1) which reduces to:

$$P(\mathbf{h}|\mathbf{c}, \mathbf{a}) = \prod_k R(\alpha_h, \beta_h) \frac{\Gamma(\alpha_h + l_k)\Gamma(\beta_h + l'_k)}{\Gamma(\alpha_h + \beta_h + l_k + l'_k)} \Big(\frac{1}{|B| - 1}\Big)^{l'_k} \tag{6.2}$$

where $\Gamma(\cdot)$ is the gamma function, $R(\alpha_h, \beta_h) = \frac{\Gamma(\alpha_h+\beta_h)}{\Gamma(\alpha_h)\Gamma(\beta_h)}$ is the normalization constant associated with $\mathrm{Beta}(\alpha_h, \beta_h)$ (which is a prior distribution for θ), $l_k = \sum_t \sum_i \mathbb{I}(h_{it} = a_{kt})\mathbb{I}(c_{it} = k)$ is the number of alleles which were not mutated with respect to the ancestral allele, and $l'_k = \sum_t \sum_i \mathbb{I}(h_{it} \neq a_{kt})\mathbb{I}(c_{it} = k)$ is the number of mutated alleles. The counting record $\mathbf{l}_k = \{l_k, l'_k\}$ is a sufficient statistic for the parameter θ_k. Note that the main difference between this derivation and the one for the HDP mixture model is that we now deal with a locus-specific indicator variable C_{it} that varies along the markers on the chromosome as the recombination is explicitly considered in this model.

The model described above can be easily generalized to un-phased genotype sequence data by introducing a genotyping model as described in Chapter 5. We assume that the observed genotype at a locus is determined by the paternal and maternal alleles of this site via the genotyping model in Equations (4.2) and (4.1).

It is noteworthy that the proposed model presents a well-defined generative model for the observed haplotypes or genotypes based on a spatial point process for stochastic recombination and also random mutations over a pool of complete ancestral chromosomes. The difference in our model compared to approaches with a similar HMM assumption (Anderson and Novembre, 2003; Daly et al., 2001; Patil et al., 2001) is that, in those models, the "ancestors" are defined independently for each block rather than as whole chromosomes, which is biologically less meaningful. Although such a generative process is still a simplification of the real biological mechanism, it enables the joint statistical characterization of a number of genetic variables of interest, via posterior inference based on well-founded statistical principles, and it strikes a reasonable tradeoff between being biologically meaningful and computationally manageable.

6.2.2 Hidden Markov Dirichlet Process for Inheritance in *open* ancestral space

So far, we have been assuming that recombination and mutation take place in a *closed* ancestral space; that is, the number of ancestral chromosomes is known *a priori*. But this assumption, which is also widely adopted in other existing approaches for LD analysis and ancestral inference, ignores the inherent uncertainty in the genetic complexity of populations. Model selection according to information theoretic score or Bayes factors is a typical solution to problems of this nature, but it can be inflexible when the hypothesis space is large. We have developed a nonparametric Bayesian framework for modeling genetic polymorphism based on the Dirichlet process (DP) mixtures and extension (Sohn and Xing, 2007a,b; Xing et al., 2004; Xing and Sohn, 2007), which allows more flexible control over the number of genetic founders.

Using an infinite Hidden Markov model which we also call the Hidden Markov Dirichlet Process (HMDP) (Sohn and Xing, 2007a,b; Xing and Sohn, 2007), we extend the HMM model proposed in Section 6.2.1 to work in an infinite ancestral space. Recall that in the HMM inheritance model described earlier, the transition probabilities can be represented as a $K \times K$ matrix, and each row of the matrix indicates the probabilities of transitioning (i.e., recombination) from the source state (e.g., founder k) to all the target states (all the founders in the pool), which sums to one. Now we do not restrict ourselves with such a K and generalize the HMM to a space with countably infinite ancestors in principal. Our generalization can be understood as modeling each row of transition probabilities from a specific founder of an HMM with a unique DP over open ancestral space, letting all these DPs (each of which is over a particular row) follow a higher level DP to ensure that they are all defined on the same open ancestral space. We have developed a hierarchical Pólya urn scheme to realize this model and facilitate sampling based posterior inference. But at a high level, the recombination probability under HMDP $P(C_{i,t+1} = k' \mid C_{i,t} = k)$ can be expressed by the same formula as in Equation (6.1), except that the $\pi_{kk'}$ now indicates the transition probability from a source state k to a target state k' in an open ancestral space under HMDP (see Xing and Sohn (2007) for the somewhat cumbersome form for this variable). This $\pi_{kk'}$ specifies the probability of ancestor chromosome k pairing with ancestor k' given that a recombination is taking place, and k' can grow arbitrarily large as needed conditioning on the given data.

The generative process described above leads naturally to an algorithm for population genetic inference. Unlike the classical coalescence models for recombination (Hudson, 1983), which have been primarily used for theoretical analysis and simulation and are not feasible for reverse ancestral inference based on observed genetic data, *Spectrum* provides a nonparametric Bayesian formalism for recombination and inheritance that is well suited for data-driven posterior inference on the latent variables that can yield rich information of the population ancestry and genetic structure of the study population. For example, using *Spectrum*, given the haplotype (or genotype) data, one can infer the ancestral structure, LD and recombination patterns of a population using the posterior distribution of inheritance variable $\mathbf{c}$ and ancestral state $\mathbf{a}$, as we will elaborate in the sequel.

6.3 MCMC Inference

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

The Gibbs sampler alternates between two stages. First it samples the inheritance variables $\{c_{it}\}$, 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_{it}'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:

$$P(c_{t+1:t+\delta} \,|\mathbf{c}^-, \mathbf{h}, \mathbf{a}) \propto \prod_{j=t}^{t+\delta} P(c_{j+1}|c_j, \mathbf{m}, \mathbf{n}) \prod_{j=t+1}^{t+\delta} P(h_j|a_{c_j,j}, \mathbf{l}_{c_j}) \tag{6.3}$$

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. Note that, naively, the sampling space of an inheritance block of length δ is $|A|^\delta$ where $|A|$ represents the cardinality of the ancestor pool. However, if we assume that the recombination rate is low and block length is not too big, then the probability of having two or more recombination events within a δ-block is very small and thus can be ignored. This approximation reduces the sampling space of the δ-block to $O(|A|\delta)$, i.e., $|A|$ possible recombination targets times δ possible recombination locations.

Accordingly, Equation (6.3) reduces to:

$$
\begin{aligned}
& p(c_{t+1:t+\delta} \mid \mathbf{c}^{-}, \mathbf{h}, \mathbf{a}) \\
& \sim p(\text{at most one recombination in}[t, t+\delta] \mid \mathbf{c}^{-}, \mathbf{h}, \mathbf{a}) \\
& \propto p(c_{t'} \mid c_{t'-1} = c_t, \mathbf{m}, \mathbf{n}) p(c_{t+\delta+1} \mid c_{t+\delta} = c_{t'}, \mathbf{m}, \mathbf{n}) \times \\
& \quad \prod_{j=t'}^{t+\delta} p(h_j \mid a_{c_{t'},j}, \mathbf{l}_{c_{t'}})
\end{aligned} \tag{6.4}
$$

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

$$
\begin{aligned}
& p(c_{t'} = k' \mid c_{t'-1} = k, \mathbf{m}, \mathbf{n}, r, d) \\
= & \begin{cases} \lambda \pi_{k,k'} + (1-\lambda)\delta(k, k') \\ \quad \text{for } k' \in \{1, ..., K\}, \text{ i.e., transition to an existing ancestor,} \\ \lambda \pi_{k,K+1} \\ \quad \text{for } k' = K+1, \text{ i.e., transition to a new ancestor,} \end{cases}
\end{aligned} \tag{6.5}
$$

where $\pi_{k,\cdot}$ represents the transition probability vector for ancestor k under HMDP.

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

$$
p(c_{t+\delta+1} \mid c_{t+\delta} = K+1, \mathbf{m}, \mathbf{n}) \quad = \quad \lambda \times \frac{n_{c_{t+\delta+1}}}{n - 1 + \alpha}. \tag{6.6}
$$

Now we can substitute the relevant terms in Equation (6.3) with Equations (6.5) and (6.6). The marginal likelihood term in Equation (6.3) can be readily computed based on Equation (4.1), by integrating out the mutation rate θ under a Beta prior (and also the ancestor a under a uniform prior if $c_{t'}$ refers to an ancestor to be newly instantiated) (Xing et al., 2004). Putting everything together, we have the proposal distribution for a block of inheritance variables. Upon sampling every c_t, we update the sufficient statistics $\mathbf{n}$, $\mathbf{m}$ and $\{\mathbf{l}_k\}$ as follows. First, before drawing the sample, we erase the contribution of c_t to these sufficient statistics. In particular, if an ancestor gets no occupancy in either the stock or the HMM urns afterwards, we remove it from our repository. Then, after drawing a new c_t, we increment the relevant counts accordingly. In particular, if $c_t = K+1$ (i.e., a new ancestor is to be drawn), we update $n = n+1$, set $n_{K+1} = 1$, $m_{c_t} = m_{c_t} + 1$, $m_{c_t,K+1} = 1$, and set up a new (empty) HMM urn with color $K+1$ (i.e. instantiating $\mathbf{m}_{K+1}$ with all elements equal to zero).

Now we move on to sample the founders $\{a_{k,t}\}$. From the mutation model in Equation (4.1), we can derive the following posterior distribution to sample the founder a_k.

$$
\begin{aligned}
p(a_{kt}|\mathbf{c}, \mathbf{h}) &\propto \int \Big(\prod_{i|c_{it}=k} p(h_{it}|a_{kt}, \theta) \Big) \text{Beta}(\theta|\alpha_h, \beta_h) d\theta \\
&= \frac{\Gamma(\alpha_h + l_{kt})\Gamma(\beta_h + l'_{kt})}{\Gamma(\alpha_h + \beta_h + l_{kt} + l'_{kt})(|B| - 1)^{l'_{kt}}} R(\alpha_h, \beta_h), \qquad (6.7)
\end{aligned}
$$

where l_{kt} is the number of allelic instances originating from ancestor k at locus t that are identical to the ancestor, when the ancestor has the pattern a_{kt}; and $l'_{kt} = \sum_i \mathbb{I}(c_{it} = k \mid a_{kt}) - l_{kt}$ represents the complement. The normalization constant of this proposal distribution can be computed by summing the right-hand side of Equation (6.7) over all possible allele states of an ancestor at the locus being sampled. If k is not represented previously, we can just set l_{kt} and l'_{kt} both to zero. Note that when sampling a new ancestor, we can only condition on a small segment of an individual haplotype. To instantiate a complete ancestor, after sampling the alleles in the ancestor corresponding to the segment according to Equation (6.7), we first fill in the rest of the loci with random alleles. When another segment of an individual haplotype needs a new ancestor, we do not naively create a new full-length ancestor; rather, we use the *empty* slots (those with random alleles) of one of the previously instantiated ancestors, if any, so that the number of ancestors does not grow unnecessarily.

6.4 Results

We validated *Spectrum* on a simulated dataset and analyzed two real datasets: the HapMap four-population data (Thorisson et al., 2005) and the single-population data from Daly et al. (2001). Although *Spectrum* can be applied to the case of genotype data as well, we primarily focus on haplotype data for simplicity. The HapMap data includes 209 individuals' haplotypes (phased by PHASE software (Stephens and Scheet, 2005; Stephens et al., 2001)) on the ENm010 region of chromosome 7. The Daly data includes 256 individuals (after excluding one person due to severe missing data), whose haplotypes (512 in total) can be recovered from trio data. For each dataset, we focus on the analysis of population structure and recombination patterns based on the ancestral origin of each SNP locus in each individual haplotype.

6.4.1 Analyzing a simulated haplotype population

We simulated a population of individual haplotypes with a fixed number K_s of randomly generated founder haplotypes, on each of which a set of recombination hotspots were pre-specified. Then we applied a recombination process, which is defined by a K_s-dimensional HMM, to the ancestor haplotypes to generate N_s individual haplotypes via sequentially recombining segments of different ancestors according to the simulated HMM states at each locus and mutating certain ancestor SNP alleles according to the emission model. All the ancestor haplotypes were set to be 100 SNPs long. The hotspots are pre-specified at every 10-th loci in the ancestor haplotypes. Overall, 30 datasets, each containing 100 individuals (i.e., 200 haplotypes) with 100 SNPs, were

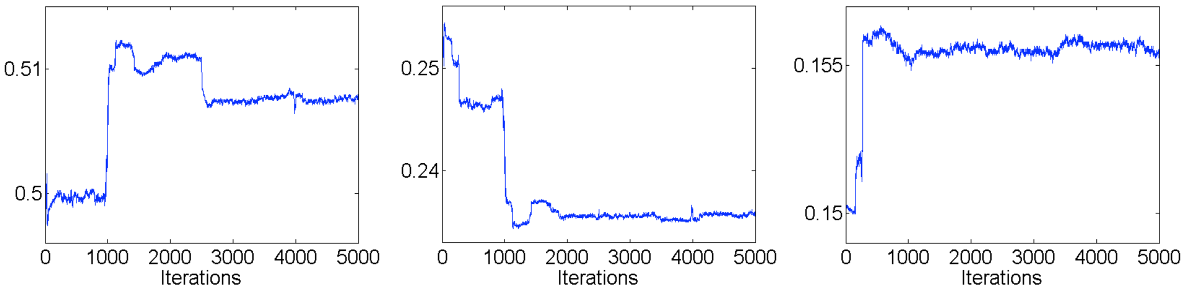


Figure 6.3: Sampling trace of the top three most occupied factors that correspond to the founder haplotypes. The x-axis represents the sampling iteration, and the y-axis represent the fraction of the occupancy (i.e., be chosen as recombination target) of each factor over total occupancy.

generated from $K_s = 5$ ancestor haplotypes. Since there is no extant method that can perform both structural analysis and recombination analysis, we compared our method with existing algorithms specialized for each of our tasks. For ancestral inference, we implemented 3 standard fixed-dimensional HMMs, with 3, 5, and 10 hidden states, respectively, where 5 corresponds to the true number of founders for the simulation. For recombination analysis, we selected the widely used *LDhat 2.0* (Fearnhead and Donnelly, 2001) for comparison. *Structure 2.1* yields a different kind of population map that is not quantitatively comparable to that from *Spectrum*. Therefore, we only show empirical comparisons on real data.

We integrated out the mutation rate θ as before, and sample variables $\{a_{k,t}\}$ and $\{c_{i,t}\}$ iteratively. We monitor convergence based on the occupancy counts of the top factors in the master DP. Typically, convergence was achieved after around 3000 samples (Figure 6.3), and the samples obtained after convergence with proper de-autocorrelation, i.e., by using samples from every 10 iterations over $5000 \sim 10000$ samples are used for computing relevant sufficient statistics. To increase the chance of proper mixing, 10 independent runs of sampling, with different random seeds, are simultaneously performed.

Founder reconstruction

Using HMDP, we successfully recovered the correct number (i.e., $K = 5$) of founders in 21 out of 30 simulated populations; for the remaining 9 populations, we inferred 6 founders, as the mode of the posterior distribution. From samples of founder states $\{a_{kt}\}$, we reconstructed the ancestral haplotypes under the HMDP model. For comparison, we also inferred the ancestors under the 3 standard HMM using an EM algorithm. We define the *ancestor reconstruction error* ϵ_a for each ancestor to be the ratio of incorrectly recovered loci over all the chromosomal sites. The average ϵ_a over 30 simulated populations under 4 different models are shown in Figure 6.4. In particular, the average reconstruction errors of HMDP for each of the five ancestors are 0.026, 0.078, 0.116, 0.168, and 0.335, respectively. There is a good correlation between the reconstruction quality and the population frequency of each ancestor. Specifically, the average (over all simulated populations) fraction of SNP loci originated from each ancestor among all loci in the population is 0.472, 0.258, 0.167, 0.068 and 0.034, respectively. As one would expect, the higher the population frequency of an ancestor is, the better its reconstruction accuracy. Interestingly, under the fixed-dimensional HMM, even when we use the correct number of ancestor states, i.e., $K = 5$, the reconstruction error is still very high (Figure 6.4), typically 2.5 times or higher than the error of HMDP. We conjecture that this is because the non-parametric Bayesian treatment of the transition rates and ancestor configurations under the HMDP model leads to a desirable adaptive smoothing effect and also less constraints on the model parameters, which allow them to be more accurately estimated. Whereas under a parametric setting, parameter estimation can easily be sub-optimal due to lack of appropriate smoothing or prior constraints, or deficiency of the learning algorithm such as local-optimality of EM.

Structural analysis

Spectrum uncovers the genetic origins of all loci of each individual haplotype in a population from Gibbs samples of the inheritance variables $\{c_{i,t}\}$. For each individual, we define an empirical *ancestor composition vector* η_e, which records the fractions of every ancestor in all the c_{it}'s of that individual. Figure 6.5 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. Five spectrums are shown in Figure 6.5, each of which corresponds to (1) true ancestor compositions, (2) ancestor compositions inferred by *Spectrum*, and (3-5) ancestor compositions inferred by HMMs with 3, 5, 10 states, respectively. To assess the accuracy of our estimation, we calculated the distance between the true ancestor compositions and the estimated ones as the mean squared distance between true and estimated η_e over all individuals in a population, and then over all 30 simulated

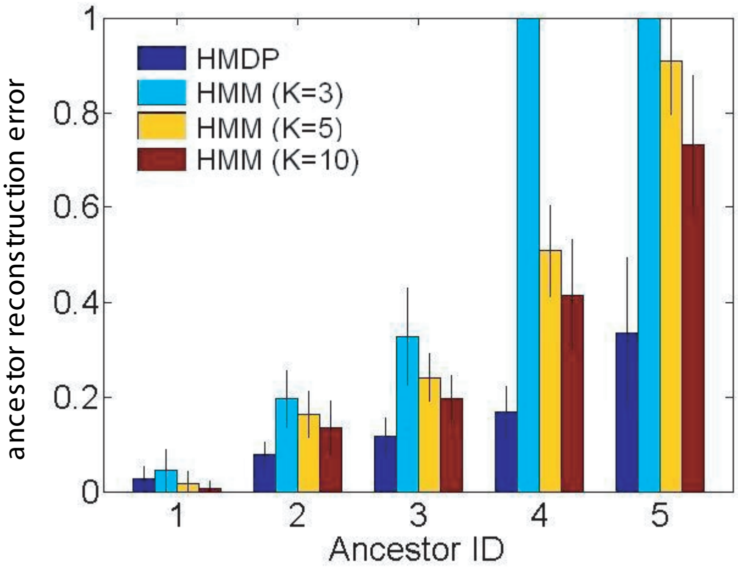


Figure 6.4: Analysis of simulated haplotype populations. A comparison of ancestor reconstruction errors for the five founders indexed along x-axis. The vertical lines show ±1 standard deviation over 30 populations.

