Bayesian Analysis of Isochores

Paul Fearnhead and Despina Vasileiou

Department of Mathematics and Statistics, Lancaster University, UK

Abstract

The statistical identification of isochore structure, the variation in large scale GC composition, of mammalian genomes is a necessary requirement for understanding both the evolution of base composition and the many genomic features such as mutation and recombination rates which covary with base composition. We have developed a Bayesian method for isochore analysis, which we demonstrate to be more accurate than the commonly used binary segmentation approach implemented within the program `IsoFinder`. The method accounts for both fine-scale and large-scale structure. We adapt direct simulation methods to allow for iid samples from the posterior distribution of our model, and provide an accurate approximation to this which can analyse data from a chromosome in a matter of seconds. We apply our method to human Chromosome 1. The resulting estimate of how GC content varies across this region is shown to be a better predictor of local recombination rates than `IsoFinder`; and we are able to detect regions consistent with the classical definition of isochores that cover 85% of the chromosome. We also show a measure of relative GC content to be particularly predictive of local recombination rates.

Keywords: Changepoint; GC Content; Direct Simulation; Hidden Markov Model; Recombination

1 Background

The genomes of many eukaryotes display striking large scale heterogeneity in base composition along their chromosomes. In particular, the distribution of G+C content (hereafter GC) along the chromosomes of mammals is highly variable (Bernardi 2000; Eyre-Walker and Hurst 2001). Initial analysis of these compositional patterns was performed using centrifugation experiments, which appeared to show discrete classes of different GC content. These results lead to the isochore model for the genomes of warm-blooded vertebrates: such genomes were thought to consist of a mosaic of isochores, defined as regions longer than 300 kilobases (kb) within which base composition is homogenous, and which belong to a number of distinct families differing greatly in GC abundance (Bernardi, 2000). The recent sequencing of many complete vertebrate genomes has led to the re-evaluation of the isochore model using in silico analyses (IHGSC, 2001). The precise meaning of the term 'isochore', particularly with regard to the notion of homogeneity within isochores implied by the prefix "iso', has been the subject of dispute. What is abundantly clear from plots of GC content across mammalian chromosomes (e.g. using UCSC Genome Browser http://genome.cse.ucsc.edu) is that there is considerable variation in base composition at large scales (hundreds of kb and above). It is the analysis of such large scale variation which we address in this study.

What are the reasons for analysing patterns of base composition? The most simple aim is descriptive: there is considerable large scale compositional

variation in mammalian genomes, but the signal is obscured by small scale noise, so we need computational tools to understand any underlying structure there may be. The second reason is for practical and predictive purposes. It appears that many features of the genome are correlated with GC content, such as gene density (Venter et al., 2001), repeat density (IHGSC, 2001), substitution rates (Hardison et al., 2003), and recombination rates (Kong et al., 2002). Much, although not all, of the covariation can be explained by GC variation (Hardison et al., 2003; Spencer et al., 2006). For example, if we are designing gene prediction methods, it is more likely that a high GC region will contain genes than a low GC region (Carpena et al., 2002), so utilising information about the isochore structure is worthwhile. Finally, there is the ultimate aim of trying to understand the evolution of genomes at the finest possible scale, that of the single nucleotide (Eyre-Walker and Hurst, 2001). What are the evolutionary forces which affect base composition? How and when were isochores created? Why do some genomes have isochores while others show far less compositional variation? To answer these sorts of questions we need to develop powerful statistical methods to infer underlying patterns of base composition and to test hypotheses concerning their evolution.

It seems clear that it is better to analyse patterns of base composition by using the underlying structure in GC variation, i.e. what we term isochores, rather than simply using windows based methods which are dependent on the choice of window size (Li et al., 2002). Current methods for the identification of isochores from sequence data can be divided into two main classes: highly efficient methods for dealing with genomic sequences many megabases (Mb) long (Nekruteno and Li, 2000; Oliver et al., 2004), and more sophisticated but slower methods that are usually applied to much smaller sequences (reviewed in Braun

and Muller, 1998). Methods for identifying isochores are segmentation methods: the aim is to divide a sequence into a number of regions, termed segments, which represent the hidden deterministic process which generates the observed compositional variation. Only at the boundaries between segments, termed changepoints, does the underlying process change, although stochastic effects generate additional variation throughout the sequence. A fast heuristic method for sequence analysis is recursive segmentation, in which binary segmentation is repeatedly applied to the data (Braun and Muller, 1998; Li et al., 2002). Binary segmentation involves choosing whether to cut a sequence into two sub-sequences and where the cut should be made. The general approach is to first find the changepoint which maximises some statistic measuring the difference in GC on either side of the changepoint. The cut is then made, provided that the evidence for segmentation is strong enough. If the sequence is cut, then binary segmentation is independently repeated on the two sub-sequences. So the recursive segmentation continues until no further changepoints can be found within any of the sub-sequences. The final set of sub-sequences becomes the identified segments. As an example, the program `IsoFinder` implement a binary segmentation for GC content (Bernaola-Galvan et al., 1996; Oliver et al., 2004).

We describe a Bayesian approach to inferring isochore structure. This approach has numerous advantages over the binary segmentation procedure implemented within `IsoFinder`: it jointly infers all changepoints, quantifies uncertainty in the underlying isochore structure, and averages over this uncertainty when producing estimates of GC content across the chromosome. It also appears to be more robust in terms of the inferred isochore structure, whereas relatively minor changes in the DNA sequence can cause comparatively large changes in the inferred isochore structure using binary segmentation.

The use of Bayesian methods for segmentation of genomic features is becoming increasingly popular. There are methods for segmentation of the DNA sequence (Liu and Lawrence, 1999; Boys et al., 2000) as well as methods for segmenting the genome based on other features (Salmenkivi et al., 2002; Fearnhead and Sherlock, 2006). Whilst our approach is based on segmenting the genome based on the DNA sequence, we focus solely on the large-scale features of the sequence. We first partition the chromosomal region of interest into 3kb windows (though the approach is robust to the choice of window size of the order of 3–5kb) and then summarise the data based on the GC content in these windows. This both filters out very fine-scale variation in GC content (such as due to CpG islands, which are $\approx$ 1kb regions of high GC content), and also leads to algorithms that are computationally more efficient, and scalable to the whole genome. Our approach is most similar to that described in Fearnhead and Liu (2007), but it is based on a more realistic model, which directly models isochore families (see Section 3) and allows for both fine and large scale structure though allowing for dependence in GC content across isochores. While there are algorithms that analyse either fine-scale variation in GC content (e.g. Fearnhead and Liu, 2007), or large-scale variation (e.g. Cohen et al., 2005; Constantini et al., 2006), our approach and that of `IsoFinder` are perhaps the only that try to infer both.

