BIOINFORMATICS

Vol. 00 no. 00 2007
Pages 1–9

# *Spectrum*: Joint Bayesian Inference of Population Structure and Recombination Events

Kyung-Ah Sohn and Eric P. Xing *

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

**ABSTRACT**

**Motivation:** While genetic properties such as linkage disequilibrium (LD) and population structure are closely related under a common inheritance process, the statistical methodologies developed so far mostly deal with LD analysis and structural inference separately, using specialized models that do not capture their statistical and genetic relationships. Also, most of these approaches ignore the inherent uncertainty in the genetic complexity of the data and rely on inflexible models built on a *closed* genetic space. These limitations may make it difficult to infer detailed and consistent structural information from rich genomic data such as populational SNP profiles.

**Results:** We propose a new model-based approach to address these issues through joint inference of population structure and recombination events under a nonparametric Bayesian framework; we present *Spectrum*, an efficient implementation based on our new model. We validated *Spectrum* on simulated data and applied it to two real SNP datasets, including single-population Daly data and the four-population HapMap data. Our method performs well relative to *LDhat 2.0* in estimating the recombination rates and hotspots on these datasets. More interestingly, it generates an *ancestral spectrum* for representing population structures which not only displays sub-structure based on population founders but also reveals details of the genetic diversity of each individual. It offers an alternative view of the population structures to that offered by *Structure 2.1*, which ignores chromosome-level mutation and combination with respect to founders.

## 1 INTRODUCTION

Single nucleotide polymorphisms, or SNPs, represent the largest class of individual differences in DNA. SNPs are remnants of ancient, (possibly) neutral DNA alterations dated back to a time measured at a genealogical scale; they contain more fine-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 [18].

Recent experimental advances have led to an explosion in SNP data from various populations. For example, the international SNP map working group [10] has reported the identification and mapping of 1.4 million single nucleotide polymorphisms (SNPs) in human genomes from four world populations. The deluge of SNP data fuels the long-standing interest in analyzing patterns of genetic variations to reconstruct the ancestral structures of modern human populations, for such genetic ancestral information can both shed light on the evolutionary history of modern populations and provide guidelines for more accurate association studies and for many other population genetics problems.

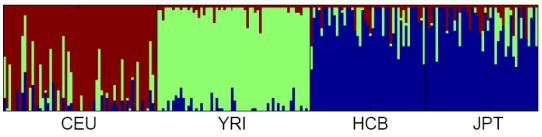
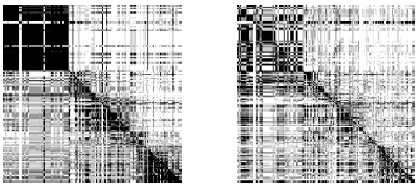



**Fig. 1.** Population structural map inferred by *Structure 2.1* on HapMap multi-population data consisting of CEU, YRI, HCB and JPT populations.

A number of variants of statistical admixture models for genetic polymorphisms have been proposed for the analysis of current population structure [16, 17, 6]. These models are instances of a more general class of hierarchical Bayesian models known as *mixed membership models* [4], which postulate that genetic markers of each individual are *iid* [16] or spatially coupled [6] samples from multiple population-specific fixed-dimensional multinomial distributions (known as *ancestry proportions* [6], 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. 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* [16, 6]. 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 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 [6]. 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

*To whom correspondence should be addressed.



(a) CEU + HCB + JPT (b) CEU + YRI

**Fig. 2.** The LD measurements, $|D'|$ (upper right), and the $p$-values for Fisher's exact test (lower left), of HapMap DB [20]. Note the LD-block structures on the mixed populations of CEU and YRI are rather opaque compared to the LD patterns of CEU+HCB+JPT populations.

at founding chromosome level [5]. 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 Fig. 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 Fig. 1 it appears to be the most homogeneous. [1]

In this paper, we present 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 alleles 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. Fig. 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 (Fig. 2(a)), some degree of block-like patterns are visible, but when CEU (European ancestry) and YRI (African ancestry) populations are mixed (Fig. 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 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. [3] and Greenspan and Geiger [9] have developed hidden Markov models for locating recombination hotspots in haplotypes; Anderson and Novembre [1] 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. [17] and for interpreting LD patterns that do not have sharp block boundaries as in Fig. 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 recently developed nonparametric Bayesian model known as a *Hidden Markov Dirichlet Process* [22] 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.

