Journal of Machine Learning Research 1 (2000) 1-48 Submitted 4/00; Published 10/00

Bayesian Nonparametric Modelling of Genetic Variations using Fragmentation-Coagulation Processes

Yee Whye Teh y.w.teh@stats.ox.ac.uk
Department of Statistics, University of Oxford, 1 South Parks Road, Oxford OX1 3TG, U.K.

Lloyd T. Elliott and Charles Blundell {elliott,c.blundell}@gatsby.ucl.ac.uk
Gatsby Computational Neuroscience Unit, UCL, 17 Queen Square, London WC1N 3AR, U.K.

Editor: Lawrence Saul

Abstract

We propose a novel class of Bayesian nonparametric models for variations in genetic data called fragmentation-coagulation processes (FCPs). FCPs model a set of sequences using a partition-valued Markov process which evolves by splitting and merging clusters. FCPs have a number of theoretically appealing properties: they are infinitely exchangeable, stationary and reversible, with equilibrium distributions given by Chinese restaurant processes. As opposed to hidden Markov models, FCPs allow for flexible modelling of the number of clusters, and they avoid label switching non-identifiability problems. We develop an efficient Markov chain Monte Carlo sampler for FCPs which uses the forward-backward algorithm, and demonstrate FCPs on genotype imputation problems, showing state-of-the-art results.

Keywords: Bayesian Nonparametrics, Fragmentation-Coagulation Processes, Genetic Variations, Genotype Imputation, Hidden Markov Models, Markov Chain Monte Carlo

1. Introduction

Driven by advances in genotyping technology, there has recently been an explosion of available data pertaining to genetic variations in human populations. The International HapMap Consortium (2003), The 1000 Genomes Project Consortium (2010), the Wellcome Trust Case Control Consortium, as well as more recent endeavours have collected whole genome data of thousands of individuals across the world, and promise to revolutionise our understanding of the genetic processes driving population change and adaptation, of the migratory histories of human populations, and of the genetic bases of various diseases and phenotypes.

The analyses of such genetic variation data using sophisticated and scalable statistical models are indispensable in unlocking their full potential. Early methods for analysing genetic sequence data (Griffiths and Marjoram, 1996; Stephens and Donnelly, 2000) are based around the celebrated coalescent model of Kingman (1982a,b), which is a model of the genealogy of genetic material that does not undergo recombination and gene conversions, so is a good model for short sequences of DNA. Over longer segments recombination and other genetic processes become more prominent and an extension of the basic coalescent accounting for recombination was proposed by Hudson (1983). Though of much theoretical interest, these early models were unfortunately not usable as statistical models for the large

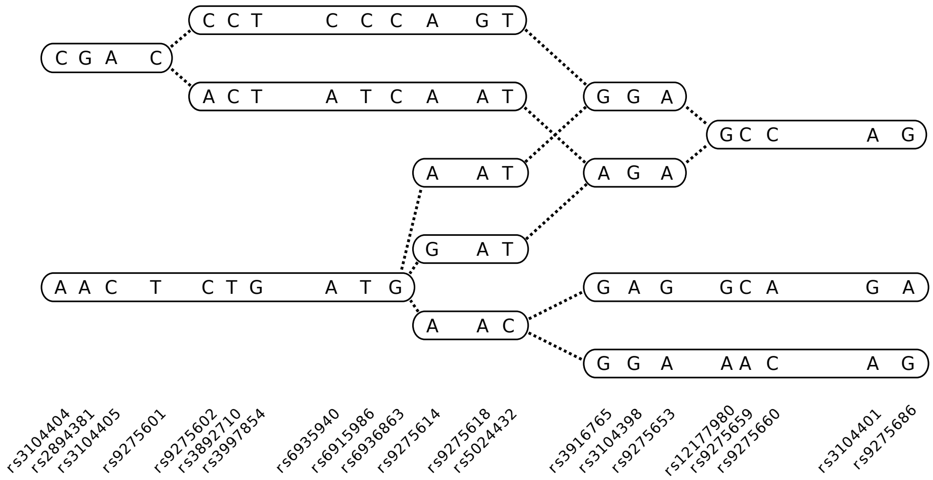

Figure 1: An example mosaic structure for SNP sequences obtained from phased trios in the CEU population in HapMap, from base pair positions 32790152 to 32795548 on Chromosome 6 (NCBI Build 36 coordinates). Each SNP sequence corresponds to a trajectory, from left to right, through the structure, passing through a number of segments. Each segment consists of a sequence of alleles, while dotted lines correspond to transitions between segments.

amounts of data now available, because of the computational difficulties associated with integrating over the complex and latent genealogical structures posited by the coalescent with recombination.

More recent and more scalable techniques are based on the insight that DNA sequences undergoing a recombination process often exhibit "mosaic" structure (Daly et al., 2001). That is, each sequence can be well approximated as a combination (or mosaic) of segments, each of which may appear in multiple sequences. An example of such mosaic structure is given in Figure 1.

Hidden Markov models (HMMs) (Rabiner, 1989; Scott, 2002) of mosaic structure, where each latent state corresponds to a different segment, have been proposed by Daly et al. (2001) and Scheet and Stephens (2006). These HMMs are learned using the expectation-maximization (EM) algorithm (Dempster et al., 1977). Alternatively, instead of using the EM algorithm, Li and Stephens (2003) and Marchini et al. (2007) proposed using observed reference sequences as HMM states, while Browning and Browning (2009) proposed a fast heuristic algorithm in which haplotypes are split and merged according a deterministic rule.

Though popular and widespread models for sequence data, HMMs suffer from two drawbacks. Firstly, they cannot flexibly adapt the number of latent states to the data, unless they use an external and often costly model selection mechanism. Secondly, as dynamical generalizations of finite mixture models, they can suffer from the label switching problem (Jasra et al., 2005), which is a non-identifiability among the HMM states resulting from the fact that typically both the prior and the likelihood model of HMMs are invariant to permutations of the latent states. The label switching non-identifiability creates exponentially many redundant modes in the posterior as well as a multitude of local optima.

Bayesian nonparametrics is a recent popular framework both to define and work with models with large support, as well as to sidestep the model selection problem of many probabilistic models (Hjort et al., 2010). For example, the Gaussian process (Rasmussen and Williams, 2006) can have large support in the space of smooth functions, while the Dirichlet process (DP) (Ferguson, 1973) is a prior whose support is the space of all probability measures. In this paper we will make use of the Chinese restaurant process (CRP) (Aldous, 1985; Pitman, 2006), a distribution over partitions related to the DP which is commonly used as a Bayesian nonparametric prior for clustering problems.

Clustering can be seen as the problem of inferring a partition of a heterogeneous data set into multiple homogeneous "clusters". A Bayesian approach to clustering starts with a prior distribution over partitions, associate each partition to the observed data using a likelihood model, and the clustering problem addressed as the computation of the posterior distribution over partitions. The CRP can be seen as a prior over partitions of the data set with sensible properties: it is exchangeable, it gives high probability to partitions with a small (relative to the data set size) number of clusters, and it has large support in the space of partitions, as opposed to finite mixture models which have support only on partitions with a fixed maximum number of clusters.

Coming back to the problem of modelling mosaic structures, while a variety of Bayesian nonparametric extensions of HMMs have been proposed (Beal et al., 2002; Teh et al., 2006; Xing and Sohn, 2007), these models nevertheless still suffer from the label switching problem. In this paper we propose a novel Bayesian nonparametric model for the mosaic structure of genetic sequences based on fragmentation-coagulation processes (FCPs). Our approach is based on the observation that a mosaic structure can be described as a sequence of partitions, one at each location of the genome. A FCP models this sequence of partitions as a Markov process on the space of partitions such that the partition at each location is marginally a CRP and whose clusters are used in the place of HMM states. FCPs do not require the number of clusters in each partition to be pre-specified (inferring them from data instead), and do not have explicit labels for clusters thus avoid the label switching problem. The partitions of FCPs evolve via a series of events, each of which involves either two clusters merging into one, or one cluster splitting into two, thus the name.

We first give a brief description of CRPs in Section 2 and set-up the problem of statistical modelling of mosaic structures in Section 3. We define FCPs and describe their properties in Section 4. Section 5 discusses in more detail how FCPs are related to other probabilistic and Bayesian nonparametric models. In Section 6 we derive a Gibbs sampler for FCPs using an augmentation scheme, and in Section 7 we describe experimental results on SNP data from the Thousand Genomes Project. Finally we conclude in Section 8.

2. The Chinese Restaurant Process

We start with a brief review of the Chinese restaurant process (CRP). Given a set S, let $\mathbf{\Pi}_S$ denote the set of unlabelled partitions of S. That is, each $\pi \in \mathbf{\Pi}_S$ is a set of disjoint non-empty subsets (which we call clusters) of S whose union is S. For each $n \in \mathbb{N} \cup \{\infty\}$ let $[n]$ denote the natural numbers $\{1, \ldots, n\}$. In this paper we are interested in random partitions of $S = [n]$. The CRP (Aldous, 1985; Pitman, 2006) is the canonical distribution on $\mathbf{\Pi}_S$, and is parameterized by a mass parameter $\gamma > 0$. For finite S, it has a probability mass function given as follows:

$$g_S(\pi; \gamma) = \frac{\Gamma(\gamma)\gamma^{|\pi|} \prod_{a \in \pi} \Gamma(|a|)}{\Gamma(n + \gamma)}, \tag{1}$$

for each $\pi \in \mathbf{\Pi}_S$, with $\Gamma(\cdot)$ the gamma function. The CRP can be described using the following metaphor: $|S|$ customers corresponding to the elements of S enter a Chinese restaurant one at a time. The first customer sits at some table, while subsequent customers sit at an already occupied table with probability proportional to the number of customers already seated there, or at a new table with probability proportional to γ. After all customers have been seated, the seating arrangement of customers around tables form a partition π of S, with occupied tables corresponding to the clusters of π. The probability mass function of the CRP given in (1) is the product of the conditional probabilities under this generative process. We write $\Pi \sim \mathrm{CRP}_S(\gamma)$ if Π is a CRP distributed random partition of S.

Given a permutation σ of S, let $\sigma(\Pi)$ be the partition of S obtained by replacing each $i \in S$ by $\sigma(i)$. The CRP is *exchangeable* in that the laws of $\sigma(\Pi)$ and Π coincide. Metaphorically, the specific order in which customers enter the restaurant does not affect the resulting distribution over seating arrangements. If $S' \subset S$, define the restriction $\pi_{|S'}$ of π onto S' to be the partition of S' obtained by removing the elements of $S \backslash S'$ as well as the resulting empty subsets from π. The CRP is also *projective* in that the restriction $\Pi_{|S'}$ is distributed as $\mathrm{CRP}_{S'}(\gamma)$. Using Kolmogorov's Extension Theorem, the exchangeable and projective nature of the CRP imply that we can extend Π to an exchangeable random partition of all of $\mathbb{N}$. The CRP is also intimately associated with the Dirichlet process via de Finetti's Theorem (Blackwell and MacQueen, 1973). In particular, define a sequence of random variables $x_1, x_2, \ldots$ by first associating each cluster $c \in \Pi$ with an i.i.d. draw θ_c from a base distribution H then assigning each $x_i = \theta_c$ for the unique subset $c \in \Pi$ containing i. Then the sequence $x_1, x_2, \ldots$ is infinitely exchangeable, and by de Finetti's Theorem has a hierarchical representation of the form

$$\begin{aligned} G &\sim \mathrm{DP}(\gamma, H) \\ x_i | G &\sim G \qquad \text{i.i.d. for each } i = 1, 2, \ldots. \end{aligned} \tag{2}$$

where the law of G, called the de Finetti measure, is given by the Dirichlet process (DP) with mass parameter γ and base distribution H. If instead of assigning each $x_i = \theta_c$ we draw each $x_i \sim F(\theta_c)$ independently from some distribution parameterised by θ_c, then $x_1, x_2, \ldots$ are still infinitely exchangeable, and the sequence has a DP mixture (with mixing kernel F) (Lo, 1984) as its de Finetti measure. In clustering terms, the CRP serves as the prior over partitions, and $F(\theta_c)$ is the likelihood model which describes the distribution of data items in cluster c. The random number of clusters under a CRP has mean $\gamma(\psi(n+\gamma) - \psi(\gamma)) \approx \gamma \log(n - \frac{1}{2} + \gamma)$ and variance $\approx \gamma \log(n - \frac{1}{2} + \gamma)$, where $\psi(x) = \frac{\partial \log \Gamma(x)}{\partial x}$ is the digamma function.