The outline of the paper is that we first give a more detailed description of the problem, followed by details of the model we use. We then describe our computational algorithm that enables us to generate samples from the posterior distribution of our model. In Section 5 we show how our method is substantially more accurate in estimating GC content than `IsoFinder`. We apply our method to analysing the GC content of human chromosome 1 in Section 6, and look at predicting local recombination rates from our estimates of the local GC content

of the region. Our paper concludes with a discussion.

2 Problem Description

The raw data consists of a single contiguous stretch of DNA data, which can be viewed as a long 'word' containing only four different 'letters': A, C, G and T. Our aim is to infer how the GC content (the proportion of letters that are G or C) varies across this contiguous region ('word'). As an example Figure 1 shows data from a 6.0 megabase (Mb) stretch of human chromosome 1. We have summarised the DNA data by partitioning the 6.0Mb region into 2000 windows, each of 3.0 kb long, and for each window plotting the proportion of letters within that window that are G or C. Overlaid on the data in Figure 1 we show the inferred isochore structure calculated by the `IsoFinder` computer program: a series of segments of homogeneous (constant mean) GC content. Throughout this paper we refer to each segment as an 'isochore', though the classical definition of isochores is usually restricted to segments whose length is of the order of 300kb or longer.

The program we used to calculate the isochore structure of the region of chromosome 1 in Figure 1 is currently perhaps the most popular program for inferring isochore structure. For example the Isochore structure detected by `IsoFinder` is displayed on the human genome web browser (`http://genome.cse.ucsc.edu`). `IsoFinder` uses a binary segmentation approach as described in the introduction.

Whilst simple and quick to implement, there are several disadvantages to `IsoFinder`. Firstly the user needs to specify a p-value to be used within the segmentation approach, which governs the amount of evidence required for a

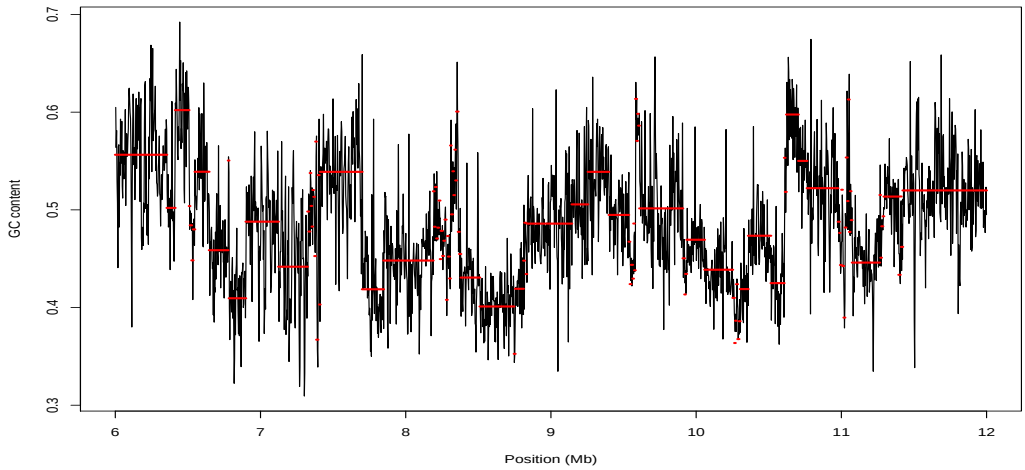

Figure 1: Plot of GC content in 3kb windows for 6Mb of human chromosome 1 (black); and `IsoFinder` segmentation (red). Data comes from positions 6Mb–12Mb of build hg17.

new breakpoint to be introduced. The results in Figure 1 were obtained with a p-value of 0.95 and give an inferred structure which contains 116 isochores. Changing the p-value to 0.99 would have produced a different structure with fewer isochores. Secondly, the method produces a single estimate of the Isochore structure, and we get no measure of the underlying uncertainty with the position of the breakpoints in this structure. The number and position of changepoints it infers can vary quite noticeably even with only minor changes to the underlying DNA sequence (see Section 5). Finally, there is also evidence that binary segmentation procedures are inferior to methods that jointly infer all changepoints (Braun et al., 2000).

Here we propose a Bayesian, model-based approach for inferring isochores. We apply our method to DNA data which is summarised by the proportion of GC content in a series of consecutive, non-overlapping windows (such as the data

presented in Figure 1). This filters out the fine-scale structure in the DNA sequence, and by summarising the data in this way we can produce a method that can scale to analysing genomewide data. The former is important in terms of the aim of analysing structure in GC content, and `IsoFinder` also filters out the structure at the 3kb–5kb level.

3 Model

Before describing the model that we chose for detecting Isochore structure, we first describe the results of some preliminary analysis and some prior knowledge of the GC structure that guided our choice of model. Throughout, we assume our data is described by $y_1, y_2, \ldots, y_n$, the average GC content in consecutive windows.

Both for simplicity and because it is known to capture the main large-scale structure in the data, we are going to assume a piecewise constant model for the underlying isochore structure. Thus a single realisation of this process will look like the output of `IsoFinder` as shown in Figure 1. This realisation can be described by changepoints and mean levels.

To specify a model we will need to describe (i) the marginal distribution of the mean level, μ, for each Isochore; (ii) the dependence structure in these mean levels across Isochores; (iii) the joint distribution of the number and position of the changepoints; (iv) the distribution of the data conditional on the Isochore structure.

It has been suggested by Bernardi (2000), that the isochore organisation of the human genome can be partitioned into four families, and the GC content for