## 2 INHERITANCE 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.

### 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

[1] In fact, the genetic diversity of each individual is captured at a higher level by the population-APs. But the AP profiles are very hard to visualize and interpret because they consist of allele frequency profiles for every locus and are independent *a priori* across loci.

recombination and mutation. This assumption corresponds to an idealized noninterference model for chromosomal crossover and a star genealogy over every inherited site. Although very simple and not realistic, this assumption has been widely adopted by statistical genetic models, such as the BLADE model for mapping [14], and numerous models for haplotype inference [23]. 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 ancestors, the transition probabilities correspond to the recombination rates between the recombining chromosome pairs, and the emission model corresponds to a mutation process that passes the chosen chromosome in the ancestors to the descendants.

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 [3, 1], 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 chromosomes, 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 $\pi$. When no recombination takes place during the inheritance process that produces the haplotype $H_i$ from an ancestor $k$, 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 $\pi_{kk}$ 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,

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

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. For simplicity, we adopt the *single-locus mutation model* in Xing et al. [21]:

$$P(h_{i,t}|a_{k,t}, \theta_k) = \theta_k^{\mathbb{I}(h_{i,t}=a_{k,t})}\Big(\frac{1-\theta_k}{|B|-1}\Big)^{\mathbb{I}(h_{i,t}\neq a_{k,t})} \quad (2)$$

where $h_{i,t}$ and $a_{k,t}$ denote the alleles at locus $t$ of an individual chromosome $i$ and its corresponding ancestor $k$, respectively; $\theta_k$ indicates the ancestor-specific mutation rate; and $|B|$ denotes the number of possible alleles. As discussed in Liu et al. [14], 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 $\theta_k$ admits a Beta prior with hyperparameter $(\alpha_h, \beta_h)$ [2], 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 $\theta$ from the joint conditional probability starting from Equation (2) which reduces to:

$$P(\mathbf{h}|\mathbf{c}, \mathbf{a}) = \prod_k R(\alpha_h, \beta_h)\frac{\Gamma(\alpha_h + l_k)\Gamma(\beta_h + l'_k)}{\Gamma(\alpha_h + \beta_h + l_k + l'_k)}\Big(\frac{1}{|B|-1}\Big)^{l'_k} \quad (3)$$

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 $\theta$), $l_k = \sum_t \sum_i \mathbb{I}(h_{i,t} = a_{k,t})\mathbb{I}(c_{i,t} = k)$ is the number of alleles which were not mutated with respect to the ancestral allele, and $l'_k = \sum_t \sum_i \mathbb{I}(h_{i,t} \neq a_{k,t})\mathbb{I}(c_{i,t} = k)$ is the number of mutated alleles. The counting record $\mathbf{l}_k = \{l_k, l'_k\}$ is a sufficient statistic for the parameter $\theta_k$ [21].

## 2.2 Genotype model for unphased data

The model described above can be easily generalized to unphased genotype sequence data by introducing a genotyping model as in Xing et al. [21]. 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:

$$P_g(g|h_{i_0,t}, h_{i_1,t}, \tau) = \xi^{\mathbb{I}(h=g)}[\mu_1(1-\xi)]^{\mathbb{I}(h\neq^1 g)}[\mu_2(1-\xi)]^{\mathbb{I}(h\neq^2 g)} \quad (4)$$

where $h \triangleq h_{i_0,t} \oplus h_{i_1,t}$ 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 $\mu_1$ and $\mu_2$ are appropriately defined normalizing constants. Again we place a beta prior $Beta(\alpha_g, \beta_g)$ on $\xi$ 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 by integrating out parameters $\xi$ and resorting to the Bayes theorem, which enables a collapsed Gibbs sampling step where necessary.