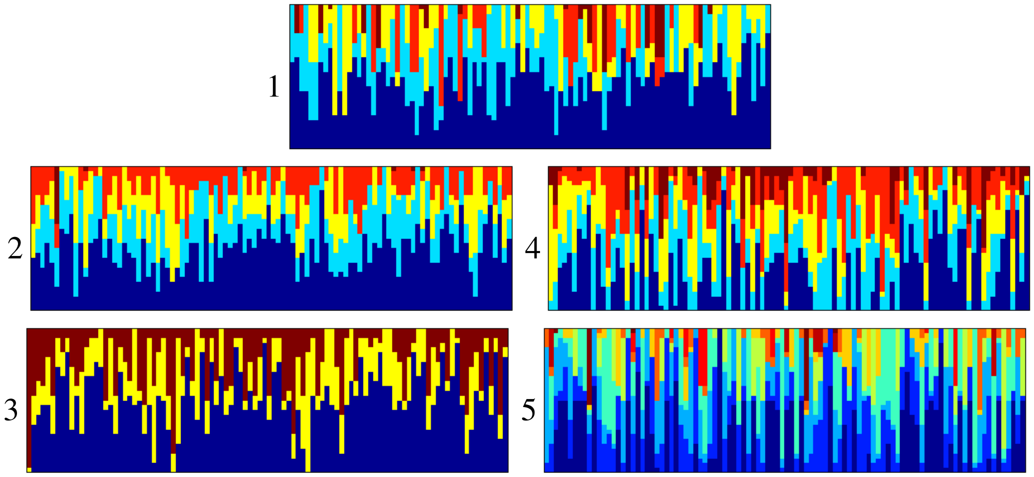


Figure 6.5: Analysis of simulated haplotype populations. The true (panel 1) and estimated (panel 2 for *Spectrum*, and panel 3-5 for 3 HMMs) population maps of ancestral compositions in a simulated population.

Table 6.1: False positive and false negative rates for recombination hotspot detection over 30 population samples. Two kinds of threshold ω's are used. The results with different tolerance windows w_{tol} are also shown.

		Spectrum			*LDhat 2.0*			HMM ($K = 5$)		
	w_{tol}	0	± 1	± 2	0	± 1	± 2	0	± 1	± 2
ω=	FPR	0.16	0.11	0.07	0.19	0.09	0.06	0.18	0.12	0.11
3rd quartile	FNR	0.11	0	0	0.22	0.11	0.11	0.33	0.11	0.11
ω s.t.	FPR	0.16	0.11	0.07	0.22	0.11	0.07	0.18	0.12	0.11
FNR$\sim$FAR	FNR	0.11	0	0	0.22	0.12	0.11	0.33	0.11	0.11

populations. We found that the distance between the *Spectrum*-derived population spectrum and the true spectrum is 0.190 ± 0.0748, whereas the distance between HMM-spectrum and true spectrum is 0.319 ± 0.0676, significantly worse than that of *Spectrum* even though the HMM is set to have the true number of ancestral states (i.e., $K = 5$). Because of dimensionality incompatibility and apparent dissimilarity to the true spectrum for other HMMs (i.e., $K = 3$ and 10), we forgo the above quantitative comparison for these two cases.

Recombination Analysis

From the Gibbs samples of $\{c_{it}\}$, we can also infer the recombination status of each locus of each haplotype. We define the *empirical recombination rates* λ_e to be the ratio of individuals who are determined to have recombinations at each locus over the total number of haplotypes in the population. We classify a locus to be a recombination hotspot if its λ_e is greater than an empirical threshold ω, which is set to be the 3rd quartile value of the estimated recombination rates. Alternatively we can set ω to be the λ_e value at which the false positive rate and the false negative rate become equal in a held-off set. Due to the stochastic nature of the recombination position in our simulation, we score a correct hit of recombination hotspot if the identified hotspot

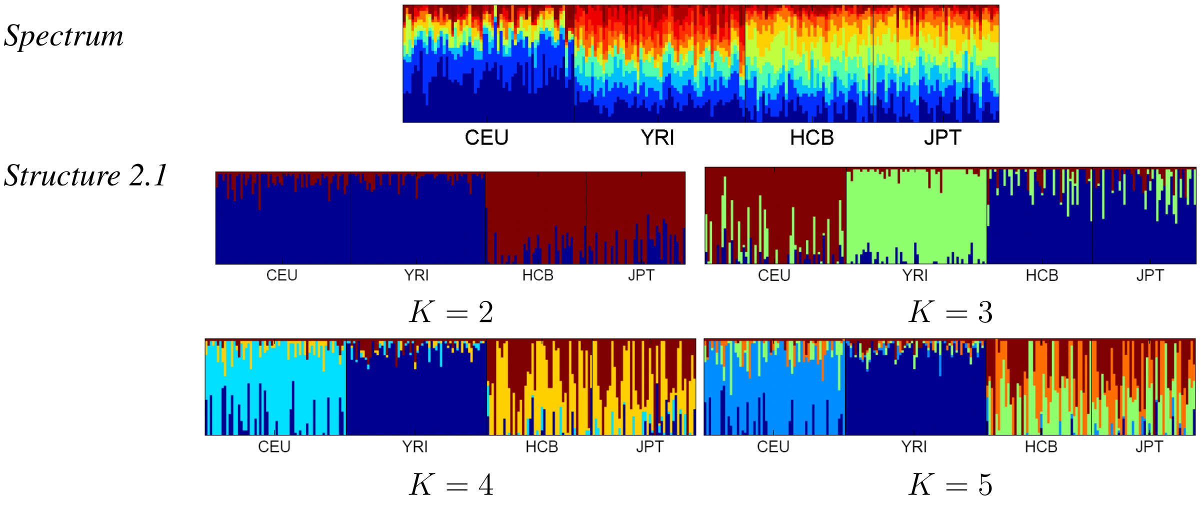


Figure 6.6: Inferred population structure of HapMap four population data from *Spectrum*, and *Structure 2.1* with different pre-specified numbers of population K.

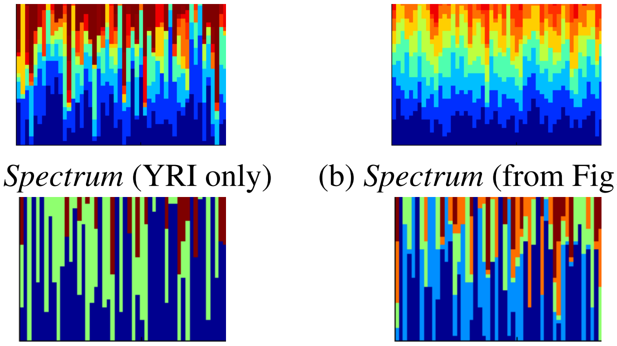
based on λ_e-thresholding falls within a small window around the true position, and the window is set to be 0, ± 1, and ± 2, respectively. Table 1 summarizes the results of the performance comparison for the recombination hotspot detection, which shows that *Spectrum* outperforms *LDhat 2.0* and HMM in most of the cases.

6.4.2 Analyzing real datasets

Population Structure Analysis

We analyzed the population structure of HapMap data (on the ENm010 region) based on the ancestor composition vector η_e. Figure 6.6 shows the results from *Spectrum* and from *Structure 2.1* with different pre-determined numbers of populations K. Both algorithms successfully identified the major geographical populations grouped as CEU, YRI, and HCB+JPT populations. However, the population map from *Structure 2.1* does not reflect the diversity of each population or similarity between populations as mentioned earlie. In contrast, the result from *Spectrum* reveals the relative diversity of each population clearly by showing the ancestral association fraction for each individual from shared ancestors.

For further comparison, we applied each method to the YRI population only. In Figure 6.7, panel (a) shows the ancestral spectrum of YRI when this population only is subject to analysis by *Spectrum*; and panel (b) re-displays the YRI spectrum extracted from Figure 6.6(a), where all four populations were analyzed together. Figure 6.7 (c) and (d) present the maps from *Structure 2.1* applied to YRI only, under three- and five-cluster assumptions, respectively. While it is not straightforward to match (a) with (b) pictorially, both maps reveal that this population is rather diverse. On the other hand, Figure 6.7 (c) and (d), both from *Structure 2.1*, show two very different structures from those in Figure 6.6, where the 4 populations were analyzed together. Since *Structure 2.1* maps each individual locus to its *origin of population* represented by a unique AP, rather than to its origin of ancestral chromosome, this result is not surprising considering the



(a) *Spectrum* (YRI only) (b) *Spectrum* (from Fig.6.6)

(c) *Structure 2.1* $(K = 3)$ (d) *Structure 2.1* $(K = 5)$

Figure 6.7: Inferred population structure of HapMap YRI population data from (a)-(b) *Spectrum* , and (c)-(d) *Structure 2.1* with different number of clusters K.

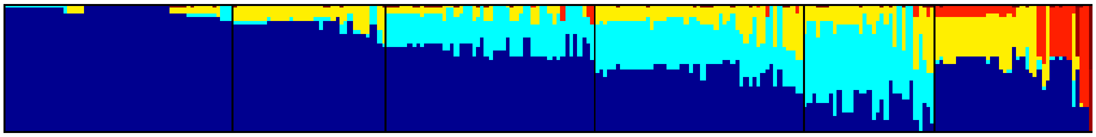


Figure 6.8: The estimated population map of the Daly dataset. The ordering of all individuals in the sample population was determined by a K-means clustering with $K = 6$, followed by a within-cluster ordering of samples based on their distances to the cluster centroid. The black vertical bars show the K-means cluster boundaries.

different level of details of the two (i.e., our *spectrum* and their *map*) representations. It seems that our method provides an arguably more robust and consistent way of showing the population structure in terms of *origin of ancestral haplotypes*, which clearly illustrates the sharing of ancestors between populations, as well as the diversities of each population. It is also noteworthy that in *Structure 2.1* the choice of K can significantly affect the result, and it is not always easy to choose the best K, as shown in Figure 6.7. In contrast, our method does not rely on a fixed number of ancestors, instead giving a flexible model for the genetic inheritance under a nonparametric Bayesian framework.

Next, we analyzed the 256 individuals (i.e., 512 haplotypes) from the Daly data set with 103 SNPs. For a more informative revelation of the underlying population structure captured by the empirical ancestor composition vector η_e, we clustered the individuals based on their η_e's and then ordered all individuals accordingly (Figure 6.8). Specifically, all individuals were clustered into 6 clusters, which is an empirical choice for illustration, using the K-means algorithm. Within each group, individual orderings were determined by their distances to the cluster centroid. Interestingly, we can see that although the Daly data were reported to be from a European-derived population that is expected to be genetically less diverse, our ancestral map suggests that in this population there exists distinct sub-structures, each with a unique ancestral composition.

Recombination analysis

For the analysis of recombination events in real datasets, rather than picking an empirical threshold, we determined the recombination hotspots as follows. We fitted the estimated λ_e's of all loci

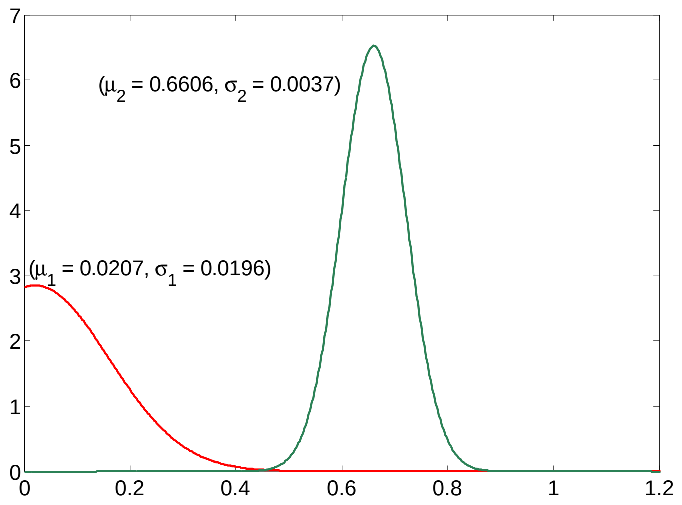


Figure 6.9: A mixture of Gaussian fitting of the estimated λ_e on HapMap data

with a one-dimensional mixture of Gaussians (Figure 6.9). Then we used the intersection point of the two Gaussian components as the threshold for determining hotspot loci. This threshold is essentially the point where the posterior probabilities of λ_e being a baseline recombination rate or a hotspot recombination rate are equal. The mass in the area where the two Gaussians overlap represents the Bayes-error of loci classification under this model. One can also employ more rigorous model-based methods for hotspot classification, and we will return to this point in the discussion section.

Figure 6.10 shows the recovered recombination rates on the ENm010 region of chromosome 7 for each population 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 6.6), 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 6.10 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 population-based recombination analysis. For comparison, the result on the mixed populations are also shown together for *Spectrum* and *LDhat 2.0* in the last column of Figure 6.10.

We also give the comparison of the recombination hotspot estimation on the Daly dataset with those reported in Daly et al. (2001) which was based on an HMM employing different numbers of states at different chromosome segments, and in Anderson and Novembre (Anderson and Novembre, 2003) which is based on a minimal description length (MDL) principle. In Figure 6.11, we show the plot of the empirical recombination rates estimated from *Spectrum*,

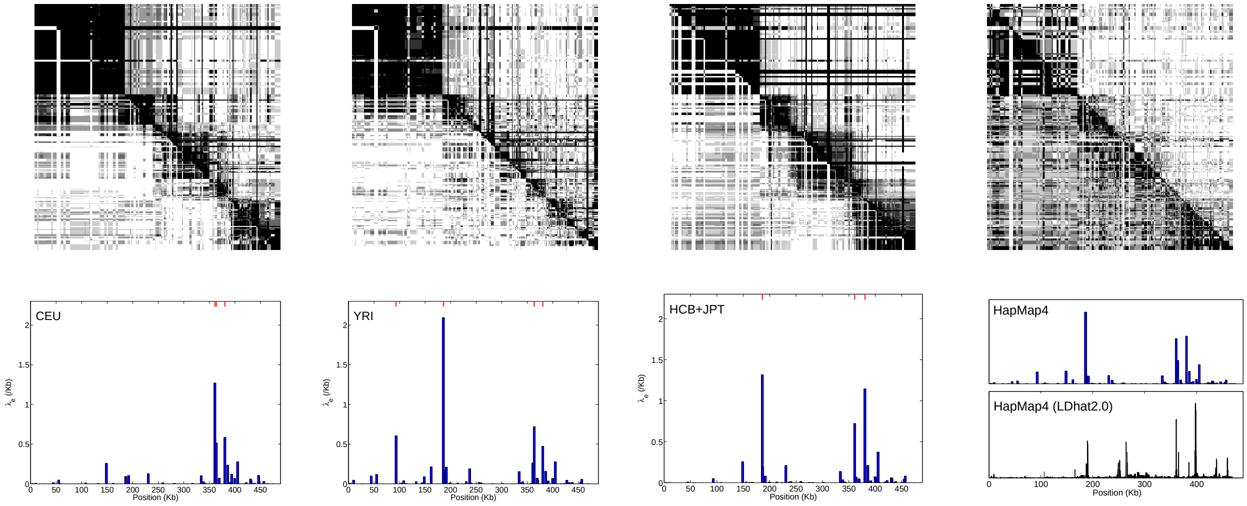


Figure 6.10: For each population of HapMap data, the LD measure with the estimated recombination rates along the chromosomal position are shown together with the detected recombination hotspots. The last column shows the result on the mixed four populations from both *Spectrum* and *LDhat 2.0*.

side-by-side with the reported recombination hotspots. We also display the LD measurements together. Note that according to *Spectrum*, certain estimated recombination hotspots are very close to each other; for example, at locus 398kb, two hotspots are right next to each other. This finding suggests that the actual LD patterns in a population sample may not simply fall into blocks with sharp boundaries universal to all individuals, as assumed in Daly's HMM model. It is more appropriate to define "hotspot regions" (i.e., stretches of consecutive hotspot loci) rather than point "hotspot loci", where necessary, to delineate haplotype blocks, as discussed in Li (2003). For example, according to the estimated λ_e's shown in Figure 6.11, 15 hotspot loci/regions (represented as thick solid vertical bars in Figure 6.11) were identified, and they divide the entire study region into 16 haplotype blocks of low diversity. Note that in Figure 6.11, the x-axis represents the actual genetic locations of the SNP loci (starting from 274kb at the leftmost with respect to a genetic reference). Since the SNPs of interest are not located uniformly in this region, the spatial-intervals as seen from Figure 6.11 between hotspots may not reflect the "lengths" of the haplotype blocks. For example, the block between 445-518kb contains 15 SNPs. At the same time, the seemingly longest interval between 738-877kb contains only 3 SNPs, two of which have high recombination rates, which render this interval to be a hotspot region as explained below. Biologically, this is not surprising because the probability of recombination between adjacent SNPs increases with their physical distance, in addition to depending on the intrinsic recombination rate. This "hotspot region" between 738-877kb is more likely to be merely a consequence of sparse location-sampling of SNPs in this region, rather than a biologically meaningful hotspot region.

For more quantitative comparison of the results, we computed information-theoretic (IT) scores based on the estimated within-block haplotype frequencies and the between-block transition probabilities under each model for a comparison. Figure 6.12 shows a comparison of these scores for haplotype blocks obtained from HMDP and the other two sources. The left panel of

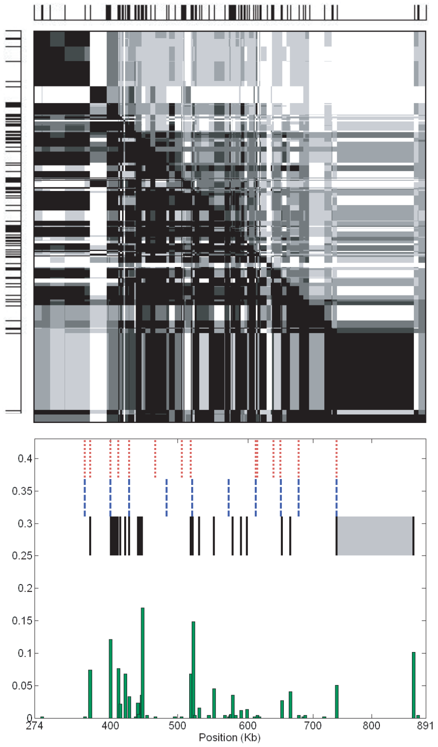


Figure 6.11: Analysis of the Daly data. Upper panel: the LD-map of the data. Lower panel: a plot of λ_e estimated via *Spectrum*; and the haplotype block boundaries according to *Spectrum* (black solid line), HMM (Daly et al., 2001) (red dotted line), and MDL (Anderson and Novembre, 2003) (blue dashed line). Note that the thickness of the black solid lines delineating the haplotype blocks is proportional to the width of the hotspot regions between adjacent blocks.