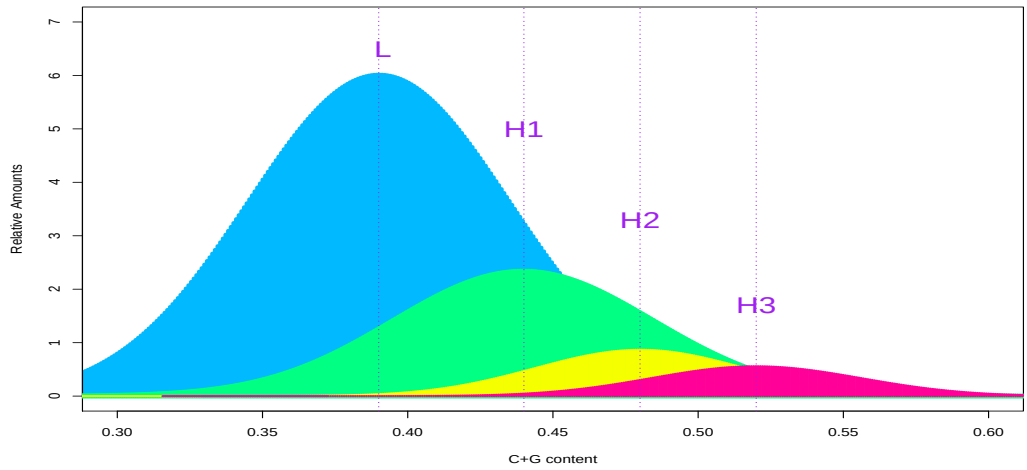

Figure 2: The Isochore Organisation into four Families according to the GC content level, one GC poor the L ( ) and three GC rich, namely the H1 ( ), H2 ( ), H3 ( )

isochores within each family can be described by a normal distribution, as shown in Figure 2 (alternatively five family models have been suggested, e.g. Constantini et al., 2006). Therefore this motivates fitting a normal mixture model for the marginal distribution of mean GC content of the isochores, with each component associated with a different family.

While previous models have assumed independence of the mean GC content of different isochores (e.g. Fearnhead and Liu, 2007), this assumption is unrealistic. Smith and Lercher (2002) documents the long range correlation that exists in the human genome, and Bernaola-Galvan et al. (2002) showed that there is structure inside long relatively homogeneous regions. We allow for dependence through a hidden Markov model, where the hidden state relates to the family of the isochore. This enables us to capture some degree of dependency, whilst remaining within a computationally tractable framework.

We model the joint distribution of the number and position of changepoints through a probabilistic model for the length of each isochore. For the results that are presented, we chose an geometric distribution for the isochore length – as the results from `IsoFinder` has a range of very short to long lengths in the isochore structure that it finds (see Figure 1), though the computational methodology below can be extended to any distribution for this length.

Finally, analysing the residuals of the chromosome 1 dataset after we run `IsoFinder`, suggests that conditional on the isochore mean, the data is, at least approximately, normally distributied and is close to being independent (autocorrelation at lag 1 is 0.25; and no significant autocorrelation at lags > 2.).

3.1 Mathematical Description of Model

The specific model we use has a hierarchical structure. Firstly, denote the number of isochores by m, and the changepoint positions by $\tau_0 < \tau_1 < \cdots < \tau_m$ with $\tau_0 = 0$, $\tau_m = n$ and the ith isochore containing observations $y_{\tau_{i-1}+1:\tau_i} = (y_{\tau_{i-1}+1}, y_{\tau_{i-1}+2}, \ldots, y_{\tau_i})$. For each isochore we associate an isochore family. We allow for K different isochore families, and let $\zeta_i \in \{1, \ldots, K\}$ denote the family of the ith isochore.

The distribution of m and the position of the changepoints is specified through a model for the isochore lengths. We let this distribution depend on the family of the isochore. We assume a Markov model for the isochore families. Thus the joint probability of the number and position of the changepoints, and the families of the isochores can be factorised as

$$\Pr(m, \tau_1, \ldots, \tau_{m-1}, \zeta_1, \ldots, \zeta_m) = \Pr(\zeta_1) \left(\prod_{i=1}^{m-1} \Pr(\tau_i | \tau_{i-1}, \zeta_i) \Pr(\zeta_{i+1} | \zeta_i) \right) Pr(\tau_m | \tau_{m-1}, \zeta_m).$$

We then assume that the distribution of the length of an isochore has a

geometric distribution, with mean $1/\lambda_k$ for an isochore in family k:

$$\Pr(\tau_i = s|\tau_{i-1} = t-1, \zeta_i = k) = \lambda_k(1-\lambda_k)^{t-s},$$

for $s = t, t+1, \ldots, n-1$. The probability of no further changepoints (i.e. $s = n$), conditional on an isochore in family k, is $(1-\lambda_k)^{n-t}$. We further introduce a $K \times K$ transition matrix P so that $\Pr(\zeta_{i+1} = l|\zeta_i = k) = P_{kl}$. The distribution $\Pr(\zeta_1)$ is defined to be the stationary distribution of P.

Conditional on $\zeta_i = k$, the isochore family of the ith isochore, the mean and variance of the GC content of that isochore is drawn from the standard conjugate normal-variance prior, with

$$\sigma_j^2 \sim IG(\nu/2, \gamma/2), \text{ and } \mu_j|\sigma_j^2 \sim N(\xi_k, \sigma_j^2/\delta_k),$$

where $IG(\cdot,\cdot)$ denotes the inverse gamma distribution and $N(\cdot,\cdot)$ the normal distribution.

Finally, conditional on the changepoints, the hidden states of the Markov chain and parameters associated with the segments, the observations are considered to be independent, and normally distributed. For observation y_j that belongs to the ith isochore:

$$y_j \sim N(\mu_i, \sigma_i^2), \quad \text{where} \quad j \in \{\tau_{i-1}+1, \ldots, \tau_i\}.$$

This model has an important feature which makes it computational tractabile. If we condition on the hyperparameters of the model, then given a changepoint at time s and the isochore family of the isochore starting at time $s+1$, the data $y_{1:s}$ and $y_{s+1:n}$ are independent of one another.

4 Bayesian Inference

We perform Bayesian inference for this model, conditional on the values of hyper-parameters: K, the number of isochore families; λ_k for $k = 1, \ldots, K$ for the distribution of the isochore length; P, the transition matrix of the isochore families; ξ_k, and δ_k for $k = 1, \ldots, K$ for the prior distribution of the isochore means; and ν and γ for the prior distribution of the segment variances. Conditional on these values we can obtain iid samples from the joint distribution of the number and position of changepoints, the family of each isochore, and the GC content of each isochore using an algorithm related to the Forward-Backward algorithm Baum et al. (1970). The algorithm we present here is a new extension of previous algorithms for changepoint models (Yao, 1984; Barry and Hartigan, 1992, 1993; Liu and Lawrence, 1999; Fearnhead, 2005; Lai et al., 2005; Fearnhead, 2006) to allow for the HMM component of the model. In particular we adapt the method of Fearnhead and Liu (2007) due to its computational efficiency (see Section 4.2). Details of how we choose the hyperparameters is given in Section 4.3

4.1 Exact Inference

Our choice of conjugate priors means that we can integrate out the parameters associated with a given isochore (conditional on the isochore family). This means that given a sample from the joint distribution of the number and position of the changepoints, and the family of each isochore, it is straightforward to sample the isochore means. We thus focus on how to simulate from the posterior distribution of the changepoint positions and isochore families.