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 [3, 1, 15] 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.

## 2.3 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

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

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. Recently, we have developed a nonparametric Bayesian framework for modeling genetic polymorphism based on the Dirichlet process (DP) mixtures and extension [21, 22], which allows more flexible control over the number of genetic founders.

Under coalescence-with-mutation (but without recombination), one can treat a haplotype from a modern individual as a descendent of their most recent common ancestor. Hoppe [11] 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 an individual haplotype and each color to its corresponding ancestor, this partition is equivalent to the one resulting from the *coalescence-with-mutation* process [11], and the probability distribution of the resulting *allele spectrum*—the numbers of colors (resp. haplotypes) with every possible number of representative balls (resp. descedants)—is captured by the well-known Ewen's sampling formula [19]. Blackwell and MacQueen [2] showed that this Pólya urn model yields samples whose distributions are those of the marginal probabilities under the *Dirichlet process* [8].

Xing et al. [21] proposed a haplotype inheritance model (without recombination) as a Dirichlet process mixture (DPM), of which a DP is used as the prior over the unbounded ancestral space (of founding haplotypes and their associated mutation rates). This model can be understood in the above Pólya urn scenario as associating each individual haplotype with a ball in the Pólya urn and associating the ancestral haplotypes and their own mutation rates with the colors. Essentially a DPM defines a "clustering" of the modern individual haplotypes based on the ball color. Notice that our construction so far requires no prior specification of the number of ancestors. Thus a DPM 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 ancestral haplotype pool, and at the same time it provides a reasonable approximation to the coalescence model by utilizing the partition structure resulted thereof (but allows further mutation within each partite to introduce further diversity among descendants of the same founder).

Using a further extension of DPM known as the Hidden Markov Dirichlet Process (HMDP) [22], which models stochastic transitions among states in an open state space, we can extend the HMM model proposed in §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., ancestor $k$) to all the target states (all ancestors in the pool), which sums to 1. Now we do not restrict ourselves with such a $K$ and generalize the HMM to a space with countably infinite ancestors in principal. Without going into technical details (but see [22]), our generalization can be understood as modeling each row of transition probabilities (from a specific ancestor) 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 [22], for which we omit details due to lack of space. 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 Eq. (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 [22] 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 [12], 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.

## 3 MCMC INFERENCE

In this section, we briefly 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 $\{C_{i,t}\}$, the inheritance variables specifying the origins of SNP alleles of all loci on each haplotype, and $\{A_{k,t}\}$, the founding alleles at all loci of each ancestral haplotype. All other variables in the model, e.g., the mutation rate $\theta$, are integrated out.

The Gibbs sampler alternates between two stages. First it samples the inheritance variables $\{c_{i,t}\}$, conditioning on all given individual haplotypes $\mathbf{h} = \{h_1, \ldots, h_{2N}\}$ and the most recently sampled configuration of the ancestor pool $\mathbf{a} = \{a_1, \ldots, a_K\}$; then given $\mathbf{h}$ and current values of the $c_{i,t}$'s, it samples every ancestor $a_k$.

To improve the mixing rate, we sample the inheritance variables one block at a time. That is, every time, we sample $\delta$ 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 $\delta$-block of inheritance variables can be written as:

$$P(c_{t+1:t+\delta}\,|\mathbf{c}^-,\mathbf{h},\mathbf{a}) \propto \prod_{j=t}^{t+\delta} P(c_{j+1}|c_j,\mathbf{m},\mathbf{n}) \prod_{j=t+1}^{t+\delta} P(h_j|a_{c_j,j},\mathbf{l}_{c_j}) \quad (5)$$

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 $\delta$ 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 $\delta$-block is very small and thus can be ignored. This approximation reduces the sampling space of the $\delta$-block to $O(|A|\delta)$, i.e., $|A|$ possible recombination targets times $\delta$ possible recombination locations. Accordingly, Eq. (5) reduces to:

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

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

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

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

To sample the ancestors $\{a_{k,t}\}$, we can derive the posterior distribution from Eq. 3. We refer the reader to Xing et al. [22] for further details.