Figure 6.12 shows the total pairwise mutual information between adjacent haplotype blocks segmented by the recombination hotspots uncovered by the three methods. The right panel shows the average entropies of haplotypes within each block. The number above each bar denotes the total number of blocks. The pairwise mutual information score of the HMDP block structure is similar to that of the Daly structure, but smaller than that of MDL. Similar tendencies are observed for average entropies. Note that the Daly and the MDL methods allow the number of founder haplotypes to vary across blocks to get the most compact local ancestor constructions. Thus their reported scores are an underestimate of the true global score because certain segments of an ancestor haplotype that are not or rarely inherited are not counted in the score. Thus the low IT scores achieved by HMDP suggest that HMDP can effectively avoid inferring spurious global and local ancestor patterns. This is confirmed by the population map shown in Figure 6.8, which shows that HMDP recovered 6 ancestors and among them the 3 dominant ancestors account for 98% of all the modern haplotypes in the population.

Table 6.2 summarizes the summary statistics that characterize each haplotype block and

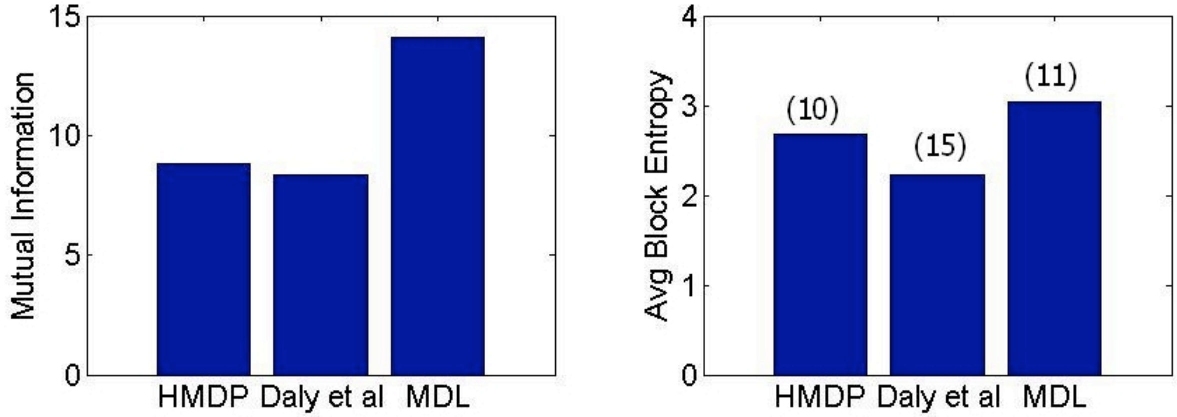


Figure 6.12: Analysis of the Daly data. Information-theoretic scores for haplotype blocks from each method. The left panel shows cross-block MI and the right shows the average within-block entropy. The total number of blocks inferred by each method are given on top of the bars.

hotspot regions. We used the threshold of 0.005 determined by the mixture of Gaussians as described above to identify recombination hotspots. The blocks were determined accordingly, with the constraint that the lengths of the identified blocks were at least three SNPs long, to avoid over-fragmenting the haplotypes. In column 1 of Table 6.2, the blocks with blockID starting with an "r" represent the hotspot regions which contain more than 2 SNPs, and others represent the haplotype blocks. The number of SNPs within the blocks varied from 3 to 15 (the second column of Table 6.2). The actual genomic region and length of each block are shown in the third and the fourth columns, respectively. The lengths of the smallest and the biggest blocks were 1.3kb and 93kb, respectively, while the average was 22kb. We also report the total number of distinct haplotypes as a reflection of diversity for each block, of which the most diverse is, not surprisingly, one of the largest blocks (which spans 71kb), which contains 17 different haplotypes. This is significantly lower than the 2^{17} possible different haplotypes one could observe had there existed no co-inheritance among loci in this block. Note that the 17 haplotypes reported here indicate the actual total observed diversity in this region among the study population, not the number of prototypes underlying these haplotypes that parsimoniously account for the majority of the observed diversity when small amounts of mutation are allowed, as reported in Daly et al. (2001). The actual demographic diversity of these blocks is much lower than that which is reflected by the total number of haplotypes, as shown by the results in columns 6-15. In columns 6-11 of Table 6.2, we report the ancestor association frequencies of haplotypes within each block, where the associations were directly estimated from the inheritance variable c_{it}'s sampled by our algorithm. We can see that, overall, 6 founders sufficed to fully account for our data, and indeed within each block, only 3-4 of them were significantly used. We present the number of necessary haplotypes to cover over 95% and 90% of the entire population, which were mostly around 3 with a few blocks with higher diversity around 10.

Table 6.2: Haplotype block structures and the summary statistics of the blocks for the Daly data. The block boundaries correspond to the x-coordinates of the λ_e peaks in Figure 6.11.

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

6.5 Discussion

We have proposed a new Bayesian method, *Spectrum*, for jointly modeling genetic recombination with mutation and population structure. Under a pool of complete founder haplotypes, *Spectrum* describes the underlying genetic process of recombination and mutation explicitly in terms of the association between founders and modern individuals. By incorporating a hidden Markov Dirichlet Process prior, which facilitates a well-defined transition process between infinite ancestor spaces, the proposed method can efficiently infer a number of important genetic variables such as recombination hotspots and ancestor patterns, jointly under a unified statistical framework.

Our model provides a new way of representing a population structure in terms of an ancestral spectrum which shows the ancestral association composition of each modern individual chromosome with the chromosomal ancestors. While the existing method based on admixture models (Falush et al., 2003) gives some degree of clear population label information, it is less informative in showing the population diversity or relationship between populations in the genetic history. In contrast, the *Spectrum* identifies the structure of sub-populations by considering the different ancestral association patterns among populations, in addition to displaying the diversity among individuals and populations, which yields a more informative representation for the population structure among shared ancestors across the populations.

Moreover, *Spectrum* allows us to recover the recombination events in each individual chromosome. Unlike other existing methods based on HMMs for recombination analysis which assume fixed recombination sites for the population and consider block-wise ancestors, we proposed a full generative model for haplotype inheritance which explicitly models the individual-level genetic recombination and mutation along the chromosome. Note that the recombination rate provided by *Spectrum* is defined with respect to the hypothetical founder pool and has not been modeled as a per-generation rate typically used in traditional recombination rate estimation models. Therefore, it is more suited for recombination hotspot analysis or for downstream applications that can benefit from the recombination block-structures such as admixture analysis or association studies.

As of now, *Spectrum* does not intrinsically capture the heterogeneity of recombination rates over loci, and the recombination rates are determined by the posterior distribution of recombination events under a universal recombination rate, rather than directly by a maximum likelihood estimation of site-specific recombination rates as in Li (2003). Also, we have not addressed the issues of threshold calculations and confidence measures of hotspot predictions as in Li (2003). These problems are of importance in various applications such as linkage-based quantitative train locus mapping and disease-gene mapping. One way of addressing these issues is to explicitly introduce more recombination states, for example, for both base-line recombination and hotspot-recombination, into the infinite HMM we proposed. Another possible extension to the existing model is to introduce priors for site-specific recombination rates for Bayesian inference.

Chapter 7

Robust estimation of local genetic ancestry in an admixed population

7.1 Introduction

The problem of inferring genetic ancestries in a population has been widely investigated for various applications such as disease gene mapping and population history inference. For example, the inferred ancestry information has been used in correcting the confounding effect by population stratification in association studies (Price et al., 2006; Wang et al., 2010). The examination of loci that have elevated probabilities of a specific ancestry has also given critical clues in selecting out potential causal variants of certain diseases in admixture mapping (Cheng et al., 2009, 2010; Zhu et al., 2011). Broadly, two different problem settings have been commonly considered for ancestral structure analysis (Alexander et al., 2009), one on the 'global ancestry' that considers the average proportion of each contributing population across the genome in an un-supervised way (Alexander et al., 2009; Falush et al., 2003; Patterson et al., 2006); and the other on the 'local ancestry' that is more concerned with a locus-by-locus ancestry given reference population data (Pasaniuc et al., 2009; Price et al., 2009; Tang et al., 2006). We consider the problem of estimating the local ancestry in an admixed population. A common scenario is to decompose the chromosomes of modern African Americans into blocks that have either African or European ancestry given the population data close to ancient African and European populations, which we call *ancestral populations*. The populations of CEU and YRI are the most typical choices for such ancestral population data when an *admixed population* of African Americans is considered. We present a new haplotype-based method for local ancestry estimation that can deal with an arbitrary number of ancestral populations in a non-parametric Bayesian framework.

A natural approach to this problem involves a Hidden Markov Model (HMM) that traces the ancestry of each individual along the markers on a chromosome. A number of different approaches have been proposed and theses methods can be largely categorized into two families depending on how they represent the ancestral populations so that the local ancestry in an admixed population can be estimated with respect to the reference information encoded by the population representation method. The first family of methods use population-specific allele frequency profiles as reference information as in traditional admixture studies (Alexander et al., 2009; Falush et al., 2003; Huelsenbeck and Andolfatto, 2007; Pritchard et al., 2000). Despite its simplicity, low computational cost, and availability of such frequency profiles in representative datasets, it is rather unnatural to model Linkage Disequilibrium (LD) under this setting because the correlations between loci are reflected only by the variation in such allele frequencies and not by the actual recombination events at the chromosome level. Therefore, either a subset of markers in low LD has to be selected in a preprocessing step, or a recombination process often needs to be indirectly embedded to utilize a denser set of markers (Pasaniuc et al., 2009; Patterson et al., 2004; Tang et al., 2006). The representation power of this family of methods thus tends to diminish when the correlations between markers are not carefully considered (Price et al., 2008).

Another family of methods are based on haplotype data that may contain richer information. These methods utilize representative haplotypes taken from each ancestral population data as reference information for the local ancestry estimation (Price et al., 2009; Sundquist et al., 2008). Each haplotype in an ancestral population, which we call an *ancestral haplotype*, constitutes a hidden state in an HMM and the basic transition mechanism involves traversing among these ancestral haplotypes. Although these approaches provide a more natural way to reflect the underlying admixing process by simulating recombinations at a real chromosome level, the inference result can be rather sensitive to the size and the choice of such ancestral haplotype data. Moreover, few existing methods make use of the genetic relatedness between ancestral populations resultant from ancient population history and therefore the populations have been typically treated as independent. To improve the robustness and the accuracy in light of these issues, HAPMIX (Price et al., 2009) introduces a 'miscopying' parameter that allows a small possibility for an allele to be copied from population 2 even when it is assumed to be originated from ancestral haplotype in population 1. In this way, it prevents unnecessary transitions among ancestral populations during inference and the allelic information in one population can be naturally borrowed by another population. However, this method is limited to two-way admixture that involves only two ancestral populations, and it is not trivial to generalize this model to consider more general demographic scenarios.

We propose a new Bayesian approach for local ancestry estimation that utilizes the multi-population haplotype data in a more systematic way. Our method is built on the assumption of a common pool of hypothetical *founder haplotypes* from which the ancestral haplotypes in multiple ancestral populations are to be inherited, and from which in turn the individuals in an admixed population are generated as well by the admixing process between ancestral populations. Motivated by the population model described in Chapter 6 and in Sohn and Xing (2007b, 2009), we represent the ancestral population data by an infinite hidden Markov model in which the hidden states correspond to the unknown number of hypothetical founder haplotypes. The recombination and mutation events are then modeled with respect to these founders as transition and emission process. For an individual in an admixed population, we extend the hidden state space to a joint space of founder haplotypes and ancestral populations. That is, we incorporate a hidden state variable consisting of two indicator variables, one for selecting the hypothetical founder haplotype and the other for selecting the ancestral population that it is originated from. The hidden state variable corresponding to the ancestral population determines the local admixing status and hence defines the local ancestry along the markers. Furthermore, population-specific time parameters are incorporated and scale the recombination rates in the corresponding populations accordingly to explain the gap between the hypothetical era of founder haplotypes and each of the ancestral populations. We observe that this also enhances the robustness of our model under scenarios that deviates from the common modeling assumption that all the populations participate in the admixture simultaneously.

A subtle issue in the proposed representation is how to choose the number of founders and how to construct them efficiently across multiple populations. Naïvely, we may assume K founders per population, but under this setting, not only one has to employ a non-trivial model selection process to determine K, but also there is in general no correspondence between the K founders in one population and another set of K founders in a different population. This problem would not only result in serious identifiability and multi-modality issue that can severely slow down inference, but also, it will restrict the information sharing across populations and hence compromise the accuracy of ancestry estimation as well. On the other hand, if we are to use one shared set of K founders, the representational power of population-specific HMM can also be limited. A non-parametric Bayesian framework using an infinite hidden Markov model gives a natural solution for this (Beal et al., 2002; Teh et al., 2010). Under an infinite HMM, an unbounded number of founder haplotypes can be systematically handled to describe a study population. If we employ multiple such infinite HMMs defined over the same set of founders, one infinite HMM per population, then it allows the founders to be shared between populations, while different populations do not have to include all these founders and can have a unique set of founders with its own frequency and recombination patterns among them. The number and the haplotypes of the founders are recovered as a result of posterior inference from data. Under a Dirichlet process prior, the posterior typically yields a parsimonious set of founders. This non-parametric Bayesian framework allows us to exploit the genetic relatedness between populations in a principled way by describing the ancestral populations in terms of a common set of founder haplotypes. In Sohn and Xing (2009), a similar approach using a hierarchical Dirichlet process has been successfully used for the problem of haplotype inference from multi-population data. However, the recombination process was not explicitly modeled in that work and a rather heuristic approach was employed to handle the linkage disequilbrium structure.

The proposed approach is fully model-based and fundamentally different from conventional haplotype-based approaches that model genetic processes directly on the given haplotypes in the ancestral population data. In our model, genetic processes such as recombination or mutation take place with respect to the hypothetical founders, and not between the ancestral haplotypes and the admixed individuals. By basing our model on the hypothetical founders that lie on top of and give rise to both ancestral and admixed population data, we utilize the multi-population data and their relatedness efficiently, unlike most existing approaches that ignore such information.

In summery, the proposed model-based approach for ancestral inference enjoys enhanced robustness and accuracy, evidenced by its substantially less sensitivity to the choice and the amount of ancestral population data comparing to other benchmark algorithms. In particular, our method shows very competitive performance even when the sample size of the ancestral population data is very small. This highlights the potential usefulness of this method in the analysis involving underrepresented populations of limited data availability. In addition, the compact population characterization by an infinite hidden Markov model improves the model flexility over existing approaches so that it can naturally handle an arbitrary number of ancestral populations instead of only two, and can be easily generalized to cases with even more complex demographic scenarios. It is also robust even under deviation from the typical modeling assumption that multiple populations participate in the admixture at the same time. Our method can utilize the whole admixed population data together to recover the model parameters such as the population proportion, or the time-scaled recombination rate, while most existing approaches require those parameters as input. The estimated parameters can reveal important clues on the history and the characteristics of the study population as will be shown in our empirical data analysis.

7.2 The statistical model

7.2.1 Problem setting

We consider an admixed population in which J ancestral populations have mixed since G generations ago. For example, if we are to recover the local ancestry of individuals in a Latino population (*admixed population*), we can incorporate $J = 3$ populations of ancient African, European, and Native American as our *ancestral populations*. In our problem setting, we assume that the haplotypes of single nucleotide polymorphisms are given for the ancestral populations and the admixed population. We will recover the pool of hypothetical founder haplotypes and their associations to individuals by statistical inference. The association of admixed individuals to the ancestral populations will be recovered along with their association to the founders, which would lead to the estimation of local ancestry.

7.2.2 Overview of admixture model based on founder haplotypes

The choice of representation about how to characterize a population is the crucial starting point in admixture modeling. Unlike most previous approaches that typically use allele frequency profiles (Pasaniuc et al., 2009; Sankararaman et al., 2008a) or representative ancestral haplotypes in their raw forms (Price et al., 2009; Sundquist et al., 2008), we employ a new haplotype-based method that builds on an assumption of hypothetical founder haplotypes of unknown cardinality. The founder-based population model with explicit recombination modeling has been introduced in Sohn and Xing (2007b) with the application to population structure and recombination analysis. Under this approach, each individual in a population is generated from the hypothetical pool of founders via a series of recombination and mutation. An individual chromosome can then be viewed as a mosaic of the founders whose pattern is determined by the association with founders. This mosaic process could be modeled as a Hidden Markov model in which the founders correspond to the hidden states, the individual haplotypes correspond to the observation sequences, the transition process is modeled by the recombination process, and the emission process by the mutation from founders to the individuals. By employing an infinite hidden Markov model, the number and the haplotypes of the founders can be recovered through posterior inference rather than being pre-specified; and the (local) inheritance association between the founders and the study individuals can also be derived.

