Recombination hotspots as a point process

Maria De Iorio¹, Eric de Silva² and Michael P. H. Stumpf^{2,*}

¹*Department of Epidemiology and Public Health, Faculty of Medicine, Imperial College London,
St Mary's Campus, Norfolk Place, London W2 1PG, UK*

²*Centre for Bioinformatics, Division of Molecular Biosciences, Imperial College London,
London SW7 2AZ, UK*

The variation of the recombination rate along chromosomal DNA is one of the important determinants of the patterns of linkage disequilibrium. A number of inferential methods have been developed which estimate the recombination rate and its variation from population genetic data. The majority of these methods are based on modelling the genealogical process underlying a sample of DNA sequences and thus explicitly include a model of the demographic process. Here we propose a different inferential procedure based on a previously introduced framework where recombination is modelled as a point process along a DNA sequence. The approach infers regions containing putative hotspots based on the inferred minimum number of recombination events; it thus depends only indirectly on the underlying population demography. A Poisson point process model with local rates is then used to infer patterns of recombination rate estimation in a fully Bayesian framework. We illustrate this new approach by applying it to several population genetic datasets, including a region with an experimentally confirmed recombination hotspot.

Keywords: haplotype blocks; genetic diversity; statistical inference

1. INTRODUCTION

Our understanding of details of the recombination process has been increasing in leaps and bounds during the last few years. The importance of the extent of recombination has been well documented for a very diverse set of problems ranging from applications in genetic epidemiology—in particular, the extent to which it determines linkage disequilibrium (LD) for association studies (Pritchard & Przeworski 2001)—to fundamental questions in evolutionary biology. Until fairly recently, both the effect on LD and the evolutionary aspects of recombination have been studied predominantly under the assumption of a more or less uniform recombination rate along the genome (Griffiths & Marjoram 1996; Wiuf & Hein 1999; Pritchard & Przeworski 2001; Nordborg & Tavaré 2002). Especially following the pioneering work of Jeffreys *et al.* (2001), who were the first to report highly resolved experimental studies of recombination hotspots, the inhomogeneity of the recombination rate along chromosomes has attracted considerable attention. The reality and importance of highly localized features in the recombination rate pattern has been verified in further experimental studies and population genetic analyses; moreover simulations have been used to evaluate the extent to which recombination rate variation is likely to affect population genetic data (Posada & Wiuf 2003; Stumpf & Goldstein 2003; Wiuf & Posada 2003).

In the theoretical analysis of the effects of recombination on patterns of genetic diversity the central quantity

is not the ‘true’ molecular per-generation recombination rate, r , but rather the so-called population recombination rate

$$\rho = 4N_e r, \quad (1.1)$$

where N_e is the effective population size. The population recombination rate ρ thus depends on (i) the molecular recombination process and (ii) the demographic process via the effective population size (Pritchard & Przeworski 2001; Nordborg & Tavaré 2002). N_e aims to subsume details of the complex demographic history of a population (from which the sample to be analysed was drawn) into a single number (Ewens 2004). The effects of demographic history on population genetic data—and in particular on LD—have been studied extensively using coalescent simulations. When making inferences from population genetic data it has to be kept in mind that what one is really estimating is some convolution of demographic and genetic processes. If the population model differs from the true demography, as is expected, results may be biased.

Incorporating variability has also become a major focus of recent theoretical work (McVean *et al.* 2004). While it is in principle possible to incorporate variable recombination rate models into a fully fledged likelihood framework, the computational costs of doing so are prohibitively expensive. In general approximations to a full likelihood treatment, notably the so-called composite likelihood estimators, allow for reasonable estimates of the population recombination rate estimators (Stumpf & McVean 2003). Because they do not use the full data but only some aspects of the data, e.g. the pattern of two-locus haplotypes, to infer recombination rates, they may suffer less from systematic bias; especially if the bias is compared to the intrinsic variability of the recombination rate process.

* Author for correspondence (m.stumpf@imperial.ac.uk).

One contribution of 12 to a Discussion Meeting Issue ‘Genetic variation and human health’.

Figure 1. Example of partition of a DNA sequence. The x-axis shows positions on the DNA sequence, with s_1 and s_n being sampled locations. On the y-axis we plot the recombination rate. The change-points t_1, t_2 split the DNA sequence in three regions: $[s_1, t_1)$, $[t_1, t_2)$, $[t_2, s_n]$. The recombination rate in each region is given by the values of μ_1, μ_2 and μ_3 .