3. Modelling Mosaic Structures using Partition-valued Stochastic Processes

Suppose we have n genetic sequences of length T, typed at m sites, which are located at positions $0 < t_1^o < \cdots < t_m^o < T$ on the sequences. Each typed value (or allele) can take on values in a discrete space $\mathbb{X}$. In the case of single nucleotide polymorphisms (SNPs) each site can take on only two alleles and we can assume $\mathbb{X} = \{0, 1\}$. For $i = 1, \ldots, n$ the ith haplotype $x_i = [x_{i1}, \ldots, x_{im}]$ consists of the sequence of m alleles on the ith sequence, and we denote the data set as $X = \{x_1, \ldots, x_n\}$.

Daly et al. (2001) noted that genetic sequences undergoing a recombination process exhibit structure where each haplotype is composed of a mosaic of contiguous segments (Figure 1). One way to model such mosaic structure is via a hidden Markov model (HMM) (Rabiner, 1989), where each sequence x_i is modelled using a Markov chain of latent state variables $c_{i1}, \ldots, c_{im}$, each of which can take on one of K states, with c_{ij} representing the local segment around site j in sequence x_i.

Looking across sequences at a fixed location j, we can equivalently interpret the latent state variables $c_{1j}, \ldots, c_{nj}$ as defining a partition of the sequences at location j which is limited to have at most K clusters. In general we can model the mosaic structure using a partition-valued stochastic process $(\Pi(t) : t \in [0, T])$, where each $\Pi(t)$ is a random partition of the index set $[n]$ that is not limited to a fixed number of clusters. At each site j, each cluster in $\Pi(t_j^o)$ is a set of sequence indices and indicates that all sequences indexed by it share the same local segment around site j. From the perspective of the ith sequence, the cluster membership of index i through the sequence of partitions in $(\Pi(t_1^o), \ldots, \Pi(t_m^o))$ describes the sequence of mosaic segments constituting x_i.

A partition-valued mosaic structure can be understood as an approximation to the ancestral recombination graph (ARG), which describes the joint genealogies of the sequences (Hudson, 1983). Embedded within the ARG is a tree-structured genealogy associated with each location on the chromosome, with locations on either side of each recombination event having different genealogies. By approximating the genealogical tree at each location by a partition, for example with each cluster representing an ancestral population, we get a partition-valued mosaic structure as described above.

Given the mosaic structure described by $(\Pi(t))$, we can model the observed sequences X using the following hierarchical model:

$$\begin{aligned} x_{ij} | \theta_{cj}, \Pi(t_j^o) &\sim F(\theta_{cj}) && \text{where } c \text{ is the unique cluster in } \Pi(t_j^o) \text{ containing } i, \\ \theta_{cj} | \omega_j &\sim H(\omega_j) && \text{for each } c \in \Pi(t_j^o). \\ \omega_j &\sim H_0 && \text{for each site } j, \end{aligned} \tag{3}$$

where $F(\theta)$ is a distribution over alleles parameterised by θ, $H(\omega)$ is a prior over the parameters at each site, with ω being a hyperparameter and H_0 its hyperprior. The model assumes that sequences in the same cluster share the same parameter thus the alleles in the cluster tend to be similar. We shall return to the specification of this hierarchy for SNP data in Section 6.2.

What is left to specify now is the prior on $(\Pi(t))$ itself, which constitutes the main objective of this paper. Before we proceed, it is prudent to consider properties that a partition-valued stochastic process $(\Pi(t))$ should have, if it were used as a prior for mosaic structures. Firstly, it should be *exchangeable* and *projective*, so that inferences drawn are invariant with respect to the number of sequences and the order in which they are numbered. Without detailed knowledge of how genetic processes affect the mosaic structure, it is also reasonable to assume that $(\Pi(t))$ is *stationary* and *reversible*. Finally, using a *Markov* process for $(\Pi(t))$ can help in the development of computationally efficient methodologies for posterior simulation that take advantage of the conditional independencies present in Markov models.

Returning to the HMM, recall that the latent variables can be equivalently understood as defining a partition-valued stochastic process. The HMM partition process is exchangeable and projective, and can be made stationary and reversible with the right choice of initial state distribution and transition probabilities. Further, the Markov property of the latent state sequences translates into a Markov property of the sequence of random partitions.

In the following, we shall introduce a fragmentation-coagulation process (FCP) as an alternative Markov process over partitions. As opposed to the HMM, which defines partitions implicitly using labelled states, the FCP defines a Markov process directly in the space of all (unlabelled) partitions of $[n]$. The FCP satisfies the desirable properties given above, and has the advantages of not needing to pre-specify the maximum number of clusters in each partition, and of not labelling the clusters so that the label switching problem is avoided (see Section 5.1).

4. A Fragmentation-Coagulation Process

We will define a homogeneous pure jump Markov process $(\Pi(t) : t \in [0, \infty))$ in which each $\Pi(t)$ is a random partition taking values in $\mathbf{\Pi}_{[n]}$. Since $\mathbf{\Pi}_{[n]}$ is finite, it is sufficient to describe the initial distribution of $\Pi(0)$ and the transition rates from each partition in $\mathbf{\Pi}_{[n]}$ to another. We will show that our Markov process is ergodic and use the unique stationary distribution as the initial distribution so that $(\Pi(t))$ is a stationary Markov process. First we describe the transitions of the Markov process, of which there are two types, *fragmentations* and *coagulations*, as the name of the process suggests. The transition rates are as follows:

- **Fragmentation:** Suppose $\pi \in \mathbf{\Pi}_{[n]}$ and let $c \in \pi$ be a cluster with $|c| \geq 2$. Let a and b be disjoint non-empty subsets such that $a \cup b = c$. Suppose $\eta = \pi - c + a + b$ is obtained by fragmenting c into a and b. We define the rate of transitioning from π to η to be given by:

$$q(\pi, \eta) = \beta \frac{\Gamma(|a|)\Gamma(|b|)}{\Gamma(|c|)} \tag{4}$$

where $\beta > 0$ is a parameter governing the rate of fragmentation. Each subset c fragments independently, and the overall rate of c fragmenting into any pair of disjoint non-empty subsets a and b can be shown to be

$$\sum_{a,b: a \cup b = c} \beta \frac{\Gamma(|a|)\Gamma(|b|)}{\Gamma(|c|)} = \beta H_{|c|-1} \tag{5}$$

where $H_p = \sum_{i=1}^{p} \frac{1}{i}$ is the pth harmonic number. The overall rate of fragmentation when the Markov process is in state π is thus $\beta \sum_{c \in \pi} H_{|c|-1}$.

- **Coagulation:** Suppose now that $a, b \in \pi$ with $a \neq b$. Let $c = a \cup b$, and let $\eta = \pi - a - b + c$ be obtained by coagulating the two subsets a and b into c. The rate of transitioning from π to η is simply given by
$$q(\pi, \eta) = \alpha, \tag{6}$$
where $\alpha > 0$ is a parameter governing the rate of coagulation. The total rate of coagulation when the Markov process is in state π is $\alpha\frac{|\pi|(|\pi|-1)}{2}$ since there are $\frac{|\pi|(|\pi|-1)}{2}$ pairs of subsets a, b that can coagulate.
- For all other pairs of partitions $\pi, \eta \in \mathbf{\Pi}_{[n]}$, the transition rate is $q(\pi, \eta) = 0$.

Note that these fragmentation and coagulation transitions are reverses of each other: If η is obtained by fragmenting a subset of π, then π can be obtained by coagulating two subsets of η and vice versa. Further, $q(\pi, \eta) > 0$ if and only if $q(\eta, \pi) > 0$. We will further elaborate on reversibility of the Markov process in Proposition 3 below.

The total rate of transition out of state $\pi \in \mathbf{\Pi}_{[n]}$ is the sum of the fragmentation and coagulation rates, given as:

$$q(\pi) = q(\pi, \cdot) = \alpha\frac{|\pi|(|\pi| - 1)}{2} + \beta \sum_{c\in\pi} H_{|c|-1} \tag{7}$$

We call $(\Pi(t))$ a *fragmentation-coagulation process* (FCP) with fragmentation rate β and coagulation rate α. Due to the simplicity of the FCP, analysis of its properties is relatively straightforward and we will describe a number of important properties in the rest of this section.

4.1 Markov Properties

Proposition 1. *The Markov process* $(\Pi(t), t \in [0, \infty))$ *is ergodic with stationary distribution given by* $\mathrm{CRP}_{[n]}(\frac{\beta}{\alpha})$.

Proof Firstly, note that $\mathbf{\Pi}_{[n]}$ is finite and that there is positive probability of $(\Pi(t))$ transiting from any $\pi \in \mathbf{\Pi}_{[n]}$ to any $\eta \in \mathbf{\Pi}_{[n]}$ within any positive amount of time, thus the Markov process is ergodic. To show that the stationary distribution is given by the CRP, we only need to demonstrate detailed balance. Since all other transition rates are zero except for the pair-wise fragmentations and coagulations, which are reverses of each other, it is sufficient to show detailed balance when $\pi, \eta \in \mathbf{\Pi}_n$ are such that η can be obtained by fragmenting a subset $c \in \pi$ into two subsets a and b:

$$\begin{aligned} g_{[n]}(\pi; \tfrac{\beta}{\alpha}) \times q(\pi, \eta) =& \frac{\Gamma(\frac{\beta}{\alpha})(\frac{\beta}{\alpha})^{|\pi|}\prod_{s\in\pi}\Gamma(|s|)}{\Gamma(n + \frac{\beta}{\alpha})} \times \beta\frac{\Gamma(|a|)\Gamma(|b|)}{\Gamma(|c|)} \\ =& \frac{\Gamma(\frac{\beta}{\alpha})(\frac{\beta}{\alpha})^{|\pi|+1}\Gamma(|a|)\Gamma(|b|)\prod_{s\in\pi, s\neq c}\Gamma(|s|)}{\Gamma(n + \frac{\beta}{\alpha})} \times \alpha \\ =& \frac{\Gamma(\frac{\beta}{\alpha})(\frac{\beta}{\alpha})^{|\eta|}\prod_{s\in\eta}\Gamma(|s|)}{\Gamma(n + \frac{\beta}{\alpha})} \times \alpha = g_{[n]}(\eta; \tfrac{\beta}{\alpha}) \times q(\eta, \pi) \end{aligned} \tag{8}$$

Thus detailed balance holds and the equilibrium distribution is $\mathrm{CRP}_{[n]}(\frac{\beta}{\alpha})$. ■

Now specifying the initial distribution to be the stationary distribution, that is,

$$\Pi(0) \sim \mathrm{CRP}_{[n]}(\tfrac{\beta}{\alpha}), \tag{9}$$

we have the following result:

Proposition 2. *The Markov process* $(\Pi(t) : t \in [0, \infty))$ *is stationary with* $\Pi(t) \sim \mathrm{CRP}_{[n]}(\frac{\beta}{\alpha})$ *marginally for each* $t \in [0, \infty)$*.*

Since the Markov process is stationary, the detailed balance argument of Proposition 1 in fact shows that it is reversible as well:

Proposition 3. *For each* $T > 0$*, define* $\Pi'(t) = \Pi(T-t)$*. Then the process* $(\Pi'(t), t \in [0, T])$ *has the same law as* $(\Pi(t), t \in [0, T])$*.*

Notice that the fragmentation events in $\Pi'(t)$ are precisely the coagulation events in $\Pi(t)$ and vice versa. As properties of a model of genetic variations, stationarity and reversibility may not necessarily be suitable, since the biological mechanisms responsible for recombination and gene conversion are likely to be non-reversible and vary across the genome. However, as simplifying modelling assumptions these are sensible approximations of the true biological process and ones that are also made in other models like the coalescent with recombination of Hudson (1983). In Section 6.2 we describe how varying recombination rates can be accommodated in the model.

4.2 The Joint Distribution

In this subsection we will consider the FCP $(\Pi(t), t \in [0, T])$ over the finite interval $[0, T]$, $T > 0$. With probability one a draw from the FCP will only contain a finite number of jump events over the finite interval, each of which is either a fragmentation or a coagulation event. It is thus possible to write the probability[1] of a sample path $(\pi(t), t \in [0, T])$ of $(\Pi(t), t \in [0, T])$.

First, we introduce an alternative description of $(\pi(t))$ based on the trajectory of clusters. For each $t \in [0, T]$, the partition $\pi(t)$ consists of a number of clusters. Each cluster $c \in \pi(t)$ was created either at time 0, or at a fragmentation or coagulation event prior to t, and persists until it terminates at another event after t or at time T. We call the trajectory of c, consisting the subset and its creation and termination times, a *path*, and overload the notation c to refer to the path as well when there is no confusion. The structure of $(\pi(t))$ can be visualized as a set of intersecting paths, where a path forks into two at a fragmentation event, and two paths merge into one at a coagulation event; see Figure 2. The path structure is discrete except for the event times. Let A be the set of paths in $(\pi(t))$.