Now we further extend this approach to model admixture events from an arbitrary number of ancestral populations. When the ancestral populations start to mix and form an admixed population, each individual chromosome in the admixed population can be decomposed into blocks with distinct ancestry. For each of these blocks, we can trace back the source of the

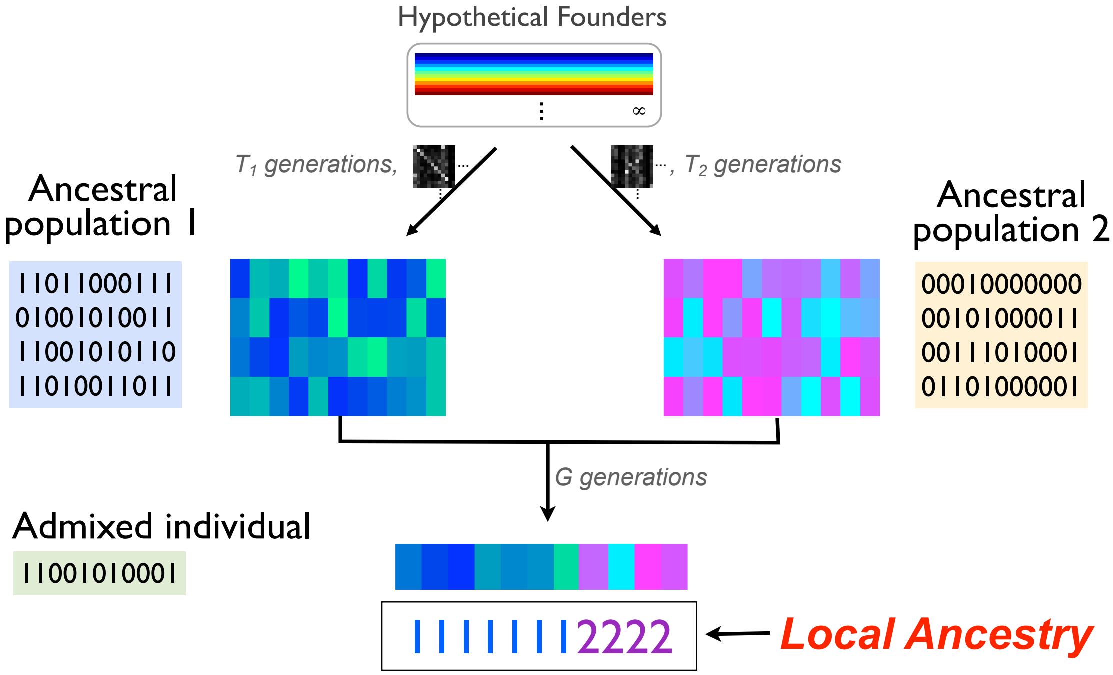


Figure 7.1: Graphical illustration of the proposed model

genetic materials to a haplotype in the corresponding ancestral population. Now, recall that this 'ancestral haplotype' is modeled as a mosaic of its founders. This means that each ancestry block in an admixed individual is further dissected into a finer-grained mosaic of founders. Therefore, the admixed inheritance process is a composite process with two different resolutions, one from the founders to ancestral haplotypes, and the other from the ancestral haplotypes to the admixed individuals. A graphical illustration of the proposed model is shown in Figure 7.1. A variant of the infinite hidden Markov model is employed to make the choice of founders and the ancestral populations at the same time along the chromosome.

7.2.3 Statistical model for generating ancestral and admixed population data

We now describe in detail the admixed inheritance model as a generative process of the individual chromosomes in ancestral populations and an admixed population with respect to a set of hypothetical founders.

Transition and emission probabilities

Our model involves a mosaic of two related jump processes of different genetic resolutions: recombination between founders, and admixture over ancestral populations. For ease of description, we assume that the individuals are haploids. Let individual haplotypes be indexed by i, ancestral populations by j, and the markers by t. And let $H_{it} \in \{0, 1\}$ and $A_{kt} \in \{0, 1\}$ represent the allele of individual i and founder k at marker t, respectively. We introduce a set of hidden state variables $S_{it} = (C_{it}, Z_{it})$ where $C_{it} \in \{1, 2, ...\}$ and $Z_{it} \in \{1, ..., J\}$ represent the indicator variables that select a founder and an ancestral population, respectively, on an i-th individual chromosome at marker t. For each ancestral population j, let ν_{jk} be the initial and background probability of founder k, and let $\pi^{j}_{k'k}$ be the transition probability that determines the recombination probability from founder k' to founder k. To take into account the different strength and pattern of the recombination across different populations, we also introduce a set of time parameters, $T_j \in (0, \infty)$, for each ancestral population j, such that T_j corresponds to the hypothetical time (generations) from the pool of founders to an ancestral population j. Similarly, $G \in [0, \infty)$ represents the time since admixture, that is, the time from ancestral populations to the admixed population. Let $\eta = (\eta_1, ..., \eta_J)$ denote the population proportion variable such that η_j is the expected proportion of ancestral population j in an admixed population. $\mathbf{r} = (r_1, r_2, ...r_T)$ and $\mathbf{d} = (d_1, \ldots, d_T)$ represent the recombination rate and the physical distance between each neighboring markers, respectively. The final transition probabilities and the emission probabilities are defined as follows:

$$\begin{align}
P(S_{i,0} = (k, j)) &= P(Z_{i,0} = j)P(C_{i,0} = k) = \nu_{jk}\eta_j \nonumber \\
P(S_{it} = (k, j) \mid S_{i,t-1} = (k', j')) &= (1 - e^{-r_t d_t G})\nu_{jk}\eta_j + \nonumber \\
&\quad e^{-r_t d_t G} e^{-r_t d_t T_j} \delta(k = k')\delta(j = j') + \nonumber \\
&\quad e^{-r_t d_t G}(1 - e^{-r_t d_t T_j})\pi^{j}_{k'k}\delta(j = j') \tag{7.1} \\
P(H_{it} \mid S_{it} = (k, j), A_{kt}) &= \theta_k^{I(H_{it} \neq A_{kt})}(1 - \theta_k)^{I(H_{it} = A_{kt})} \tag{7.2}
\end{align}$$

We assume a founder-specific mutation parameter θ_k that determines the probability of mutation from a founder k to individuals.

The overall idea underlying this representation is the two-layered inheritance framework, one from the time of hypothetical founders to ancestral populations, and the other from those ancestral populations to the admixed population. If we set $G = 0$ in Equation (7.1), this two-layered framework is reduced to the model of the first layer that characterizes the ancestral populations with respect to the founder haplotypes. Under the reduced model, each population is associated with its own hidden Markov model parameters and the recombination rate scaled by T_j. Suppose $(C_{i,t-1}, Z_{i,t-1}) = (k', j')$ which means i-th chromosome has inherited from founder k' at marker $t - 1$ in ancestral population j'. At the next marker t, it either selects a new founder k with probability $(1 - e^{-T_j r_t d_t})\pi^{j}_{k',k}$ and set $C_{it} = k$, or no recombination takes place with the remaining probability and $C_{it} = C_{i,t-1}$. If we trace the values of C_{it} across all the t, it will decompose the chromosome i into blocks with distinct associated founders. Therefore, each chromosome can be thought of as a mosaic of such founders.

Now, at the second layer which involves the admixture, this sequential process for selecting founders C_{it} occurs within the same ancestral population with probability $e^{-r_t d_t G}$ so that $Z_{it} = Z_{i,t-1}$. Or with probability $(1 - e^{-r_t d_t G})$, a new population j as well as a new founder k is chosen jointly with a probability proportional to the product of population proportion η_j and the background probability ν_{jk}. Therefore, chromosomes both in the ancestral populations and in the admixed population are modeled as mosaics of founders determined by the sequence of C_{it}. In addition, each admixed individual i is associated with another resolution of mosaic determined by the sequence of Z_{it} across t. The estimation of local ancestry can be done by tracing the posterior probability of Z_{it} along the markers.

Note that even when no admixture is assumed, we still have the flexibility of choosing a different founder chromosome. This feature helps to control the number of transitions among populations effectively so that the hidden state doesn't need to change excessively. Moreover, the population-specific time parameters scale the recombination probabilities accordingly so that more ancient populations can have higher probabilities of recombination with respect to the founders. This representation is especially useful in producing heterogenous resolution of mosaics in different ancestral populations. Although we assume the J populations participate in the admixture simultaneously, those parameters allow various resolution of recombination patterns in ancestral haplotypes and this greatly improves robustness of the model against the violations of such modeling assumption as well as the accuracy of the ancestry estimation.

The cardinality of the founder space

Instead of fixing the number of hypothetical founders by doing statistical model selection, we adapt a more flexible non-parametric approach by employing an infinite hidden Markov model (Beal et al., 2002; Teh et al., 2010) so that the number of hidden states does not need to be prespecified. Recall that if we consider a finite, say K, hidden states, the transition probabilities will be represented as a $K \times K$ matrix. Each row k of this matrix sums to one and defines the probabilities of switching from a source state k to all the target states.

Now, if we consider an infinite hidden state space, each row of the transition matrix would be an infinite dimensional vector which sums to one and the Dirichlet Process (DP) (Blackwell and MacQueen, 1973; Ferguson, 1973) has been effectively used to describe such probability distributions. To ensure all the row-specific DPs are built on the same state space, another Dirichlet Process is shared as a common base measure at a top level, which actually corresponds to a hierarchical Dirichlet Process model (Teh et al., 2010). Basically, (k, k')-element of the transition matrix π^j defines the transition probability from state k to state k' in population j, and for a given source state k, the target state index k' can increase as large as needed by the given data. Infinite-dimensional vector of initial probabilities ν_j can be defined in a similar way under the same hierarchical Dirichlet process framework. Since we consider multiple such infinite HMMs for multiple populations, we let the same base measure shared across all the populations. This infinite HMM-based framework leads to a very simple solution to how many founders to consider and how to construct the founder space across multiple populations. The HMM parameters of our admixture model thus can be summarized as follows:

$$\begin{aligned} \beta &\sim GEM(\gamma) \\ \nu_j &\sim DP(\tau, \beta) \\ \pi^j_k &\sim DP(\tau, \beta) \end{aligned}$$

where τ and γ define the scale parameters for the population-specific DPs and the top level DP, respectively.

Other parameter description

We assume Beta prior for each of the mutation parameters θ_k, and Dirichlet distribution prior for the population proportion parameter $\eta \sim Dirichlet(\xi_1, ..., \xi_J)$.

For simplicity of inference, we transform the variables such that r_t and T_j are combined as $g^r_{jt} = r_t T_j$. Similarly, we use the notation $G^r_t := r_t G$. We assume these variables are *i.i.d* under Gamma prior. Then Equation (7.1) is transformed as follows:

$$\begin{aligned} P(S_{it} = (k,j) \mid S_{i,t-1} = (k',j')) \;=\; & e^{-G^r_t d_t} e^{-g^r_{jt} d_t} \delta(k = k')\delta(j = j') + \\ & e^{-G^r_t d_t}(1 - e^{-g^r_{jt} d_t})\delta(j = j')\pi^j_{k'k} + \\ & (1 - e^{-G^r_t d_t})\nu_{jk}\eta_{ij} \end{aligned} \qquad (7.3)$$

In summary, infinite hidden Markov model parameters combined with population genetics parameters are used to capture different characteristics in populations and to describe admixture event from an arbitrary number of populations in a unified framework. While we assume an infinite number of founders a priori, the posterior inference usually produces a small number of founders and this gives a compact representation of populations for the admixture analysis.

7.2.4 Posterior Inference

To overcome the drawbacks of slow convergence in traditional Gibbs sampling, we employ a variant of beam sampling proposed for infinite HMMs (Van Gael et al., 2008). Basically, it extends the well-known dynamic programming technique of the forward-backward algorithm in a finite state HMM to an infinite state space case. It exploits the property that in an observation sequence of finite length, the number of actually realized hidden states is finite at each iteration step. Therefore, the number of states to be considered in forward-backward algorithm can be adaptively changed over iterations.

Forward-backward algorithm for the proposed infinite HMM

We introduce auxiliary variables u_t for $t = 0, ..., T-1$ with the following distribution:

$$\begin{aligned} u_{i0} \mid S_{i0} = (k,j) \;&\sim\; \text{Uniform}(0, \nu_{jk}\eta_{ij}) \\ u_{it} \mid S_{it} = (k,j), S_{i,t-1} = (k',j') \;&\sim\; \text{Uniform}(0, q_{it}) \qquad \text{for } t = 1, ..., T-1 \end{aligned}$$

where

$$\begin{aligned} q_{it} \;=\; & e^{-G^r_t d_t} e^{-g^r_{jt} d_t}\delta(k = k')\delta(j = j') + \\ & \quad e^{-G^r_t d_t}(1 - e^{-g^r_{jt} d_t})\delta(j = j')\pi^j_{k'k} + (1 - e^{-G^r_t d_t})\nu_{jk}\eta_j \end{aligned}$$

For notational convenience, we omit the notation i. Let the forward probabilities be

$$\alpha_t(k,j) = P(S_t = (k,j) \mid H_{0:t}, u_{0:t})$$

Then

$$\begin{aligned}
\alpha_0(k,j) &\propto P(S_0=(k,j),H_0,u_0) \propto P(S_0=(k,j))P(u_0 \mid S_0=(k,j))P(H_0 \mid C_0=k) \\
&= \delta(u_0 < \nu_{jk}\eta_{Z_0})P(H_0 \mid C_0=k) \\
\alpha_t(k,j) &\propto \sum_{k',j'} P(S_t=(k,j),S_{t-1}=(k',j'),H_t,u_t \mid H_{0:t-1},u_{0:t-1}) \\
&\propto P(H_t \mid C_t=k)\sum_{k',j'} P(u_t \mid S_t=(k,j),S_{t-1}=(k',j')) \times \\
&\quad\; P(S_t=(k,j) \mid S_{t-1}=(k',j'))\alpha_{t-1}(k',j') \\
&\propto P(H_t \mid C_t=k) \times \\
&\quad\; \sum_{j'=0}^{J-1}\sum_{k'=0}^{\infty} \delta(u_t < P(S_t=(k,j) \mid S_{t-1}=(k',j')))\alpha_{t-1}(k',j') \qquad (7.4)
\end{aligned}$$

Given $u_0, ..., u_{T-1}$, the number of states k such that $\alpha_t(k,j) > 0$ for $t = 0, ..., T-1$ is finite: for $t = 0$, the number of k such that $\nu_{jk} > u_0$ is finite for any j since $\sum_k \nu_{jk} = 1$ with $\nu_{jk} \geq 0$, and recursively, we can see the number of k with $\alpha_t(k,j) > 0$ is finite. Therefore, the infinite sum over the previous states in the calculation of forward probability reduces to a finite sum.

C_{T-1} and Z_{T-1} can be sampled from $\alpha_{T-1}(k,j)$. Then for $t = T-2, ..., 0$, we sample C_t and Z_t using

$$\begin{aligned}
&P(C_t, Z_t \mid H_{0:T-1}, u_{0:T-1}, C_{t+1}, Z_{t+1}) \propto \\
&\qquad P(C_{t+1}, Z_{t+1} \mid C_t, Z_t)\alpha_t(C_t, Z_t)P(u_{t+1} \mid S_t, S_{t+1})
\end{aligned}$$

Since the entire inheritance process from founders to ancestral populations and then the admixed population is modeled in a single Bayesian framework, it allows the exact posterior inference by putting the ancestral and admixed population data together in a single series of beam sampling iterations described above. However, this is not optimal in terms of time complexity as we often favor to run multiple test sets after we get reference information about the ancestral populations. Therefore, we split the whole inference process into two phases: 1) training phase where the model parameters about ancestral populations are learned, and 2) ancestry estimation phase that actually recovers the ancestry of admixed individuals.

One caveat of this decomposition is that we may not fully take advantage of the flexibility of the infinite model. This is because we need to constrain the hidden state space somehow as a finite space when the output from the training phase is returned. As an n-th posterior sample from Bayesian inference of the training phase, we get a finite number $K^{(n)}$ of founder haplotypes and the related HMM parameters of $\pi^{(n)}$ and $\nu^{(n)}$ with $g_j^{r(n)}$ for each j. Averaging these results as one training output is not straightforward as $K^{(n)}$ can be different across different n. A plausible approach would be to keep multiple, say N posterior samples $\mathbb{S} = \{\mathbf{A^{(n)}}, \pi^{\mathbf{(n)}}, \nu^{\mathbf{(n)}}\}_{n=1,...,N}$ and run the ancestry estimation routine N times using each of these parameters in $\mathbb{S}$. Then the N posterior distributions of the ancestry indicator variable Z can be easily averaged to form the final posterior distribution since Z is defined over a fixed number of populations J unlike C or other parameters that depend on K. Note that $g_j^{r(n)}$ does not depend on K, so we can use the posterior mean of $g_j^{r(n)}$ as the final estimate for it. Another practical approach would be to select a single output from the training phase such as a MAP solution, and estimate the local ancestry based on the single set of parameters. Empirically, we observe that the performance degradation by this MAP solution with respect to the first approach is relatively small.

Training phase

For an individual in an ancestral population j, we can set the time since admixture G to be zero and the population indicator variables Z to be observed as constant. Then the hidden state variable $S_{it} = (C_{it}, Z_{it})$ can be replaced with a C_{it} indicating the founder and Equation (7.3) is reduced to the followings :

$$\begin{aligned} P(C_{i0} = k) &= \nu_{Z_{i0}k} \\ P(C_{it} = k \mid C_{i,t-1} = k') &= e^{-g^r_{Z_{i0}t}d_t}\delta(k = k') + (1 - e^{-g^r_{Z_{i0}t}d_t})\pi^{Z_{i0}}_{k'k} \end{aligned}$$

We infer the variable C through the Beam sampling algorithm, and the other variables through the standard Gibbs sampling. If we reduce the model to the training phase, we can treat the variable Z as observed. Therefore, the forward probabilities are written as follows:

$$\begin{aligned} \alpha_0(k) &\propto P(C_0 = k, H_0, u_0) \propto P(C_0 = k)P(u_0 \mid C_0 = k)P(H_0 \mid C_0 = k) \\ &= \delta(u_0 < \nu_{Z_0k}\eta_j)P(H_0 \mid C_0 = k) \\ \alpha_t(k) &\propto \sum_{k'} P(C_t = k, C_{t-1} = k', H_t, u_t \mid H_{0:t-1}, u_{0:t-1}) \\ &\propto P(H_t \mid C_t = k) \sum_{k'} P(u_t \mid C_t = k, C_{t-1} = k')P(C_t = k \mid C_{t-1} = k')\alpha_{t-1}(k') \\ &\propto P(H_t \mid C_t = k) \sum_{k'=0}^{\infty} \delta(u_t < P(C_t = k \mid C_{t-1} = k'))\alpha_{t-1}(k') \end{aligned} \tag{7.5}$$

Note that the contribution of transition at each neighboring loci $t-1$ and t to the parameter π and g^r_{jt} is not all equal because of the self-transition probability forced by the recombination model in Equation (7.3). We handle this by sampling auxiliary binary variables $M_{it} \sim Bernoulli(1 - e^{-g^r_{Z_{i0}t}d_t})$ to indicate whether the jump occurs in the transition or not. The transition probability can be decomposed as follows:

$$P(C_{it} \mid C_{i,t-1}) = P(M_{it} = 0)\delta(C_{it} = C_{i,t-1}) + P(M_{it} = 1)\pi^j_{C_{i,t-1},C_{it}}$$

Then we sample M_{it} given C_{it} and $C_{i,t-1}$ backward in forward-backward process from

$$P(M_{it}|C_{it} = (k, j), C_{i,t-1}) \propto P(M_{it})P(C_{it} = k \mid C_{i,t-1} = k', M_{it})$$

Now, π can be sampled as in Van Gael et al. (2008), but conditional on M, which involves the transitions with $M_{it} = 1$ only. g^r_{jt} can also be sampled conditional on M using $P(g^r_{jt} \mid \{C_{:t}, C_{:,t-1}, M_{:t}\}) \propto P(g^r_{jt}) \prod_{i \in Pop\ j} P(C_{i,t} \mid C_{i,t-1}, M_{it})$. The overall sampling procedure is summarized in Algorithm 1.