Firstly define the marginal likelihood of observations $y_{t:s}$ conditional on them

belonging to a single isochore, and the isochore belongs to family k by $R(t, s, k)$. We allow for missing data in one or more windows. For notational simplicity define $y_i = 0$ if there is no observation for window i, so that $\sum_{i=t}^{s} y_i$ is the sum of observations for windows t to s inclusive; and let n_t be the cumulative sum of observations up to an including window t (so $n_t = t$ if there is no missing data). Then the marginal likelihood is

$$\begin{aligned} R(t,s,k) &= \int_{\mu}\int_{\sigma^2}\left(\prod_{i=t}^{s} f(y_i|\mu,\sigma)\right) p(\mu|\sigma,\zeta=k)p(\sigma)\ \mathrm{d}\sigma\ \mathrm{d}\mu. \\ &= \frac{\pi^{-(n_s-n_t+1)/2}\ \ \gamma^{\nu/2}}{\left(\gamma\ \ +\sum_{i=t}^{s} y_i^2+\xi_k^2\delta_k-\frac{\sum_{i=t}^{s} y_i\ +\xi_k^2\delta_k}{n_s-n_t+1+\delta_k}\right)^{(n_s-n_t+1+\nu)/2}}\frac{\Gamma\big((n_s-n_t+1+\nu)/2\big)}{\Gamma(\nu/2)}. \end{aligned} \tag{1}$$

We introduce a 2-dimensional state at time t, (C_t, Z_t), where C_t is defined as the time of the most recent changepoint prior to time t, and Z_t is the family of the current isochore at time t. Under our model (C_t, Z_t) is a Markov chain with transition probabilities:

$$\Pr(C_{t+1}=j, Z_{t+1}=l|C_t=i, Z_t=k) = \begin{cases} (1-\lambda_k) & \text{if } j=i \text{ and } l=k, \\ \lambda_k P_{kl} & \text{if } j=t, \end{cases}$$

with all other transitions having zero probability.

We can then write down recursions for the filtering probabilities $\Pr(C_t, Z_t|y_{1:t})$ (See Appendix A for derivation). For $j < t$ we have

$$\Pr(C_{t+1}=j, Z_{t+1}=l|y_{1:t+1}) \propto \frac{R(j+1,t+1,l)}{R(j+1,t,l)}\Pr(C_t=j, Z_t=l|y_{1:t})(1-\lambda_k), \tag{2}$$

while

$$\Pr(C_{t+1}=t, Z_{t+1}=l|y_{1:t+1}) \propto R(t+1,t+1,l)\sum_{i=0}^{t-1}\sum_{k=1}^{K}\Pr(C_t=i, Z_t=k|y_{1:t})\lambda_k P_{kl}. \tag{3}$$

The recursions are initalised with $P(C_1 = 0, Z_{t+1} = l|y_1) = \Pr(\zeta_1 = l)R(1, 1, l)$ for $l = 1, ldots, K$. Note that the normalising constant of these equations is $\Pr(y_{t+1}|y_{1:t})$ and thus a by-product of solving them is that we can calculate the likelihood $\Pr(y_{1:n})$ (see Fearnhead, 2008).

These recursions can be solved for $t = 1, \ldots, n$. Once calculated, the last changepoint and isochore family can be simulated directly from $\Pr(C_n, Z_n|y_{1:n})$. Then if we condition on a changepoint at t with isochore family l ($C_{t+1} = t$, $Z_{t+1} = l$) we have that

$$\Pr(C_t = i, Z_t = k|y_{1:n}, C_{t+1} = t, Z_{t+1} = l) \propto \Pr(C_t = i, Z_t = k|y_{1:t})\lambda_k P_{kl},$$

for $i = 0, \ldots, t-1$ and $k = 1, \ldots, K$. Simulating from this distribution gives us both the family and the position of the beginning of the isochore that ends at t. Thus we can recursively simulate the changepoints and isochores backwards in time.

4.2 A Computationally Efficient Algorithm

The above algorithm enables iid samples to be drawn from the posterior distribution of the number and position of the changepoints and the families of the isochores. Once these have been sampled, it is trivial to sample the parameters (mean GC content, and variance) for each isochore. However, the algorithm suffers from the disadvantage that its computational and storage costs is quadratic in the number of observations, n.

To have an algorithm that scales linearly with n, and can be applied to data from whole chromosomes, we applied resampling ideas from Fearnhead and Liu (2007). Their idea is to approximate $\Pr(C_t, Z_t|y_{1:t})$ by a discrete distribution with fewer than tK support points: by stochastically removing support points

that have small probability. We used their Stratified Rejection Control algorithm (see also Liu et al., 1998), which was shown to perform well both theoretically and in simulation studies. The basic idea is as follows. Assume that we have a discrete distribution with N support points ($c_t^{(i)}, z_t^{(i)}$ and associated probabilities $w_t^{(i)}$ (for $i = 1, \ldots, N$) that approximate $\Pr(C_t, Z_t|y_{1:t})$. We can produce an approximation with fewer support points using the following resampling algorithm:

(i) Choose an arbitrary cut-off α. Order the support points so that if $i < j$ then either $c_t^{(i)} < c_t^{(j)}$ or $c_t^{(i)} = c_t^{(j)}$ and $z_t^{(i)} < z_t^{(j)}$.

(ii) Simulate u a realisation of a uniform random variable on $[0, \alpha]$. Set $i = 1$.

(iii) If $w_t^{(i)} \geq \alpha$ goto (iv); else let $u = u - w_t^{(i)}$. If $u \leq 0$ then let $u = u + \alpha$ and set $w_t^{(i)} = \alpha$, otherwise set $w_t^{(i)} = 0$.

(iv) Let $i = i + 1$; if $i < N$ return to (ii); otherwise remove all support points for which $w_t^{(i)} = 0$, and renormalise the probabilities of the remaining support points.

The idea is that support points with probability, $w_t^{(i)}$ less than α are probabilistically removed. The probability of removing a support point is $w_t^{(i)}/\alpha$. Those support points that are kept have their probabilities increased in step (iii) so that the algorithm is unbiased (before normalisation, expected probability of a support point after resampling is equal to its probability before). The ordering in step (i) ensures that the support points removed are evenly spread over the support of $\Pr(C_t, Z_t|y_{1:t})$. Note that α governs the trade-off between smaller approximation (smaller α) and speed (larger α).

In practice we have found $\alpha = 10^{-6}$ introduces negligible error, but can greatly

increase the speed of the overall algorithm. In the application in Section 6, the resulting algorithm approximated the filtering densities, $\Pr(C_t, Z_t|y_{1:t})$, by distributions with an average of around 200 support points. The true distributions had an average of 80,000 support points, so this led to a 400-fold reduction in CPU and memory cost.

