A Nonparametric Bayesian Approach for Haplotype Reconstruction from Single and Multi-Population Data

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

April 2007
CMU-ML-07-107

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Abstract

Uncovering the haplotypes of single nucleotide polymorphisms and their population demography is essential for many biological and medical applications. Methods for haplotype inference developed thus far –including those based on approximate coalescence, finite mixtures, and maximal parsimony– often bypass issues such as unknown complexity of haplotype-space and demographic structures underlying multi-population genotype data. In this paper, we propose a new class of haplotype inference models based on a nonparametric Bayesian formalism built on the *Dirichlet process*, which represents a tractable surrogate to the coalescent process underlying population haplotypes and offers a well-founded statistical framework to tackle the aforementioned issues. Our proposed model, known as a hierarchical Dirichlet process mixture, is exchangeable, unbounded, and capable of coupling demographic information of different populations for posterior inference of individual haplotypes, the size and configuration of haplotype ancestor pools, and other parameters of interest given genotype data. The resulting haplotype inference program, *Haploi*, is readily applicable to genotype sequences with thousands of SNPs, at a time-cost often two-orders of magnitude less than that of the state-of-the-art PHASE program, with competitive and sometimes superior performance. *Haploi* also significantly outperforms several other extant algorithms on both simulated and realistic data.

Keywords: haplotype inference, Dirichlet Process, Hierarchical Dirichlet process, mixture model, population genetics

1 Introduction

Recent experimental advances have led to an explosion of data which document genetic variation at the DNA level within and between populations. For example, the international SNP map working group Group[1] has reported the identification and mapping of 1.4 million single nucleotide polymorphisms (SNPs) from the genomes of 4 different human populations in the world. These kinds of data lead to challenging inference problems whose solutions could lead to greater understanding of the genetic basis of disease propensities and other complex traits[2,3].

SNPs represent the largest class of individual differences in DNA. A SNP refers to the existence of two specific nucleotides chosen from $\{A, C, G, T\}$ at a single chromosomal locus in a population; each variant is called an *allele*. A *haplotype* refers the joint allelic identities of a list of SNPs at contiguous sites in a local region of a single chromosome. Assuming no recombination in this local region, a haplotype is inherited as a unit. For diploid organisms (such as humans), each individual has two physical copies of each chromosome (except for the Y chromosome) in his/her somatic cells; one copy is inherited from the mother, and the other from the father. Thus during each generation of inheritance when chromosomes come in pairs, two haplotypes, for example, $h_1 \equiv (1, 1, 0, 0)$ and $h_2 \equiv (0, 0, 1, 1)$ of a 4-loci region, go together to make up a *genotype*, which is the list of *unordered* pairs of alleles in the attendant region, e.g., $g \equiv (1/0, 1/0, 1/0, 1/0)$ in case of the aforementioned two haplotypes. That is, a genotype is obtained from a pair of haplotypes by omitting the specification of the association of each allele with one of the two chromosomes—its *phase*. Indeed, phase is in general ambiguous when only the genotypes of a SNPs sequence are given[4,5]. For example, in the above example, given the g, an alternative configuration of the haplotypes, $h_1' \equiv (1, 1, 1, 1)$ and $h_2' \equiv (0, 0, 0, 0)$, is also consistent with the genotype; but observing multiple genotypes in a population can help to bias the phase reconstruction toward the true haplotypes. The problem of inferring SNP haplotypes from genotypes is essential for the understanding of genetic variations and linkage disequilibrium patterns in a population. For example, accurate inferences concerning population structures or quantitative trait locus maps usually demand the analysis of the genetic states of possibly non-recombinant segments of the subject's chromosome(s)[6]. Thus, it is advantageous to study haplotypes, which consist of several closely spaced (hence linked) phase-known SNPs and often prove to be more powerful discriminators of genetic variations within and among populations.

Common biological methods for assaying genotypes typically do not provide phase information for individuals with heterozygous genotypes at multiple autosomal loci; phase can be obtained at a considerably higher cost via molecular haplotyping[7]. In addition to being costly, these methods are subject to experimental error and are low-throughput. Alternatively, phase can also be inferred from the genotypes of a subject's close relatives[5]. But this approach is often hampered by the fact that typing family members increases the cost and does not guarantee full informativeness. It is desirable to develop automatic and robust *in silico* methods for reconstructing haplotypes from genotypes and possibly other data sources (e.g., pedigrees).

Key to the inference of individual haplotypes based on a given genotype sample, is the formulation and tractability of the marginal probability of the haplotypes of a study population. Consider the set of haplotypes, denoted as $H = \{h_1, h_2, \ldots, h_{2n}\}$ (where $h_i \in \mathcal{P}^T$, $\mathcal{P}$ denotes the allele space of the polymorphic markers and T denotes the length of the marker sequence), of a random

sample of $2n$ chromosomes of n individuals taken from a population at stationarity of some inheritance process, e.g., an infinitely-many-allele (IMA) mutation model. Under common genetic arguments, the ancestral relationships amongst the sample back to its most recent common ancestor (MRCA) can be described by a genealogical tree, and computing $p(H)$ involves a marginalization over all possible genealogical trees consistent with the sample, which is widely known to be intractable. As discussed in Stephens and Donnelly[8], write $P(H)$ as a product of conditionals based on the chain rule, i.e.,

$$P(h_1, h_2, \ldots, h_{2n}) = P(h_1)P(h_2|h_1)\ldots P(h_{2n}|h_1\ldots h_{2n-1}), \tag{1}$$

then the generation of a haplotype sample H can be viewed as a sequential process that draw one haplotype at a time conditioning on all the previously drawn haplotypes, e.g., by introducing random mutations to the latter. (This is equivalent to sampling from a genealogy evolving in non-overlapping generations.) Therefore, one can develop tractable approximation to $P(H)$ by appropriately approximating the conditionals in Eq. (1). Stephens and Donnelly[8] suggested an approximation to $P(h_i|h_1\ldots h_{i-1})$ that captures, among several desirable genetic properties, the *parental-dependent-mutation* (PAM) property*, by modeling h_i as the progeny of a randomly-chosen existing haplotype through a geometric-distributed number of mutations. This model, referred to as the PAC (for *Product of Approximate Conditionals*) model, forms the basis of the PHASE program[9], which has set the state-of-the-art benchmarks in haplotype inference.

However, one caveat of the PAC model, as acknowledged in Li and Stephens[10], is that it implicitly assumes existence of an ordering in the haplotype sample, therefore the resulting likelihood does not enjoy the property of exchangeability that we would expect to be satisfied by the true $p(H)$. Although empirically this pitfall appears to be inconsequential after some heuristic averaging over a moderate number of random orderings, it is difficult to associate this approximation to an explicit assumption about the population demography and genealogy underlying the sample. For example, the genealogy of haplotypes with possibly common ancestry is replaced by asymmetric pairwise relationships (induced by the conditional mutation model) between the haplotypes. The resulting "flattening" of the latent genealogical history makes it difficult to use the PAC method to discover and exploit latent demographic structures such as estimating the number and pattern of prototypical haplotypes (i.e., founders), which may be indicative of genetic bottlenecks and divergence time of the study population, or to make use of the ethnic identities of the sample to improve haplotyping accuracy in multi-population haplotype inference.

The finite mixture models adopted by programs such as HAPLOTYPER represent another class of haplotype models that rely very little on demographic and genetic assumptions of the sample[11–14]. Under such a model, haplotypes are treated as latent variables associated with specific frequencies, and the probability of a genotype is given by:

$$p(g) = \sum_{h_1, h_2 \in \mathcal{H}} p(h_1, h_2) f(g\,|h_1, h_2), \tag{2}$$

*The parental-dependent-mutation posit 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.

where $f(g|h_1, h_2)$ is a noisy channel relating the observed genotype to the unobserved true underlying haplotypes[†], and $\mathcal{H}$ denote the set of all possible haplotypes of a given region. Under the assumption of Hardy-Weinberg equilibrium (HWE), an assumption that is standard in the literature and will also be made here, the mixing proportion $p(h_1, h_2)$ is assumed to factor as $p(h_1)p(h_2)$. Given this basic statistical structure, the haplotype inference problem can be viewed a *missing value inference* and *parameter estimation* problem. Numerous statistical models and statistical inference approaches have been developed for this problem, such as the maximum likelihood approaches via the EM algorithm[11,15–17], and a number of parametric Bayesian inference methods based on Markov Chain Monte Carlo (MCMC) techniques[12,14].

The finite mixture model defines an exchangeable $P(H)$. But since such models are data-driven rather than genetically motivated, they offer no insight of the genealogical history underlying the sample. Furthermore, these methodologies have 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, which reflects the diversity of the genome and its evolutionary history, is unknown for any given population data; thus we have a mixture model problem in which a key aspect of the inferential problem involves inference over the number of mixture components, i.e., the haplotypes. There is a plethora of combinatorial algorithms based on various deterministic hypothesis such as the "parsimony" principles that offer control over the complexity of the inference problem[4,18–20], and these methods have demonstrated effectiveness in certain settings and provided important insights to the problem (see Gusfield[21] for an excellent survey). But compared to the statistical approaches, they offer less flexibility and/or scalability in handling missing value, typing error, evolution modeling and more complex scenarios on the horizon in haplotype modeling (e.g., recombinations, genetic mapping, etc.). Most current statistical methods for haplotype inference bypass the issue of ancestral-space uncertainty via an *ad hoc* specification of the number of haplotypes needed to account for the given genotypes. Although certain coalescent-based models[14] or model-selections methods can partially address these issues.