Algorithm 1 Procedure for training iHMMs in reference populations

Input: Haplotype data H for ancestral populations

Output: N posterior samples of founders and the related HMM parameters $\{\mathbf{A^{(n)}}, \pi^{(\mathbf{n})}, \nu^{(\mathbf{n})}, \mathbf{g^{r(n)}}\}$ for $n = 1, \ldots, N$

```
repeat
    for each individual chromosome $i$ do
        Sample the auxiliary variables $u_{it}$ for $t = 0, ..., T-1$.
        Sample $C_{it} \mid u, H, A$ using the beam sampling algorithm
        Sample $A_{k,t}$ and $\theta_k$
        Sample parameters $\nu, \pi, \beta$ and $g^r$.
    end for
until convergence
```

Algorithm 2 Procedure for estimating local ancestry in an admixed individual

Input: Haplotype data H for an admixed population, estimated parameters $\{\mathbf{A^{(n)}}, \pi^{(\mathbf{n})}, \nu^{(\mathbf{n})}, \mathbf{g^{r(n)}}\}$

Output: Posterior distribution of $Z = (Z_{it})$.

```
for $n = 1, \ldots, N$ do
    repeat
        for each individual chromosome $i$ do
            Sample $S_{it} = (C_{it}, Z_{it}) \mid H, A$ using the forward-backward algorithm
            Sample $\theta$, $\eta$, and $G^r$ .
        end for
    until convergence
    Keep $S$ posterior samples of $Z$
end for
Average $N \cdot S$ posterior samples and return the final posterior distribution of $Z$
```

Ancestry estimation phase

As the variables A, g^r, ν, π are returned in the training stage, the unknown variables now are the admixture proportion η, the generations since admixture G, the mutation rate θ of founders, and $S = (C, Z)$ for the admixed individuals. We re-sample θ in the ancestry estimation phase instead of getting it from the training step because θ can reflect additional information about the admixed population by describing it in terms of the discrepancy between founders and the population. As we now deal with a finite number of hidden states obtained from the training phase, it is not necessary to incorporate the auxiliary variable u to sample S in the ancestry estimation phase. The variables S_{it} thus are sampled through a standard forward-backward algorithm. As in the training stage, the transition probability at each marker can be decomposed into two parts, depending on whether the jump process for admixture occurs or not. We use the similar technique to sample G^r by introducing an auxiliary variable $L_{it} \sim Bernoulli(1 - e^{-G_t^r d_t})$. The overall sampling scheme is summarized in Algorithm 2.

If the time since admixture G, population proportion η, and the recombination rate r is assumed to be known as is often the case in typical admixture analysis, we can omit the second step of parameter sampling (line 5 in Algorithm 2) and re-use θ that can be returned from the training stage. Then it is also possible to get an approximate solution by use of a posterior decoding from forward-backward steps in a finite dimensional HMM.

7.3 Result

7.3.1 Simulation design

To validate the proposed method, we simulated admixed individuals using Human Genome Diversity Project (HGDP) data genotyped on Illumina Infinium HumanHap550 BeadChips (Jakobsson et al., 2008). Considering previous results that have revealed distinct genetic characteristics across different continents, we selected reference populations that would serve as putative ancestral populations: YRI for African ancestry, CEU for European, JPT and CHB for East Asia, and Maya for Native American ancestry. Each of the resulting ancestral populations contained 30, 30, 28, and 13 individuals, respectively. We first focus on chromosome 22 in the simulation study.

To take into account the discrepancy between real ancestral populations and those used in training, we generated admixed individuals using populations which are similar but not identical to those used as ancestral populations. For example, individuals in Russian and BantuKenya populations are mixed to simulate an admixed population and then the local ancestries of these individuals are estimated with respect to CEU (European) and YRI (African) populations. For each simulation scenario below, we generate 30 admixed individuals.

The performance is measured as the mean squared error rate of ancestry probabilities along the loci. Specifically, let p_{ijt} denote the probability of ancestry j at a locus t in an individual i. The average error rate of $\sum_{j=1}^{J}\sum_{t=1}^{T}(p_{ijt}^{true} - p_{ijt}^{est})^2/T$ across all the individuals is reported. We compare our results with the two state-of-the-art methods that stem from different population representation methods : LAMP (Pasaniuc et al., 2009; Sankararaman et al., 2008b), the method based on allele frequency profiles as reference information, and HAPMIX (Price et al., 2009) that uses representative individual haplotypes in the ancestral populations, which appear to outperform other available methods such as HAPPA (Sundquist et al., 2008), SABER (Tang et al., 2006) or ANCESTRYMAP (Patterson et al., 2004) from previous studies. Since these benchmark algorithms require the parameters for recombination r, the admixture time G, and the population proportion η to be specified as input, we provided the true values of these parameters to all the algorithms in the simulation study. Additionally, each haplotype data for ancestral populations were converted to allele frequency profiles and then LAMP was run with these frequency data as input. For the analysis below, we used the MAP solution as our parameter estimation from the training phase.

7.3.2 Performance on two-way admixture

The first simulation scenario considers two-way admixture of ancient European and African populations. The proportion of each ancestral population was set to be equal ($\eta = (0.5, 0.5)$).

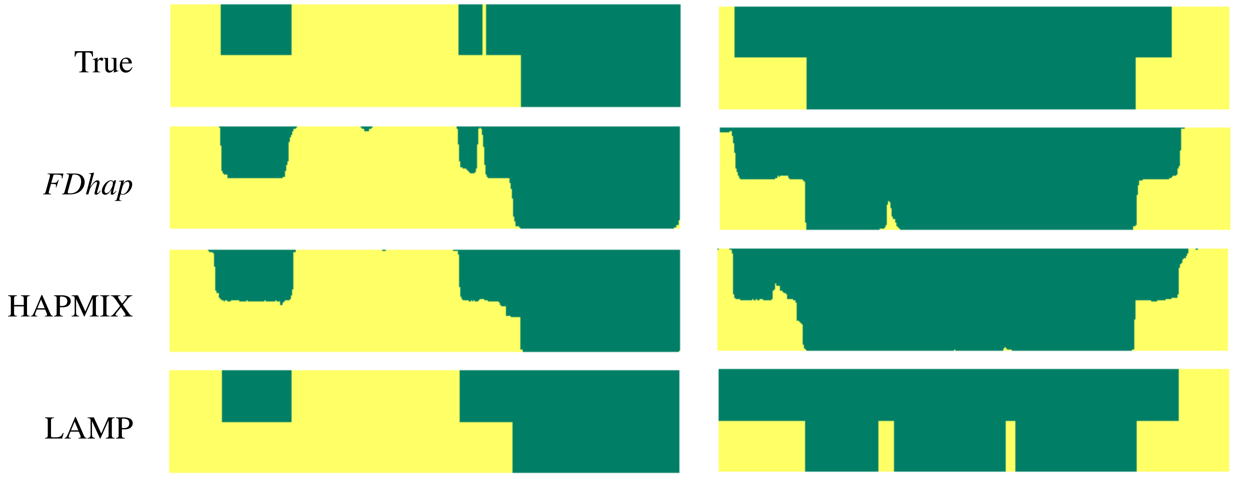


Figure 7.2: True and estimated local ancestries of two sample individuals in an admixed population from African and European populations. The x-axis corresponds to chromosomal position and the y-axis corresponds to the ancestry probability (yellow: African, dark grean: European)

The local ancestries of the admixed individuals were estimated based on the trained model using two ancestral populations of African (YRI) and European (CEU). In Figure 7.2, we first display the true and the estimated local ancestry probabilities of two sample individuals in an admixed population. The yellow color corresponds to African ancestry, and the dark green corresponds to European ancestry. The length of the vertical color bar at each chromosomal location along the x-axis is proportional to the corresponding ancestry probability. While all the algorithms produce reasonable results in general, the proposed method denoted by FDhap (FounDer haplotype based admixture model) is especially effective in picking out fine details of ancestry changes as can be seen in the example.

The overall performance of each algorithm across all the generated samples are shown in Figure 7.3. Roughly, we can see that FDhap and HAPMIX perform comparably to each other and tend to outperform LAMP. Still, all the three algorithms perform reasonably well as can be seen in the small overall error rates. For example, the average error rates for $G = 10$ were 0.0077, 0.0086, and 0.0116 in FDhap, LAMP, and HAPMIX, respectively.

7.3.3 Performance as a function of data size in training set

To further evaluate each method in terms of its performance with respect to the training data size, we varied the number of available individual samples per ancestral population. Specifically, we trained the model using 3, 5, 10, 20, 30 individuals, hence, 6, 10, 20 40, 60 haplotypes, per ancestral population and estimated the ancestries based on each of the trained model. The same two-way admixture scenario from African and European populations is considered of which the result on the full dataset is shown in Figure 7.3. The performance of each algorithm is presented as a function of training data size in Figure 7.4. It is clearly seen that the proposed method substantially outperforms the other benchmark algorithms, especially when the data size is small. Even when only a few ancestral haplotypes are available, it still gives very good estimates of the

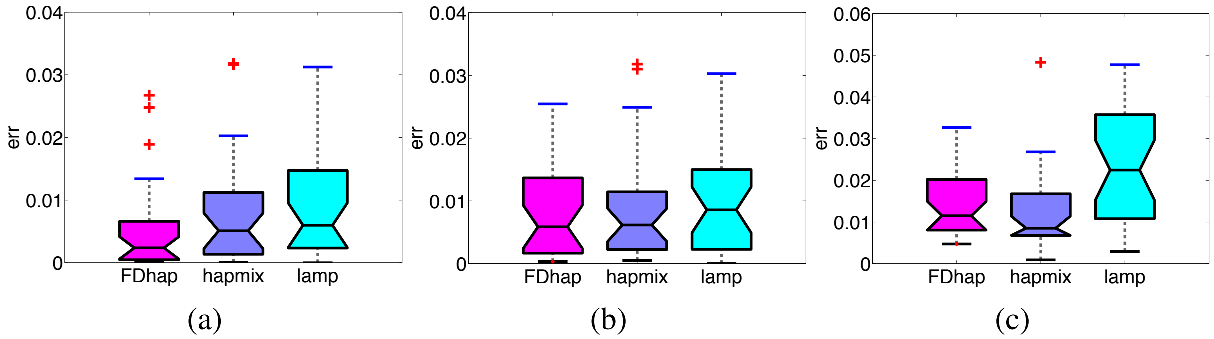


Figure 7.3: Boxplot for mean squared error rates of ancestry estimation for two-way admixture of African and European populations since G generations ago with (a)$G = 5$, (b) $G = 10$, and (c) $G = 20$.

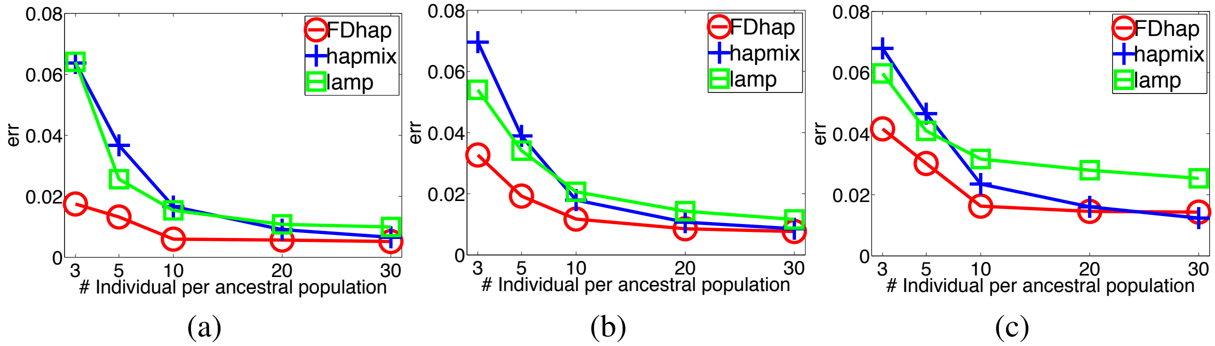


Figure 7.4: Error rate as a function of the number of individuals per train population. Two-way admixture of African and European popualtions since G generations ago with (a) $G = 5$, (b) $G = 10$, and (c) $G = 20$.

local ancestries compared to the others. Therefore, our method can be especially useful in the analysis of admixture effect involving non-traditional populations where the amount of available genotypes is still limited.

7.3.4 Performance on three-way admixture

We now consider the admixture that involves more than two ancestral populations. Analogous to the formation of Puerto Rican population (Tang et al., 2007), we included CEU, YRI, and Maya populations as ancestral populations for African, European, and Native American ancestry, and generated an admixed population using Russian, BantuKenya, and Pima with admixing proportion of 0.66, 0.18, and 0.16, respectively. Figure 7.5 shows the resulting error rates across different values of G. Since HAPMIX cannot handle more than two ancestral populations directly, we ran it in three different modes such that each run tries to estimate the targeted ancestry versus the other two ancestries as was done its original paper (Price et al., 2009). For this reason, we compare the performance on each ancestry separately. Overall, our method performs significantly better than the other two in most of the analyzed cases.

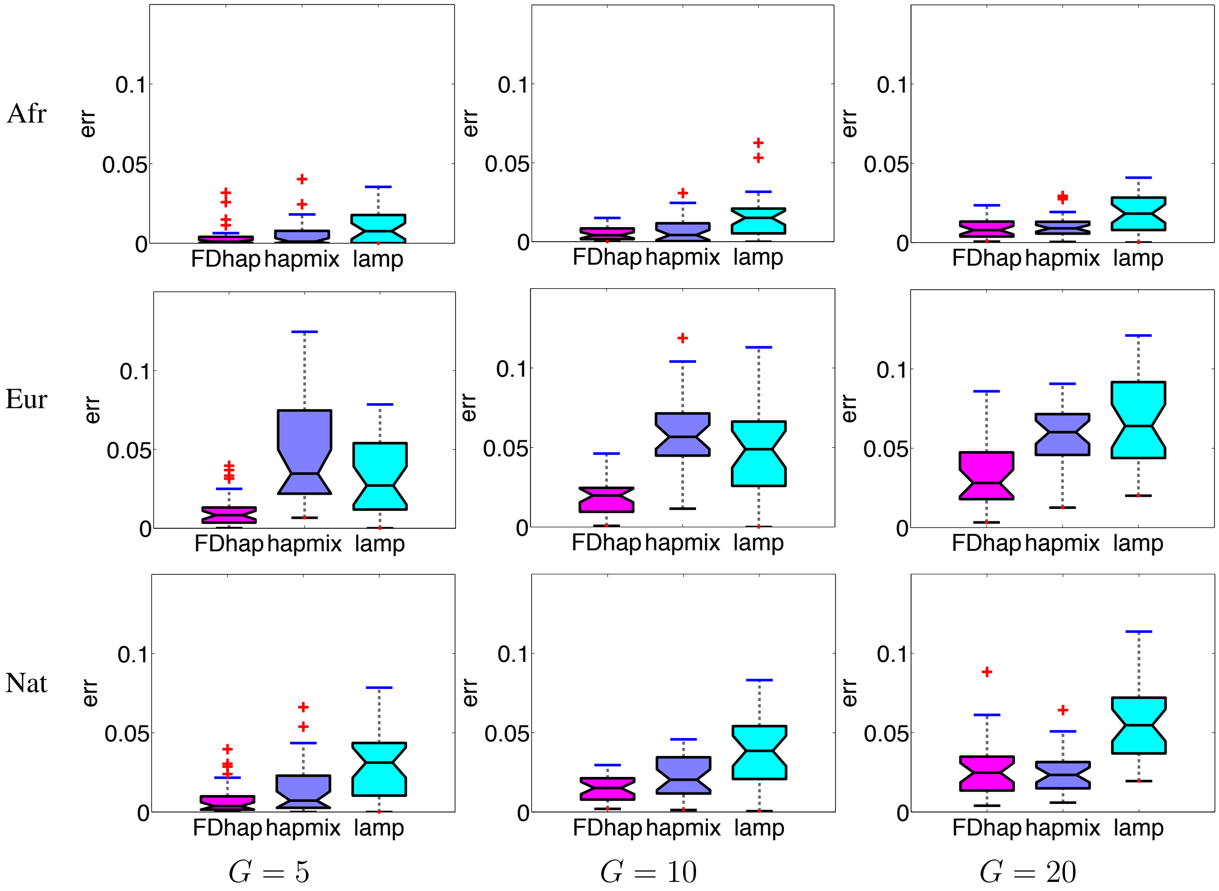


Figure 7.5: Boxplot for mean squared error rates of ancestry estimation. Three-way admixture of African, European, and Native American populations since G generations ago. Since HAPMIX is applicable to only two-way admixture case and was run to estimate each ancestry versus the other two, we report the error rate on each ancestry separately.

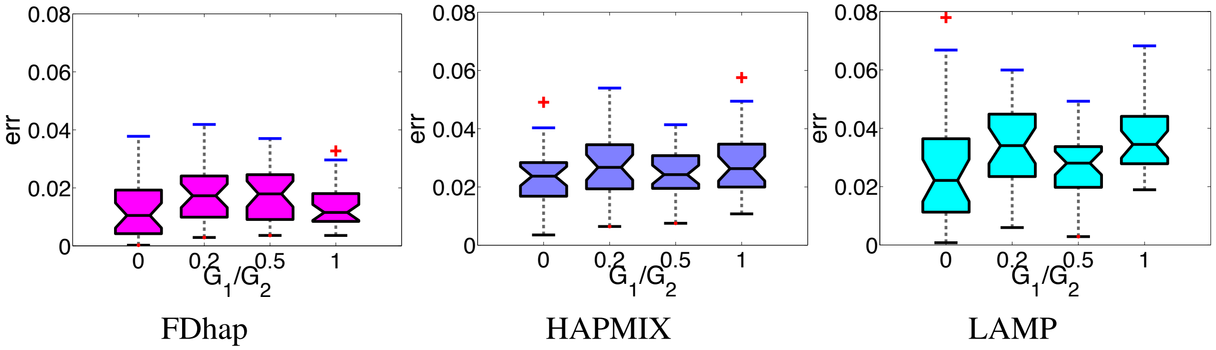


Figure 7.6: Robustness under deviation from the modeling assumption. The x-axis represents the ratio G_1/G_2, where G_1 denotes the number of generations for which the first two populations had mixed and G_2 means the additional number of generations since the third population joined and have further mixed together.