[1] In this paper we interpret probabilities $\mathbb{P}(\cdot)$ as densities with respect to Lebesgue measure for continuous quantities like the times of fragmentations and coagulations, and probability mass functions for discrete quantities like partitions.

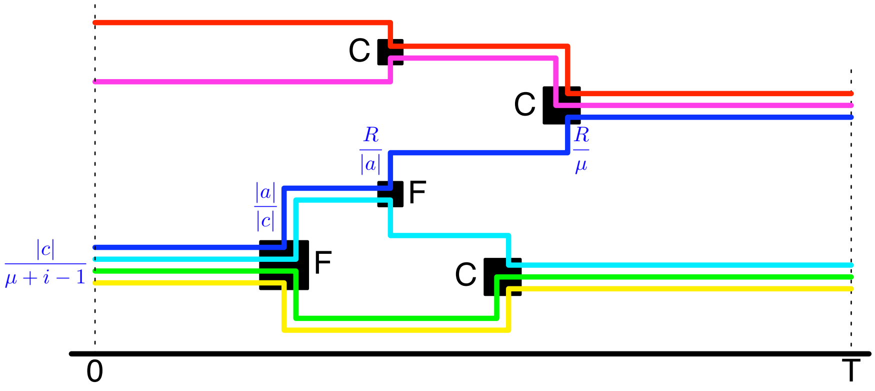

Figure 2: Visualizing a fragmentation-coagulation process as a set of trajectories through paths (segments). Each line is a sequence trajectory and bundled lines form paths. C: coagulation event. F: fragmentation event. Fractions are, for the dark blue trajectory leading from the lower path to the upper path, from left to right: probability of joining cluster c at time 0, probability of following cluster a at a fragmentation event, rate of starting a new path (a fragmentation event occurring), and rate of joining with an existing path (a coagulation event occurring).

Proposition 4. *The probability of a sample path $(\pi(t))$ under the law of the FCP is:*

$$\mathbb{P}((\Pi(t)) = (\pi(t))) = \frac{\beta^{|A|-F-C}}{\alpha^{|A|-2F-2C}} \frac{\Gamma(\frac{\beta}{\alpha})}{\Gamma(n+\frac{\beta}{\alpha})} \exp\left(-\int_0^T q(\pi(t))dt\right) \frac{\prod_{c\in A_{<>}} \Gamma(|c|)}{\prod_{c\in A_{><}} \Gamma(|c|)} \tag{10}$$

where $A_{><}$ is the set of paths created by a coagulation and terminated by a fragmentation, $A_{<>}$ is the set of paths created either at time 0 or fragmentation, and terminated at time T or a coagulation, and F and C are the numbers of fragmentation and coagulation events in $(\pi(t))$ respectively.

The proof is given in the appendix and consists simply of collecting terms in the probabilities associated with events and waiting times. Notice that the probability (10) is unchanged by reversing $(\pi(t))$ as in Proposition 3, since fragmentations become coagulations and vice versa, while $A_{<>}$ and $A_{><}$ are unchanged by reversing (except that the creation and termination times are reversed). This gives a direct proof of Proposition 3. It is also unchanged by permuting each partition $\pi(t)$ by the same permutation of $[n]$, thus exchangeability (Proposition 6 below) follows as well.

4.3 The Conditional Distribution

In this subsection we provide an alternative incremental construction of $(\Pi(t))$ which can be seen as a dynamical generalization of the Chinese restaurant metaphor. Such a construction will be useful in showing projectivity as well as in developing the Markov chain Monte Carlo sampling algorithm described in Section 6.

For each $i = 1, \ldots, n$, let $\Pi_{|[i]}(t)$ be the restriction of $\Pi(t)$ onto $[i]$. The law of $(\Pi(t))$ can be equivalently described as the sequence of conditional laws of $(\Pi_{|[i]}(t))$ given $(\Pi_{|[i-1]}(t))$, with $\Pi_{|[0]}(t)$ understood as the empty partition. Since $\Pi_{|[i]}(t)$ differs from $\Pi_{|[i-1]}(t)$ only through the addition of the ith element, its conditional distribution can in turn be described by the distribution of the trajectory of the ith element. At each time t, let $c_i(t) = c \backslash \{i\}$ where c is the unique cluster of $\Pi_{|[i]}(t)$ containing i. Then $(c_i(t) : t \in [0, T])$ is a stochastic process which describes the trajectory of the ith element through the existing mosaic structure given in $(\Pi_{|[i-1]}(t))$. Each $c_i(t)$ is either a cluster in $\Pi_{|[i-1]}(t)$ (if the ith element is in an existing cluster) or the empty set (if the ith element is in its own cluster in $\Pi_{|[i]}(t)$). We claim that the law of $(c_i(t))$ is that of a Markov process, with initial distribution and transition rates as described below:

- At time 0, since $\Pi_{|[i]}(0)$ is CRP distributed, we have the initial distribution,

$$\mathbb{P}(c_i(0)|(\Pi_{|[i-1]}(t))) = \begin{cases} \frac{|c|}{i-1+\frac{\beta}{\alpha}} & \text{if } c_i(0) = c \in \Pi_{|[i-1]}(0), \\ \frac{\frac{\beta}{\alpha}}{i-1+\frac{\beta}{\alpha}} & \text{if } c_i(0) = \emptyset. \end{cases} \tag{11}$$

- Subsequently, suppose the ith element is in an existing path, say $c_i(t-) = c \in \Pi_{|[i-1]}(t-)$. Here $t-$ denotes an infinitesimal time prior to t. Then there are three possible scenarios:
  - If c fragmented into two clusters $a, b \in \Pi_{|[i-1]}(t)$ at time t, then $c_i(t)$ will follow one of the two resulting paths a or b with probabilities proportional to their sizes:

$$\mathbb{P}(c_i(t)|(\Pi_{|[i-1]}(t)), c_i(t-) = c) = \begin{cases} \frac{|a|}{|c|} & \text{for } c_i(t) = a, \\ \frac{|b|}{|c|} & \text{for } c_i(t) = b. \end{cases} \tag{12}$$

  - If c coagulated with another cluster in $\Pi_{|[i-1]}(t-)$ at time t, forming a new cluster $c' \in \Pi_{|[i-1]}(t)$, then the ith element will follow path c', i.e. $c_i(t) = c'$ with probability 1.
  - If no fragmentation or coagulation involving c occurs at time t, then the ith element will fragment out from c to form its own cluster, that is, $c_i(t) = \emptyset$, at rate $\frac{\beta}{|c|}$.
- Finally, if $c_i(t-) = \emptyset$, then the ith element will be in a path by itself so will not be involved in any fragmentation events. On the other hand, the path will coagulate with another path $c \in \Pi_{|[i-1]}(t)$ with rate α, i.e. $c_i(t)$ will transition to c with rate α. The total coagulation rate is $\alpha|\Pi_{|[i-1]}(t)|$ since there are $|\Pi_{|[i-1]}(t)|$ other paths.

The structure of $(c_i(t))$ can be described using a Chinese restaurant metaphor as follows. At time 0 the ith customer starts by sitting at either one of the tables in $\Pi_{|[i-1]}(0)$ or a new table. If she sits in an existing table, which splits into two tables at some time, she chooses one of the two resulting tables. If the table merges with another one, she continues sitting at the merged table. In addition, if she is sitting at an existing table, she can decide to start a new table through a new fragmentation event at any time. Or, if she is sitting at a table by herself, she could merge with an existing table through a new coagulation event.

Proposition 5. *The conditional distribution of* $(c_i(t))$ *given* $(\Pi_{|[i-1]}(t))$ *under the FCP is Markov and given by the above description.*

Proof By multiplying the probabilities of the various events above, it is possible to derive the conditional probability of a sample path of $(c_i(t))$ given $\Pi_{|[i-1]}$. The product of these over $i = 1, \ldots, n$ can be shown to be equal to the joint probability (10) of a sample path for the whole FCP $(\Pi(t))$, thus the conditional distribution of $(c_i(t))$ under the FCP is as given above. ■

4.4 Exchangeability and Projectivity

The discussions in the preceding two subsections on the joint and conditional distributions already indicate that the FCP is both exchangeable and projective:

Proposition 6. *The process* $(\Pi(t), t \in [0, \infty))$ *is exchangeable: if* σ *is a permutation of* $[n]$*, then* $(\sigma(\Pi(t)))$ *has the same FCP law as* $(\Pi(t))$*.*

Proposition 7. *The process* $(\Pi(t), t \in [0, \infty))$ *is projective: the restriction* $(\Pi_{|[n']}(t))$ *to* $[n'] \subset [n]$ *is a FCP defined over partitions of* $[n']$ *with the same parameters* α *and* β*.*

Alternatively, we can see that the FCP is exchangeable and projective because both its initial distribution $\mathrm{CRP}_{[n]}(\frac{\beta}{\alpha})$ and its rates of fragmentation (4) and coagulation (6) are. Specifically, they are invariant to permutations of $[n]$, and if $n' < n$ and $\pi', \eta' \in \mathbf{\Pi}_{[n']}$, then it can be shown that

$$q_{n'}(\pi', \eta') = \sum_{\substack{\pi,\eta \in \mathbf{\Pi}_{[n]} \\ \pi_{|[n']} = \pi', \eta_{|[n']} = \eta'}} q_n(\pi, \eta) \tag{13}$$

where $q_{n'}$ and q_n are the jump rates of the FCP over partitions of $[n']$ and $[n]$ respectively. For example, in the context of Section 4.3, consider a cluster $c \in \Pi_{|[i-1]}(t-)$ which fragments into two cluster $a, b \in \Pi_{|[i-1]}(t)$ at time t. Suppose that element i was in cluster c just prior to t: $c_i(t-) = c$. Then at the fragmentation event i can follow either a or b, and the rates of the resulting fragmentation satisfy the equality

$$\beta \frac{\Gamma(|a \cup \{i\}|)\Gamma(|b|)}{\Gamma(|c \cup \{i\}|)} + \beta \frac{\Gamma(|a|)\Gamma(|b \cup \{i\}|)}{\Gamma(|c \cup \{i\}|)} = \beta \frac{\Gamma(|a|)\Gamma(|b|)}{\Gamma(|c|)} \tag{14}$$

As a model for mosaic structures in genetic sequences, these are sensible properties for the FCP to have: exchangeability implies that the model should be invariant to the indexing of the sequences $x_1, \ldots, x_n$, while projectivity implies that inference based on the model about some observed sequences will not be affected by knowledge of an additional but unknown number of unobserved sequences. These properties are not shared by the PHASE model of Li and Stephens (2003). Their comments regarding the non-exchangeability of PHASE served as inspiration for this present work.

Just as the exchangeability and projectivity of the CRP implies the existence of an exchangeable random partition over $\mathbb{N}$, we have, by Kolmogorov's Extension Theorem,

Corollary 8. *There is a Markov process* $(\Pi_\infty(t) : t \in [0, \infty))$ *such that each* $\Pi_\infty(t)$ *is an exchangeable random partition of* $\mathbb{N}$ *with law* $\mathrm{CRP}_\mathbb{N}(\frac{\beta}{\alpha})$*, and the restriction of* $(\Pi_\infty(t))$ *to* $[n]$ *is* $(\Pi(t))$*.*

In fact the process $(\Pi_\infty(t), t \geq 0)$ is an example of the class of *exchangeable fragmentation-coagulation processes*[2] studied by Berestycki (2004). This is a large class of exchangeable Markov processes over partitions of $\mathbb{N}$ that evolves via fragmentations and coagulations as above. Berestycki (2004) showed that the evolution of all such processes can be described using an *erosion* component where an item $i \in \mathbb{N}$ splits off from the cluster it belonged to, forming a singleton cluster by itself, a *Kingman* component involving coagulations of pairs of clusters as above (Kingman, 1982a,b), a fragmentation component involving the fragmentation of a cluster into two or more (possibly an infinite number of) clusters, and a multiple coagulation component involving the simultaneous coagulation of an infinite number of clusters into one or more (possibly an infinite number of) clusters.