The coalescent with recombination, often also referred to as the ancestral recombination graph, has become the canonical model to study the effects of recombination in population genetics (see the recent text by [Hein *et al.* 2004](#) for an excellent discussion). This genealogical framework has proved useful as a basis for inferential procedures. Crucially, considering the genealogical process shows that many recombination events will leave no trace in the data. This was confirmed in extensive simulation studies ([Stumpf 2004](#)).

In addition to the genealogical view of recombination, [Wiuf & Hein \(1999\)](#) have introduced the notion of recombination as a Poisson point process along a DNA sequence. Below we will develop a formalism which is based on this concept. The data used to infer the recombination rate are the observed recombination events in a population sample. While the number of observable recombination events is considerably smaller than the number of recombination events, regions with higher recombination rate will, for a large range of values of ρ , also have a higher number of observable recombination events. The particular attraction of the current approach—apart from its speed and ease of implementation—is that it treats the recombination rate as a ‘physical’ property of the sequence. It is thus, in some sense, related to the rate determined in sperm-typing experiments. It does nevertheless also depend on the population history through the data: demography can influence patterns of nucleotide diversity as well as the shape of the genealogy (with recombination) quite profoundly.

Predominantly, therefore, this method aims to detect regions where there is evidence for a significantly increased recombination rate, i.e. we are scanning genomic regions for recombination hotspots. Below we will show that results obtained from the present approach are (i) in good agreement with experimental

studies and (ii) yield consistent results when applied to different populations.

2. RECOMBINATION AS A POINT PROCESS

Recombination hotspots can be defined as small regions along a DNA sequence where the recombination rate is increased significantly relative to the local background rate. The literature is rich with proposals on how to estimate recombination rates and to identify recombination hotspots ([Hudson 2001](#); [Fearnhead & Donnelly 2002](#); [McVean *et al.* 2002](#); [Li & Stephens 2003](#); [McVean *et al.* 2004](#)). As a full likelihood approach is impracticable due to the computational burden, most of these methods rely on an approximation of the coalescent likelihood. Here we present statistical methods for hotspot detection based on the number of recombination events that have occurred in the genealogy describing the joint history of a sample. We assume that recombination events occur along a DNA sequence according to a Poisson process whose intensity depends on the local recombination rate. Let s_i and s_j be any two positions on the DNA sequence and, without loss of generality, let $s_i < s_j$. Then, the number of recombination events between each pair of sites s_i, s_j , $\tilde{N}_{(s_i, s_j)}$, has a Poisson distribution

$$\tilde{N}_{(s_i, s_j)} \sim \text{Poisson}(R(s_i, s_j)), \quad (2.1)$$

where $R(s_i, s_j)$ denotes the rate of the Poisson distribution (and therefore the expected number of recombination events) in the region (s_i, s_j)

$$R(s_1, s_2) = \int_{s_1}^{s_2} \rho(s) ds, \quad (2.2)$$

where $\rho(s)$ is the local recombination rate. Some recombination events are essentially undetectable, therefore it is usually difficult to estimate from the available data the true number of recombination events that have occurred in a specific region of a DNA sequence. Consider, for example, the case in which both parents of the recombinant are of the same type in the region sequenced. Therefore, it makes sense to use the minimum number of recombination events in the history of the sample ([Hudson & Kaplan 1985](#); [Myers & Griffiths 2003](#)) and $\tilde{N}_{(s_i, s_j)}$ can be thought of as missing value. Let $N_{(s_i, s_j)}$ be the minimum number of recombination events between positions s_i and s_j . We assume that $N_{(s_i, s_j)}$ and $\tilde{N}_{(s_i, s_j)}$ are stochastically related

$$N_{(s_1, s_2)} | \tilde{N}_{(s_1, s_2)} \sim \text{Binomial}(\tilde{N}_{(s_1, s_2)}, \theta),$$

and $N_{(s_i, s_j)}$ is observed. In other words, we assume that each recombination event has a probability θ of being detected and that event detections are independent. It is easy to show that

$$N_{(s_1, s_2)} \sim \text{Poisson}(\theta R(s_1, s_2)). \quad (2.3)$$

In the remainder of the paper we will consider only a value of θ fixed for all the DNA sequence, but what follows can be easily extended to the case where different θ are assumed for different regions of the sequence.

Recombination events tend to cluster around specific positions along the DNA sequence. The methodology aims to identify the locations of such

clusters. The proposed statistical model is an extension of Bayesian partition models (BPM) to this context (Denison *et al.* 2002). BPM have been successfully used in various areas, including grouping of binary experiments (Consonni & Veronese 1995; Green 1995), semiparametric regression and classification (Denison *et al.* 1998) and spatial modelling of disease risk (Denison & Holmes 2001; Green & Richardson 2002). Note that the model will automatically scale for the distance between sites as the Poisson rate in equation (2.2) is defined as the integral of the recombination rate between s_i and s_j .