7.3.5 Robustness under deviation from admixture assumption

We investigate the robustness under deviation from the modeling assumption that all the ancestral populations participate in the admixing simultaneously. We generated admixed populations from three ancestral populations that started to mix at two different time points. More specifically, Russian and BantuKenya populations are mixed for G_1 generations with 50%/50% proportion. Then this admixed population is mixed with the third population of Pima for G_2 generations with 50%/50%, resulting in the overall proportion of 0.5, 0.25, 0.25. We fixed G_2 to be 10 and varied G_1 to be 0, 2, 5, and 10 where $G_1 = 0$ corresponds to the case in which the modeling assumption holds. The result is summarized in Figure 7.6. In each plot for each algorithm, x-axis corresponds to the values of G_1/G_2 and y-axis shows the error rates. The proposed method resulted in not only the lowest error rates, but also the most stable performance across different values of G_1/G_2. For more quantitative comparison of robustness across different algorithms, we calculated the linear regression coefficient of G_1/G_2 versus the error rates. The resulting slopes were -0.0011, 0.0029, and 0.0074 for FDhap, HAPMIX, and LAMP, which again supports the superior robustness of the proposed method.

7.3.6 Sensitivity analysis on model parameters

Since the parameters of η and G were assumed to be known in our simulation study in parallel with other methods, we also examine how the performance of FDhap is affected by incorrectly specifying these parameters. The performance is shown for the dataset simulated with $G = 10$ and $\eta = (0.5, 0.5)$ in Figure 7.7. In each plot, x-axis shows the specified parameters where the values are shown in log scale in case of G. We could see that there was almost no effect when η was incorrectly set in the range from 0.2 to 0.8. When we examined the result on G, the algorithm had the general tendency to favor a specified value G smaller than the true value. The effect of mis-specified value of G was minimal when the discrepancy was within a factor of 2. Even in the extreme case such as G varied by a factor of 5, the error still remained within the twice of the error rates when the true value was given.

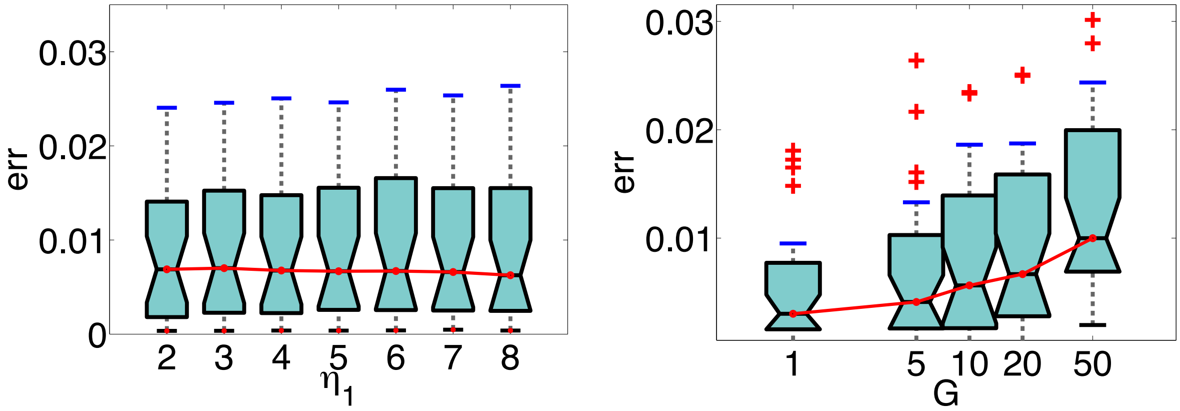


Figure 7.7: Sensitivity analysis: boxplot for error rates as a function of specified parameter values (a) η_1 and (b) G when the true values are $\eta_{true} = (0.5, 0.5), G_{true} = 10$.

7.3.7 Empirical analysis of HGDP data

To illustrate our method on real data, we applied it to 22 autosomes of the HGDP dataset (Jakobsson et al., 2008). Four ancestral populations of YRI, CEU, JPT+CHB, and Maya were chosen as in the simulation study to represent African, European, East Asian, and Native American ancestries. We then recovered the local ancestries in the remaining 28 populations. Since the time since admixture is not available for real data, we let our program estimate the parameters by posterior inference.

The mean ancestry proportion of each population estimated from our algorithm is summarized in Table 7.1. Overall, the ancestry vector agrees very well with their geographical locations or known history. For example, populations such as Yoruba, Mandenka, BiakaPygmy, or BantuSouthAfrica recovered pure African ancestries, Druze, Basque, Russian and Adygei populations had dominant European ancestries (≥ 0.978), and Pima or Colombian populations resulted in almost pure Native American ancestries (≥ 0.983).

More interestingly, the result also identifies the populations that have strong evidence of admixing effect among multiple ancestries. For instance, the proportion of European ancestry in Uygur population was 0.35, that of East Asian ancestry was 0.41, and the remaining proportion of 0.24 in Native American ancestry. Previous analysis in Xu et al. (2008) and Xu and Jin (2008) claimed that Uygur had roughly 50–60% of European ancestry and 40–50% of East Asian ancestry from the analysis based on two-way admixture. More recent study in Li et al. (2009) showed evidences that the estimation of European ancestry in these studies appear to be biased and suggested a newly estimated proportion of around 30%. Our result largely agrees with these results in that the estimated East Asian ancestry (41%) is similar to that in Xu et al. (2008) and in addition the estimation of European ancestry (35%) is closer to the more recent result in Li et al. (2009) than Xu et al. (2008). Considering its geographical location and the resulting population history, we suggest that Uygur population has about 35% of European ancestry, 41% of East Asian ancestry, and the remaining proportions of ancestries in other contributing populations that have greater similarity to the Native American population. Although only one or two populations are selected to serve as each putative ancestral population in our study, these populations have

Table 7.1: Estimated ancestry proportions of populations in HGDP dataset with respect to four ancestral populations of African, European, East Asian, and Native American.

	African	European	East Asian	Native Amer
Yoruba	1.000	0.000	0.000	0.000
Mandenka	1.000	0.000	0.000	0.000
BiakaPygmy	1.000	0.000	0.000	0.000
BantuSouthAfrica	1.000	0.000	0.000	0.000
San	0.999	0.001	0.000	0.000
MbutiPygmy	0.999	0.000	0.000	0.001
BantuKenya	0.998	0.001	0.000	0.000
Mozabite	0.141	0.818	0.013	0.028
Bedouin	0.035	0.941	0.006	0.018
Palestinian	0.013	0.966	0.006	0.015
Basque	0.000	0.998	0.000	0.001
Russian	0.000	0.990	0.003	0.007
Druze	0.002	0.989	0.002	0.006
Adygei	0.000	0.978	0.008	0.014
Kalash	0.000	0.930	0.027	0.043
Balochi	0.015	0.888	0.031	0.066
Burusho	0.000	0.741	0.088	0.170
Uygur	0.000	0.348	0.414	0.239
Yakut	0.000	0.045	0.848	0.106
Mongola	0.000	0.006	0.960	0.034
Daur	0.000	0.004	0.972	0.024
Cambodian	0.000	0.004	0.977	0.019
Lahu	0.000	0.000	0.987	0.013
Yi	0.000	0.001	0.991	0.009
Melanesian	0.001	0.039	0.821	0.140
Papuan	0.002	0.081	0.733	0.185
Pima	0.001	0.012	0.004	0.983
Colombian	0.002	0.001	0.001	0.996

shown to be close surrogates of the distinct ancestral components in a large number of studies so far. Since our result is obtained by utilizing a denser set of markers as well, we believe our estimates reveal more meaningful admixture proportions than the previous analyses.

To further analyze each population data, we examined the estimated parameters of $\hat{G}^r$, the admixture time scaled by the recombination rate, and the empirical mutation parameter $\tilde{\theta}$ computed as an average discrepancy between individuals and corresponding founders within each of the populations. Note that we can think of $\hat{G}^r$ as showing the relative strength of admixing effect in the population. Moreover, $\tilde{\theta}$ can be interpreted as the relative age of the population because it describes the gap between the common founders to the population. The result is displayed in Figure 7.8. We colored the bars based on the geographic location of the corresponding population. The result has nice correspondence with the known geographic labels as well. For example, all the populations in African continent showed the smallest values of $\hat{G}^r$, indicating the lowest levels of admixing effect. The top three populations suggesting the strongest admixing effect include Mozabite, Uygur, and Burusho. Most populations in the continents of East Asia and America showed medium levels of admixing effect.

The empirical $\hat{\theta}$ estimated for the data reveals more striking pattern. We find that the ordering of populations by their parameter values almost exactly agrees with the geographic locations out of Africa. That is, all the populations in African continent had the largest values of $\hat{\theta}$ implying their oldest ages, populations in Eurasia came next, and Oceanian populations were the third. Populations in East Asian region formed the fourth cluster and then Pima and Colombian populations showed the smallest values of $\hat{\theta}$ which implies later formation of the populations than others. It is noteworthy that Yoruba, which appears to be the closest to the training population of YRI, recovers a much larger value of mutation rate $\hat{\theta}$ than all the populations in geographic locations other than African continent. This comes from the nice property of our model that we do not directly use the training haplotype data as our reference, we rather infer the corresponding common founders across all the population data together and then work in a framework dealing with founders and admixed individuals. Otherwise, it would be impossible to obtain such a result because the discrepancy of Yoruba and its reference data would be much smaller than most of the other populations.

Based on the analysis above, we selected 11 populations that showed elevated signals of admixture and show their local ancestry proportions in Figure 7.9 that can be used for further downstream analysis.

7.4 Discussion

We have proposed a new haplotype-based framework for modeling the admixing events over ancestral populations and estimating the local ancestries of the individuals in the admixed population. By modeling hypothetical founders and their inheritance processes, and by using population-specific hidden Markov models to represent the ancestral populations with an unbounded number of founders, our method can lead to accurate stratification of local ancestry of individual chromosomes in an admixed population.

Previous admixture studies have suggested that the world populations are not independent of each other, but rather are structured through population admixing history and the resulting

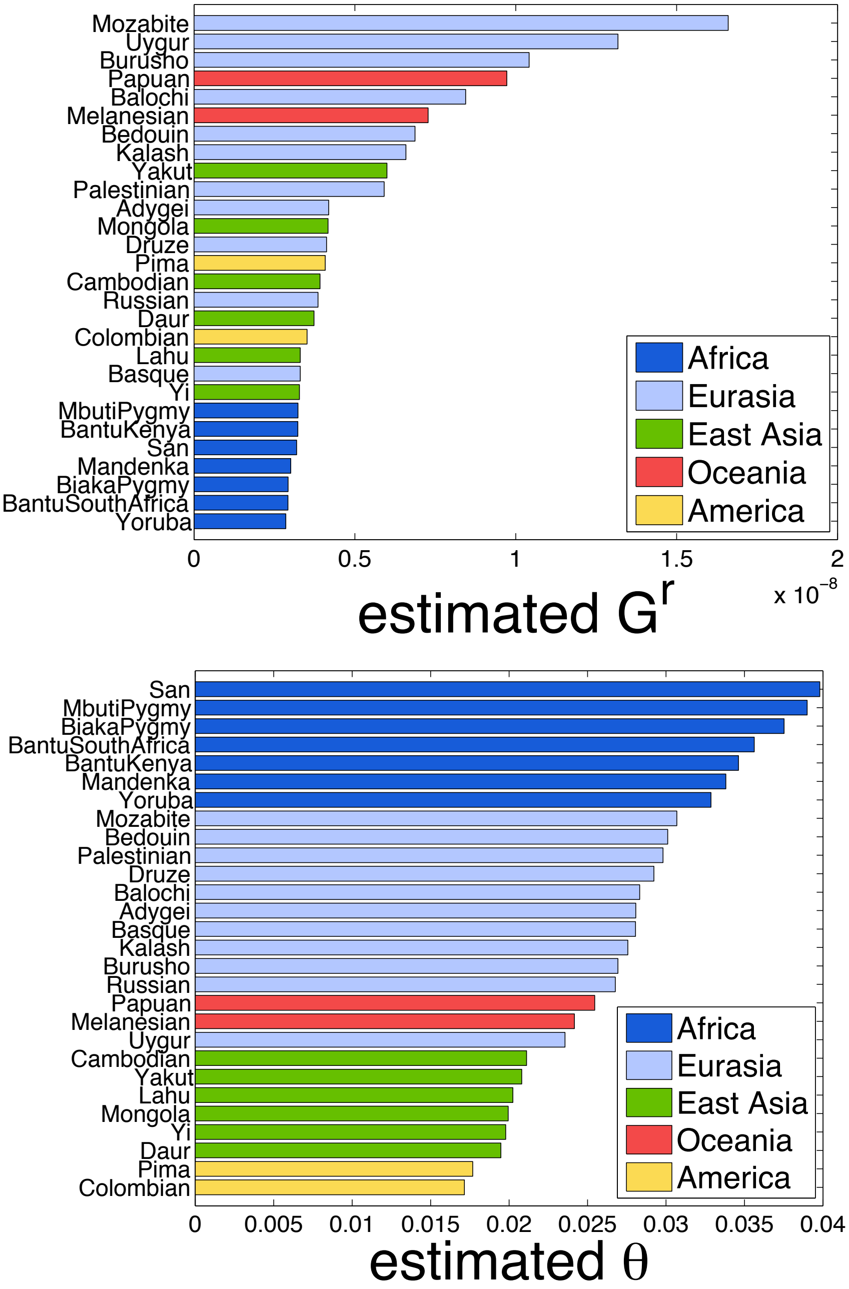


Figure 7.8: Estimated parameters sorted in decreasing order. Top: estimated G^r, time since admixture scaled by recombination rate. Bottom: empirical mutation rate θ computed as the average discrepancy between individuals and their founders.

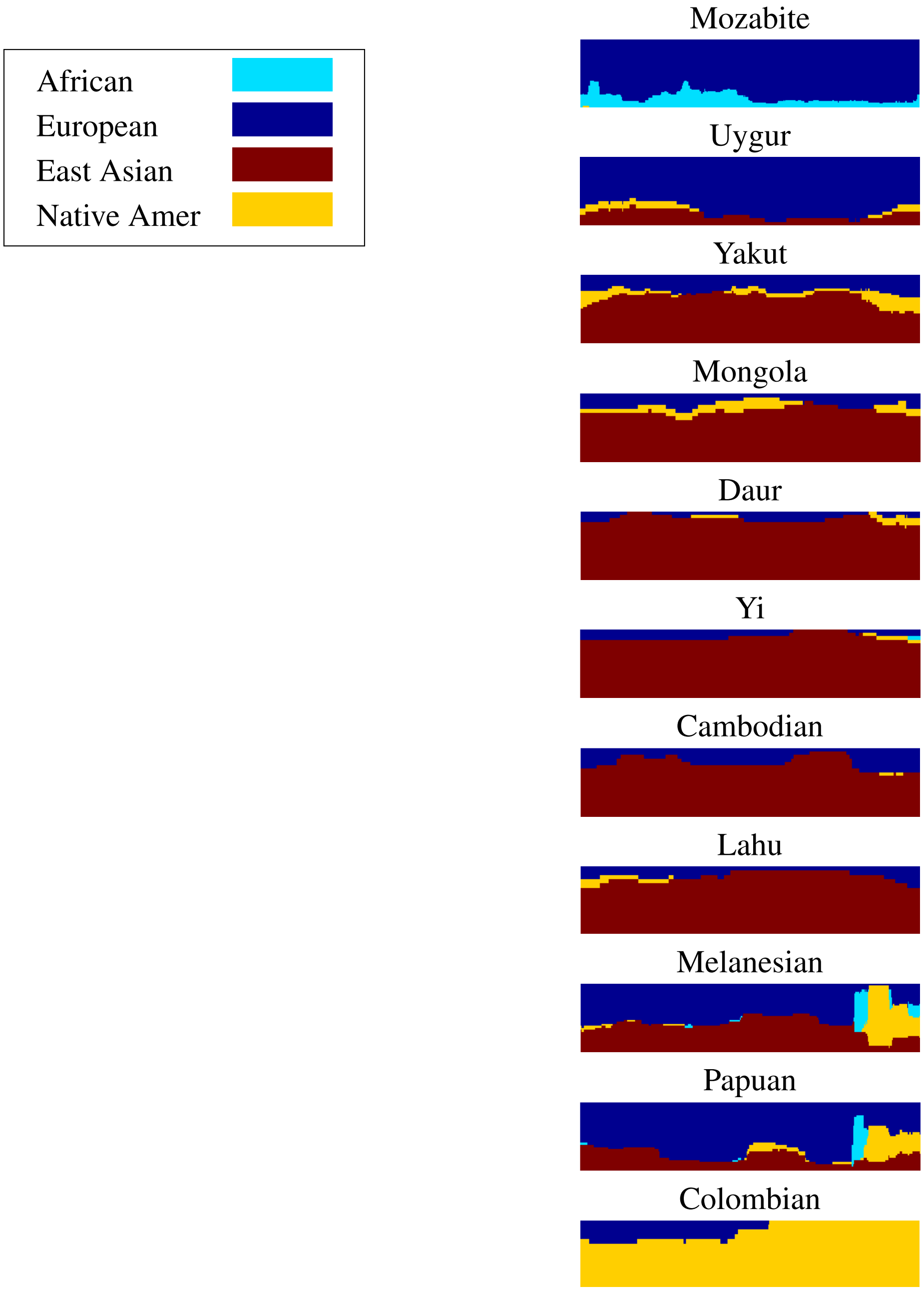


Figure 7.9: Map of ancestry proportions along chromosome 22 on admixed populations from HGDP data. The x-axis corresponds to chromosome positions and the y-axis denotes the ancestry proportion. We selected 11 admixed populations based on the estimated ancestry proportions such that the largest ancestry proportion is less than 90%

gene flow. Most existing approaches for local ancestry estimation have ignored such relatedness and treated the populations as unrelated. We explore this dependency among populations and efficiently utilize it by building a unified model that covers all the ancestral populations and the admixed population together. As shown in our Results, this modeling strategy is especially helpful when only a limited amount of data is available to represent the ancestral populations. Since genetic information in one population can be naturally shared by another population in such a framework, it effectively enhances the robustness of the proposed model regarding the choice of the ancestral population data. We expect that various types of practical analysis dealing with non-typical study populations would benefit from the proposed method.