Our Markov process has no erosion, no multiple coagulations, a Kingman coagulation rate of α, and a fragmentation component where each cluster can only fragment into exactly two clusters. As such, the process described here is perhaps the simplest non-trivial example of an exchangeable fragmentation-coagulation process, though we have not found references in the literature describing it specifically. The closest work is by Bertoin (2007), who described a class of reversible exchangeable fragmentation-coagulation processes that can be thought of as generalizations of our process, along similar lines as the two-parameter Pitman-Yor process (Perman et al., 1992; Pitman and Yor, 1997) is a generalisation of the Dirichlet process.

The fascinating mathematical developments in Berestycki (2004) pertain to the properties of the projective limit $(\Pi_\infty(t), t \geq 0)$. Although the importance of this endeavour cannot be neglected, our interests in this paper are in using the finite n version of these processes to model variations in genetic data. As such, we refer the interested reader to Berestycki (2004), and will concentrate on investigating and using properties of the finite n case in the rest of this paper.

We will end this section with a discussion of the splitting rate of the fragmentation component of our process. Given a $p \in (0, 1)$, consider a random binary partition of $\mathbb{N}$ consisting of only two subsets where each $i \in \mathbb{N}$ is independently in one subset with probability p and in the other subset otherwise. Let ϱ_p denote the resulting law of the random partition. The *splitting rate* of the fragmentation is a measure μ over $\mathbf{\Pi}_\mathbb{N}$ given by the mixture:

$$\mu(d\rho) = \frac{\beta}{2} \int_0^1 \varrho_p(d\rho) p^{-1}(1-p)^{-1} dp. \tag{15}$$

The fragmentation rates in (4) can now be obtained as

$$\mu(\{\rho \in \mathbf{\Pi}_\mathbb{N} : \rho_{|c} = \{a, b\}\}) = \beta \frac{\Gamma(|a|)\Gamma(|b|)}{\Gamma(|c|)}, \tag{16}$$

[2] Berestycki (2004) referred to these as fragmentation-coalescent processes rather than fragmentation-coagulation processes. To avoid confusion, here we use the word coagulation to describe the evolution of partition-valued Markov processes along the genetic sequences, and reserve the word coalescent to describe the genetic process of genealogies coalescing backwards in time.

with the factor of 2 cancelled by the $\varrho_p = \varrho_{1-p}$ symmetry in the binary partition construction. This fragmentation component coincides with that in the beta-splitting model of Aldous (1996) with beta equalling -1. It has been explored independently as a Bayesian nonparametric model for densities by Neal (2003) who called it the Dirichlet diffusion tree.

4.5 Mean Statistics

Let $n \in \mathbb{N}$ and $T > 0$ both be finite, and $(\Pi(t), t \in [0, T])$ be a FCP defined on the finite interval $[0, T]$. In this section we will derive the expectations of a number of useful statistics of $(\Pi(t))$ under the FCP law, which will give further insights into the law of the FCP and how it depends on α and β. The first is a direct consequence of the fact that the marginal distribution of $\Pi(t)$ is $\mathrm{CRP}_{[n]}(\frac{\beta}{\alpha})$:

Proposition 9. *The expected number of paths crossing time t is*

$$\mathbb{E}[|\Pi(t)|] = \frac{\beta}{\alpha}\left(\psi(n + \tfrac{\beta}{\alpha}) - \psi(\tfrac{\beta}{\alpha})\right) \tag{17}$$

where $\psi(x) = \frac{\partial \log \Gamma(x)}{\partial x}$ is the digamma function. For large n, this is $\mathbb{O}(\frac{\beta}{\alpha}\log(n - \frac{1}{2} + \frac{\beta}{\alpha}))$.

If we define the length of a path to be the difference between the time the path was created and when it was terminated, then the expected total length of all paths is simply

$$\int_0^T \mathbb{E}[|\Pi(t)|]dt = T\frac{\beta}{\alpha}\left(\psi(n + \tfrac{\beta}{\alpha}) - \psi(\tfrac{\beta}{\alpha})\right) \tag{18}$$

The second mean statistics can be derived (in the appendix) by noting that the transition rate of a Markov process is the expected number of events per unit time:

Proposition 10. *The expected number of fragmentation and coagulation events is:*

$$\mathbb{E}[F + C] = T\frac{\beta^2}{\alpha}\left(\left(\psi(n + \tfrac{\beta}{\alpha}) - \psi(\tfrac{\beta}{\alpha})\right)^2 + \psi'(n + \tfrac{\beta}{\alpha}) - \psi'(\tfrac{\beta}{\alpha})\right) \tag{19}$$

where $\psi'(x)$ is the trigamma function. For large n, this is $\mathbb{O}(T\frac{\beta^2}{\alpha}\log(n - \frac{1}{2} + \frac{\beta}{\alpha})^2)$.

In fact, since the process is reversible, $\mathbb{E}[F] = \mathbb{E}[C]$ so the expected numbers of fragmentation events and coagulation events are both exactly half that in (19). Since the number of paths is $|A| = |\Pi(0)| + 2F + C$, we get that the expected number of paths is

$$\mathbb{E}[|A|] = \frac{\beta}{\alpha}\left(\psi(n + \tfrac{\beta}{\alpha}) - \psi(\tfrac{\beta}{\alpha})\right) + \frac{3}{2}T\frac{\beta^2}{\alpha}\left(\left(\psi(n + \tfrac{\beta}{\alpha}) - \psi(\tfrac{\beta}{\alpha})\right)^2 + \psi'(n + \tfrac{\beta}{\alpha}) - \psi'(\tfrac{\beta}{\alpha})\right) \tag{20}$$

The expected length of each path can now be approximated by dividing (18) by (20), which for large n becomes $\mathbb{O}((\beta \log(n - \frac{1}{2} + \frac{\beta}{\alpha}))^{-1})$. In particular, note that as $n \to \infty$ we expect path lengths to reduce to 0.

5. Related Work

In this section we will how fragmentation-coagulation processes relate to a number of existing statistical and Bayesian nonparametric models. In particular, we will discuss relationships to hidden Markov models, dependent Dirichlet processes, and Bayesian nonparametric models for hierarchical clustering.

5.1 Hidden Markov Models and the Label Switching Problem

As noted in Section 2, the latent variables of hidden Markov models (HMMs) can be equivalently expressed as Markov chains over partitions as in the case of FCPs. A subtle but important distinction between FCPs and HMMs is that while the clusters in FCPs are unlabelled, those in HMMs are labelled by the corresponding latent state. If both the HMM prior and likelihood are invariant to permutations of the latent states then the posterior will have exponentially many modes, each corresponding to a permutation of the labels. This is called the label switching problem (Jasra et al., 2005; Spezia, 2009).

An even more subtle difficulty arises in applications of HMMs to genetic sequences, due to fact that the observation models (3) at different sites are independent. In particular, the likelihood is invariant to permuting the states at different positions using different permutations, while the prior typically gives higher probabilities to self-transitions and so is not invariant under such separate permutations. This implies that each such permutation gives rise to a different sub-optimal local mode in the posterior distribution. As there are a much larger number of such permutations, it is very easy for any posterior simulation technique to get trapped in one of the local modes.

To demonstrate this difficulty, we compare FCPs against HMMs on a very simple data set comprising 160 sequences. Half of the sequences consist of 16 observations of ‘0’ and the other half consist of 16 observations of ‘1’. We use a Bayesian HMM with two latent states and posterior simulations are obtained by alternating between the forward-filtering-backward-sampling procedure to sample the latent state variables and Gibbs sampling for the parameters. The probability τ of self-transitions is given a $\text{Beta}(10, 0.1)$ prior, while uniform priors are used for the emission probabilities. The FCP model uses the same emission model as the HMM but the FCP prior over partition sequences. The global posterior optima under both models assign all ‘0’ observations to one state (cluster) and all ‘1’ observations to another state (cluster).

Figure 3 shows the convergence of MCMC samplers for both models, with the HMM requiring more iterations to converge than the FCP. This is because the random initialisation of the HMM parameters gives higher probabilities under a state to ‘0’ or ‘1’ randomly at each location in the sequence, and it takes a while before the prior preference for self-transitions switches these around so that one state gives preference for one observed value. On the other hand, the FCP simply assigns the ‘0’ sequences to one path and ‘1’ sequences Varying the strength of the priors changes the speed of convergence but does not affect the qualitative difference in convergence speeds of posterior simulations in the models. On the other hand, increasing the amount of data actually worsens the convergence since the likelihood gets peakier resulting in less chance for the MCMC sampler to traverse across modes unless specialised label switching moves are implemented. We expect that this difficulty with

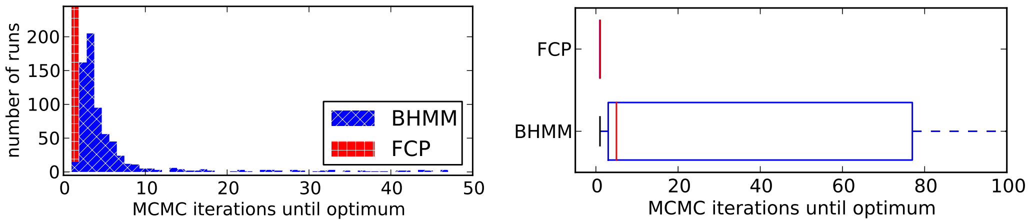

Figure 3: Empirical comparison of convergence speeds for Bayesian HMM and FCP. We used the number of MCMC iterations before each model first encountered its respective optimal states as a test for convergence. 1000 independent runs were conducted for each model. *Left:* Histogram of the number of iterations required for convergence. *Right:* Box plot showing quartiles and median.

label switching will be worse in more complex models consisting of larger numbers of latent states and more complex mosaic structures.

5.2 Dependent Dirichlet Processes

Since the fragmentation-coagulation process is exchangeable, projective, and is marginally distributed as a CRP, the de Finetti measure of our model can be expressed as $(G(t) : t \in [0, \infty))$ where each $G(t)$ is marginally DP. In other words, it is a dependent Dirichlet process (DDP) (MacEachern, 1999). These are a wide class of measure-valued stochastic processes with DP marginal laws. Among other applications, they have been used for spatial models (Gelfand et al., 2005; Griffin and Steel, 2006), text models (Rao and Teh, 2009), time series models (Caron et al., 2007; Griffin, 2011), density regression (Griffin and Steel, 2008) and in biostatistics (Dunson, 2010).

5.3 Hierarchical Clustering Models based on Fragmentation and Coagulation Processes

The theory of fragmentation and coagulation processes have recently received much attention in the probability literature (Bertoin, 2006), though these have so far not significantly impacted research on Bayesian nonparametrics. Nevertheless, in hindsight the theory can serve to unify a number of previously investigated Bayesian nonparametric models for trees.

The first such model is the Dirichlet diffusion tree of Neal (2003), which combines a fragmentation process with fragmentation rates given by (4) (allowing β to vary with position t) and a Brownian diffusion on the paths forming the branches of the tree. Neal (2003) showed that observations at the leaves of the tree are exchangeable and so by de Finetti's Theorem there is an underlying random distribution corresponding to the process. The properties of this random distribution, including its structure and continuity, have not been well explored. Recently Knowles and Ghahramani (2011) extended the model to a Pitman-Yor diffusion tree, which produces a multifurcating tree rather than the binary Dirichlet diffusion tree. While the Dirichlet diffusion tree is related to the beta-splitting cladogram of Aldous (1996), the Pitman-Yor diffusion tree is related to the Gibbs fragmentation tree of

McCullagh et al. (2008). The nested CRP of Blei et al. (2010) (see also the related nested DP (Rodríguez et al., 2008)) can also be construed as a fragmentation process, albeit in discrete steps instead of continuous time as here. It is possible to derive the Dirichlet diffusion tree as a continuous time limit of the nested CRP, by taking the mass parameters to scale with the step sizes as they go to zero.

While fragmentation processes construct trees in a top down manner, coagulation processes construct these from the bottom up. The canonical example is the coalescent (Kingman, 1982a,b), which has coagulation rates (6) with $\alpha = 1$. The coalescent is a cornerstone in mathematical genetics, with a large variety of extensions and generalisations (Hein et al., 2005). It has also been used as a prior for Bayesian hierarchical clustering in machine learning (Teh et al., 2008), who suggested that the bottom up construction is more amenable to efficient sequential Monte Carlo posterior simulations. Ho et al. (2006) noted that the Chinese restaurant franchise, a CRP representation of the hierarchical Dirichlet process (Teh et al., 2006), is a discrete time coagulation process, thus making a connection between coagulation processes and certain classes of hierarchical Bayesian nonparametric models (Teh and Jordan, 2010). Wood et al. (2009) made use of this connection, along with a fragmentation-coagulation duality (Pitman, 1999) related to the FCPs described here, to develop an efficient Bayesian nonparametric model for natural language sequences.