Besides the ancestral-space uncertainty issue discussed above, the haplotype models developed so far are still limited in their flexibility and are inadequate for addressing many realistic problems. Consider for example a genetic demography study, in which one seeks to uncover ethnic- and/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 we may also wish to discover which haplotypes are shared between subpopulations, and what are their frequencies. Empirical and theoretical evidence suggests that an early split of an ancestral population following a populational bottleneck (e.g., due to sudden migration or environmental changes) may lead to ethnic-group-specific populational diversity, which features both ancient haplotypes (that have high variability) shared among different ethnic groups, and modern haplotypes (that are more strictly conserved) uniquely present in different ethnic groups[22]. This structure is analogous to a hierarchical clustering setting in which different groups comprising

[†]A prevalent form of f in the literature is $f \equiv \mathbb{I}(h_1 \oplus h_2 = g)$, which is a deterministic indicator function of the event that haplotypes h_1 and h_2 are consistent with g. More desirable forms of f would model errors in the genotyping or data recording process, a point we will return to later in the paper.

multiple clusters may share clusters with common centroids.

In this paper, we describe a new class of haplotype inference models based on a nonparametric Bayesian formalism built on the *Dirichlet process* (DP)[23,24], which offers a well-founded statistical framework to tackle the problems discussed above more efficiently and accurately. As we discuss in the sequel, the Dirichlet process can induce a partition of an unbounded population in a way that is closely related to the Ewens sampling formula[25], thus it can be viewed as an exchangeable approximation to the joint distribution of population haplotypes under a coalescent process. On the other hand, in the setting of mixture models, the Dirichlet process is able to capture uncertainty about the number of mixture components[26]. A hierarchical extension of DP also leads to an elegant model that couples the demographic information in different populations for solving multi-population haplotype inference problems.

Our model differs from the extant models in the following important ways: 1) Instead of resorting to *ad hoc* parametric assumptions or model selection over the number of population haplotypes, as in many parametric Bayesian models, we introduce a nonparametric prior over haplotypes ancestors, which facilitates posterior inference of the haplotypes (and other genetic properties of interest) in an "open" state space that can accommodate arbitrary sample size. 2) Our model captures similar genetic features as those emphasized in Stephens et al.[9], including the parent-dependent-mutation property, but with an *exchangeable* likelihood function rather than an order-dependent one as in the PAC model. 3) The hierarchical Bayesian framework of our model explicitly captures ancestral/population structures and incorporates demographic/genetic parameters so that they can be inferred or estimated along with the haplotype phase based on given genotype data. 4) Our model can explicitly exploit the ethnic labels, and potentially latent sub-population structures of the sample, to improve haplotyping accuracy. Some fragments of the technical aspects of the proposed model was announced before in conferences in the machine learning community[27,28], but to our knowledge the full statistical model and its population genetic interpretations are new to the genetics community, and a computer program based on this model for haplotype reconstruction from large genotype data is not yet available. In this paper, we describe this new nonparametric Bayesian approach for haplotype modeling in detail, and we present an efficient Monte Carlo algorithm, *Haploi*, for haplotype inference based on the proposed model, which is readily applicable to genotype sequences with thousands of SNPs, at a time-cost often at least two-orders of magnitude less than that of the state-of-the-art PHASE program, with competitive and sometimes superior performance (mostly in long sequences). We also show that *Haploi* significantly outperforms several other extant haplotype inference algorithms on both simulated and realistic data. A C++ implementation of *Haploi* can be obtained from the authors via email request, and will be soon made public on world wide web once interface and GUI development are completed.

2 The Statistical Model

As motivated in the introduction, it is desirable to explicitly explore the demographic characteristics such as population structure and ethnic label, and the genetic scenarios such as coalescence and mutation, underlying the study populations, while performing haplotype inference on complex population samples. In the following, we present two novel nonparametric Bayesian models for

haplotype inference that facilitate this desire. We begin with a basic model for the simplest demographic and genetic scenario, in which we ignore individual ethnic labels in the sample (as in most extant haplotyping methods), and assume absence of recombination in the sample. Then we generalize this model to a multi-population scenario. Finally we deal with long genotypes with recombinations with an algorithmic approach motivated by the *partition-ligation* scheme[12].

2.1 A Dirichlet process mixture model for haplotypes

2.1.1 Dirichlet process mixture

Given a sample of n chromosomes, under neutrality and random-mating assumptions, the distribution of the *genealogy trees* of the sample can be approximated by that of a random tree known as the n-coalescent[29]. Additionally, on each lineage there is a point process of mutation events. The best understood, and also the simplest instances of such mutation processes is the *infinitely-many-alleles* (IMA) model, in which each mutation in the lineage produces a novel type, independent of the parental allele. IMA can be understood as an independent Poisson process with rate, say, $\alpha/2$, which is determined by the size of the evolving population N (usually $N >> n$) and the per-generation mutation rate μ (i.e., $\alpha = 4N\mu$). Although easy to analyze, IMA is unrealistic because it fails to capture dependencies among haplotypes. On the other hand, to date no closed-form expression of $P(H)$ is known for the more realistic *parent-dependent mutation* (PDM) model under the n-coalescent; approximations such as the PAC model has been used as a tractable surrogate.

In the sequel, we describe an alternative approach for modeling $P(H)$ based on a nonparametric Bayesian formalism known as the *Dirichlet process*, which leads to a new class of models for haplotype distribution that approximately captures major properties that would result from a *coalescent-with-PDM* model.

We begin with a brief genetic motivation of the proposed approach. Rather than focusing on a complete random genealogy up to the MRCA, we instead consider a sample of n individuals from a population characterized by an unknown set of *founding haplotypes*, with unknown *founder frequencies*. For now we focus attention on a small chromosomal region within which the possibility of recombination over relevant time-scales is negligible. As a consequence, we postulate that each individual's genotype is formed by drawing two random haplotype *founders* from an ancestral pool, one for each of the two actual haplotypes of this individual, which can be mutated version of their corresponding founders. We further assume that we are given noisy observations of the resulting genotypes. Below we show how this setting relates to the *coalescent-with-IMA* and *coalescent-with-PDM* models.

Under Kingman's *coalescent-with-IMA* model, one can treat a haplotype from a modern individual as a descendent of a most resent common ancestor with unknown haplotype via random mutations that alter the allelic states of some SNPs[29]. Hoppe[30] observed that a coalescent process in an infinite population leads to a partition of the population at every generation that can be succinctly captured by the following Pólya urn scheme.

Consider an urn that at the outset contains a ball of a single color. At each step we either draw a ball from the urn and replace it with two balls of the same color, or we are given a ball of a new color which we place in the urn. One can see that such a scheme leads to a partition of

the balls according to their color. Mapping each ball to a haploid individual and each color to a possible haplotype, this partition is equivalent to the one resulted from the *coalescence-with-IMA* process[30], and the probability distribution of the resulting *allele spectrum*—the numbers of colors (i.e., haplotypes) with every possible number of representative balls (i.e., decedents)—is captured by the well-known Ewens' sampling formula[25].

Letting parameter α define the probabilities of the two types of draws in the aforementioned Pólya urn scheme, and viewing each (distinct) color as a sample from Q_0, and each ball as a sample from Q, Blackwell and MacQueen[24] showed that this Pólya urn model yields samples whose distributions are those of Q_0 the marginal probabilities under the *Dirichlet process*[23]. Formally, a random probability measure Q is generated by a DP if for any measurable partition $A_1, \ldots, A_k$ of the sample space (e.g., the partition of an unbounded haploid population according to common haplotype patterns), the vector of random probabilities $Q(A_i)$ follows a Dirichlet distribution: $(Q(A_1), \ldots, Q(A_k)) \sim \text{Dir}(\alpha Q_0(A_1), \ldots, \alpha Q_0(A_k))$, where α denotes a *scaling parameter* and Q_0 denotes a *base measure*. The Pólya urn construction of DP makes explicit an order-independent sequential sampling scheme to draw samples from a DP. Specifically, having observed n samples with values $(\phi_1, \ldots, \phi_n)$ from a Dirichlet process $DP(\alpha, Q_0)$, the distribution of the value of the $(n+1)$th sample is given by:

$$\phi_{n+1}|\phi_1, \ldots, \phi_n, \alpha, Q_0 \sim \sum_{k=1}^{K} \frac{n_k}{n+\alpha}\delta_{\phi_k^*}(\cdot) + \frac{\alpha}{n+\alpha}Q_0(\cdot), \tag{3}$$

where $\delta_{\phi_k^*}(\cdot)$ denotes a point mass at a unique value ϕ_k^*, n_k denotes the number of samples with value ϕ_k^*, and K denotes the number of unique values in the n samples drawn so far. This conditional distribution is useful for implementing Monte Carlo algorithms for haplotype inference under DP-based models. We will return to this point in the Appendix.

Under a DP distribution described above, the sampled haplotypes follow an IMA model, meaning that all different haplotypes (i.e., ball colors) are independent. How can we take into consideration the fact that, in a real haplotype sample one would expect that some haplotypes differ only slightly (i.e., at a few SNP loci) whereas some differ much more significantly—a phenomenon caused by possibly *parent-dependent mutations*. Now we describe a *DP mixture* (DPM) model that approximate this effect.