## 4 RESULTS

We validated *Spectrum* on a simulated dataset and analyzed two real datasets: the HapMap four-population data [20] and the single-population Daly data [3]. Although *Spectrum* can be applied to both haplotype and genotype data, in this paper we focus on haplotype data for simplicity. The HapMap data includes 209 individuals' haplotypes (phased by PHASE software [20]) 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.

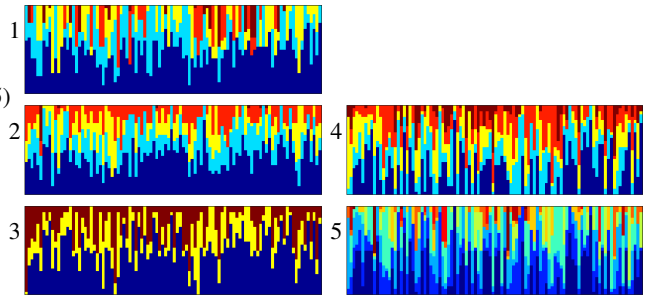



**Fig. 3.** Analysis of simulated haplotype populations. The true (panel 1) and estimated (panel 2 for *Spectrum*, and panel 3-5 for 3 HMMs) population maps of ancestral compositions in a simulated population.

**Table 1.** False positive and false negative rates for recombination hotspot detection over 30 population samples. Two kinds of threshold $\omega$'s are used. The results with different tolerance windows $w_{tol}$ are also shown.

| | | *Stepcrum* | | | *LDhat 2.0* [7] | | | HMM ($K=5$) | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $w_{tol}$ | 0 | $\pm 1$ | $\pm 2$ | 0 | $\pm 1$ | $\pm 2$ | 0 | $\pm 1$ | $\pm 2$ |
| $\omega$= | 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 |
| $\omega$ 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 |

### 4.1 Analyzing a simulated haplotype population

We simulated a population of individual haplotypes with a fixed number $K_s$ (unknown to *Spectrum*) of randomly generated ancestor haplotypes, on each of which a set of recombination hotspots were pre-specified. Then we applied a hand-specified recombination process, which is defined by a $K_s$-dimensional HMM, to the ancestor haplotypes to generate $N_s$ individual haplotypes via sequentially recombining segments of different ancestors according to the simulated HMM states at each locus and mutating certain ancestor SNP alleles according to the emission model. All the ancestor haplotypes were set to be 100 SNPs long. 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 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 (the true number of ancestors for the simulation) and 10 hidden states, respectively. For recombination analysis, we selected the widely used *LDhat 2.0* [7] for comparison. *Structure 2.1* yields a different kind of population map that is not quantitatively comparable to that from *Spectrum*; thus we only show empirical comparisons on real data.

*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* $\eta_e$, which records the fractions of every ancestor in all the $c_{i,t}$'s of that individual. Fig. 3 displays an *ancestral spectrum* constructed from the $\eta_e$'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 $\eta_e$.