4.3 Choosing the Hyper-parameters

The above exact simulation method requires the specification of the hyper-parameters. Based on Figure 2, in our analysis below we specify the number of isochore families to be 4, and the mean GC content to be $(\xi_1, \xi_2, \xi_3, \xi_4) = (0.39, 0.44, 0.48, 0.53)$. The other hyper-parameters were estimated via maximum likelihood using a Monte Carlo EM algorithm. Details of this are given in Appendix B.

5 Comparison with `IsoFinder`

We compared our new Bayesian approach with that of `IsoFinder` on a series of simulated data sets. In order to simulate data that captures the structure we observe in real data, we based all simulated data sets around the inferred isochore structure in the region of chromosome 1 that is shown in Figure 1. This inferred structure was taken to be the underlying truth that we wish to estimate.

Our three simulated data sets differed in how the observations relate to the underlying isochore structure:

(A) Observations are independent and normally distributed with common variance; isochore mean given by the sample mean of the observations

within that isochore.

(B) Same as (A) except that we introduce extra changepoints within the longer isochores (greater than 90kb), and recalculate the mean of the observations based on these. Extra changepoints were added approximately every 60kb.

(C) Same marginal distribution as (A), but we introduce dependence into the observations through an AR(1) model for the observation, with 1-lag correlation of 0.25.

The idea of (B) and (C) is to introduce extra structure in the observation process within each isochore; (C) allows for dependencies that are greater than those inferred in the true data (see Section 3).

We compare the results of our method and `IsoFinder` in terms of estimating the underlying GC content. For our method we estimate the GC content for any 3kb window via the posterior mean of the GC content for that window. We measure the accuracy of a method by averaging the square error of its estimated GC content from the true GC content across the 2000 3kb windows.

To run `IsoFinder` on the simulated data, we had to simulate sequence data. We did this by simulating the order of nucleotides for each window at random subject to the constraint on the number of GC nucleotides for that window. Whilst this approach does not adequately reflect the true fine-scale structure in DNA sequences, this should not affect `IsoFinder` much as it filters out fine-scale structure at less than the 3kb level. However we did run `IsoFinder` on two simulated sequences for each data set, and found a noticeable difference in the segmentation and hence the performance of the method, which suggests the segmentation approach is relatively sensitive to minor changes in the DNA sequence being analysed. The reason for this is that `IsoFinder` gives a single

Data	Bayesian Approach	`IsoFinder` $p = 0.95$		`IsoFinder` $p = 0.99$	
	MSE	MSE	MSE	MSE	MSE
(A)	1.7×10^{-4}	3.1×10^{-4}	3.0×10^{-4}	3.6×10^{-4}	3.4×10^{-4}
(B)	2.0×10^{-4}	2.5×10^{-4}	2.7×10^{-4}	3.4×10^{-4}	3.7×10^{-4}
(C)	2.6×10^{-4}	4.5×10^{-4}	4.1×10^{-4}	3.7×10^{-4}	3.7×10^{-4}

Table 1: Results of our method (Bayesian Approach) and `IsoFinder`, for two different significance levels, at inferring GC structure for 3 different simulation scenarios (see text for details). We give mean square error (MSE) for our method, and MSE for two sequences (simulated for same GC content per 3kb window) for each scenario for the `IsoFinder` analyses.

segmentation of the data. As there are often a range of segmentations that are plausible for a given data-set, minor changes can mean that `IsoFinder` jumps between two relatively dissimilar segmentations. One of advantage of the Bayesian approach we take is that we average over possible segmentations, and thus our algorithm is much more robust to minor changes in the sequence data (which for the Bayesian approach correspond to small changes in the GC content of each window).

Our results are given in Table 1. Our method is substantially more accurate at inferring GC content that `IsoFinder`, regardless of whether `IsoFinder` is run with a significance value of $p = 0.95$ or $p = 0.99$. Mean square error is reduced by between 40% and 50% for data set (A); between 20% and 40% for (B); and between 30% and 40% for (C).

6 Analysis of Chromosome 1 DNA sequence

We applied our method to data from human Chromosome 1, with the DNA sequence taken from build hg17 (available from `http://hgdownload.cse.ucsc.edu/goldenPath/hg17`). There was substantial missing data around the centromere of the chromosome, and so we analysed the two arms of the chromosome separately: the first 120,408kb; and positions 146,328kb–245,217kb. In total this accounts for over 219MB of sequence (approx 7% of the human genome), with over 73,000 3kb windows. We counted as missing data any window for which the complete DNA sequence was not known, and this resulted in 431 missing data points.

Analysis of the data for a given set of hyper-parameters took substantially less than a minute on a desktop PC. Convergence of the EM algorithm was achieved within around 30 iterations. Figure 3 shows the estimates of mean GC content. The estimated parameters suggest that there is less heterogeneity within the L isochore family (mean number of windows between changepoints is 18 for the L family, and 5 for each of H1–H3). For all isochore families we expect between 15 and 25 consecutive isochores from that family. For each family, the modal transition is to a neighbouring isochore family.

For comparison we also analysed the data using the method of Fearnhead and Liu (2007), which corresponds to the special case of $K = 1$ in our model (i.e. it ignores isochore families). We compared the models based on the change in log-likelihood. There was an increase of 1,530 for the $K = 4$ model; while the number of estimated hyper-parameters increased by only 16. Thus using either information criteria such as AIC or BIC, or a chi-squared test based on the likelihood ratio statistic, there is over-whelming evidence in favour of choosing

the model with $K = 4$.

6.1 Detecting Classical Isochores

One ongoing question is to what extent "classical isochores" exist within the human chromosome, and whether they are detectable. Cohen et al. (2005) define a classical isochore as a region of a chromosome which is (i) longer than 300kb; (ii) is more homogeneous in its composition than the chromosome on which it resides; and (iii) can be classified into an isochore family . They suggest testing (ii) for an isochore based on whether there is significant evidence for the variance in GC content within a segment is smaller than the variance in GC content across the whole chromosome. Results based on the method of Cohen et al. (2005) suggest that only 41% of the human genome lies within such classical isochores.

We considered whether our analysis enables us to detect classical isochores more effectively. Our idea was to use the posterior distribution on the isochore family for each window. We first calculated the modal isochore family for each window, and then partitioned the chromosome into contiguous regions of the same modal family. For each contiguous region we then tested whether they satisfied conditions (i) and (ii) above (by definition they trivially satisfy (iii)). The regions that satisfied these conditions are shown in Figure 3. In total these covered 85% of the chromosome.