3. RECOMBINATION RATE AS A PIECE-WISE POLYNOMIAL

The general idea behind partition models is to split the DNA sequence into a number of disjoint regions, where the data points in each region are assumed to come from a common distribution indexed by some region-specific parameters ϕ and the distributions of data in different regions are independent of each other. In one dimension partition models are just change-point models and the disjoint regions are defined by the location of the change-points. We develop a Bayesian multiple change-point analysis of recombination event data by assuming the recombination rate on the DNA sequence is a piece-wise constant polynomial or a step function, where the number of change-points and their locations are treated as unknown parameters and need to be inferred. Between each change-point we fit a constant level for the recombination rate ρ . In other words, we partition the DNA sequence in intervals of possibly different length. In each region the data are assumed to be generated independently from locally parameterized models. This allows us to place in a single region those points relating to an unusual cluster, and these points do not necessarily influence the response function in nearby locations. Moreover, assuming independence among the intervals implies that the response function at the cluster centre tends not to be over-smoothed (Denison & Holmes 2001; Denison *et al.* 2002). In conclusion, the partitioning idea applied to the number of recombination events along a DNA sequence implies a clustering of segments of a DNA sequence into larger regions with similar intensity. The model lets the data dictate the nature of the recombination rate, allowing it the flexibility to capture unusual features of the data (such as hotspots).

Let $S = \{s_1, \dots, s_n\}$ denote the set of n DNA sampled positions, with

$$s_1 < s_2 < \dots < s_n.$$

We define a partition of dimension k via the random parameters $t_1, \dots, t_k$ that give the locations of the k change-points, i.e. $\mathcal{T}_k = \{t_1, \dots, t_k\}$. We assume $t_1 < t_2 < \dots < t_k$ with $s_1 < t_1$ and $t_k < s_n$. For computational convenience, we restrict the t_i to coincide with one of the sampled locations in S . This assumption can be easily relaxed, with slight changes to the Markov chain Monte Carlo (MCMC) scheme

described below. The k change-points split the DNA region in $k+1$ regions A_ℓ

$$[s_1, t_1), [t_1, t_2), \dots, [t_{k-1}, t_k), [t_k, s_n].$$

If μ_ℓ denotes the parameters of the distributions in region A_ℓ , then the likelihood of a single observation in the region is given by

$$\begin{aligned} p(N_{(s_i, s_j)} | (s_i, s_j) \in A_\ell, \mu_\ell) &= \text{Pois}(\theta R(s_i, s_j)) \\ &= \text{Pois}(\theta \mu_\ell (s_j - s_i)) \\ &= \frac{(\theta \mu_\ell (s_j - s_i))^{N_{(s_i, s_j)}} \exp(-\theta \mu_\ell (s_j - s_i))}{N_{(s_i, s_j)}!}. \end{aligned} \quad (3.1)$$

Note that μ_ℓ is interpreted as the recombination rate in region A_ℓ , i.e. $\rho(s) = \mu_\ell$ for $s \in A_\ell$. An example of a partition of a DNA sequence is shown in figure 1.

The likelihood for the complete data set is given by

$$\prod_{\ell=1}^{k+1} \prod_{r=1}^{N_\ell} p(N_{(s_{i_r}, s_{j_r})} | (s_{i_r}, s_{j_r}) \in A_\ell, \mu_\ell), \quad (3.2)$$

where N_ℓ is the number of data points in region A_ℓ and (s_{i_r}, s_{j_r}) , $r=1, \dots, N_\ell$ denote pairs of sampled location contained in region A_ℓ . Since θ is simply a scaling factor we prefer to reparametrize the model in terms of $\lambda_\ell = \theta \mu_\ell$.

We now discuss the prior distribution on the partition $(k, \mathcal{T}_k)$. As we have little idea on how many change-points are required to adequately approximate the truth we assume that k can take any value in $\{1, \dots, n-1\}$ with the same probability, i.e. we assume a discrete uniform prior distribution for k . Given k , we then assume that each partition structure $\mathcal{T}_k$ with the same number of change-points is equally likely. There is a natural hierarchical structure in this set-up which we formalize as $p(k, \mathcal{T}_k) = p(k)p(\mathcal{T}_k | k)$.