In our comparative study, HAPMIX appears to perform very well when enough data for ancestral populations are given and also for older admixture events. However, this method does not allow one to analyze the admixing effect from more than two ancestral populations. Instead, one ancestry versus all the other ancestries should be estimated. While this setting may be fine for some applications, this constraint limits its applicability to complex admixture scenarios and may compromise its ability to deal with older admixtures. LAMP has a slightly different focus: while its performance was shown to be worse than the other two in general in our simulation study, it can deal with multiple ancestral populations as our model. And computationally this method was significantly faster than the other two haplotype-based methods. However, LAMP seems to be more suited for very recent admixture case, and its performance tends to drop quite sharply as we consider more ancient admixture events. On the other hand, in a very recent admixture case, LAMP tends to be less sensitive to the amount of training data than HAPMIX as shown in Figure 7.4. Our approach is more general and of more practical utility in that it can incorporate an arbitrary number of ancestral populations with comparable or superior performance than HAPMIX under various scenarios. In comparison of computation time with HAPMIX, our method requires additional, but off-line computation time for model training, which is linear in the number of individuals and the number of markers. For the ancestry estimation phase, we would additionally need a series of MCMC iteration times if we want to estimate the parameters of interest such as admixture time or mutation rates.

It is worth mentioning some of previous approaches for global ancestry analysis as well to position our method in context. STRUCTURE (Pritchard et al., 2000) has been one of the most widely used softwares for admixture analysis, and more recently, other softwares such as EIGENSTRAT (Patterson et al., 2006) and ADMIXTURE (Alexander et al., 2009) have also gain great popularity especially for their computational efficiency. In global ancestry estimation problems, typically no prior information is provided for the ancestral populations and the ancestries of given individuals are recovered as mean proportions of each possible ancestry. Therefore, it can be considered as an unsupervised problem. In contrast, local ancestries are mostly estimated based on the given reference information such as allele frequencies or genotypes of putative ancestral populations. There has been more recent work that bridges the gap between these two approaches. For example, LAMP can also run in an ‘unsupervised mode’ such that it recovers the allele frequency profiles of ancestral populations as well as the local ancestries. Also, ADMIXTURE, which is for the global ancestry estimation, recently added a new feature that the known ancestries of some reference individuals can be exploited (Alexander and Lange, 2011). For haplotype-based approaches, this extension is not straightforward in general because one needs to deal with a set of hidden haplotypes that results in a large number of parameters. Regarding this aspect, our model for the local ancestry has the desirable property that it integrates out the ancestral population data during the inference and works with the hypothetical founders and the admixed population data. Therefore, we expect that the extension of the model to an unsupervised case would also be a promising direction to pursue.

In this approach, we assumed that phased haplotype data are given. In practice, a number of softwares are available for haplotype phasing (Browning and Browning, 2009; Li et al., 2010; Scheet and Stephens, 2006), so the phase information can be readily available in processing step. It would also be quite easy to extend the model to deal with unphased genotypes. For example, we may assume that the haplotypes of ancestral populations are given, and then we allow unphased genotypes for admixed individuals, as in the setting considered in Price et al. (2009). Then the only additional computation would be one more step in our posterior sampling to recover the phasing of genotypes as well as the hidden states in the ancestry estimation phase.

While our model is built on a non-parametric Bayesian framework involving an infinite model, we approximated the resulting distribution by splitting the inference into training and test phases for computational purpose. We used a MAP solution from the training phase which appears to work reasonably well compared to the case using multiple trained samples. Roughly, we observed that the performance gain by multiple trained samples tended to be more significant when the training was less stable, for example, the case with small-sized training data or when the convergence has not reached yet during the training phase. It would be worth investigating the effect of different approximations to the exact distribution on the performance and further improve the computational complexity.

Chapter 8

Conclusion

In this thesis, we have presented non-parametric Bayesian models that allow efficient and robust inference of ancestral genetic processes from SNP data. Using a Dirichlet process based haplotype inheritance model as a building block, we have extended the model to incorporate various genetic processes and hence to uncover important genetic quantities from given data. The non-parametric Bayesian models using a hierarchical Dirichlet process and infinite hidden Markov models have been applied to the three specific problems of haplotype inference from multi-population data, joint inference of population structure and the recombination events, and local ancestry estimation in admixed populations.

As mentioned in Introduction of this thesis, the proposed models put a particular focus on exploiting the shared structural information contained in the data from multiple groups. By use of a hierarchical model that covers the grouped data systematically in a non-parametric Bayesian framework, the models could utilize the latent and shared information underlying the data effectively and have been shown to perform significantly better than previous methods under various applications and scenarios. Furthermore, Dirichlet process based mixture models have allowed us to model the inherent uncertainty about the genetic components such as the number of founder haplotypes. The resulting models can also reveal interesting characteristics of the study populations such as the mutation rate that can be interpreted as the relative age of the population with respect to the hypothetical founder pool, the recombination rate, the time since admixture, or population sub-structure, through the model parameters that can be inferred jointly in the proposed framework

Interesting future directions that connect the model developed under this thesis with the downstream analysis includes the combination of admixture analysis and the disease association study. The local ancestry information that is returned from the proposed model in Chapter 7 can play an important role in selecting out the SNPs that are associated with the phenotypic traits of which distributions show significant differences in multiple contributing populations. In Zhu et al. (2011), the admixture mapping followed by the association study have identified a novel genetic locus affecting the blood pressure. We could propose a model-based approach that combines the ancestry model in this thesis and the association model such that the ancestry information plays as a prior for the association strength.

The computational complexity is another issue that commonly arises in nonparametric Bayesian analysis. While the Beam sampling describes in Chapter 7 could improve the computation time significantly over the traditional MCMC sampling schemes, it is still slower than many other parametric methods. We may further improve the computational complexity by exploiting the redundant or parallel structure in the computation for posterior inference. Actually, the overall algorithm allows a parallelization along various dimensions, especially for training phase, for example, along different ancestral populations or along different individuals in each of the populations. A parallel framework for machine learning such as GraphLab (Low et al., 2010) or Bratieres et al. (2010) would serve as a viable option to start with.

The proposed framework can also be applied to the problem of detecting signatures of selective sweeps on the chromosome. Conventional methods for selective sweeps have mostly relied on different forms of summary statistics often defined heuristically (Teshima and Przeworski, 2006). However, it is not obvious how to evaluate the performance of different summary statistics, and moreover, it is hard to capture the complex structural information contained in the genome using such summary statistics. More recently, model-based approaches such as in Kim and Stephan (2002) have employed a hidden Markov model to detect the selective sweeps based on allele frequency spectrum as observations. The Bayesian inheritance framework we have developed can be adopted for this purpose, for example, we may assume the alleles are determined by some hidden states (e.g. *sweep* and *neutral*) and then a different haplotype distribution can be modeled to generate the individual allele at a specific site depending on the hidden state where *sweep* sites follow a more skewed founder distribution.

Appendix A

Supplementary materials

A.1 Details of the PL Procedure for haplotype inference

This section describes the detailed procedure of partition-ligation algorithm used in *Haploi*, which can be divided into three steps: 1) atomic block typing; 2) bottom-level pairwise ligation to generate overlapping blocks; and 3) hierarchical ligation of overlapping blocks until only one block is left. In step 1, we partition given genotype sequences into L short blocks of length T and phase each atomistic block using the proposed HDPM. From this step, we obtain all the individual haplotypes and also the population haplotype pool for each block. Let $\mathcal{A}_i^T \equiv \{A_{k,\ (i-1)T+1\ :\ iT} \mid k = 1, \ldots, K_i^T\}$ denote the population haplotype pool for T SNPs in the i-th block which ranges from locus $(i-1)T+1$ to iT.

In the next step, we ligate every pair of neighboring blocks: $\mathcal{A}_i^T \& \mathcal{A}_{i+1}^T \rightarrow \mathcal{A}_i^{2T}, i = 1, \ldots, L-1$. Specifically, for each pair of neighboring blocks i and $i+1$, given $\mathcal{A}_i^T$ and $\mathcal{A}_{i+1}^T$, we can impute at most four new stitched haplotypes from an individual since each individual has only two possible haplotypes within each block. In practice, we often have fewer because an individual can be homozygous or the stitched haplotype may already have been imputed from earlier individuals. We pool such stitched haplotypes from all the individuals to form $\mathcal{A}_i^{2T}$, which usually leads to only a small subset of $\mathcal{A}_i^T \times \mathcal{A}_{i+1}^T$. Then based on a finite dimensional Dirichlet prior over $\mathcal{A}_i^{2T}$, we do Gibbs sampling as in Niu *et al.*'s PL scheme to obtain individual haplotypes for each overlapping $2T$ region. To compensate possible ill-ligated blocks, we can redo the direct haplotype inference based on HDPM on those merged blocks whose entropy of haplotype distribution is above some threshold (Figure 5.2 step 2-1). This is computationally affordable since the length of the ligated block at this stage is not yet too big and we can start with better initialization than random assignment. The output from step 2 are $L-1$ sets of length $2T$ population haplotypes, $\{\mathcal{A}_i^{2T}\ :\ i = 1 \ldots L-1\}$, overlapping on T loci for each adjacent pair; and all individual haplotypes in these length $2T$ overlapping segments.

In step 3, we hierarchically ligate overlapping adjacent blocks from the previous iteration, until the full sequence is covered (Figure 5.2, step 3). Specifically, as in step 2, we build the candidate population haplotype pool by adding every unique stitched-haplotype resulted from ligating the haplotypes of the two shorter blocks in every individual. When the overlapping regions of a pair of atomistic haplotypes in an individual are consistent, ligation to a longer haplotype is trivially a merging of the two overlapping haplotypes, and this avoids generating all combinations of the atomistic haplotypes from each block. Only when the overlapping regions in an individual are inconsistent, we grow the haplotype space of the merged blocks by including all possible ligations consistent with the atomistic haplotypes and the individual genotype. For example, suppose a particular individual's haplotypes were recovered as 000**100/** 100**010** at loci 1 to 6 for the first block, and **110**000/ **000**100 at loci 4 to 9 for the next block, and three SNPs are overlapping in the two blocks. Then to accommodate the discrepancy on the 4th and 5th SNPs, we have four possible haplotypes, 10, 01, 00, 11, for these two loci; for the remaining parts of the region covered by these two blocks, i.e., loci 1-3 and loci 6-9, we have two haplotypes (which are from the atomistic haplotypes determined in the previous iteration) for each of them. So a combination of all these possibilities will add the following sixteen haplotypes to the population haplotype space for the ligated segment:

000**100**000/010**010**100, 000**110**000/100**000**100, 000**010**000/100**100**100,
000**000**000/100**110**100, 000**100**100/100**010**000, 000**110**100/100**000**000,
000**010**100/100**100**000, 000**000**100/100**110**000.

Under the newly formed population haplotype space at each ligation iteration, we again apply a Gibbs sampler as in step 2 to determine the individual haplotypes of all remaining unphased individuals over the ligated block under a fixed-dimensional Dirichlet prior of the haplotype frequencies in this trimmed haplotype space. We continue this process hierarchically until there is only one block left. Since each time we only employ overlapping regions of size T, the number of steps needed to complete the ligation of a whole sequence is comparable to Niu et al. (2002)'s hierarchical PL scheme.

A.2 Sensitivity analysis under hierarchical Dirichlet process mixture

Table A.1 summarizes the sensitivity analysis result on the hyper-parameters of HDP scale parameters for the diverse dataset ($\theta = 0.05$) as in Table 5.1 which was on the conserved dataset. While the number of ancestors within each ethnic group is recovered less stably compared to the conserved dataset case, the total number of ancestors could be restored in much more stable pattern. Also, we observe that K and θ compromises with each other: when K is larger compared to the true value of 17, θ is recovered as a bigger value, while the resulting haplotyping error remains quite similar except for a few extreme values of priors. This would imply that even when we cannot recover the exact number of ancestors, our model still gives good estimates of the inheritance process for haplotype phasing and related parameter recovery.

A.3 A summary of HapMap ENCODE regions

The description of 10 HapMap ENCODE regions used in Figure 5.7 is summarized in Table A.2.

Table A.1: A sensitivity analysis to the hyper-parameters of HDP on diverse dataset

κ	ι	κ/ι	K_1	K_2	K_3	K_4	K_5	total K (17)	θ (0.045)	err_s
0.1	0.1	1	7.8	5.3	5.6	6.7	6.7	17.1	0.035	0.1012
	0.5	0.2	7.1	5.2	6.8	7.7	7.4	17.9	0.038	0.0952
	1	0.1	7.1	5.2	7.4	6.0	6.3	17.1	0.036	0.0923
	10	0.01	7.4	5.1	6.1	5.8	6.6	17.0	0.035	0.0744
	100	0.001	5.0	5.0	6.0	5.0	5.1	15.3	0.049	0.1042
	1000	0.0001	5.0	5.0	5.0	5.0	4.0	15.0	0.045	0.0714
0.5	0.1	5	8.6	5.3	7.3	5.8	7.6	18.2	0.035	0.0774
	0.5	1	7.9	5.2	7.3	6.6	7.2	17.6	0.035	0.0923
	1	0.5	8.5	5.1	6.1	6.7	7.0	17.7	0.036	0.0774
	10	0.05	6.9	5.1	5.6	5.1	6.8	17.0	0.037	0.0595
	100	0.005	5.4	5.0	5.0	5.1	5.4	16.0	0.041	0.0952
	1000	0.0005	5.0	5.0	5.0	5.0	4.0	15.0	0.045	0.0952
1	0.1	10	7.1	5.2	7.0	5.6	7.3	16.5	0.045	0.0952
	0.5	2	5.9	5.7	8.1	6.1	7.7	18.4	0.036	0.1131
	1	1	7.5	5.1	6.2	7.3	6.2	17.1	0.032	0.0982
	10	0.1	7.4	5.1	6.4	5.2	6.1	17.1	0.041	0.0833
	100	0.01	5.8	5.0	5.9	5.0	5.6	16.0	0.041	0.0595
	1000	0.001	5.0	5.0	5.0	5.0	5.0	15.0	0.044	0.0923
10	0.1	100	7.1	5.2	7.0	6.3	7.0	17.4	0.041	0.0833
	0.5	20	7.1	5.5	9.7	7.5	7.0	17.4	0.035	0.1250
	1	10	6.8	5.3	5.8	7.3	7.8	15.4	0.043	0.1071
	10	1	7.5	5.0	6.6	6.0	6.5	17.1	0.032	0.0923
	100	0.1	7.0	5.0	7.0	5.7	5.0	17.0	0.039	0.1012
	1000	0.01	5.0	5.0	5.0	4.0	5.0	14.0	0.049	0.0863
100	0.1	1000	11.0	9.6	13.6	13.1	12.3	30.6	0.041	0.1696
	0.5	200	9.0	6.4	10.1	8.9	8.7	20.2	0.044	0.1518
	1	100	8.7	6.1	9.0	7.6	9.2	19.2	0.045	0.0893
	10	10	7.9	5.4	9.2	7.1	7.8	19.2	0.036	0.0744
	100	1	6.7	5.0	5.5	5.0	6.7	17.1	0.037	0.0804
	1000	0.1	5.0	5.0	5.6	5.0	5.0	15.0	0.046	0.0923
1000	0.1	10000	16.8	13.7	19.6	19.4	15.8	63.4	0.016	0.1548
	0.5	2000	14.2	11.3	18.5	16.2	15.8	49.5	0.035	0.1607
	1	1000	12.7	11.6	16.9	15.1	14.1	43.6	0.030	0.1488
	10	100	9.1	8.4	12.3	9.4	9.5	24.5	0.038	0.1042
	100	10	8.2	6.3	8.5	7.3	9.3	19.7	0.034	0.0714
	1000	1	7.1	5.1	5.1	6.1	7.0	16.0	0.039	0.1310

Table A.2: A summary of the 10 HapMap ENCODE regions used in this study.