In the context of mixture models, one can associate common data centroids, i.e., *haplotype founders* rather than all distinct haplotypes, with colors drawn from the Pólya urn model and thereby define a "clustering" of the (possibly noisy) data $\{h_i\}$ (e.g., modern haplotypes that are "recognizable" variants of their corresponding founders) via likelihood function $p(h_i|\phi_i)$. As obvious from Eq. (3), the samples (i.e., ball-draws) $\{\phi_i\}$ from a DP (i.e., the urn) tend to concentrate themselves around some unique values $\{\phi_k^*\}$ (i.e., colors); thus conditioning on each such unique value ϕ_k^*, we have a mixture component $p(h_i|\phi_k^*)$ for the data. Such a mixture model is known as the DP mixture[26,31]. Note that a DP mixture requires no prior specification of the number of components, which is typically unknown in genetic demography problems. It is important to emphasize that here DP is used as a *prior distribution* of mixture components. Multiplying this prior by a likelihood that relates the mixture components to the actual data yields a *posterior distribution* of the mixture components, and the design of the likelihood function is completely up to the

modeler based on specific problems. This nonparametric Bayesian formalism forms the technical foundation of the haplotype modeling and inference algorithms to be developed in this paper.

2.1.2 DPM for haplotype inference

Now we briefly recapitulate the basic DPM for haplotypes first proposed in Xing et al.[27]. In the next section we generalize this model to multi-population haplotypes, and describe specific Bayesian treatments of all relevant model parameters.

Write $H_{i_e} = [H_{i_e,1}, \ldots, H_{i_e,T}]$, where the sub-subscript $e \in \{0, 1\}$ denotes the two possible parental origins (i.e., paternal and maternal), for a haplotype over T contiguous SNPs from individual i; and let $G_i = [G_{i,1}, \ldots, G_{i,T}]$ denote the genotype these SNPs of individual i. For diploid organisms such as human, we denote the two alleles of a SNP by 0 and 1; thus each $G_{i,t}$ can take on one of four values: 0 or 1, indicating a homozygous site; 2, indicating a heterozygous site; and '?', indicating missing data. (A generalization to polymorphisms with k-ary alleles is straightforward, but omitted here for simplicity.) Let $A_k = [A_{k,1}, \ldots, A_{k,T}]$ denote an ancestor haplotype (indexed by k) and θ_k denote the *mutation rate* of ancestor k; and let C_i denote an *inheritance variable* that specifies the ancestor of haplotype H_i. We write $P_h(H|A)$ for the *inheritance model* according to which individual haplotypes are derived from a founder, and $P_g(G|H_0, H_1)$ for the *genotyping model* via which noisy observations of the genotypes are related to the true haplotypes. Under a DP mixture, we have the following Pólya urn scheme for sampling the genotypes, G_i, $i = 1, \ldots, n$, of a sample with n individuals:

- Draw first haplotype:

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

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

- for subsequent haplotypes:

  – sample the founder indicator for the ith haplotype:

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

  where n_{c_i} is the *occupancy number* of class c_i—the number of previous samples generated from founder a_{c_i}.

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

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

– sample the haplotype according to its founder:

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

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

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

Under appropriate parameterizations of the base measure Q_0, the inheritance model P_h, and the genotyping model P_g, which will be described in detail shortly, the problem of phasing individual haplotypes and estimating the size and configuration of the latent ancestral pool under a DPM model can be solved via posterior inference given the genotypes from a (presumably ethnically homogeneous) population descending from a single pool of ancestors, using, for example, a Gibbs sampler as we will outline in the Appendix.

As mentioned earlier, treating haplotype distribution as a mixture model, where the set of mixture components correspond to the pool of ancestral haplotypes, or *founders*, of the population, can be justified by straightforward statistical genetics arguments. Crucially, however, the size of this pool is unknown; indeed, knowing the size of the pool would correspond to knowing something significant about the genome and its history. In most practical population genetic problems, usually the detailed genealogical structure of a population (as provided by the coalescent trees) is of less importance than the population-level features such as pattern of common ancestor alleles (i.e., founders) in a population bottleneck, the age of such alleles, etc. In this case, the DP mixture offers a principled approach to generalize the finite mixture model for haplotypes to an infinite mixture model that models uncertainty regarding the size of the ancestor haplotype pool, and at the same time it provides a reasonable approximation to the coalescence-with-PDM model by utilizing the partition structure resulted thereof, but allowing further mutations within each partite to introduce further diversity among descents of the same founder.

2.2 A Hierarchical DP mixture model for multi-population haplotypes

Now we consider the case in which there exist multiple sample populations (e.g., ethnic groups), each modeled by a distinct DP mixture. Note that now we have multiple ancestor pools, one for each attendant population; instead of modeling these populations independently, we place all the population-specific DPMs under a common prior, so that the ancestors (i.e., mixture components) 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. Genetically, this means that for every possible ancestral haplotype, it can either be a founder in only one of the populations, or be a founder shared in some or all attendant populations; in the latter case, the frequencies of this haplotype founder being inherited are different in different populations.

To tie population-specific DP mixtures together in this way, we use a hierarchical DP mixture model[32], in which the base measure associated with each population-specific DP mixture is itself drawn from a higher-level Dirichlet process $\mathrm{DP}(\gamma, F)$. Since a draw from this higher-level DP is a discrete measure with probability 1, atoms drawn by different population-specific DPs from $\mathrm{DP}(\gamma, F)$—the haplotype founders and its mutation rate $\phi_k \equiv \{A_k, \theta_k\}$, which are used as the

mixture components in each of the population-specific DP mixtures—are not going to be all distinct (i.e., the same (A_k, θ_k) can be drawn by two different population-specific DPMs). This allows sharing of components across different mixture models.

2.2.1 Hierarchical Dirichlet Process

Before presenting the HDP mixture for haplotypes, we digress with a brief description of the HDP formalism. As with the DP, it is useful to describe the marginals induced with an HDP using the more concrete representation of Pólya urn models. Imagine we set up a single "stock" urn at the top level, which contains balls of colors that are represented by at least one ball in one or multiple urns at the bottom level. At the bottom level, we have a set of *distinct* urns which are used to define the DP mixture for each population. Now let's suppose that upon drawing the m_j-th ball for urn j at the bottom, the stock urn contains n balls of K distinct colors indexed by an integer set $\mathcal{C} = \{1, 2, \ldots, K\}$. Now we either draw a ball randomly from urn j, and place back two balls both of that color, or with some probability we return to the top level. From the stock urn, we can either draw a ball randomly and put back two balls of that color in the stock urn and one in j, or obtain a ball of a new color $K+1$ with probability $\frac{\gamma}{n-1+\gamma}$ and put back a ball of this color in both the stock urn and urn j of the lower level. Essentially, we have a master DP (the top urn) that serves as a source of atoms for J child DPs (bottom urns).

Associating each color k with a random variable ϕ_k whose values are drawn from the base measure F, and recalling our discussion in the previous section, we know that draws from the stock urn can be viewed as marginals from a random measure distributed as a Dirichlet Process Q_0 with parameter (γ, F). From Eq. (3), for n random draws $\phi = \{\phi_1, \ldots, \phi_n\}$ from Q_0, the conditional prior for $(\phi_n | \phi_{-n})$, where the subscript "$-n$" denotes the index set of all but the n-th ball, is

$$\phi_n | \phi_{-n} \sim \sum_{k=1}^{K} \frac{n_k}{n-1+\gamma} \delta_{\phi_k^*}(\phi_n) + \frac{\gamma}{n-1+\gamma} F(\phi_i), \tag{4}$$

where $\phi_k^*, k = 1, \ldots, K$ denote the K distinct values (i.e., colors) of ϕ (i.e., all the balls in the stock urn), and n_k denote the number of balls of color k in the top urn.

Conditioning on Q_0 (i.e., using Q_0 as an atomic base measure of each of the DPs corresponding to the bottom-level urns), the m_j-th draws from the jth bottom-level urn are also distributed as marginals under a Dirichlet measure:

$$\begin{aligned} \phi_{m_j} | \phi_{-m_j} &\sim \sum_{k=1}^{K} \frac{m_{j,k} + \tau \frac{n_k}{n-1+\gamma}}{m_j - 1 + \tau} \delta_{\phi_k^*}(\phi_{m_j}) + \frac{\tau}{m_j - 1 + \tau} \frac{\gamma}{n-1+\gamma} F(\phi_{m_j}) \\ &= \sum_{k=1}^{K} \pi_{j,k} \delta_{\phi_k^*}(\phi_{m_j}) + \pi_{j,K+1} F(\phi_{m_j}), \end{aligned} \tag{5}$$

where $\pi_{j,k} := \frac{m_{j,k} + \tau \frac{n_k}{n-1+\gamma}}{m_j - 1 + \tau}$, $\pi_{j,K+1} = \frac{\tau}{m_j - 1 + \tau} \frac{\gamma}{n-1+\gamma}$, and $m_{j,k}$ denotes the number of balls of color k in the j-th bottom urn. In our case, $\pi_j = (\pi_{j,1}, \pi_{j,2}, \ldots)$ gives the *a priori* frequencies (i.e., mixture weighs) of haplotype founders in population j.

2.2.2 HDPM for multi-population haplotype inference

Using the HDP construction described above, we now define an HDP mixture model for the genotypes in J populations. Elaborating on the notational scheme used earlier, let $G_i^{(j)} = [G_{i,1}^{(j)}, \ldots, G_{i,T}^{(j)}]$ 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 basic generative structure of multi-population genotypes under an HDPM is as follows, which is also illustrated graphically in Figure 1.

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

where in practice the first three sampling steps follow the nested Pólya urn scheme described above. Note that in the HDP the base measure of each lower-level DP is a draw from the root $\mathrm{DP}(\gamma, F)$. From this description, it is apparent that the totality of all atomic samples, i.e., all instantiated haplotype founders and their associated mutation rates, from the base measure F form the common support (i.e., candidate founder patterns) of both the root DP and all the population-specific DPs. The child DPs place different mass distributions, i.e., *a priori* frequencies of haplotype founders, on this common support, in a population-specific fashion.