By comparison, applying the same method to the isochores detected by `IsoFinder`, gave only 60% coverage of the chromosome. Our results are comparable with those of Constantini et al. (2006), who also found classical isochores that cover 85% of the genome, though they were using a somewhat

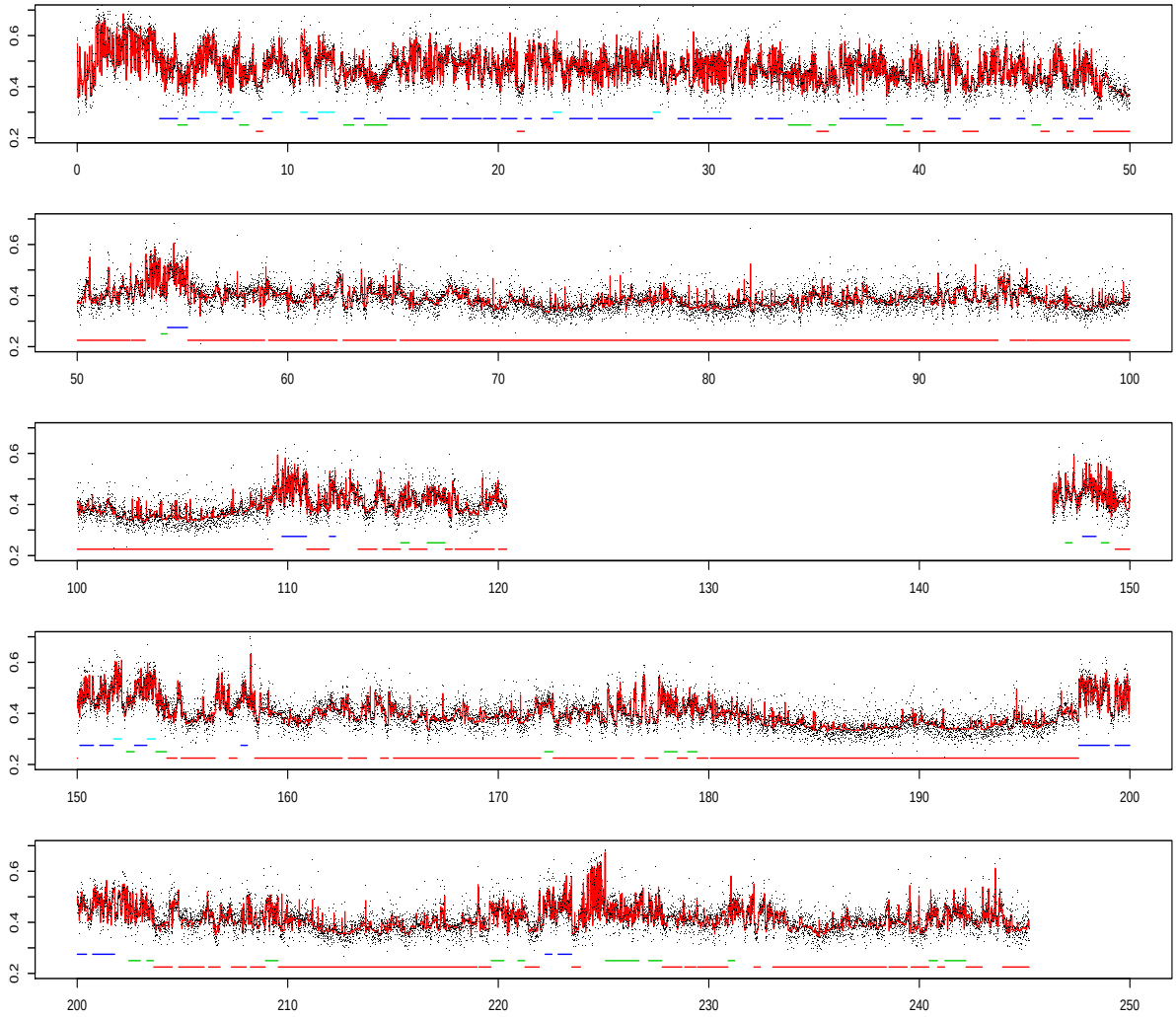

Figure 3: Results of our model on the Human Chromosome 1 dataset: the raw data is given by black dots; the inferred GC content is given by the red line. At the bottom of the plots we show the regions of contiguous modal isochore family that are consistent with the features of classical isochores (different levels of the lines correspond to different families; from highest to lowest: H3, H2, H1 and L.

adhoc algorithm specifically designed to detect such classical isochores.

6.2 Fine-Scale Correlation with Recombination Rates

One motivation for inferring GC content is to look for correlation with other genomic features. Galtiera et al. (2001) have established a causal relationship between recombination and GC content by examining the variation on the levels of polymorphism in the genome, and we decided to investigate this relationship on chromosome 1. We downloaded the fine-scale recombination map (Myers et al., 2005) from phase II of the HapMap project (available from `http://www.hapmap.org/`) and analysed the correlation between these estimates of recombination (for each 3kb window) and GC content. Our rationale is that a better method at estimating the local variation in GC content is likely to have greater correlation with local recombination rate.

We first log-transformed the recombination rate estimates, so that their marginal distribution was close to normal. We then calculated the correlation between log-recombination and GC content as inferred by (i) GC content within the 3kb window; (ii) `IsoFinder`; and (iii) our method. The resulting correlations were (i) 0.30, (ii) 0.30, and (iii) 0.32. An alternative approach is to infer GC content using a larger region centered on each window. If we implement this, and choose the optimal size of region (27kb), we obtain correlation that is almost identical to that obtained by our method (0.32).