6. Posterior Simulation

In this section, we describe a Markov chain Monte Carlo inference algorithm for obtaining posterior samples of the partition structure $(\Pi(t))$ and the hyperparameters α, β and ω. Sampling the hyperparameters given $(\Pi(t))$ uses standard techniques, and the main difficulty is in sampling $(\Pi(t))$ given the observations and hyperparameters. Although $(\Pi(t))$ is just a Markov process, using the standard forward-backward inference is infeasible because of the exponentially large space over partitions and because $(\Pi(t))$ is a continuous-time Markov process. We take a Gibbs sampling approach by making use of exchangeability: each step of the algorithm treats a sequence i as the last sequence to be added into the partition structure of $(\Pi(t))$, resampling its trajectory $(c_i(t))$ given $(\Pi_{|[n]\backslash i}(t))$ and data x_i.

6.1 Gibbs Sampling the Mosaic Structure

Recall that the ith sequence x_i consists of observations x_{ij} at positions t^o_j, $j \in [m]$. The conditional distribution of the trajectory $(c_i(t))$ given the trajectories $(\Pi_{|[n]\backslash\{i\}}(t))$ of the other sequences is a Markov jump process as described in Section 4.3, suitably altered to condition on $(\Pi_{|[n]\backslash\{i\}}(t))$ instead of on $(\Pi_{|[i-1]}(t))$. The conditional probability of x_i given the other sequences and the trajectory $(c_i(t))$ is derived from emission model (3). We will assume that $H(\omega)$ is conjugate to $F(\theta)$, so that this can be computed efficiently. The product of these two conditional probabilities gives the desired conditional distribution of $(c_i(t))$ given x_i and the rest of the model. Since the observations are conditionally independent given the trajectory, which is Markov, this has a structure similar to a HMM, and hence it is conceivable to develop a similar forward-backward algorithm to sample $(c_i(t))$ from its conditional distribution. The main difficulty is that $(c_i(t))$ is a continuous-time

Markov jump process. One approach is to discretise time[3] t, which is approximate and can be computationally expensive if a fine discretisation is used. Another is to compute the transition probability of $c_i(t^o_{j+1})$ given $c_i(t^o_j)$ by matrix exponentiation, which is also expensive since this has to be computed anew for each iteration, each i and each j separately.

Instead, we will make use of a recently proposed algorithm for continuous-time Markov jump processes (Rao and Teh, 2011, 2012). This algorithm is based on the idea of *uniformisation* (Jensen, 1953), which is an alternative generative process for Markov jump processes different from the more well-known Gillespie (1977) algorithm. It operates by first generating a set of potential jump times J^{pot} according to a Poisson process whose rate dominates that of the Markov jump process. Given J^{pot}, the trajectory of the Markov jump process is then generated using a Markov chain which can only transition at the potential jump times. This Markov chain is allowed to stay in the same state and not transition, which amounts to rejecting the potential jump, an idea similar to that of thinning a Poisson process (Lewis and Shedler, 1979).

The idea of Rao and Teh (2011, 2012) is to treat J^{pot} as an auxiliary variable to the Markov jump process, and simply apply Gibbs sampling to the augmented system. Conditioned on J^{pot}, the Markov jump process may only transition at the times in J^{pot}, reducing it to a discrete-time Markov chain, and the normal forward-filtering-backward-sampling algorithm can be applied to get a new trajectory conditioned on observed data as well. On the other hand, conditioned on the trajectory of the Markov jump process, J^{pot} can be shown to be independent of the observed data, and consists only of the transition times of the trajectory along with a random finite set of times drawn according to a Poisson process with piecewise constant rates.

We shall adapt this algorithm to our situation, which is somewhat more complicated than in Rao and Teh (2011, 2012) as our Markov jump process $(c_i(t))$ is not itself homogeneous. For each time t, let $S(t) = \Pi_{|[n]\backslash\{i\}}(t) \cup \{\emptyset\}$ be the state space for $c_i(t)$, and let $Q_t(s', s)$ be the rate of $c_i(t)$ transiting from state $s' \in S(t-)$ to $s \in S(t)$ as described in Section 4.3. The diagonal terms of the transition rate matrices are

$$Q_t(s, s) = -\sum_{s' \neq s} Q_t(s, s'). \tag{21}$$

Let U_t be a dominating rate with

$$U_t > \max_{s \in S(t)} |Q_t(s, s)| \tag{22}$$

We use $U_t = 2 \max_{s \in S(t)} -Q_t(s, s)$ in our simulations. Let J^{fixed} consist of the start time 0, end time T, times of fragmentation and coagulation events in $(\Pi_{|[n]\backslash\{i\}}(t))$, and observation times $\{t^o_1, \ldots, t^o_m\}$, all of which are distinct with probability one. These constitute the fixed event times of $(c_i(t))$ when the forward-backward algorithm needs to account for the probabilities of these events. Outside of these fixed event times, $(c_i(t))$ can also transition at additional jump times, which are defined as follows. Let $(c_i^{\text{prev}}(t))$ be the previous sample trajectory of $(c_i(t))$ and J^{prev} the set of jump times in $(c_i^{\text{prev}}(t))$ beside those in J^{fixed}. These

[3] In this section we will refer to t as *time* to be consistent with continuous-*time* Markov jump processes, although in the genetics context t is a location on the chromosome.

consist of the times when the ith trajectory split off from an existing cluster to form its own cluster, and when it merged back to another existing cluster in $(\Pi_{|[n]\backslash\{i\}}(t))$. Let J^{aux} be an additional set of jump times drawn according to a Poisson process with piecewise constant rates

$$U_t - |Q_t(c_i^{\text{prev}}(t), c_i^{\text{prev}}(t))|. \tag{23}$$

The potential jump times now consist of $J^{\text{pot}} = J^{\text{prev}} \cup J^{\text{aux}}$, and it can be shown that the above process describes precisely the conditional distribution of J^{pot} given $(c_i^{\text{prev}}(t))$. On the other hand, conditioned on J^{pot} the discrete-time Markov chain $(c_i(t) : t \in J \cup J^{\text{fixed}})$ has initial probabilities (11), transition probabilities given in Section 4.3 at the fixed event times $t \in J^{\text{fixed}}$, and transition probabilities given by

$$\mathbb{P}(c_i(t) = s | c_i(t-) = s') = \delta(s', s) + \frac{Q_t(s', s)}{U_t} \qquad \text{for } s' \in S(t-), s' \in S(t), \tag{24}$$

at the potential jump times $t \in J^{\text{pot}}$, where $t-$ now denotes the last time step prior to t in the Markov chain, and $\delta(s', s)$ is the Kronecker delta function.

We can now define the forward messages at the above event times $t \in J^{\text{fixed}} \cup J^{\text{pot}}$ as follows:

$$\begin{aligned}\lambda(s, t-) &= \mathbb{P}\big(\{x_{ij} : t_j^o < t\} \,\big|\, c_i(t-) = s, J^{\text{pot}}\big) &\qquad& \text{for } s \in S(t-),\\ \lambda(s, t) &= \mathbb{P}\big(\{x_{ij} : t_j^o \le t\} \,\big|\, c_i(t) = s, J^{\text{pot}}\big) && \text{for } s \in S(t),\end{aligned} \tag{25}$$

where we have suppressed the conditioning on $(\Pi_{|[n]\backslash\{i\}}(t))$ and $\{x_{i'j} : i' \neq i\}$ for simplicity. Note that the above forward messages are not the standard forward messages for HMMs. They were defined in a reversible manner, making use of the reversibility of the FCP, so that the same algorithmic implementation can be applied to compute the backward messages as well. The messages can be computed using a forward filtering phase (which operates on the reverse of the process in Section 4.3 so that fragmentation events appear as coagulation events and vice versa):

- At the starting position $t = 0$, we have simply that

$$\lambda(s, t) = 1 \qquad\qquad \text{for } s \in S(0). \tag{26}$$

- At an observed site $t = t_j^o$, we incorporate the local likelihood term:

$$\lambda(s, t) = \lambda(s, t-)\ell(x_{ij} | \{x_{i'j} : i' \in s\}, \omega_j) \quad \text{for } s \in S(t). \tag{27}$$

  where $\ell(x_{ij} | \{x_{i'j} : i' \in s\}, \omega_j)$ is the conditional probability of x_{ij} given the other observations in subset s:

$$\ell(x_{ij} | \{x_{i'j} : i' \in s\}) = \frac{\int f(x_{ij}|\theta_{sj}) \prod_{i' \in s} f(x_{i'j}|\theta_{sj}) h(\theta_{sj}|\omega_j) d\theta_{sj}}{\int \prod_{i' \in s} f(x_{i'j}|\theta_{sj}) h(\theta_{sj}|\omega_j) d\theta_{sj}} \tag{28}$$

  where $f(x|\theta)$ is the probability of x under $F(\theta)$, and $h(\theta|\omega)$ is the probability of θ under the prior $H(\omega)$.

- At an existing fragmentation event where a cluster $c \in S(t-)$ fragments into $a, b \in S(t)$,

$$\lambda(a,t) = \lambda(b,t) = \lambda(c,t-), \qquad \lambda(s,t) = \lambda(s,t-) \quad \text{for } s \in S(t)\backslash\{a,b\}. \tag{29}$$

- At an existing coagulation event where subsets $a, b \in S(t-)$ merges into $c \in S(t)$,

$$\lambda(c,t) = \lambda(a,t-)\tfrac{|a|}{|c|} + \lambda(b,t-)\tfrac{|b|}{|c|}, \qquad \lambda(s,t) = \lambda(s,t-) \quad \text{for } s \in S(t)\backslash\{c\}. \tag{30}$$

- At a potential jump time $t \in J^{\text{prev}} \cup J^{\text{aux}}$,

$$\lambda(s,t) = \lambda(s,t-) + \sum_{s' \in S(t-)} \lambda(s',t-)\tfrac{Q_t(s',s)}{U_t} \quad \text{for } s \in S(t). \tag{31}$$

- Finally, at $t = T$, we make use of the stationarity of the FCP to compute,

$$\mathbb{P}\big(\{x_{i1},\dots,x_{im}\}, c_i(t) = s \,\big|\, J^{\text{pot}}\big) = \lambda(s,t-) \times \begin{cases} \frac{|s|}{n-1+\frac{\beta}{\alpha}} & \text{if } s \in \Pi_{|[n]\backslash\{i\}}(t), \\ \frac{\frac{\beta}{\alpha}}{n-1+\frac{\beta}{\alpha}} & \text{if } s = \emptyset. \end{cases} \tag{32}$$

A backward-sampling phase can be easily derived corresponding to the forward-filtering phase, giving a new sample path $(c_i^{\text{next}}(t))$.

In our experiments the mosaic structure $(\Pi(t))$ was initialised by starting with an empty structure, and adding sequences in the dataset into the structure one at a time using the uniformised forward-backward method described above.

6.2 Hyperparameter and Emission Model Specification and Updates

We parameterise the FCP using an alternative parameterisation to α, β which is more intuitive and amenable for more informed prior specification. In particular, consider $\mu = \frac{\beta}{\alpha}$ and $\nu = \alpha$. Then Proposition 2 shows that μ controls the marginal behaviour of the FCP with the marginal distribution of each $\Pi(t)$ simply $\text{CRP}_{[n]}(\mu)$. On the other hand, both rates of fragmentation (4) and coagulation (6) are linear in ν, so that ν controls the evolution rate of the Markov process. In our simulations we allow ν to vary with location, taking on value ν_j between sites t_j^o and t_{j+1}^o. This allows for the effective distances between sites to be adapted to account for possibly varying recombination rates and hotspots on the chromosome. The joint probability for a sample path $(\pi(t))$ is then:

$$P((\Pi(t)) = (\pi(t))) = \mu^{|A|-F-C} \frac{\Gamma(\mu)}{\Gamma(n+\mu)} \frac{\prod_{c \in A_{<>}} \Gamma(|c|)}{\prod_{c \in A_{><}} \Gamma(|c|)}$$
$$\times \prod_{j=1}^{m-1} \nu_j^{F_j+C_j} \exp\left(- \int_{t_j^o}^{t_{j+1}^o} \nu_j \frac{|\pi(t)|(|\pi(t)|-1)}{2} + \nu_j \mu_j \sum_{c \in \pi(t)} H_{|c|-1} dt \right) \tag{33}$$

where F_j, C_j are the numbers of fragmentation and coagulation events between locations t_j^o and t_{j+1}^o respectively. In our experiments we use a log normal prior for $\log \mu$ centered at

10 and spanning a six orders of magnitude range, and an independent log normal prior for each ν_j centered at a value such that the approximate expected length of each path (see Section 4.5) spans 100 sites.