Recall that the base measure F in the above generative process is defined as a distribution from which haplotype founders $\phi_k := \{A_k, \theta_k\}$ are drawn. Thus it is a joint measure on both A and θ. 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 PDM mutation rate. For generality, we assume each $A_{k,t}$ (and also each $H_{i,t}$) takes its value from an allele set $\mathcal{P}$. For other building blocks of the HDPM model, we propose the following specifications.

Haplotype inheritance model: Omitting all but the locus index t, we can define our inheritance model to be a *single-locus mutation model* as follows[27]:

$$P_h(h_t|a_t, \theta) = \theta^{\mathbb{I}(h_t \neq a_t)} \left(\frac{1-\theta}{|\mathcal{P}|-1}\right)^{\mathbb{I}(h_t = a_t)} \tag{6}$$

where $\mathbb{I}(\cdot)$ is the indicator function. This model corresponds to a star genealogy resulting from infrequent mutations over a shared ancestor, and is widely used as an approximation to a full coalescent genealogy starting from the shared ancestor (e.g., Liu et al.[33]).

Given this inheritance model, it can be shown that the marginal conditional distribution of a haplotype sample $\mathbf{h} = \{h_{i_e} : e \in \{0, 1\}, i \in \{1, 2, ..., I\}\}$ takes the following form resulted from

an integration of θ in the joint conditional:

$$p(\mathbf{h}|\mathbf{a},\mathbf{c}) = \prod_{k=1}^{K} R(\alpha_h,\beta_h)\frac{\Gamma(\alpha_h+l_k)\Gamma(\beta_h+l'_k)}{\Gamma(\alpha_h+\beta_h+l_k+l'_k)}\left(\frac{1}{|\mathcal{P}|-1}\right)^{l'_k}, \tag{7}$$

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

Genotype observation model: Next, we assume that the observed genotype at a locus is determined by the paternal and maternal alleles of this site via the following genotyping model[27]:

$$P_g(g|h_{i_0,t},h_{i_1,t},\tau) = \xi^{\mathbb{I}(h=g)}[\mu_1(1-\xi)]^{\mathbb{I}(h\neq^1 g)}[\mu_2(1-\xi)]^{\mathbb{I}(h\neq^2 g)} \tag{8}$$

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

Hyperprior for coalescent rate: Lastly, to capture uncertainty over the scaling parameters γ (or α in the single-layer DPM model), which is twice the mutation rate in the coalescent over the haplotype founders, we use a vague inverse Gamma prior:

$$p(\gamma^{-1}) \sim \mathcal{G}(1,1) \Rightarrow p(\gamma) \propto \gamma^{-2}\exp(-1/\gamma)). \tag{9}$$

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:

$$p(\gamma|k,n) \quad \propto \quad \frac{\gamma^{k-2}\exp(1/\gamma)\Gamma(\gamma)}{\Gamma(n+\gamma)}. \tag{10}$$

The distribution $p(\log(\gamma)|k,n)$ is log-concave, so we may efficiently generate independent samples from this distribution using adaptive rejection sampling[34].

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.

Putting everything together, we have constructed a HDPM model for multi-population haplotypes. The two-level nested Pólya urn schemes described above motivates an efficient and easy-to-implement MCMC algorithm to sample from the posterior associated with HDPM, which is similar to the MCMC algorithms developed for DPM. We will give details of this algorithms in an Appendix that is available in the electronic version of this article.

3 Partition-ligation and the *Haploi* program

As for most of the 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 can not 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.[12] proposed a divide-and-conquer heuristic known as Partition-Ligation (PL), which was adopted by a number of haplotype inference algorithms including PL-EM[35], PHASE[9,36], and CHB[14]. We equipped our haplotype inference algorithm based on the HDPM model with a variant of the PL heuristic, and present a new tool, *Haploi*, for *haplo*type *i*nference of either single or multiple population genotype data containing thousands of SNPs.

Unlike the original PL-scheme in Niu et al.[12], which works on disjoint blocks and then recursively ligate the phased blocks into larger (non-overlapping) haplotypes via Gibbs sampling in the *product space* of all the "atomistic haplotypes" of every attendant pair of blocks to be ligated, under a fixed-dimensional Dirichlet prior of the frequencies of the ligated haplotype ; our PL-scheme generate partially overlapping intermediate blocks from smaller blocks phased at the lower level, and the pairs of overlapping blocks are recursively merged into larger ones by leveraging the redundancy of information from overlapping regions, as well as an overall parsimonious criteria. Empirically we found that these strategy can lead to a significant reduction of the size of the haplotype search space for long genotypes such as those with thousands of SNPs, and therefore facilitate a more efficient inference algorithm.

Figure 2 outlines the PL-procedure adopted by our program *Haploi*. 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.[12]). Then we determine the "atomistic haplotypes" of each block using HDPM. In the first ligation step, we ligate every neighboring pairs of nonoverlapping blocks, $B_1\&B_2, B_2\&B_3, \ldots, B_{L-1}\&B_L$, into $L-1$ overlapping blocks $\{B'_j : j = 1 : L-1\}$, each of length $2T$, using the Gibbs sampling method used in Niu et al.[12]. To compensate the obviously ill-ligated blocks, we do additional HDPM inference for those blocks whose entropy of haplotype distribution is above some threshold. 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. For subsequent ligations of partially overlapping blocks, when the overlapping regions of a pair of atomistic haplotypes in the attendant (adjacent) SNP genotype blocks are consistent, ligation to a longer haplotype is trivially a merging of the overlapping haplotypes. Only when the overlapping regions are inconsistent, we grow the haplotype space of the ligated blocks by including the "product" of the two inconsistent "atomistic haplotypes", i.e., all possible ligations consistent with either of the atomistic haplotypes and the overall genotype. Specifically, suppose there are discrepancies in an estimated individual

haplotype on the overlapping region between B'_i and B'_{i+1}. We do not discard any estimated haplotypes from B'_i and B'_{i+1}, but instead add all possible haplotypes consistent with the genotype formed by combining the segment in B'_i on positions $1, \ldots, (T + \tau)$ and the segment in B'_{i+1} on positions $(\tau + 1)$,...,$(2T)$ where $T + \tau$ represents the location of discrepancy. When the overlapping regions are homozygous, then there would not be any discrepancy, but we cannot resolve the phase of the legated blocks. In this case we again include into our haplotype space for the ligated blocks all possible combinations of haplotype pairs from B'_i and B'_{i+1}. This heuristic would typically result in a haplotype space for an ligated-block of length 3T that is much smaller than the trivial product-space of nonoverlapping lower-level blocks. Then we apply a Gibbs sampler as in Niu et al.[12] to determine all individual haplotypes of the ligated-block under a fixed-dimensional Dirichlet prior of the haplotype frequencies in the trimmed haplotype space. Since each time we only employ overlapping regions of size T, the number of steps needed to complete the ligation of a long sequence is roughly the same as needed in the original hierarchical PL-scheme in Niu et al.[12].

4 Results

In this section, we present a comparison of *Haploi*, with PHASE 2.1.1[9,36], fastPHASE[37], HAPLOTYPER 1.0[12], and CHB 1.0[14].

We run each program using its default parameter settings. Three different error measures were used for evaluation: err_s, the ratio of incorrectly phased SNP sites over all non-trivial heterozygous SNPs (excluding individuals with a single heterozygous SNP), and err_i, the ratio of incorrectly phased individuals over all non-trivial heterogeneous individuals (i.e., those with at least two heterogeneous SNPs), and d_w, the switch distance, which is the number of phase flips required to correct the predicted haplotypes over all non-trivial heterogeneous SNPs. Both short (~ 10 SNPs) and long ($10^2 \sim 10^3$) sequences were tested when the program permits[‡]. For short SNP sequences, we primarily use err_s and err_i; whereas for long sequences we compare d_w according to common practice, since it is regarded as a more sensible indicator of performance for longer SNP sequences. On short SNPs, we test on a large number of samples and report summary statistics of errors over all samples; whereas for long SNPs, we present error over each of samples (of different lengths) we tested due to heavy computational cost.

We have also estimated other population genetic metrics of interest, such as the haplotype frequencies, the mutation rates θ, and the number of reconstructed haplotype founders K, under the HDP and DP models. We will present some of these results to illustrate consistency of our model (on simulation), and the characteristics of some real data set being studied.

[‡]Specifically, we applied fastPHASE only on long SNP sequences since it is tailored for fast processing of long sequences at the cost of reduced accuracy; and we applied both HAPLOTYPER 1.0 and CHB 1.0 only on short SNP sequences since they can not reach convergence within acceptable test time (> 40 hours) on samples with more than 200 SNPs.

4.1 Simulated Multi-Population SNP Data

For simulation-based tests, we used a pool of haplotypes taking from the coalescent-based synthetic dataset in Stephens et al.[9], 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 100 individuals consisting of five populations, each with 20 individuals. Each population is derived from 5 founders, two of which are shared among all populations, and the other three are population-specific. Thus overall the total number of founders across the five populations is 17. We test our algorithm on two data sets with different degree of sequence diversity. In the *conserved* data set, we assume the mutation rate θ to be 0.01 for all populations and all loci; 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.