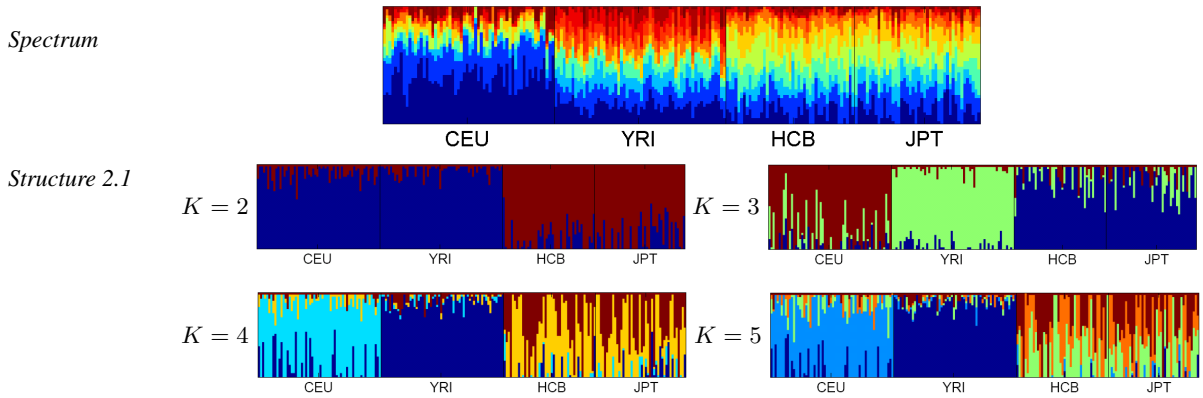



**Fig. 4.** Inferred population structure of HapMap four population data from *Spectrum*, and *Structure 2.1* with different pre-specified numbers of population $K$.

Five spectrums, corresponding 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, are shown in Fig. 3. 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 $\eta_e$ over all individuals in a population, and then over all 30 simulated populations. We found that the distance between the *Spectrum*-derived population spectrum and the true spectrum is $0.190 \pm 0.0748$, whereas the distance between HMM-spectrum and true spectrum is $0.319 \pm 0.0676$, 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.

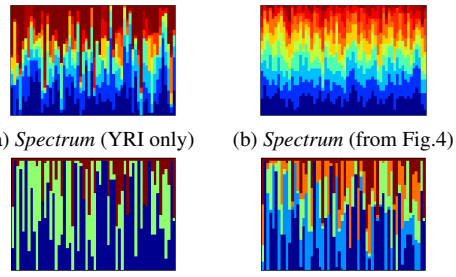
*Recombination Analysis* From the Gibbs samples of $\{c_{i,t}\}$, we can also infer the recombination status of each locus of each haplotype. We define the *empirical recombination rates* $\lambda_e$ to be the ratio of individuals who are determined to have recombinations at each locus over the total number of haplotypes in the population. We classify a locus to be a recombination hotspot if its $\lambda_e$ is greater than an empirical threshold $\omega$, which is set to be the 3rd quartile value of the estimated recombination rates. Alternatively we can set $\omega$ to be the $\lambda_e$ 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 based on $\lambda_e$-thresholding falls within a small window around the true position, and the window is set to be 0, $\pm 1$, and $\pm 2$, respectively. Table 1 summarizes the results of the performance comparison, which show that *Spectrum* outperforms *LDhat 2.0* and HMM significantly in most of the cases.

## 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 $\eta_e$. Fig. 4 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 earlier in this paper. 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 Fig. 5, 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 Fig. 4(a), where all four populations were analyzed together. Fig. 5 (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, Fig. 5 (c) and (d), both from *Structure 2.1*, show two very different structures from those in Fig. 4, 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 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 chromosome*, 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 Fig. 5. 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 $\eta_e$, we clustered the individuals based on their $\eta_e$'s and then ordered all individuals accordingly (Fig 6). Specifically, all individuals were clustered into 6 clusters (which is an empirical choice just 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.



(a) *Spectrum* (YRI only) (b) *Spectrum* (from Fig.4)

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

**Fig. 5.** Inferred population structure of HapMap YRI population data from (a)-(b) *Spectrum* , and (c)-(d) *Structure 2.1* with different number of clusters $K$.