For the emission model, we use a very simple model where each cluster parameter θ_{cj} is binary, indicating whether sequences in cluster c at location j exhibit a 0 or 1 allele. The prior $H(\omega_j)$ for θ_{cj} is Bernoulli with mean ω_j, and ω_j is in turn given a conjugate beta prior with mean 0.5 and mass γ_j. Finally, $\log \gamma_j$ is given a uniform prior over the range $[\log(10^{-4}), \log(1)]$, corresponding to the observation that most allele frequencies are skewed to either extremely. We have found this to perform significantly better than the obvious hierarchy with $F(\theta)$ a beta distribution. This is because the true mosaic structure is non-Markov and exhibits long range dependencies which is not easily captured by the Markov FCP. Using a beta emission allows the model to "cop out" by merging multiple mosaic segments into single paths, and using the beta to model the resulting variations. Using the Bernoulli emission forces the FCP to model each distinct sequence of alleles using a different path, creating more and longer paths in the posterior.

In our simulations all hyperparameters were initialised at the respective means in the log domain, and updated using slice sampling (Neal, 2003), while ω_j is integrated out. The parameters are updated ten times for each Gibbs sweep over all sequences to update $(\Pi(t))$.

7. Experimental Evaluation

We conducted empirical studies on 20 datasets generated from Phase 1 Release v3 of the Thousand Genomes Project acquired on 17/5/2012. The datasets were obtained from the non-pseudoautosomal region of 524 male X chromosomes. Each dataset consists of a non-overlapping segment of 500 contiguous SNPs, spanning an average length of about 10^5 base pairs. Software for fastPHASE, one of the alternatives we compared against, requires all sequences to be paired, so we could only compare an even number of X chromosomes. For each dataset we randomly discarded one of the 525 male X chromosomes that were included in Phase 1 of the Thousand Genomes Project, and randomly paired up the remaining 524 chromosomes (preserving the phase information).

Datasets were generated with some alleles masked, and various methods are tested on their accuracies in imputing masked alleles. The masking is carried out under two scenarios. In the first scenario, a proportion of the alleles were masked uniformly at random from among all pairs of sequences and SNP sites. Assuming that the same assaying protocol were applied to all of the individuals in a study, this condition models noise intrinsic to that assaying protocol. The proportion of missing alleles was varied between 10% and 50%. Large genetic datasets such as the 1000 Genomes Project can be used as a reference panel against which noisy and sparsely assayed individuals are registered (Howie et al., 2009). The second scenario was designed to simulate this study/reference framework. A portion of the sequence pairs in each dataset were randomly chosen to be in the study panel, and a randomly chosen portion of the SNP sites were masked in the study panel. The remaining sequence pairs were not masked and were used as the reference panel. Both proportions were varied between 10% and 50%. In both scenarios, the masking was done such that at each site at least one minor allele was observed. Each method is used to impute all masked

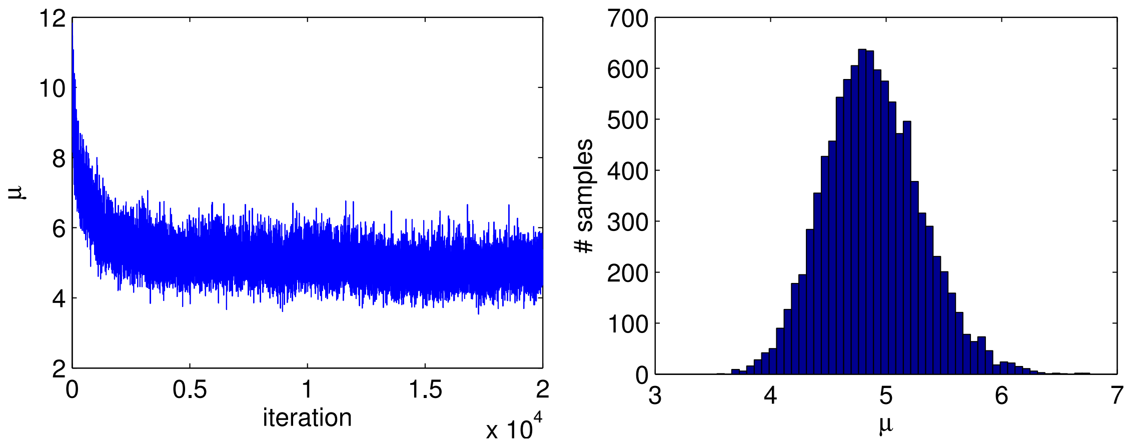

Figure 4: Left: Posterior μ samples over the course of the MCMC run. Right: histogram of samples 10001 to 20000.

alleles and tested on its imputation accuracy. The baseline accuracy, found by predicting the major allele for every missing entry, averaged over all scenarios and segments, was 93.88%.

7.1 Posterior Distribution

We first investigate the convergence of the proposed MCMC algorithm and the posterior distribution obtained on one of the datasets and under the first scenario with 25% of the alleles masked. The MCMC algorithm was run for 40000 iterations, from which we collected 20000 samples. Figure 4 shows the posterior samples of μ, which indicates convergence after about 20000 iterations. In the top two rows of Figure 5 we show the same plots for three ν_j parameters. One was chosen with lowest mean value of ν_j among the second 10000 samples, one with highest mean, and one with a medium value. There are significantly higher auto-correlations in the ν_j samples with slower mixing (particularly the middle ν_j). In the bottom row we plot the corresponding numbers of fragmentation and coagulation events $F_j + C_j$. Higher values of ν_j are associated with larger $F_j + C_j$, which is to be expected from (33), and is the cause of the higher auto-correlations in ν_j.

Figure 6 summarises some posterior statistics of the mosaic structure. The mosaic structure exhibits clear spatial variations, with different numbers of clusters and events present in different parts of the chromosome. The numbers of fragmentation and coagulation events in each segment between consecutive sites varies widely, with most segments having very small number of events, while a few has significantly larger numbers.

In Figure 7 we plot the accuracy of the method over the course of the MCMC run. For each S, after S samples were collected, these were used to impute the missing alleles and the accuracy on predicting the true alleles was computed. We find that the imputation accuracy of the sampler converges much more rapidly than the parameters would indicate. In fact we see that the accuracy already achieves its highest value after only approximately 500 iterations.

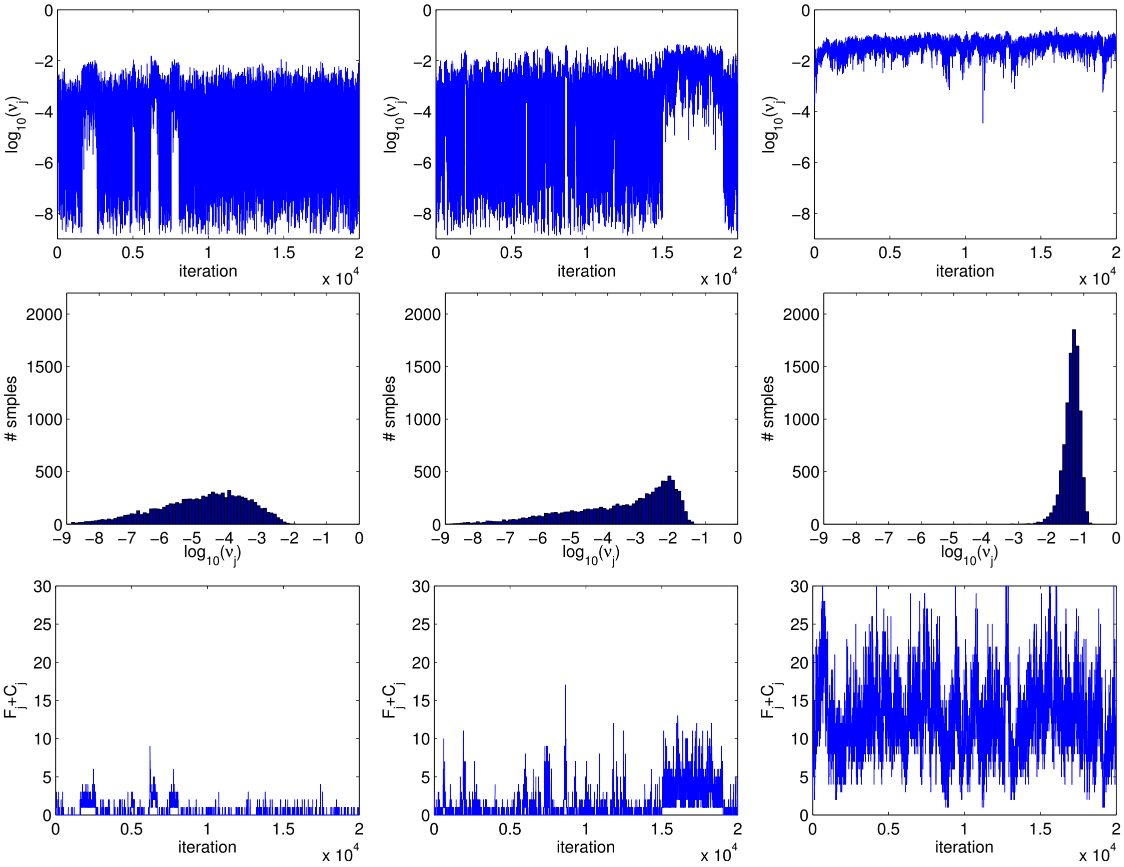

Figure 5: Top row: Posterior ν_j samples over the course of the MCMC run. Middle row: Histogram of samples 10001 to 20000. Bottom row: corresponding $F_j + C_j$ samples. Left column shows the ν_j with lowest posterior mean as estimated by the second half of the samples, right column shows the ν_j with highest mean, and middle shows a ν_j with medium mean.

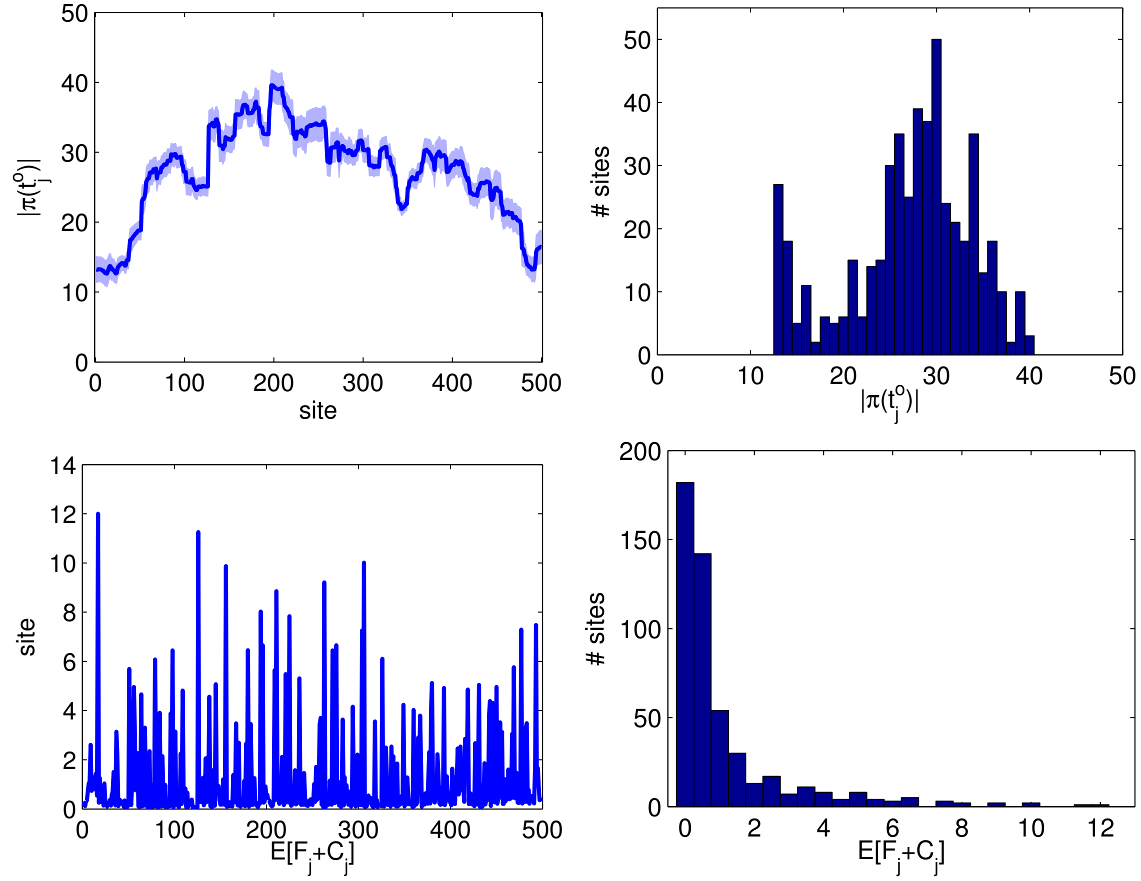

Figure 6: Top Left: posterior mean and standard deviation of the number of clusters at each SNP site. Top Right: Histogram of the mean number of clusters. Bottom Left: posterior mean of the number of fragmentation and coagulation events between each pair of consecutive sites. Bottom Right: Histogram of the mean number of fragmentation and coagulation events.