4.1.1 Haplotype Accuracy

We compare *Haploi*, implemented either based on a DPM model or an HDPM model (dubbed as *Haploi*-HDP and *Haploi*-DP, respectively, when distinction is needed; otherwise, we simply use *Haploi* for *Haploi*-HDP), with extant phasing methods applied in two modes on the synthetic data sets. As mentioned in the introduction, given multi-population genotype data, to use an extant method, 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*-HDP takes a different approach, by making explicit use of the population labels (if available) and jointly solve multiple coupled haplotype inference problems. (Note that when only a single population is concerned, or no population label is available, *Haploi*-HDP is still valid, and is equivalent to a baseline *Haploi*-DP with one more layer of DP hyper-prior). We apply our method to the simulated multi-population data, and compare its overall performance on the whole data with the outcomes of other algorithms run in mode-I, and compare the score of our method on each population with the outcome of extant methods run in mode-II.

Figure 3 summarizes the overall performance of *Haploi* on 50 conserved simulated samples and 50 diverse simulated samples, along with those of the reference algorithms run in mode-I. On the conserved samples, which are presumably easier to phase, *Haploi*-HDP outperforms all the other algorithms appreciably. On the diverse samples, which are more challenging to phase (due to more severe inconsistencies among individual haplotypes in the samples caused by high mutation rate), *Haploi*-HDP outperforms all other algorithms with a significant margin.

[§]Here our simulation scheme assumes a star-genealogy with uniform mutation rate for each sub-population. This simulation scheme is simple to implement under our multi-population scenario for testing the bias/variance of a number of estimators of interest, and it reasonably approximates the genetic demography of samples under an IMA model on a coalescent tree. We have tested an early version of *Haploi*, which employed an DPM (rather than HDPM) without partition-ligation, on the single-population, coalescent-based simulated data used in Stephens et al.[9] in comparison with PHASE (see Xing et al.[27]), and the results were similar.

Haploi-HDP also dominates other methods when the latter are run in mode-II on the simulated data. Table 1 shows a comparison of the accuracy of *Haploi*-HDP on each sub-population (directly extracted from results obtained in a single run of *Haploi*-HDP on all populations) in a simulated data set with results from separate runs of the other algorithms on each sub-population of genotypes. Note that, here K is expected to be 5 for each group, and the estimation by *Haploi*-HDP is quite close to this value. On the conserved data set, CHB shows the best result while all algorithms performed comparably. On the more difficult diverse dataset, the HDP approach outperformed other algorithms and inferred the founders of each group more robustly than the group-specific runs using the baseline *Haploi*-DP (which employ independent DPs for each sub-population).

4.1.2 MCMC and parameter estimation

Typically, the Pólya urn Gibbs sampler for *Haploi* converges within 1000 iteration (figure not shown) on the synthetic data. This contracts sampling algorithms used in some of the other haplotype inference algorithms, which typically needs tens of thousands of samples 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 algorithms. In Figure 4 (a) and (b), we show the histogram of the estimated K— the number of recovered ancestors, across the 50 datasets via both of our algorithms. Recall that we expect K to be 17, and the estimated K under both the DP and HDP models turns out to be very close to this number on the *conserved* datasets (i.e., those with a small mutation rate); from the diverse data sets, *Haploi*-HDP can still offer a good estimate of the number of ancestors, whereas *Haploi*-DP recovered more ancestors (around 25 on average) than the true one. This is not surprising since a haplotype which appears in more than one population can have different frequencies in different populations, the baseline *Haploi*-DP can not capture such sub-population structure, and the higher divergence due to both mutation and population diversification can make it generate more templates to describe the given dataset. Note also that the parametric methods (PHASE, HAPLOTYPER and CHB) cannot provide an estimate of K. Here we report K', the number of distinct haplotypes present in the population for these algorithms: for the conserved data sets, the average number of distinct haplotypes across different data sets is 30.24, 29.86, 30.8 for PHASE, HAPLOTYPER and CHB, respectively; and for the diverse data sets, the averages were 63.28, 54.68, 56.38, respectively. Since for these methods the assignment of plausible haplotypes for a new sample must be made among all these candidates (rather than based on a smaller number of prototypes represented by the founders), the large magnitude of K' may partially explain the much higher sampling cost incurred by these methods.

Our Gibbs sampler also provides reasonable estimates of the mutation rates underlying the sample. Figure 4 (c) and (d) show the histograms of the estimated θ across the 50 datasets by *Haploi*-HDP and *Haploi*-DP for both the conserved and diverse cases. We observe that for the conserved data sets, *Haploi*-HDP yields highly consistent and low variance estimations of θ, and the quality of the estimates due to *Haploi*-DP is slightly worse. For the diverse data sets both algorithms tend to slightly underestimate the mutation rates, and 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 an-

cestors. So in fact K and θ can not be independently determined, possibly following a similar argument of the un-identifiability of the evolution time and population size under IAM model. But empirically, *Haploi*-HDP appears to struck a reasonable balance between K and θ, and offer plausible estimates of both.

Finally, we compare the accuracies of population haplotype frequencies estimated by each algorithm (Fig. 5). The discrepancy between the true frequencies and estimated ones is measure by two distance measures commonly used in the literature: the $L1$-norm $D_1(p,q) = \sum_x |p(x) - q(x)|/2$, as used in Stephens et al.[9], Excoffier and Slatkin[11]; and the KL-Divergence $D_{KL}(p||q) = \sum_x p(x) \log \frac{p(x)}{q(x)}$. The left column of Figure 5a reports the D_1 and D_{KL} computed on ALL haplotypes frequencies estimated by different algorithms from the conserved data sets. As shown there, $Haploi$-HDP achieves the lowest discrepancy by a significant margin over all other algorithms been compared. The runner-up, PHASE, beats the baseline $Haploi$-DP (in the third place) by a small margin. When measured only on the frequent haplotypes (i.e., with frequencies ≥ 0.05, shown in the right column of Figure 5a), the discrepancies decrease significantly, but the relative ordering of all the compared algorithms remain the same, except that now PHASE and $Haploi$-DP are almost tied at the second lowest. For the more difficult diverse data sets, the same tendency can be observed (Fig 5b).

4.2 The HapMap Data

We test our algorithms on both short SNP segments (i.e., ~ 10 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: CEPH (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 cases, is) performed on all populations, we evaluate the phasing outcome from all algorithms on only the CEPHs and Yorubas since their true haplotypes can be (almost) unambiguously deduced from trios. The loci that can not be unambiguously phased from the trios were excluded from our evaluation. We selected 10 ENCODE regions ranging each spanning roughly 500 Kb from the HapMap DB. Since each population may have SNPs in different chromosomal positions, we extracted the common SNPs across all the populations for our experiments, the resulting segments each contain from 254 to 972 SNPs (Table 2).

4.2.1 Short SNP sequences

Phasing short SNPs is the basic operation of large-scale haplotype inference problems which rely either on a partition-ligation heuristics, or on a model-based methods such as recombination process, to integrate short phased haplotype segments into long haplotypes. Figure 6 shows a comparison of the phasing accuracy on all 7-SNP segments (following a recommendation in Niu et al.[12] on the optimal size-range of basic units for subsequent ligation) from the ten ENCODE regions by 5 algorithms. Overall, in all ten regions, with significant margins, *Haploi*-HDP exhibits the lowest median error rates, and the smallest performance variance (as can be assessed from the range spanned by the upper and lower quartiles).

Recall that *Haploi*-HDP can exploit the population structure when available to form more reliable estimates of the haplotype founders, and thereby more accurate inference of the individual haplotypes. This is confirmed in our empirical experiment summarized in Figure 6a (left) and b (right). In Fig 6a, all algorithms were applied to genotype data from two populations, CEPH and Yoruba, whose true phase are known from trios. In Fig 6b, all algorithms were applied to genotype data from all 4 populations, although the outcomes were only validated only on CEPH and Yoruba. Thus in the second scenario we solve a bigger haplotype inference problem, on data that contain richer population information. Comparing the left and right panels of Figure 6, on the four-population phasing task, *Haploi*-HDP achieved lower median error rates in 7 out of the 10 ENCODE regions than on the two-population phasing task. On the contrary, the three parametric methods PHASE, HAPLOTYPER, and CHB, all appear not able to benefit from increased population diversity, and performed significantly worse in the four-population scenario than in the two-population scenario. This means that in reality, if the genotype data are collected from a highly heterogeneous population, these methods may offer compromised results.

Interestingly, although the baseline *Haploi*-DP is not doing well in the two population scenario, its performance is not compromised by the increased population diversity, and even improved in 4 of the 10 regions. As a result, it emerged as the second best method in the four-population phase task, dominating over all the three parametric methods. While *Haploi*-DP does not explicitly use sub-population structure, it is possibly that its underlying DP model are less confounded by the increased population diversity due to its parsimonious nature (i.e., maintaining a compact set of founders that explains the observations), and can exploit the increased abundance of data to obtain better estimates of the population metrics such founder types and haplotype frequencies.

4.2.2 Long SNP sequences

Figure 7 shows a comparison of haplotype reconstruction quality on the entire ENCODE regions described in Table 2, using PHASE, fastPHASE, and *Haploi* equipped with the PL heuristic[¶]; and as in the last section, we performed haplotype inference using each method under both a two-population scenario, and a four-population scenario. The lengths of these regions range from 254 to 972 SNPs, and as a result for three of the 20 experiments (10 regions and two scenarios) we could not get the output from PHASE after a 31-day run, so we omit the corresponding results in our summary figure.