To gain a greater insight into the correlation between GC content and recombination rate, we repeated the analysis for overlapping 5Mb windows across chromosome 1 (see Figure 4). The correlation of recombination rate with the estimates from our method is greater than that with the estimate of

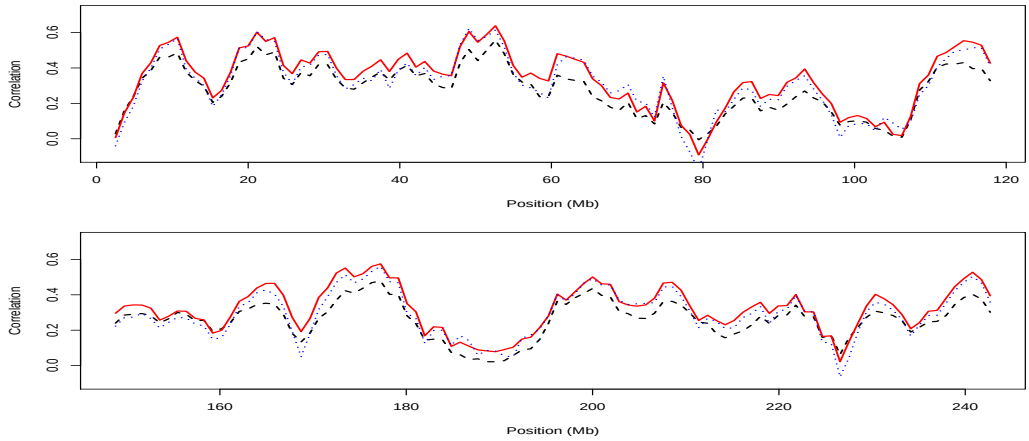

Figure 4: Correlation of GC content with log recombination rate for 5Mb windows of chromosome 1. The lines correspond to 3 different measurements of GC content: raw 3kb data (black dashed-line); `IsoFinder` estimate (blue, dotted-line); our estimate (red, full-line).

`IsoFinder` in over 80% of the windows (and greater than correlation with raw GC in 95% of the windows). However more striking is the fact that correlation between GC content and recombination rate varies considerably across the chromosome, from < 0 to > 0.6.

One reason for this is that the auto-correlation structure in GC content extends much further than auto-correlation of recombination rates. Thus a relative measure of GC content, which considers the difference between the local estimate of GC content and the average GC content over a Mb scale may be a better predictor of local recombination rate. For a 3kb window let r_i denote the log recombination-rate, G_i the inferred GC content from our method, and $\bar{G}_i$ the average GC content for a 1Mb region centered on the region. Linear regression suggests the predictor $G_i - 0.65\bar{G}_i$ for r_i. This predictor has correlation 0.36 with recombination rate.

7 Discussion

We have presented a novel Bayesian method for inferring isochore structure from DNA sequence data. Both our simulation results, and the correlation analysis of the chromosome 1 data suggest that it is a more accurate approach than the binary segmentation procedure used by the program `IsoFinder`. We have derived a new direct simulation algorithm that enables iid draws from the posterior distribution, and shown how using resampling ideas we can implement an approximate algorithm that can analysis data from a whole chromosome in a matter of seconds.

Inference for the hyper-parameters of our model are possible using an EM algorithm. For our analysis we fixed the number of isochore families, K, and their mean GC content based on prior information (Bernardi, 2000). It would have been possible to perform inference for the mean GC content of the families within the EM algorithm, though maximisation becomes more difficult due to multiple local maximum. Inference for the number of families is also possible using our algorithm, as the algorithm calculates the marginal likelihood of the data, which can then be used to compare models with different values of K.

An important feature of our model is that it captures both fine-scale variation in GC content, which can be used to look at correlation of GC content with other features; and also large-scale variation which helps define regions of the chromosome that fit within the classical idea of isochores. Our analysis of GC content on chromosome 1 demonstrated the accuracy of both the inferences for fine-scale and large-scale variation in GC content. Furthermore we showed that relative measures of GC content may be the most predictive of local recombination-rates; something that has not been considered previously (see e.g.

Spencer et al., 2006).

The program implementing our method is available from the first author.

Acknowledgements: This work was supported by Engineering and Physical Sciences Research Council grant GR/T19698/01 to the first author, and a Lancaster University 40th Anniversary PhD studentship to the second author. We thank Danny Wilson for helpful comments. We dedicate the paper to the memory of Nick Smith, who first motivated us to look at this problem.

Appendix A: Derivation of Recursions (2) and (3)

By Bayes theorem

$$\Pr(C_{t+1}=j, Z_{t+1}=l|y_{1:t+1}) \propto \Pr(y_{t+1}|y_{1:t}, C_{t+1}=j, Z_{t+1}=l)\Pr(C_{t+1}=j, Z_{t+1}=l|y_{1:t})$$

For recursion (2) we use the fact that if $j < t$

$$\Pr(C_{t+1}=j, Z_{t+1}=l|y_{1:t}) = \Pr(C_t=j, Z_t=l|y_{1:t})(1-\lambda_k).$$

Furthermore by the conditional independence property of the model

$$\begin{aligned}\Pr(y_{t+1}|y_{1:t}, C_{t+1}=j, Z_{t+1}=l) &= \Pr(y_{t+1}|y_{j+1:t}, C_{t+1}=j, Z_{t+1}=l) \\ &= \frac{\Pr(y_{j+1:t+1}|C_{t+1}=j, Z_{t+1}=l)}{\Pr(y_{j+1:t}|C_{t+1}=j, Z_{t+1}=l)} \\ &= \frac{R(j+1,t+1,l)}{R(j+1,t,l)}.\end{aligned}$$

A similar derivation applies for (3), but here

$$\Pr(C_{t+1}=t, Z_{t+1}=l|y_{1:t}) = \sum_{i=0}^{t-1}\sum_{k=1}^{K}\Pr(C_t=i, Z_t=k|y_{1:t})\lambda_k P_{kl}.$$

Appendix B: Monte Carlo EM Algorithm

Within the EM algorithm, the full data consists of the number and position of the changepoints, the family of each isochore, and the mean GC content and

variance associated with each isochore. From this we define summary statistics: m_k the number of complete isochores of family k (i.e. excluding the final isochore); l_k the number of windows contained in isochores of family k; n_{ij} the number of transitions from family i to family j. Let $\mathcal{S}_k$ denote the set of isochores from family k and μ_i and $\beta_i = 1/\sigma_i^2$ be the mean and precision of isochore i. Let $m = 1 + \sum_{k=1}^{K} m_k$ denote the number of isochores as before.

For analytical simplicity we drop the contribution to the likelihood from the family of the first isochore. Due to the length of the data sets analysed, the effect of this is negligible. The resulting full-data log-likelihood is:

$$\sum_{i=1}^{K}\sum_{j=1}^{K} n_{ij}\log(P_{ij}) + \sum_{k=1}^{K}(m_k\log(\lambda_k) + (l_k - m_k)\log(1-\lambda_k)) + \sum_{i=1}^{K}\left(-n_k\log(\delta_k)/2 - \frac{\delta_k}{2}\sum_{j\in\mathcal{S}_k}\beta_j(\xi_k-\mu_j)^2\right) + \nu\sum_{j=1}^{m}\log(\beta_j)/2 + m\nu\log\gamma - \gamma\sum_{j=1}^{m}\beta_j/2 - m\log(\Gamma(\nu/2)).$$

The EM algorithm proceeds by taking expectation of the log-likelihood (with respect to their conditional distribution given the current values of the hyper-parameters). Within our Monte Calo EM algorithm we estimate this expectation using the simulated realisations from the current posterior distribution. Thus for example if we denote the Monte Carlo estimates of the expectations of n_{ij} for $i, j = 1, \ldots, K$ by $\hat{n}_{ij}$, then our new estimates of P_{ij} are $\hat{n}_{ij}/(\sum_{k=1}^{K}\hat{n}_{ik})$. Updates for all hyper-parameters except ν can be performed analytically; we maximise ν numerically.