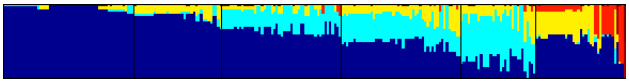


**Fig. 6.** The estimated population map of the Daly dataset. The ordering of all individuals in the sample population was determined by a K-means clustering with $K = 6$, followed by a within-cluster ordering of samples based on their distances to the cluster centroid. The black vertical bars show the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

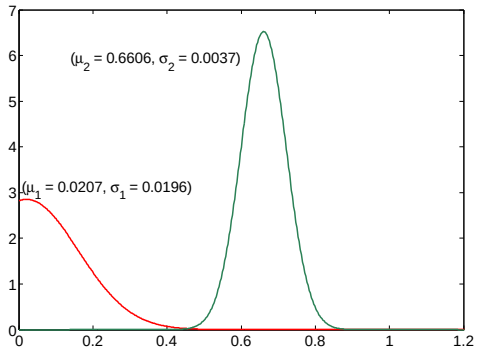



**Fig. 7.** A mixture of Gaussian fitting of the estimated $\lambda_e$ on HapMap data

*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 $\lambda_e$'s of all loci with a one-dimensional mixture of Gaussians (Fig 7). Then we used the intersection point of the two Gaussian components as the threshold for determining hotspot loci. This threshold is essentially the point where the posterior probabilities of $\lambda_e$ 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.

Fig. 8 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 (Fig. 4), 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. Fig. 8 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 Fig. 8.

Finally, we give the comparison of the recombination hotspot estimation on the Daly dataset with those reported in Daly et al. [3] (which is based on an HMM employing different numbers of states at different chromosome segments) and in Anderson and Novembre [1] (which is based on a minimal description length (MDL) principle). In Fig. 9, we show the plot of the empirical recombination rates estimated from *Spectrum*, 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 and Stephens [13]. For example, according to the estimated $\lambda_e$'s shown in Fig. 9, 15 hotspot loci/regions (represented as thick solid vertical bars in Fig. 9) were identified, and they divide the entire study region into 16 haplotype blocks of low diversity. Note that in Fig 9, 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 Fig 9 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.

Table 2 summarizes the summary statistics that characterize each haplotype block (and hotspot regions). We used the threshold of 0.005 determined by the mixture of Gaussians as described above to identify recombination hotspots. The blocks were determined accordingly, with the 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 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 2). The actual genomic region and length of each block are shown in the third and the fourth columns, respectively.