Overall, with its sophisticated recombination model suitable in particular for long sequences, PHASE dominates *Haploi* with a small margin under the two-population scenario when it converges; and *Haploi* dominates fastPhase in most cases under the two-population scenario, also with a small margin (Fig. 7a). But in terms of computational cost, fastPhase was the fastest, it 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 (Fig. 8a). Under the four-population scenario, *Haploi* outperformed or virtually tied with PHASE in 6 of the 10 regions, and lost to PHASE in the

[¶]We could not get output of Haplotyper and CHB for these long sequences. Instead we included fastPhase result, which is said to be much faster than PHASE with a slight performance degradation[37].

remaining cases by small margins; but it outperformed fastPhase in almost all cases significantly (Fig. 7b). Time-wise, all methods took longer, but the overall trend is the same as in the two-population scenario (Fig. 8b).

In summary, our results shows that *Haploi* is competent and robust for phasing very long SNP sequences from diverse genetic origins at reasonable time cost, even though it has not yet employed any sophisticated way for processing long sequences, such as the recombination process, which was used by both PHASE and fastPhase. Since *Haploi* appeared to dominate these two methods over short SNPs, we believe that an upgrade that incorporates explicit recombination models in conjunction with the HDP model for long SNPs are likely to lead to more accurate haplotype reconstructions, as we will discuss in the Discussion section.

4.2.3 Mutation rates and population diversity underlying the HapMap data

As for the simulated data, we estimated the mutation rates at all different sites and different ENCODE regions with respect to their corresponding haplotype founders inferred under the proposed HDPM model. Figure 9 shows the histograms of these estimations for each of the four populations. We estimated the mean mutation rate for each population by fitting the histograms with an exponential distribution. Interestingly, the estimated mutation rate of the Yoruba population with African ancestry, which is around 0.010, is significantly higher than those of the other three populations, which are similar to each other (i.e., around 0.005). The Yoruba population also exhibits the highest ancestral diversity among all four population, reflected by the average number of haplotype founders uncovered for all 7-SNP segments of the ENCODE regions, whereas the Han Chinese and the Japanese populations are equally much less diverse (Fig 9e-h). Overall, although each population on average has only 4-6 founder for a DNA segment spanned by 7 SNPs, the total number of founders for regions of the same size across all population is over 10 (Fig 9i), indicating that on average 2 founder of each population are unique, while 3 are shared across all four populations. Of course, we would like to point out that such estimates should not be taken as the actual bottle-neck sizes of the attendant populations; they are merely the statistically inferred most parsimonious hypothesis based on all short chromosome regions (i.e., 7-SNP segments), which can statistically explain the observed genotype data.

5 Discussion

5.1 Demographic and statistical properties of DP and HDP mixtures

We have proposed a new Bayesian approach to haplotype inference for single and 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., haplotype founders) 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.

As in the PAC model[10], the generative process of a haplotype sample from a haplotype distribution $P(H)$ under the Dirichlet process mixture can be viewed as a sequential process that draw one haplotype at a time conditioning on all the previously drawn haplotypes; and our model also achieves the following four desirable properties[‖] captured in PAC, albeit in a very different way:

(1) the next haplotype is more likely to match a previously drawn haplotype;

(2) the probability of seeing a novel haplotype in the new draw increases as the rate of mutation increases;

(3) the probability of seeing a novel haplotype in the new draw decreases as the number of distinct haplotypes increases in the previous draws;

(4) 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.

The first three properties are obvious from the Pólya urn construction of DP. To see the fourth property, note that when a next haplotype is to be sampled, we pick an ancestor (with probability proportional to the number of progenies it has) of some previous 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 amongst haplotypes, by relating them to their corresponding ancestor (aka, haplotype 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 on top of its corresponding founders rather than been completely random, thereby achieve the PDM effect. Above these founders, we model their genealogy and type history by a *coalescent-with-IMA* model, whose resulting marginal (of the ancestors) is equivalent to that of the Dirichlet process. Here a new founder can be sampled independent of the type-history in the coalescent (rather according to a PDM) from the base measure, with probability proportional to the IAM 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-IAM model (i.e., a DP); 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 ancestor-dependent Poisson mutation process, to approximate the coalescent-with-PAM model for haplotype samples. The time of the bottleneck depends on the value of the scaling parameter α of the DP (which is twice the value of the IAM mutation rate). One can introduce a prior to this parameter (as described in our methodology section) 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

[‖] Indeed Li and Stephens[10] listed five desirable properties in any approximation to an intractable coalescent model of haplotypes, but the last one is for recombinant samples, a scenario not explicitly modeled here, but we will discuss both heuristic and principled treatment of this issue later in the paper.

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 explore in this paper.

5.2 Extensions

Unlike the models underlying PHASE and fastPhase, the HDP model underlying the *Haploi* program does not explicitly model the recombination process that shape the LD patterns of long SNP sequences. Since *Haploi* appeared to dominate PHASE and fastPhase over short SNPs (as shown in Figure 6), we believe that an upgrade that incorporates explicit recombination models in conjunction with the HDP model for long SNPs is likely to lead to more accurate haplotype reconstructions. The hidden Markov Dirichlet process recently developed by us to model recombination in open ancestral space offers a promising path for such an upgrade[38].

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. For example, given incomplete genotype data, one can define the unobserved genotypes as hidden variables, and process with the same Gibbs sampling algorithm given in the Appendix for haplotype inference with the addition of one more sampling step that imputes values for these hidden variables based on a proposal defined by the conditional distribution of genotypes given relevant haplotypes. In another possible extension, although in the present study we have assumed that the population structure—the ethnic 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.

Acknowledgments

Web Resources

The URLs for data presented herein are as follows:
International HapMap Project, http://www.hapmap.org/

A C++ implementation of *Haploi* can be obtained from the authors via email request, and will be soon made public at http://www.cs.cmu.edu/∼epxing/ once the interface and GUI development are completed.

References

1. The International SNP Map Working Group (2001). A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature*, 409:928 – 933.

2. A. Chakravarti (2001). Single nucleotide polymorphisms: . . .to a future of genetic medicine. *Nature*, 409:822–823.

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

4. A. Clark (1990). Inferences of haplotypes from PCR-amplified samples of diploid populations. *Mol Biol Evol*, 7:111–122.

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

6. M. W. Kenneth and A. G. Clark (2002). Linkage disequilibrium and the mapping of complex human traits. *Trends Genet*, 18(1):19–24.

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

8. M. Stephens and P. Donnelly (2000). Inference in molecular population genetics. *J Roy Statist Soc Series B*, 62:605–655.

9. M. Stephens, N. Smith, and P. Donnelly (2001). A new statistical method for haplotype reconstruction from population data. *Am J Hum Genet*, 68:978–989.

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

11. L. Excoffier and M. Slatkin (1995). Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Mol Biol Evol*, 12(5):921–7.

12. T. Niu, S. Qin, X. Xu, and J. Liu (2002). Bayesian haplotype inference for multiple linked single nucleotide polymorphisms. *Am J Hum Genet*, 70:157–169.

13. G. Kimmel and R. Shamir. Maximum likelihood resolution of multi-block genotypes. In *Proc 8th Ann Int Conf Res Comp Mol Biol (RECOMB)*, pages 2–9, 2004.

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

15. M.E. Hawley and K.K. Kidd (1995). Haplo: a program using the EM algorithm to estimate the frequencies of multi-site haplotypes. *J Hered*, 86(5):409–11.

16. J.C. Long, R.C. Williams, and M. Urbanek (1995). An EM algorithm and testing strategy for multiple-locus haplotypes. *Am J Hum Genet*, 56(3):799–810.

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

18. A. Clark, K. M. Weiss, D. A. Nickerson, S. L. Taylor, A. Buchanan, J. Stengard, V. Salomaa, E. Vartiainen, M. Perola, E. Boerwinkle, and C. F. Sing (1998). Haplotype structure and population genetic inferences from nucleotide-sequence variation in human lipoprotein lipase. *Am J Hum Genet*, 63:595–612.

19. D. Gusfield. Haplotyping as perfect phylogeny: Conceptual framework and efficient solutions. In *Proc 6th Ann Int Conf Res Comp Mol Biol (RECOMB)*, pages 166–175, 2002.

20. E. Eskin, E. Halperin, and R.M. Karp (2003). Efficient reconstruction of haplotype structure via perfect phylogeny. *J Bioinf Comput Biol*, 1:1–20.

21. D. Gusfield. An overview of combinatorial methods for haplotype inference. Technical Report, UC Davis, 2004. URL `http://wwwcsif.cs.ucdavis.edu/~gusfield/hapreview.pdf`.

22. J. K. Pritchard (2001). Are rare variants responsible for susceptibility to complex disease? *Am J Hum Genet*, 69:124–137.

23. T. S. Ferguson (1973). A Bayesian analysis of some nonparametric problems. *Ann Statist*, 1: 209–230.

24. D. Blackwell and J. B. MacQueen (1973). Ferguson distributions via Pólya urn schemes. *Ann Statist*, 1:353–355.

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

26. M. D. Escobar and M. West (2002). Bayesian density estimation and inference using mixtures. *J Amer Statist Assoc*, 90:577–588.

27. E.P. Xing, R. Sharan, and M.I Jordan. Bayesian haplotype inference via the Dirichlet process. In *Proc 21th Int Conf on Machine Learning*, pages 879–886, New York, 2004. ACM Press.

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

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

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

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

32. Y. Teh, M. I. Jordan, M. Beal, and D. Blei (2006). Hierarchical Dirichlet processes. *J Amer Statist Assoc* (to appear).

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

34. C. E. Rasmussen. The infinite Gaussian mixture model. In *Advances in Neural Information Processing Systems 12*, pages 554–560, Cambridge, MA, 2000. MIT Press.

35. Zhaohui S. Qin, Tianhua Niu, and Jun S. Liu (2002). Partition-ligationexpectation-maximization algorithm for haplotype inference with single-nucleotide polymorphisms. *Am J Hum Genet*, 71:1242–1247.