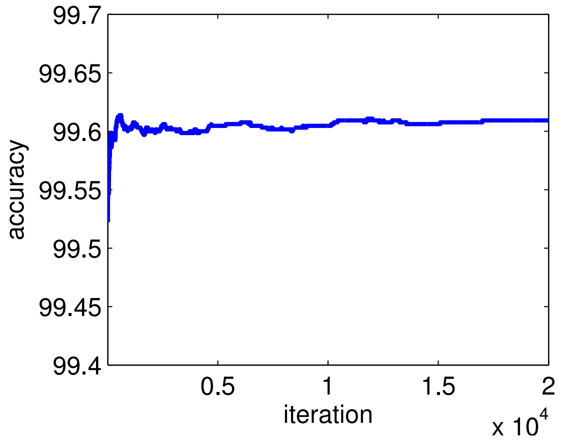

Figure 7: Imputation accuracy over the course of the MCMC run.

7.2 Imputation Results

In view of the above observation regarding imputation accuracies, we used the following MCMC scheme for the second set of experiments. The MCMC chain is initialised with the log hyperparameters at their prior means and the mosaic structure constructed sequentially. 3 sweeps of Gibbs updates to the mosaic structure are performed with fixed hyperparameters, followed by 3 more sweeps where hyperparameters are updated too. We have found that the first sweeps with fixed hyperparameters are important because the sequential initialisation tends to produce larger numbers of clusters and events, so if hyperparameters are updated right afterward they will become too large and the sampler will take longer to converge. These constitute the burn-in period, and are followed by 40 MCMC iterations, from which we collect 20 samples. The above is repeated 5 times, resulting in 100 samples which are used to impute missing alleles. For each subsequent repeat the hyperparameters are kept at the previously sampled values, but the mosaic structure is reinitialised as before. The 5 repeats increase the chance that the MCMC sampler will explore multiple modes of the posterior over mosaic structures.

In the following, we compared imputation accuracies obtained by our method (FragCoag) to that by FastPHASE (Scheet and Stephens, 2006) and Beagle (Browning and Browning, 2009), two state of the art methods for genotype imputation. In preliminary experiments we have found that Impute2 (Howie et al., 2009) does not work well on phased data as in our situation, and produced significantly worse results than FastPHASE and Beagle. In addition we have also compared against a FragCoag implementation with a beta emission model (see Section 6.2) which we refer to as FragCoag-beta.

Figure 8 shows the imputation accuracies obtained by the four methods on the first scenario, where alleles were masked uniformly at random. We see that FragCoag produces consistently and significantly better accuracies than both FastPHASE and Beagle, while FastPHASE and Beagle produced similar results. We also see that FragCoag-beta performed significantly worse than the other methods, as discussed in Section 6.2. Accuracy differences among the 20 datasets at each percentage of alleles withtheld were significant according to

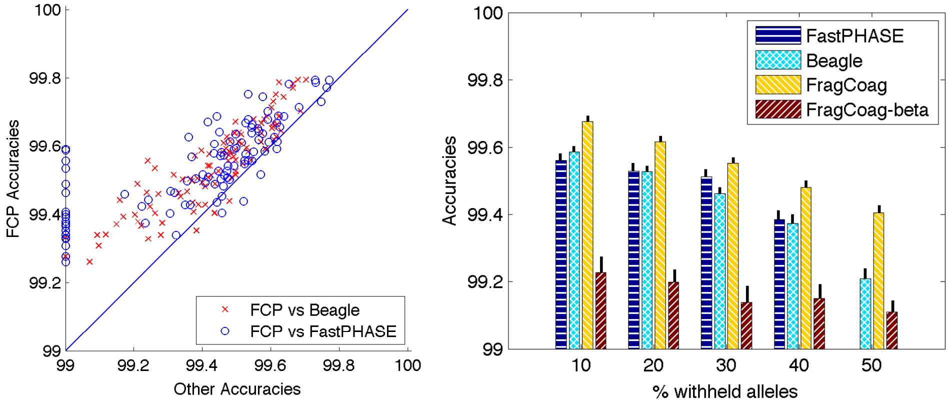

Figure 8: Imputation accuracies on the first scenario where alleles are missing uniformly at random. Left: Each point compares the accuracy obtained by FragCoag against that obtained by FastPHASE and Beagle on one dataset. Sometimes FastPHASE and Beagle obtained accuracies less than 90%, in which case they were clipped at 90% (points on y-axis of figure). FragCoag accuracies were always above 90%. Right: Accuracies for each proportion of masked alleles, averaged over the 20 datasets, with vertical bars showing the standard errors.

a sign test at 95% confidence level. The accuracies of all methods degrade as the proportion of masked alleles increases. Above 40% masked alleles the FastPHASE software produced abnormal behaviour and reported accuracies plummet below 90%. We believe this is an issue with the released software and not with the method itself.

In Figure 9 we compare the imputation accuracies of the four methods on the second study/reference panel scenario. The qualitative trends are similar, with FragCoag significantly better than both fastPHASE and Beagle. One difference from the previous scenario is that FastPHASE is better than Beagle here, and did not produce the runtime errors observed in the previous scenario. Accuracy differences were significant except for p% withheld sites, q% withheld sequences, for $(p, q) \in \{(30, 10), (40, 30), (50, 30)\}$, where differences between FastPHASE and FragCoag were insignificant at 95% confidence.

Figure 10 shows the calibration of the probabilities imputed by FragCoag. For each dataset, the missing alleles are binned in terms of the imputed probabilities of the major allele. The proportion of major alleles for each bin is then plotted against the average imputed probability. The imputation probabilities are slightly under-estimated but reasonably well-calibrated. In preliminary experiments we have found that imputed probabilities become better calibrated for longer MCMC runs. Here we have reported only the calibration of the short runs described above.

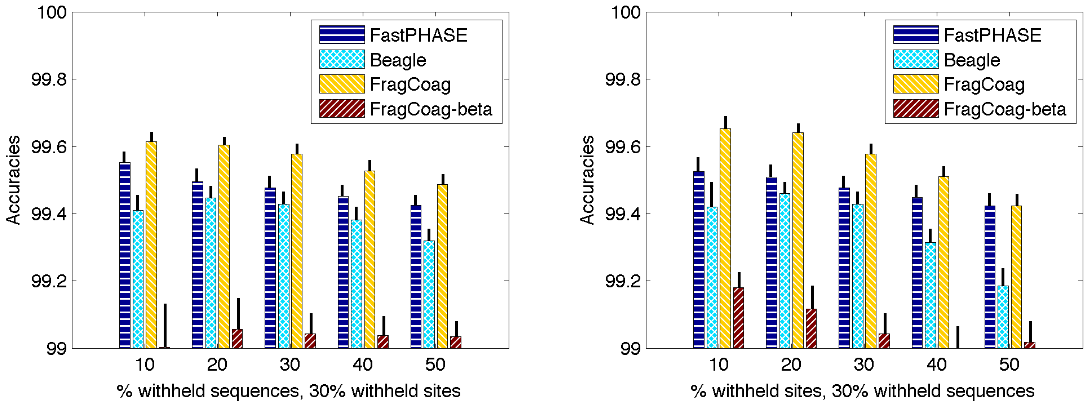

Figure 9: Imputation accuracies on study/reference panel scenario. Left: Accuracies of methods are plotted against the proportion of sequence pairs in the study panel. The proportion of masked sites in the study panel is held fixed at 30%. Right: Accuracies of methods as the proportion of masked sites in study panel is varied. The proportion of sequence pairs in the study panel is fixed at 30%.

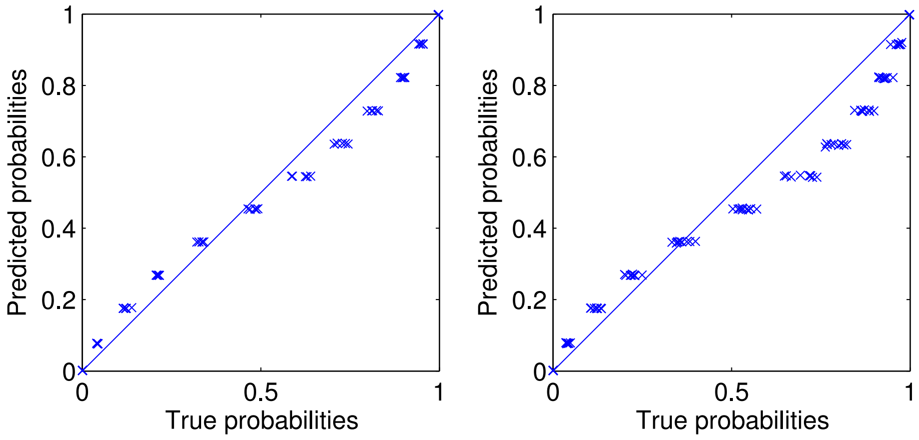

Figure 10: Calibration of allele imputation probabilities. Left plot shows the uniformly missing at random scenario, while right plot shows the study/reference panel scenario.

8. Discussion

We have described a novel model of genetic variations accounting for linkage disequilibrium based on fragmentation-coagulation processes. Our Bayesian nonparametric model can automatically adapt the number of clusters it uses to model data, and does not suffer from the label switching problems of HMMs. We described a scalable Markov chain Monte Carlo inference algorithm, and showed in experiments that our model achieves state-of-the-art performance in imputation accuracy.

We are currently extending the research in a few directions. We are experimenting with an extension to handle unphased data, and in using fragmentation-coagulation processes for phasing such data. We are also exploring ways to scale up our method to handle chromosome scale datasets, and are intending to release the software for public use.

Other future directions include MCMC methods to improve the mixing of the loci distance parameters ν_j, and in investigating the relationship between the ν_j parameters and recombination rates and hotspots. We have found positive correlation between these, which is unsurprising given that hotspots are where there are breaks in haplotypes.

Our work is the first application of fragmentation-coagulation processes to the statistical analysis of sequence data. We expect such processes will find more statistical applications beyond modelling genetic variations in the future. For example, FCPs can be used to model community structure in evolving social networks, where communities can split and merge with time.

Acknowledgements

We thank Gil McVean and Chris Holmes for discussions and for generously hosting Yee Whye Teh when he was at Oxford for a short sabbatical in 2011, during which part of the research reported here was carried out. We also thank Andriy Mnih and Vinayak Rao for discussions. We gratefully acknowledge generous funding from the Gatsby Charitable Foundation.

Appendix A. Proof of Proposition 4

The probability of $(\Pi(t)) = (\pi(t))$ can be written as the product of the probability of the initial state $\Pi(0) = \pi(0)$, times the probabilities of each subsequent jump. Since the initial distribution is $\mathrm{CRP}_{[n]}(\frac{\beta}{\alpha})$, the probability of the initial state is:

$$\mathbb{P}(\Pi(0) = \pi(0)) = g_{[n]}(\pi(0); \tfrac{\beta}{\alpha}) = \frac{\Gamma(\frac{\beta}{\alpha})(\frac{\beta}{\alpha})^{|\pi(0)|} \prod_{a\in\pi(0)} \Gamma(|a|)}{\Gamma(n + \frac{\beta}{\alpha})} \tag{34}$$

For each subsequent jump, let t be the time of the jump, and t' the time of the previous jump (or 0 for the first jump). The distribution of the holding time in state $\pi(t')$ is exponential with rate given by (7), so

$$\begin{aligned}&\mathbb{P}(\Pi(s) = \pi(s) = \pi(t') \forall s \in [t', t), \Pi(t) \neq \pi(t') | \Pi(t') = \pi(t'))\\ =&q(\pi(t')) \exp(-(t - t')q(\pi(t')))\end{aligned} \tag{35}$$

At the jump, the probability of a fragmentation of $c \in \pi(t')$ into two clusters $a, b \in \pi(t)$ is:

$$\mathbb{P}(\Pi(t) = \pi(t) | \Pi(t-) = \pi(t-)) = \frac{\beta}{q(\pi(t-))} \frac{\Gamma(a)\Gamma(b)}{\Gamma(c)} \tag{36}$$

where $t-$ denotes an infinitesimal time prior to t. On the other hand, the probability of a coagulation of $a, b \in \pi(t')$ into $c \in \pi(t)$ is:

$$\mathbb{P}(\Pi(t) = \pi(t) | \Pi(t-) = \pi(t-)) = \frac{\alpha}{q(\pi(t-))} \tag{37}$$

Multiplying (34), (35), (36) and (37) over all jumps as well as the final hold state until time T, we get the probability of $(\Pi(t)) = (\pi(t))$ in Proposition 4. The first term of (10) uses the identity $|A| = |\pi(0)| + 2F + C = |\pi(T)| + F + 2C$, which is a result of each path being created either at time 0 or by a fragmentation (which creates two paths) or coagulation, and similarly terminated at time T or by a fragmentation or coagulation (which terminates two paths). The exponential integral is from (35) summed over all hold states between events. The $q(\pi(t'))$ factors in (35) cancels the $q(\pi(t-))$ factors in (36) or (37). Finally, the last ratio of gamma factors results from the ratios of gammas in (36). Only the gamma factors associated with paths in $A_{<>}$ and $A_{><}$ are left, all other gamma factors cancel off.