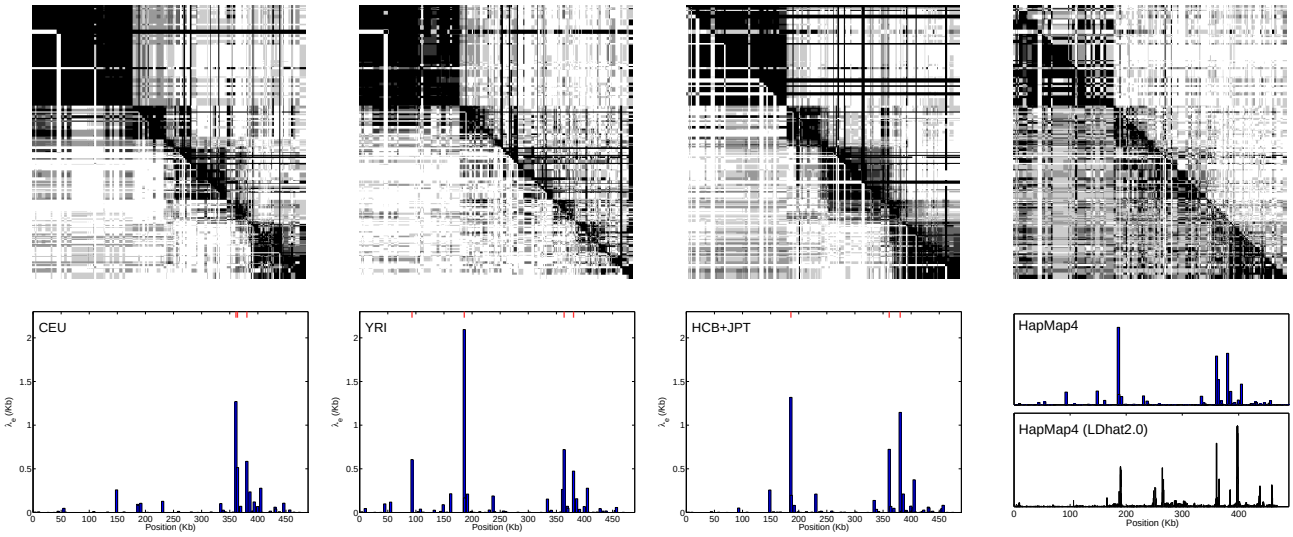



**Fig. 8.** For each population of HapMap data, the LD measure with the estimated recombination rates along the chromosomal position are shown together with the detected recombination hotspots. The last column shows the result on the mixed four populations from both *Spectrum* and *LDhat 2.0*.

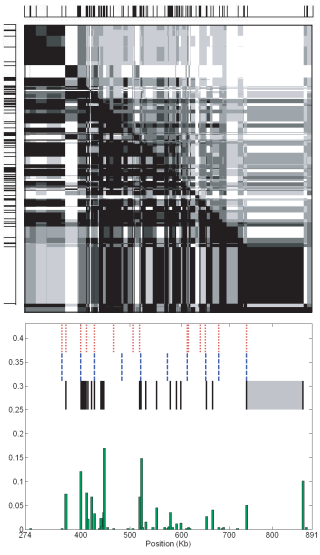



**Fig. 9.** Analysis of the Daly data. Upper panel: the LD-map of the data. Lower panel: a plot of $\lambda_e$ estimated via *Spectrum*; and the haplotype block boundaries according to *Spectrum* (black solid line), HMM [3] (red dotted line), and MDL [1] (blue dashed line). Note that the thickness of the black solid lines delineating the haplotype blocks is proportional to the width of the hotspot regions between adjacent blocks.

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. [3]. 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 2, we report the ancestor association frequencies of haplotypes within each block, where the associations were directly estimated from the inheritance variable $c_{i,t}$'s sampled by our algorithm. We can see that, overall, 6 founders sufficed to fully account for our data, and indeed within each block, only 3-4 of them were significantly used. 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.

## 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 ancestral chromosomes, *Spectrum* describes the underlying genetic process of recombination and mutation explicitly in terms of the association between ancestors 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 [6] 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

**Table 2.** Haplotype block structures and the summary statistics of the blocks for the Daly data. The block boundaries correspond to the x-coordinates of the $\lambda_e$ peaks in Fig. 9.

| 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 |

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.

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 and Stephens [13]. Also, we have not addressed the issues of threshold calculations and confidence measures of hotspot predictions as in Li and Stephens [13]. These problems are of importance in various applications such as linkage-based quantitative trait locus mapping and disease-gene mapping. One way of addressing these issues is to explicitly introduce more recombination states (e.g., for both base-line recombination and hotspot-recombination) into the infinite HMM we proposed and/or to introduce priors for site-specific recombination rates for Bayesian inference.

## ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under Grant No. 0523757 and by the Pennsylvania Department of Health's Health Research Program under Grant No. 2001NF-Cancer Health Research Grant ME-01-739. E.P.X. is also supported by a NSF CAREER Award under Grant No. DBI-0546594.

## REFERENCES

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

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

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

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

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

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

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

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

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

[10]International SNP Map Working Group. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature*, 409:928–933.

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

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

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

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

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

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

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

[18]M. Stoneking. Single nucleotide polymorphisms: From the evolutionary past. . . *Nature*, 409:821–822, 2001.

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

[20]Gudmundur A. Thorisson, Albert V. Smith, Lalitha Krishnan, and Lincoln D. Stein. The international hapmap project web site. *Genome Research*, 15:1591–1593, 2005.

[21]E.P. Xing, R. Sharan, and M.I Jordan. Bayesian haplotype inference via the Dirichlet process. In *Proceedings of the 21st International Conference on Machine Learning*, pages 879–886, New York, 2004. ACM Press.

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

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