36. M. Stephens and P. Scheet (2005). Accounting for decay of linkage disequilibrium in haplotype inference and missing data imputation. *Am J Hum Genet*, 76:449–462.

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

38. Eric P. Xing and Kyung-Ah Sohn (2007). Hidden Markov Dirichlet process: Modeling genetic recombination in open ancestral space. *Bayesian Analysis* (to appear).

Table 1: Comparison of *Haploi* and other algorithms run in mode-II on the synthetic multi-population data.

		Haploi-HDP			*Haploi*-DP			PHASE		HAPLOTYPER		CHB	
θ	pop	err_s	err_i	K	err_s	err_i	K	err_s	err_i	err_s	err_i	err_s	err_i
0.01	(1)	0.0159	0.0556	5	0.0159	0.0556	5	0.0000	0.0000	0.0159	0.0556	0.0238	0.0714
	(2)	0.0000	0.0000	5	0.0175	0.0590	5	0.0000	0.0000	0.0526	0.0588	0.0152	0.0625
	(3)	0.0141	0.0625	4	0.0000	0.0000	5	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	(4)	0.0366	0.1765	4	0.0244	0.0590	5	0.0366	0.1765	0.0244	0.1176	0.0448	0.2143
	(5)	0.0000	0.0000	5	0.0244	0.0710	7	0.0488	0.0714	0.0732	0.1429	0.0000	0.0000
	avg	0.0133	0.0589		0.0164	0.0489		0.0171	0.0496	0.0332	0.0749	0.0167	0.0696
0.05	(1)	0.0758	0.2780	5	0.0758	0.3330	6	0.1970	0.6111	0.0758	0.2222	0.1429	0.4118
	(2)	0.1640	0.5000	5	0.1640	0.5560	8	0.1148	0.3333	0.1967	0.4444	0.1250	0.3529
	(3)	0.0886	0.4120	5	0.1140	0.5290	5	0.1013	0.4706	0.1139	0.5294	0.0877	0.3333
	(4)	0.0455	0.2110	5	0.0568	0.3680	10	0.1705	0.6316	0.1136	0.4737	0.1167	0.4000
	(5)	0.1640	0.4120	7	0.2180	0.4120	6	0.1818	0.4706	0.1273	0.4118	0.0921	0.3125
	avg	0.1076	0.3626		0.1257	0.4396		0.1531	0.5034	0.1255	0.4163	0.1129	0.3621

Table 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

Figure 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 $P_h(\cdot|a_k)$ that allows mutations with respect to the founders, as indicated by the underscores at the mutated loci in the individual haplotypes. Finally, the genotype observations are related to the haplotype pairs of every individual via a noisy channel $P_g(\cdot|\cdot)$.

Figure 2: A hierarchical partition-ligation scheme used in *Haploi*.

Figure 3: Performance on simulated datasets. Two kinds of datasets with different mutation rates θ were tested. Each dataset includes 100 individuals from 5 groups (20 from each). The sequence length was fixed to 10. The performance of each algorithm is represented in terms of err_s, err_i and d_w. Each bar represents the average error rate across 50 different datasets where the standard deviation is shown as a vertical line.

Figure 4: Top row: Histograms of the number of recovered ancestors, K, across the 50 conserved data sets (panel (a)), and across the 50 diverse data sets (panel (b)). Bottom row: Histograms of the estimated mutation rates over the 50 conserved data sets (panel (c)), and over the 50 diverse data sets (panel (b)). The left graph in each panel shows the result from HDP, and right one shows that from DP.

Figure 5: A comparison of the accuracies of haplotype frequencies estimated by five algorithms. (a) Box-plots of D_1's (top) and D_{KL}'s (bottom) 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.

Figure 6: A comparison of haplotyping accuracies of all 7-SNP segments from 10 ENCODE regions. (a) Box-plots of error rates when data only from the CEPH and Yoruba population are used. (b) Error rates under the four-population scenario. The top row shows the summary statistics of err_s, and the bottom row shows that of err_i.

Figure 7: Performance on the full sequences of the selected ten ENCODE regions. (a) Under the two-population scenario. (b) Under the four-population scenario. For cases of which the computation did not finish within a tolerable duration (i.e., 800 hours), we cap the bar with a "$\approx$" to indicate that the results are not available (NA).

Figure 8: Time complexity on the full sequences of the selected ENCODE regions. (a) Under the two-population scenario. (b) Under the four-population scenario. We cap the bars corresponding to not-completed (NC) cases with a "$\approx$".

Figure 9: Histograms of the estimated mutation rates over all SNP loci (top row, (a)-(d)) and the number of founders of all 7-SNP segments (bottom row, (e)-(h)) of the 10 ENCODE region analyzed in §5.2.1 in each of the four attended populations. In panel (e) we show a histogram of the number of founders over all population.

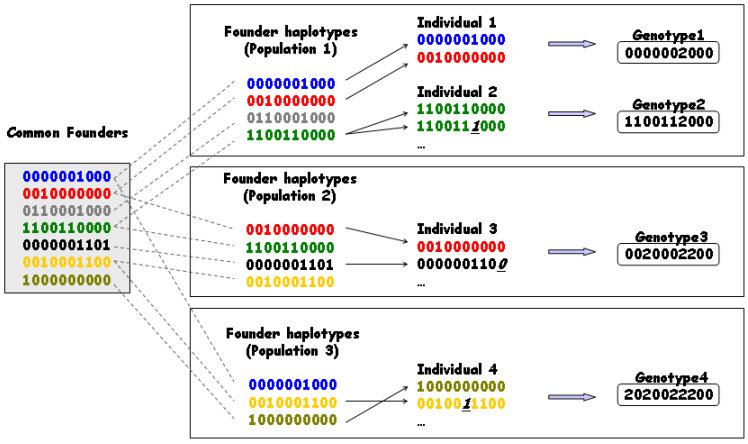


Figure 1:

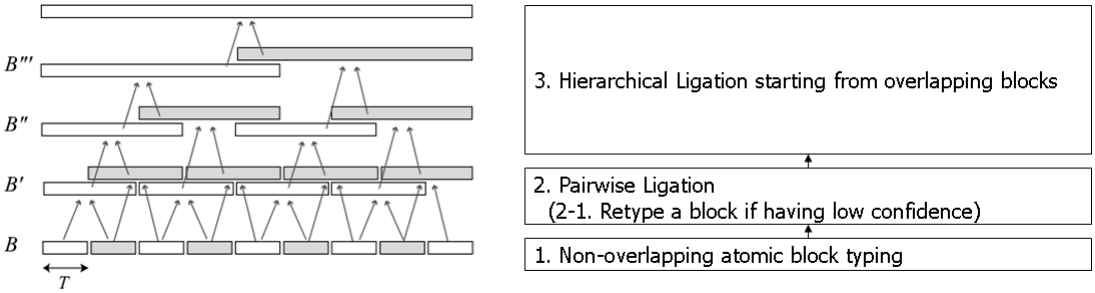


Figure 2:

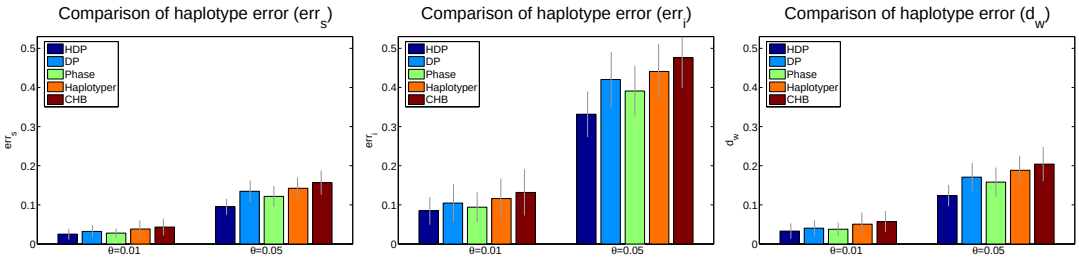
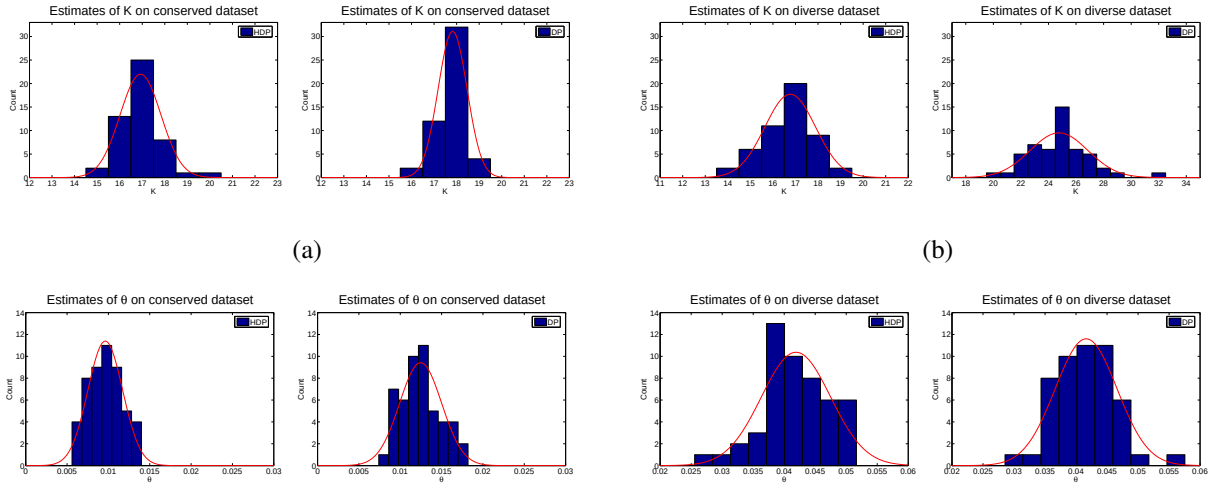