	Region name	#SNPs	Chrs.	start–end (Mb)	length (Kb)
1	ENm010	254	7	26.7 – 27.2	497
2	ENr232	379	9	127.1 – 127.6	496
3	ENr123	391	12	38.6 – 39.1	499
4	ENr321	495	8	118.8 – 119.3	498
5	ENm013	548	7	89.4 – 89.9	494
6	ENr213	565	18	23.7 – 24.2	565
7	ENm014	694	7	126.1 – 126.6	497
8	ENr112	728	2	51.6 – 52.1	498
9	ENr131	857	2	234.8 – 235.3	499
10	ENr113	972	4	118.7 – 119.2	498

Bibliography

Airoldi, E., Blei, D., Xing, E., and Fienberg, S. (2006). Mixed membership stochastic block models for relational data, with applications to protein-protein interactions. *Proceedings of International Biometric Society-ENAR Annual Meetings*. 5.5

Alexander, D. H. and Lange, K. (2011). Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. *BMC Bioinformatics*, 12:246. 7.4

Alexander, D. H., Novembre, J., and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, 19(9):1655–1664. 2.3, 7.1, 7.4

Anderson, E. and Novembre, J. (2003). Finding haplotype block boundaries by using the minimum-description-length principle. *American journal of human genetics*, 73(2):336–354. (document), 1.1, 2.1, 6.1, 6.2.1, 6.2.1, 6.4.2, 6.11

Beal, M. J., Ghahramani, Z., and Rasmussen, C. E. (2002). The infinite hidden Markov model. *Advances in Neural Information Processing Systems*, 14. 3.3, 7.1, 7.2.3

Blackwell, D. and MacQueen, J. B. (1973). Ferguson Distributions Via Polya Urn Schemes. *Annals of Statistics*, 1(2):363–355. 3.1, 5.1, 7.2.3

Blei, D., Ng, A., and Jordan, M. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022. 5.5

Bratieres, S., Van Gael, J., and Vlachos, A. (2010). Scaling the iHMM: Parallelization versus Hadoop. *... (CIT)*. 8

Browning, B. L. and Browning, S. R. (2009). A Unified Approach to Genotype Imputation and Haplotype-Phase Inference for Large Data Sets of Trios and Unrelated Individuals. *American Journal of Human Genetics*, 84(2):210–223. 1.1, 7.4

Browning, S. R. and Browning, B. L. (2007). 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. 2.2, 5.1, 5.4

Chakravarti, A. (2001). To a future of genetic medicine. *Nature*, 409(6822):822–823. 1.1

Cheng, C.-Y., Kao, W. H. L., Patterson, N., Tandon, A., Haiman, C. A., Harris, T. B., Xing, C., John, E. M., Ambrosone, C. B., Brancati, F. L., Coresh, J., Press, M. F., Parekh, R. S., Klag, M. J., Meoni, L. A., Hsueh, W.-C., Fejerman, L., Pawlikowska, L., Freedman, M. L., Jandorf, L. H., Bandera, E. V., Ciupak, G. L., Nalls, M. A., Akylbekova, E. L., Orwoll, E. S., Leak, T. S., Miljkovic, I., Li, R., Ursin, G., Bernstein, L., Ardlie, K., Taylor, H. A., Boerwinckle, E., Zmuda, J. M., Henderson, B. E., Wilson, J. G., and Reich, D. (2009). Admixture mapping of 15,280 African Americans identifies obesity susceptibility loci on chromosomes 5 and X. *PLoS Genetics*, 5(5):e1000490. 7.1

Cheng, C.-Y., Reich, D., Wong, T. Y., Klein, R., Klein, B. E. K., Patterson, N., Tandon, A., Li, M., Boerwinkle, E., Sharrett, A. R., and Kao, W. H. L. (2010). Admixture mapping scans identify a locus affecting retinal vascular caliber in hypertensive African Americans: the Atherosclerosis Risk in Communities (ARIC) study. *PLoS Genetics*, 6(4):e1000908. 7.1

Clark, A. (2003). Finding genes underlying risk of complex disease by linkage disequilibrium mapping. *Current opinion in genetics & development*. 1.1

Daly, M., Rioux, J., Schaffner, S., and Hudson, T. (2001). High-resolution haplotype structure in the human genome. *Nature Genetics*. (document), 1.1, 2.1, 6.1, 6.2.1, 6.2.1, 6.4, 6.4.2, 6.11, 6.4.2

Erosheva, E., Fienberg, S., and Lafferty, J. (2004). Mixed-membership models of scientific publications. *Proc Natl Acad Sci U S A*, 101 (Suppl 1):5220–5227. 6.1

Escobar, M. D. and West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 90:577–588. 3.1

Excoffier, L. and Hamilton, G. (2003). Comment on genetic structure of human populations. *Science*, 300(5627):1877b–. 6.1

Excoffier, L. and Slatkin, M. (1995). Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Molecular biology and evolution*, 12(5):921–927. 1.1, 4, 5.1

Fallin, D. and Schork, N. J. (2000). Accuracy of haplotype frequency estimation for biallelic loci, via the expectation-maximization algorithm for unphased diploid genotype data. *American journal of human genetics*, 67(4):947–959. 5.1

Falush, D., Stephens, M., and Pritchard, J. K. (2003). Inference of Population Structure Using Multilocus Genotype Data: Linked Loci and Correlated Allele Frequencies. *Genetics*, 164(4):1567–1587. 1.1, 2.3, 6.1, 6.1, 6.5, 7.1

Fearnhead, P. and Donnelly, P. J. (2001). Estimating recombination rates from population genetic data. *Genetics*, 159:1299–1318. 2.1, 6.4.1

Ferguson, T. S. (1973). A Bayesian Analysis of Some Nonparametric Problems. *Annals of Statistics*, 1(2):209–230. 3.1, 5.1, 7.2.3

Fisher, R. A. (1930). The genetical theory of natural selection. *Clarendon Press, Oxford*. 4.2

Gay, J., Myers, S., and McVean, G. (2007). Estimating meiotic gene conversion rates from population genetic data. *Genetics*, 177(2):881–894. 2.2

Greenspan, G. and Geiger, D. (2004). High density linkage disequilibrium mapping using models of haplotype block variation. *Bioinformatics (Oxford, England)*, 20 Suppl 1:i137–44. 6.1

Gusfield, D. (2004). An overview of combinatorial methods for haplotype inference.... *Methods for SNPs and Haplotype Inference*. 5.1

Hawley, M. E. and Kidd, K. K. (1995). HAPLO: a program using the EM algorithm to estimate the frequencies of multi-site haplotypes. *The Journal of heredity*, 86(5):409–411. 5.1

Hodge, S., Boehnke, M., and Spence, M. (1999). Loss of information due to ambiguous haplotyping of SNPs. *Nature Genetics*, 21(4):360–361. 2.2

Hoppe, F. M. (1984). Pólya-like urns and the ewens' sampling formula. *Journal of Math. Biol.*, 20(1):91–94. 4, 4.2

Hudson, R. R. (1983). Properties of a neutral allele model with intragenic recombination. *Theoretical population biology*, 23(2):183–201. 6.2.2

Huelsenbeck, J. P. and Andolfatto, P. (2007). Inference of Population Structure Under a Dirichlet Process Model. *Genetics*, 175(4):1787–1802. 7.1

Ishwaran, H. and James, L. F. (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, 90:161–173. 3.1

Jakobsson, M., Scholz, S. W., Scheet, P., Gibbs, J. R., VanLiere, J. M., Fung, H.-C., Szpiech, Z. A., Degnan, J. H., Wang, K., Guerreiro, R., Bras, J. M., Schymick, J. C., Hernandez, D. G., Traynor, B. J., Simon-Sanchez, J., Matarin, M., Britton, A., van de Leemput, J., Rafferty, I., Bucan, M., Cann, H. M., Hardy, J. A., Rosenberg, N. A., and Singleton, A. B. (2008). Genotype, haplotype and copy-number variation in worldwide human populations. *Nature*, 451(7181):998–1003. 7.3.1, 7.3.7

Kim, Y. and Stephan, W. (2002). Detecting a local signature of genetic hitchhiking along a recombining chromosome. *Genetics*, 160:765–777. 8

Kimmel, G. and Shamir, R. (2004). Maximum likelihood resolution of multi-block genotypes. In *RECOMB '04: Proceedings of the eighth annual international conference on Resaerch in computational molecular biology*. ACM Request Permissions. 4, 5.1

Kingman, J. F. C. (1982). On the Genealogy of Large Populations. *Journal of Applied Probability*, 19:27–43. 2.4, 4.2

Li, H., Cho, K., Kidd, J. R., and Kidd, K. K. (2009). Genetic landscape of Eurasia and "admixture" in Uyghurs. *American Journal of Human Genetics*, 85(6):934–7; author reply 937–9. 1.1, 7.3.7

Li, N. (2003). Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics*. 2.2, 2.4, 5.1, 5.3, 5.5, 6.4.2, 6.5

Li, Y. and Abecasis, G. (2006). Mach 1.0: Rapid haplotype reconstruction and missing genotype inference. *Am J Hum Genet.*, S79:2290. 5.1, 5.4

Li, Y., Willer, C. J., Ding, J., Scheet, P., and Abecasis, G. R. (2010). MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genetic Epidemiology*, 34(8):816–834. 1.1, 2.2, 7.4

Liu, J., Sabatti, C., Teng, J., and Keats, B. (2001). Bayesian analysis of haplotypes for linkage disequilibrium mapping. *Genome Research*. 4.1, 6.2.1

Long, J. and Williams, R. (1995). An EM algorithm and testing strategy for multiple-locus haplotypes. *American Journal of Human ...*. 5.1

Low, Y., Gonzalez, J., Kyrola, A., Bickson, D., Guestrin, C., and Hellerstein, J. M. (2010). GraphLab: A New Framework for Parallel Machine Learning. *In the 26th Conference on*

Uncertainty in Artificial Intelligence (UAI). 8

Muller, P., Quintana, F., and Rosner, G. (2004). A method for combining inference across related nonparametric bayesian models. *Journal of the Royal Statistical Society, Series B*, 66(3):735–749. 5.5

Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *J. Computational and Graphical Statistics*, 9(2):249–256. 3.1

Niu, T., Qin, Z. S., Xu, X., and Liu, J. S. (2002). Bayesian haplotype inference for multiple linked single-nucleotide polymorphisms. *American journal of human genetics*, 70(1):157–169. 5.1, 5.3, 5.3, 5.4.4, A.1

Pasaniuc, B., Sankararaman, S., Kimmel, G., and Halperin, E. (2009). Inference of locus-specific ancestry in closely related populations. *Bioinformatics (Oxford, England)*, 25(12):i213–21. 1.1, 2.3, 7.1, 7.2.2, 7.3.1

Patil, N., Berno, A., Hinds, D., Barrett, W., Doshi, J., Hacker, C., Kautzer, C., Lee, D., Marjoribanks, C., McDonough, D., Nguyen, B., Norris, M., Sheehan, J., Shen, N., Stern, D., Stokowski, R., Thomas, D., Trulson, M., Vyas, K., Frazer, K., Fodor, S., and Cox, D. (2001). Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *science*, 294(5547):1719–1723. 1.1, 2.1, 2.2, 6.2.1

Patterson, N., Hattangadi, N., Lane, B., Lohmueller, K. E., Hafler, D. A., Oksenberg, J. R., Hauser, S. L., Smith, M. W., O'Brien, S. J., Altshuler, D., Daly, M. J., and Reich, D. (2004). Methods for High-Density Admixture Mapping of Disease Genes. *American Journal of Human Genetics*, 74(5):979–1000. 1.1, 7.1, 7.3.1

Patterson, N., Price, A. L., and Reich, D. (2006). Population Structure and Eigenanalysis. *PLoS genetics*, 2(12):e190. 2.3, 7.1, 7.4

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, 38(8):904–909. 1.1, 7.1

Price, A. L., Tandon, A., Patterson, N., Barnes, K. C., Rafaels, N., Ruczinski, I., Beaty, T. H., Mathias, R., Reich, D., and Myers, S. (2009). Sensitive Detection of Chromosomal Segments of Distinct Ancestry in Admixed Populations. *PLoS genetics*, 5(6):e1000519. 1.1, 2.2, 2.3, 7.1, 7.2.2, 7.3.1, 7.3.4, 7.4

Price, A. L., Weale, M. E., Patterson, N., Myers, S. R., Need, A. C., Shianna, K. V., Ge, D., Rotter, J. I., Torres, E., Taylor, K. D., Goldstein, D. B., and Reich, D. (2008). Long-Range LD Can Confound Genome Scans in Admixed Populations. *American Journal of Human Genetics*, 83(1):132–135. 7.1

Pritchard, J. K. (2001). Are rare variants responsible for susceptibility to complex diseases? *American journal of human genetics*, 69(1):124–137. 5.1

Pritchard, J. K., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945. 2.3, 6.1, 7.1, 7.4

Qin, Z. S., Niu, T., and Liu, J. S. (2002). Partition-ligation-expectation-maximization algorithm for haplotype inference with single-nucleotide polymorphisms. *American journal of human* *genetics*, 71(5):1242–1247. 1.1, 4, 5.1, 5.3

Rasmussen, C. E. (2000). The infinite Gaussian mixture model. In *Advances in Neural Information Processing Systems 12*. 5.2.2

Rodriguez, A., Dunson, D. B., and Gelfand, A. E. (2006). The nested dirichlet process. *Technical report, Institute of Statistics and Decision Sciences, Duke University*. 5.5

Rosenberg, N. A., Pritchard, J. K., Weber, J. L., Cann, H. M., Kidd, K. K., Zhivotovsky, L. A., and Feldman, M. W. (2002). Genetic structure of human populations. *Science*, 298:2381–2385. 2.3, 6.1, 6.1

Sankararaman, S., Kimmel, G., Halperin, E., and Jordan, M. I. (2008a). On the inference of ancestries in admixed populations. In *RECOMB'08: Proceedings of the 12th annual international conference on Research in computational molecular biology*. Springer-Verlag. 7.2.2

Sankararaman, S., Sridhar, S., Kimmel, G., and Halperin, E. (2008b). Estimating Local Ancestry in Admixed Populations. *American Journal of Human Genetics*, 82:290–303. 7.3.1

Scheet, P. and Stephens, M. (2006). A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *American Journal of Human Genetics*, 78(4):629–644. 1.1, 2.2, 5.4, 1, 7.4

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica*, 1(4):639–50. 3.1, 3.1

Sohn, K.-A., Ghahramani, Z., and Xing, E. P. (2011). Robust estimation of local genetic ancestry in admixed populations using a non-parametric Bayesian approach. *In review*. 1.2

Sohn, K.-A. and Xing, E. P. (2007a). Hidden Markov Dirichlet process: modeling genetic recombination in open ancestral space. *Advances in Neural Information Processing Systems*. 1.2, 6.2.2

Sohn, K.-A. and Xing, E. P. (2007b). Spectrum: Joint bayesian inference of population structure and recombination events. *Bioinformatics (Oxford, England)*, 23(13):i479–i489. 1.2, 6.2.2, 7.1, 7.2.2

Sohn, K.-A. and Xing, E. P. (2009). A hierarchical Dirichlet process mixture model for haplotype reconstruction from multi-population data. *Annals of Applied Statistics*, 3(2):791–821. 1.2, 4, 7.1

Stephens, M. and Scheet, P. (2005). Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. *American Journal of Human Genetics*, 76:449–462. 1.1, 2.1, 5.4, 6.4

Stephens, M., Smith, N., and Donnelly, P. (2001). A new statistical method for haplotype reconstruction from population data. *American journal of human genetics*, 68(4):978–989. 2.2, 4, 5.1, 5.3, 5.4, 5.4.1, 6.4

Stoneking, M. (2001). Single nucleotide polymorphisms. From the evolutionary past... *Nature*, 409(6822):821–822. 6.1

Sundquist, A., Fratkin, E., Do, C. B., and Batzoglou, S. (2008). Effect of genetic divergence in identifying ancestral origin using HAPAA. *Genome Research*, 18(4):676–682. 1.1, 2.3, 7.1, 7.2.2, 7.3.1

Tang, H., Choudhry, S., Mei, R., Morgan, M., Rodriguez-Cintron, W., Burchard, E. G., and Risch, N. J. (2007). Recent genetic selection in the ancestral admixture of Puerto Ricans. *American Journal of Human Genetics*, 81(3):626–633. 7.3.4

Tang, H., Coram, M., Wang, P., Zhu, X., and Risch, N. (2006). Reconstructing Genetic Ancestry Blocks in Admixed Individuals. *American Journal of Human Genetics*, 79:1–12. 2.3, 7.1, 7.3.1

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2010). Hierarchical Dirichlet Processes. *Journal of American Statistical Association*, 101(476):1566–1581. 3.1, 3.2, 3.2, 3.2, 3.3, 5.1, 5.2.1, 7.1, 7.2.3

Teshima, K., G. C. and Przeworski, M. (2006). How reliable are empirical genomic scans for selective sweeps. *Genome Res.*, 16:702–712. 8

Thorisson, G. A., Smith, A. V., Krishnan, L., and Stein, L. D. (2005). The international hapmap project web site. *Genome Research*, 15:1591–1593. (document), 6.2, 6.4

Van Gael, J., Saatci, Y., Teh, Y. W., and Ghahramani, Z. (2008). Beam sampling for the infinite hidden Markov model. In *ICML '08: Proceedings of the 25th international conference on Machine learning*. ACM. 3.3, 7.2.4, 7.2.4

Wang, X., Zhu, X., Qin, H., Cooper, R. S., Ewens, W. J., Li, C., and Li, M. (2010). Adjustment for local ancestry in genetic association analysis of admixed populations. *Bioinformatics (Oxford, England)*, 27(5):670–677. 1.1, 7.1

Wright, S. (1931). Evolution in mendelian populations. *Genetics*, 16:97–157. 4.2

Xing, E., Sharan, R., and Jordan, M. (2004). Bayesian haplotype inference via the Dirichlet process. In *Proceedings of the 21st International Conference on Machine Learning*. 4, 6.2.2, 6.3

Xing, E., Sohn, K.-A., Jordan, M., and Teh, Y. W. (2006). Bayesian multi-population haplotype inference via a hierarchical dirichlet process mixture. In *Proc 23th Int Conf on Machine Learning*, pages 1049–1056, New York. ACM Press. 1.2, 5.1, 5.2.1

Xing, E. P., Jordan, M. I., and Sharan, R. (2007). Bayesian haplotype inference via the Dirichlet process. *Journal of computational biology : a journal of computational molecular cell biology*, 14(3):267–284. 1.1, 4.1

Xing, E. P. and Sohn, K.-A. (2007). Hidden Markov Dirichlet process: Modeling genetic recombination in open ancestral space. *Bayesian Analysis*. 1.2, 5.5, 6.2.2

Xu, S., Huang, W., Qian, J., and Jin, L. (2008). Analysis of genomic admixture in Uyghur and its implication in mapping strategy. *American Journal of Human Genetics*, 82(4):883–894. 1.1, 7.3.7

Xu, S. and Jin, L. (2008). A genome-wide analysis of admixture in Uyghurs and a high-density admixture map for disease-gene discovery. *American Journal of Human Genetics*, 83(3):322–336. 1.1, 7.3.7

Zhang, K., Deng, M., Chen, T., Waterman, M. S., and Sun, F. (2002). A Dynamic Programming

Algorithm for Haplotype Block Partitioning. *Proceedings of the National Academy of Sciences of the United States of America*, 99(11):7335–7339. 1.1, 2.1

Zhang, Y., Niu, T., and Liu, J. S. (2006). A coalescence-guided hierarchical bayesian method for haplotype inference. *Am. J. Hum. Genet.*, 79:313–322. 4.1, 5.1, 5.3

Zhu, X., Young, J. H., Fox, E., Keating, B. J., Franceschini, N., Kang, S., Tayo, B., Adeyemo, A., Sun, Y. V., Li, Y., Morrison, A., Newton-Cheh, C., Liu, K., Ganesh, S. K., Kutlar, A., Vasan, R. S., Dreisbach, A., Wyatt, S., Polak, J., Palmas, W., Musani, S., Taylor, H., Fabsitz, R., Townsend, R. R., Dries, D., Glessner, J., Chiang, C. W. K., Mosley, T., Kardia, S., Curb, D., Hirschhorn, J. N., Rotimi, C., Reiner, A., Eaton, C., Rotter, J. I., Cooper, R. S., Redline, S., Chakravarti, A., and Levy, D. (2011). Combined admixture mapping and association analysis identifies a novel blood pressure genetic locus on 5p13: contributions from the CARe consortium. *Human Molecular Genetics*, 20(11):2285–2295. 7.1, 8