References

Barry, D. and Hartigan, J. A. (1992). Product partition models for change point problems. *The Annals of Statistics*, 20:260–279.

Barry, D. and Hartigan, J. A. (1993). A Bayesian analysis for change point problems. *Journal of the American Statistical Society*, 88:309–319.

Baum, L. E., Petrie, T., Soules, G., and Weiss, N. (1970). A maximisation technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Ann. Math. Stats.*, 41:164–171.

Bernaola-Galvan, P., Carpena, P., Román-Roldán, R., and Oliver, J. L. (2002). Study of statistical correlations in DNA. *Gene*, 300:105–115.

Bernaola-Galvan, P., Roman-Roldan, R., and Oliver, J. L. (1996). Compositional segmentation and long-range fractal correlations in dna sequences. *Physical Review E*, 53:5181–5189.

Bernardi, G. (2000). Isochores and evolutionary genomics of vertebrates. *Gene*, 241:3–17.

Boys, R. J., Henderson, D. A., and Wilkinson, D. J. (2000). Detecting homogeneous segments in DNA sequences by using hidden Markov models. *Journal of the Royal Statistical Society, Series C*, 49:269–285.

Braun, J. V., Braun, R. K., and Muller, H. G. (2000). Multiple changepoint fitting via quasilikelihood, with application to DNA sequence segmentation. *Biometrika*, 87:301–314.

Braun, J. V. and Muller, H. G. (1998). Statistical methods for DNA sequence segmentation. *Statistical Science*, 13:142–162.

Carpena, P., Bernaola-Galvan, P., Roman-Roldan, R., and Oliver, J. L. (2002). A simple and species-independent coding measure. *Gene*, 300:97–104.

Cohen, N., Dagan, T., Stone, L., and Graur, D. (2005). GC composition of the human genome: In search of isochores. *Molecular Biology and Evolution*, 22:1260–1272.

Constantini, M., Clay, O., Auletta, F., and Bernardi, G. (2006). An isochore map of human chromosomes. *Genome Research*, 16:536–541.

Eyre-Walker, A. and Hurst, L. D. (2001). The evolution of isochores. *Nature Review Genetics*, 2:549–555.

Fearnhead, P. (2005). Exact Bayesian curve fitting and signal segmentation. *IEEE Transactions on Signal Processing*, 53:2160–2166.

Fearnhead, P. (2006). Exact and efficient inference for multiple changepoint problems. *Statistics and Computing*, 16:203–213.

Fearnhead, P. (2008). Computational methods for complex stochastic systems: A review of some alternatives to MCMC. *To appear in Statistics and Computing*.

Fearnhead, P. and Liu, Z. (2007). Online inference for multiple changepoint problems. *Journal of the Royal Statistical Society Series B*, 69:589–605.

Fearnhead, P. and Sherlock, C. (2006). Bayesian analysis of Markov modulated Poisson processes. *Journal of the Royal Statistical Society, Series B*, 68:767–784.

Galtiera, N., Piganeaub, G., Mouchiroudb, D., and L., D. (2001). GC-content evolution in mammalian genomes: The biased gene conversion hypothesis. *Genetics*, 159:907–911.

Hardison, R. C., Roskin, K. M., Yanf, S., Diekhans, M., Kent, W. J., Weber, R., Elnitski, L., and Li et al., J. (2003). Covariation in frequencies of substitution, deletion, transposition, and recombination during eutherian evolution. *Genome Research*, 13:13–26.

IHGSC (2001). Initial sequencing and analysis of the human genome. *Nature*, 409:860–921.

Kong, A., Gubdjartsson, D. F., Sainz, J., Jonsdottir, G. M., Gudjonson, S. A., Richardsson, B., Sigurdardottir, S., Barnard, J., Hallbeck, B., Masson, G., Shlien, A., Palsson, S. T., Frigge, M. L., Thorgeirsson, T. E., Gulcher, J. R., and Stefansson, K. (2002). A high-resolution recombination map of the human genome. *Nature Genetics*, 31:241–247.

Lai, T. L., Liu, H., and Xing, H. (2005). Autoregressive models with piecewise constant volatility and regression parameters. *Statistica Sinica*, 15:279–301.

Li, W. T., Bernaola-Galvan, P., Haghighi, F., and Grosse, I. (2002). Applications of recursive segmentation to the analysis of DNA sequences. *Gene*, 276:57–72.

Liu, J. S., Chen, R., and Wong, W. H. (1998). Rejection control and sequential importance sampling. *Journal of the American Statistical Society*, 93:1022–1031.

Liu, J. S. and Lawrence, C. E. (1999). Bayesian inference on biopolymer models. *Bioinformatics*, 15:38–52.

Myers, S., Bottolo, L., Freeman, C., McVean, G. A. T., and Donnelly, P. (2005). A fine-scale map of recombination rates and hotspots across the human genome. *Science*, 310:321–324.

Nekruteno, A. and Li, W. H. (2000). Assessment of compositional heterogeneity within and between eukaryotic genomes. *Genome Research*, 10:1986–1995.

Oliver, J. L., Carpena, P., Hackenberg, M., and Bernaola-Galvan, P. (2004). IsoFinder: computational prediction of isochores in genome sequences. *Nucleic Acids Research*, 32:W287–W292.

Salmenkivi, M., Kere, J., and Mannila, H. (2002). Genome segmentation using piecewise constant intensity models and reversible jump MCMC. *Bioinformatics*, 18:S211–S218.

Smith, N. G. and Lercher, M. J. (2002). Regional similarities in polymorphism in the human genome extended for many megabases. *Trends in Genetics*, 18(281–283).

Spencer, C., Deloukas, P., Hunt, S., Mullikin, J., Myers, S., Silverman, B., Donnelly, P., Bentley, D., and McVean, G. (2006). The influence of recombination on human genetic diversity. *Plos Genetics*, 2:e148.

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., and Sutton et al., G. G. (2001). The sequence of the human genome. *Science*, 291:1304–1351.

Yao, Y. (1984). Estimation of a noisy discrete-time step function: Bayes and empirical Bayes approaches. *The Annals of Statistics*, 12:1434–1447.