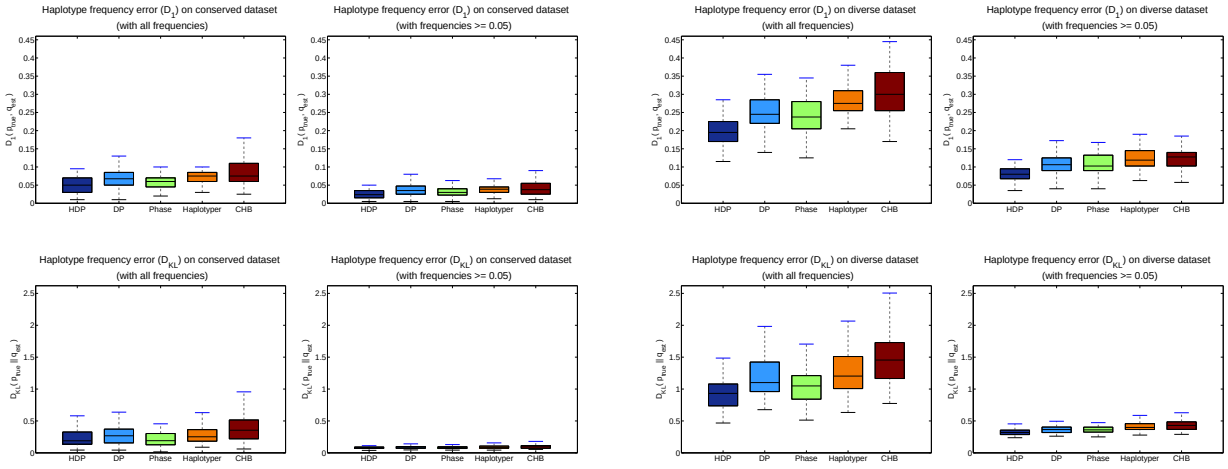
Figure 3:



(a) (b)

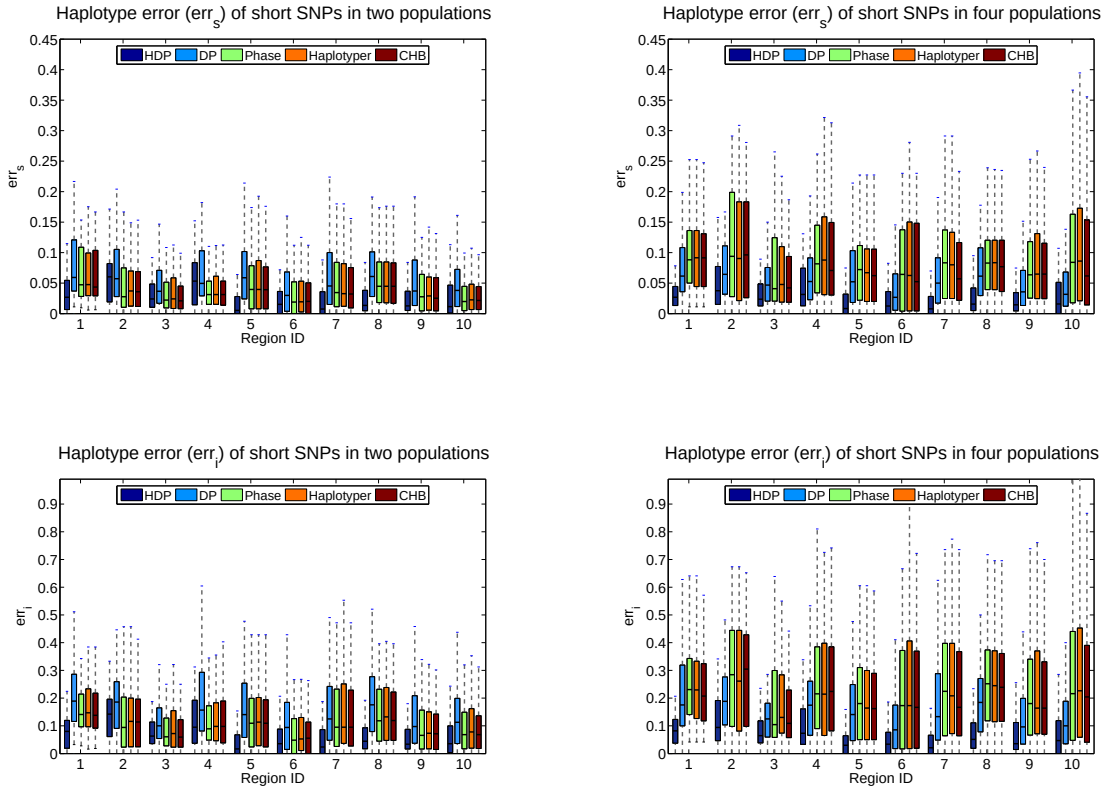
(c) (d)

Figure 4:



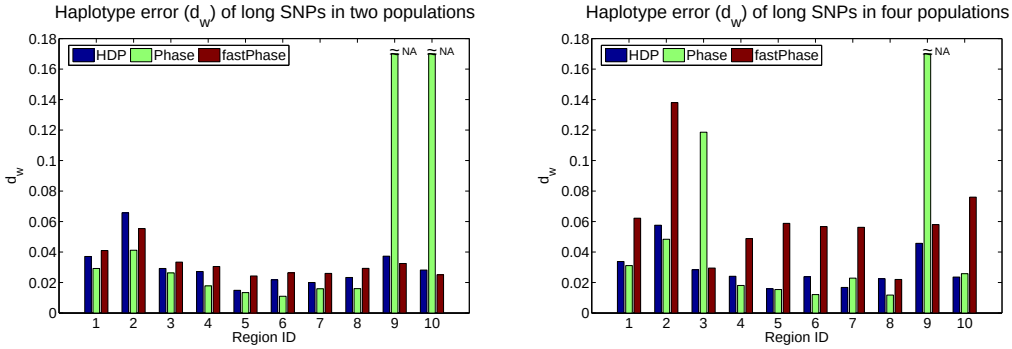
(a) (b)

Figure 5:



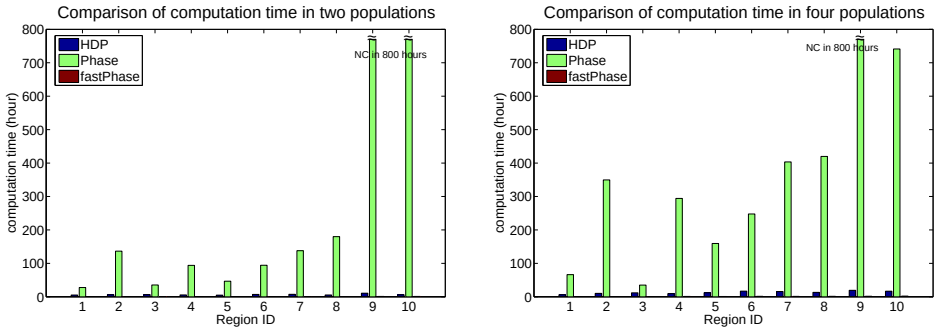
(a) (b)

Figure 6:



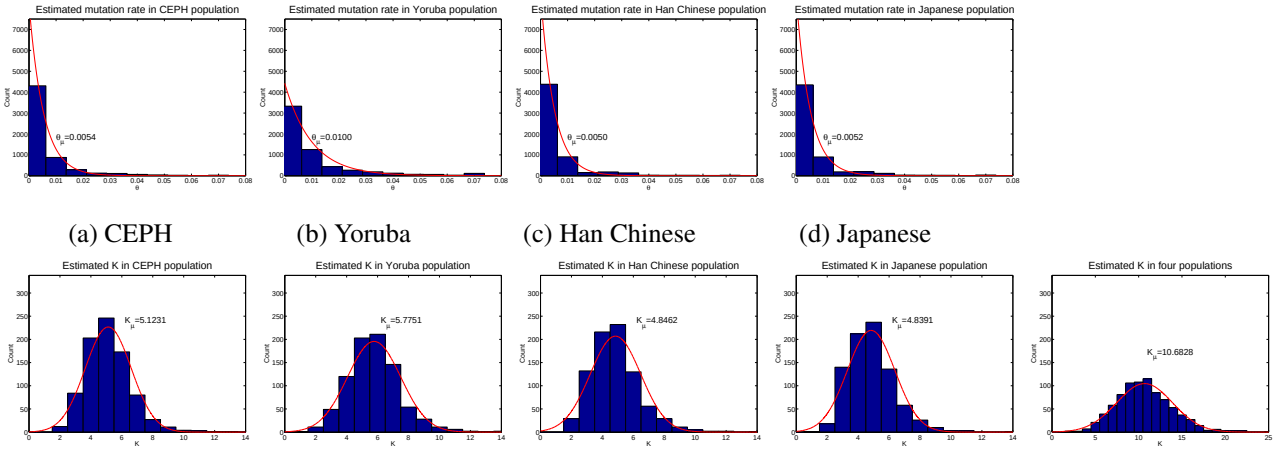
(a) (b)

Figure 7:



(a) (b)

Figure 8:



(a) CEPH (b) Yoruba (c) Han Chinese (d) Japanese

(e) CEPH (f) Yoruba (g) Han Chinese (h) Japanese (i) Overall

Figure 9:

Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213

Carnegie Mellon