Appendix B. Proof of Proposition 10

The number of events can be expressed as a sum over each subsequent trajectory, of the number of new fragmentation and coagulation events introduced by that trajectory. Noting a transition rate is the expected number of events per unit time, the expected total number of events is:

$$\mathbb{E}\left[\sum_{i=2}^{n} \int_0^T \mathbb{P}(c_i(t) = \emptyset | \Pi_{|[i-1]}) \alpha |\Pi_{|[i-1]}(t)| + \sum_{c \in \Pi_{|[i-1]}(t)} \mathbb{P}(c_i(t) = c | \Pi_{|[i-1]}) \frac{\beta}{|c|} dt\right]$$

Using Fubini's Theorem,

$$= \sum_{i=2}^{n} \int_0^T \mathbb{E}\left[\mathbb{P}(c_i(t) = \emptyset | \Pi_{|[i-1]}) \alpha |\Pi_{|[i-1]}(t)| + \sum_{c \in \Pi_{|[i-1]}(t)} \mathbb{P}(c_i(t) = c | \Pi_{|[i-1]}) \frac{\beta}{|c|}\right] dt$$

Using Proposition 2, that the marginal distributions are $\mathrm{CRP}_{[n]}(\frac{\beta}{\alpha})$,

$$\begin{aligned}
&= \sum_{i=2}^{n} \int_0^T \mathbb{E}\left[\frac{\frac{\beta}{\alpha}}{i-1+\frac{\beta}{\alpha}} \alpha |\Pi_{|[i-1]}(t)| + \sum_{c \in \Pi_{|[i-1]}(t)} \frac{|c|}{i-1+\frac{\beta}{\alpha}} \frac{\beta}{|c|}\right] dt \\
&= \beta \sum_{i=2}^{n} \frac{1}{i-1+\frac{\beta}{\alpha}} \int_0^T \mathbb{E}\left[|\Pi_{|[i-1]}(t)| + \sum_{c \in \Pi_{|[i-1]}(t)} 1\right] dt \\
&= 2\beta \sum_{i=2}^{n} \frac{1}{i-1+\frac{\beta}{\alpha}} \int_0^T \mathbb{E}\left[|\Pi_{|[i-1]}(t)|\right] dt
\end{aligned}$$

with $\Pi_{|[i-1]}(t) \sim \mathrm{CRP}_{[i-1]}(\frac{\beta}{\alpha})$. The expectation of the number of clusters under a CRP is well-known, and gives,

$$
\begin{aligned}
&=2\beta T \sum_{i=2}^{n} \frac{1}{i-1+\frac{\beta}{\alpha}} \sum_{j=1}^{i-1} \frac{\frac{\beta}{\alpha}}{j-1+\frac{\beta}{\alpha}} \\
&=\frac{2\beta^2 T}{\alpha}\frac{1}{2}\left(\left(\sum_{i=1}^{n} \frac{1}{i-1+\frac{\beta}{\alpha}}\right)^2 - \sum_{i=1}^{n}\left(\frac{1}{i-1+\frac{\beta}{\alpha}}\right)^2\right) \\
&=\frac{\beta^2 T}{\alpha}\left(\left(\psi(n+\tfrac{\beta}{\alpha}) - \psi(\tfrac{\beta}{\alpha})\right)^2 + \psi'(n+\tfrac{\beta}{\alpha}) - \psi'(\tfrac{\beta}{\alpha})\right)
\end{aligned}
\tag{38}
$$

where $\psi(\cdot)$ is the digamma function and $\psi'(\cdot)$ its derivative the trigamma function.

References

D. Aldous. Exchangeability and related topics. In *École d'Été de Probabilités de Saint-Flour XIII–1983*, pages 1–198. Springer, Berlin, 1985.

D. Aldous. Probability distributions on cladograms. In D. Aldous and R. Pemantle, editors, *Random Discrete Structures*, pages 1–18. Springer, 1996.

M. J. Beal, Z. Ghahramani, and C. E. Rasmussen. The infinite hidden Markov model. In *Advances in Neural Information Processing Systems*, volume 14, 2002.

J. Berestycki. Exchangeable fragmentation-coalescence processes and their equilibrium measures. *Electronic Journal of Probability*, 9:770–824, 2004.

J. Bertoin. *Random Fragmentation and Coagulation Processes*. Cambridge University Press, 2006.

J. Bertoin. Two-parameter Poisson-Dirichlet measures and reversible exchangeable fragmentation-coalescence processes. *Combinatorics, Probability and Computing*, 17:329–337, 2007.

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

D. M. Blei, T. L. Griffiths, and M. I. Jordan. The nested Chinese restaurant process and Bayesian nonparametric inference of topic hierarchies. *Journal of the Association for Computing Machines*, 57(2):1–30, 2010.

B. L. Browning and S. R. Browning. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *American Journal of Human Genetics*, 84:210–223, 2009.

F. Caron, M. Davy, and A. Doucet. Generalized Polya urn for time-varying Dirichlet process mixtures. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, volume 23, 2007.

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

A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 39:1–38, 1977.

D. B. Dunson. Nonparametric Bayes applications to biostatistics. In N. Hjort, C. Holmes, P. Müller, and S. Walker, editors, *Bayesian Nonparametrics*. Cambridge University Press, 2010.

T. S. Ferguson. A Bayesian analysis of some nonparametric problems. *Annals of Statistics*, 1(2):209–230, 1973.

A. E. Gelfand, A. Kottas, and S. N. MacEachern. Bayesian nonparametric spatial modeling with Dirichlet process mixing. *Journal of the American Statistical Association*, 100(471): 1021–1035, 2005.

D. T. Gillespie. Exact stochastic simulation of coupled chemical reactions. *Journal of Physical Chemistry*, 1977.

J. E. Griffin. The Ornstein-Uhlenbeck Dirichlet process and other time-varying processes for Bayesian nonparametric inference. *Journal of Statistical Planning and Inference*, 141: 3648–3664, 2011.

J. E. Griffin and M. F. J. Steel. Order-based dependent Dirichlet processes. *Journal of the American Statistical Association, Theory and Methods*, 101:179–194, 2006.

J. E. Griffin and M. F. J. Steel. Bayesian nonparametric modelling with the Dirichlet process regression smoother. Technical report, University of Kent and University of Warwick, 2008.

R. C. Griffiths and P. Marjoram. Ancestral inference from samples of DNA sequences with recombiation. *Journal of Computational Biology*, 3(4):479–502, 1996.

J. Hein, M. H. Schierup, and C. Wiuf. *Gene Genealogies, Variation and Evolution*. Oxford University Press, 2005.

N. Hjort, C. Holmes, P. Müller, and S. Walker, editors. *Bayesian Nonparametrics*. Number 28 in Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2010.

M. W. Ho, L. F. James, and J. W. Lau. Coagulation fragmentation laws induced by general coagulations of two-parameter Poisson-Dirichlet processes. http://arxiv.org/abs/math.PR/0601608, 2006.

B. N. Howie, P. Donnelly, and J. Marchini. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genetics*, 5(6), 2009.

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

A. Jasra, C. C. Holmes, and D. A. Stephens. Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling. *Statistical Science*, 20(1):50–67, 2005.

A. Jensen. Markoff chains as an aid in the study of Markoff processes. *Scandinavian Actuarial Journal*, 1953.

J. F. C. Kingman. The coalescent. *Stochastic Processes and their Applications*, 13:235–248, 1982a.

J. F. C. Kingman. On the genealogy of large populations. *Journal of Applied Probability*, 19:27–43, 1982b. Essays in Statistical Science.

D. Knowles and Z. Ghahramani. Pitman-Yor diffusion trees. In *Proceedings of the International Conference on Uncertainty in Artificial Intelligence*, 2011.

P. A. W. Lewis and G. S. Shedler. Simulation of nonhomogeneous Poisson processes by thinning. *Naval Research Logistics Quarterly*, 26(3):403–413, 1979.

N. Li and M. Stephens. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics*, 165(4):2213–2233, 2003.

A.Y. Lo. On a class of Bayesian nonparametric estimates: I. density estimates. *Annals of Statistics*, 12(1):351–357, 1984.

S. MacEachern. Dependent nonparametric processes. In *Proceedings of the Section on Bayesian Statistical Science*. American Statistical Association, 1999.

J. Marchini, B. Howie, S. Myers, G. McVean, and P. Donnelly. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nature Genetics*, 39(7): 906–913, 2007.

P. McCullagh, J. Pitman, and M. Winkel. Gibbs fragmentation trees. *Bernoulli*, 14(4): 988–1002, 2008.

R. M. Neal. Slice sampling. *Annals of Statistics*, 31:705–767, 2003.

M. Perman, J. Pitman, and M. Yor. Size-biased sampling of Poisson point processes and excursions. *Probability Theory and Related Fields*, 92(1):21–39, 1992.

J. Pitman. Coalescents with multiple collisions. *Annals of Probability*, 27:1870–1902, 1999.

J. Pitman. *Combinatorial Stochastic Processes.* Lecture Notes in Mathematics. Springer-Verlag, Berlin, 2006.

J. Pitman and M. Yor. The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *Annals of Probability*, 25:855–900, 1997.

L. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77:257–285, 1989.

V. Rao and Y. W. Teh. Spatial normalized gamma processes. In *Advances in Neural Information Processing Systems*, volume 22, pages 1554–1562, 2009.

V. Rao and Y. W. Teh. Fast MCMC sampling for Markov jump processes and continuous time Bayesian networks. In *Proceedings of the International Conference on Uncertainty in Artificial Intelligence*, 2011.

V. Rao and Y. W. Teh. Fast MCMC sampling for Markov jump processes and extensions. Submitted, 2012.

C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

A. Rodríguez, D. B. Dunson, and A. E. Gelfand. The nested Dirichlet process. *Journal of the American Statistical Association*, 103(483):1131–1154, 2008.

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

S.L. Scott. Bayesian methods for hidden Markov models. *Journal of the American Statistical Association*, 97(457):337–351, 2002.

L. Spezia. Reversible jump and the label switching problem in hidden markov models. *Journal of Statistical Planning and Inference*, 139(7):2305 – 2315, 2009.

M. Stephens and P. Donnelly. Inference in molecular population genetics. *Journal of the Royal Statistical Society*, 62:605–655, 2000.

Y. W. Teh and M. I. Jordan. Hierarchical Bayesian nonparametric models with applications. In N. Hjort, C. Holmes, P. Müller, and S. Walker, editors, *Bayesian Nonparametrics*. Cambridge University Press, 2010.

Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

Y. W. Teh, H. Daume III, and D. M. Roy. Bayesian agglomerative clustering with coalescents. In *Advances in Neural Information Processing Systems*, volume 20, pages 1473–1480, 2008.

The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature*, 467:1061–1073, 2010.

The International HapMap Consortium. The international HapMap project. *Nature*, 426: 789–796, 2003.

F. Wood, C. Archambeau, J. Gasthaus, L. F. James, and Y. W. Teh. A stochastic memoizer for sequence data. In *Proceedings of the International Conference on Machine Learning*, volume 26, pages 1129–1136, 2009.

E. P. Xing and K. Sohn. Hidden Markov Dirichlet process: Modeling genetic recombination in open ancestral space. *Bayesian Analysis*, 2(2), 2007.