A major advantage of BPM is that conjugate priors can be assigned to the data within each interval. In this way we are able to marginalize over many of the model parameters, reducing the effective dimension of the model space during simulation and providing a simple analytic expression for the marginal likelihood of any proposed partition. We assume that the prior distributions for each λ_ℓ are independent

$$p(\lambda_1, \dots, \lambda_{k+1} | k, \mathcal{T}_k) = \prod_{\ell=1}^{k+1} p(\lambda_\ell | k, \mathcal{T}_k).$$

For the Poisson model given in equation (3.1), the conjugate prior distribution is a Gamma distribution. Therefore, we choose $p(\lambda_\ell | k, \mathcal{T}_k) = \text{Gamma}(\alpha, \beta)$ for some suitable choice of the hyper-parameters α and β . Let $\mathcal{N}$ denote the set of all the data and let $\mathcal{N}_\ell$ the subset of the data assigned to region A_ℓ . The conditional posterior distribution for each λ_ℓ is then a Gamma distribution

$$\begin{aligned} p(\lambda_\ell | \mathcal{N}_\ell, k, \mathcal{T}_k) \\ = \text{Gamma} \left(\alpha + \sum_{r=1}^{N_\ell} N_{(s_{i_r}, s_{j_r})}, \beta + \sum_{r=1}^{N_\ell} (s_{j_r} - s_{i_r}) \right), \\ \ell = 1, \dots, k+1 \end{aligned} \quad (3.3)$$

Figure 2. Recombination rates inferred from sperm-typing studies (top) and statistically inferred recombination intensities (bottom) for the HLA. Dots indicate SNP locations.

The marginal likelihood for this model can be written as

$$p(\mathbf{N}|k, \mathcal{T}_k) = \prod_{\ell=1}^{k+1} \left\{ \frac{\beta^\alpha \prod_{r=1}^{N_\ell} (s_{j_r} - s_{i_r})^{N(s_{i_r}, s_{j_r})}}{\Gamma(\alpha) \prod_{r=1}^{N_\ell} N(s_{i_r}, s_{j_r})!} \times \frac{\Gamma\left(\alpha + \sum_{r=1}^{N_\ell} N(s_{i_r}, s_{j_r})\right)}{\left(\beta + \sum_{r=1}^{N_\ell} (s_{j_r} - s_{i_r})\right)^{\alpha + \sum_{r=1}^{N_\ell} N(s_{i_r}, s_{j_r})}} \right\} \quad (3.4)$$

(see Denison *et al.* 2002). The posterior distribution for the partition $(k, \mathcal{T}_k)$ is then given by

$$p(k, \mathcal{T}_k | \mathbf{N}) \propto p(\mathbf{N} | k, \mathcal{T}_k) p(k, \mathcal{T}_k).$$

4. COMPUTATIONAL STRATEGY

A main advantage of the proposed methodology is efficiency and simplicity of computations. We now describe how to generate samples from the posterior distribution of $(k, \mathcal{T}_k)$. We use a MCMC sampler (Gilks *et al.* 1996; Robert & Casella 1999) which constructs a Markov chain whose stationary distribution is $p(k, \mathcal{T}_k)$. The algorithm is reasonably straightforward: we need to devise moves around

the space of the partitions structure which ensure that the chain is aperiodic and irreducible. We propose the simple scheme given below but others are possible.

Algorithm:

- (i) Initialize the partition structure, i.e. initialize the number of change-points k and their locations $\mathcal{T}_k = \{t_1, \dots, t_k\}$.
- (ii) Propose a new partition structure k^* , $\mathcal{T}_{k^*}$ by randomly drawing a sampled location s from $\{s_2, \dots, s_{n-1}\}$. If $s \in \mathcal{T}_k$ then $\mathcal{T}_{k^*} = \mathcal{T}_k \setminus \{s\}$ and $k^* = k - 1$. If $s \notin \mathcal{T}_k$ then $\mathcal{T}_{k^*} = \mathcal{T}_k \cup \{s\}$ and $k^* = k + 1$. The proposed change is then accepted with probability

$$\gamma = \min \left\{ 1, \frac{p(\mathbf{N} | k^*, \mathcal{T}_{k^*})}{p(\mathbf{N} | k, \mathcal{T}_k)} \right\},$$

otherwise the current partition structure is retained.

- (iii) Repeat step 2 until convergence.

Note that the acceptance probability γ is just the Bayes factor (Kass & Raftery 1995) in favour of the proposed model over the current one. Therefore, the posterior distribution from which we simulate samples only depends on the partition structure $k, \mathcal{T}_k$. Inference about $\lambda_1, \dots, \lambda_{k+1}$ is straightforward, as their conditional posterior distributions are readily available (cf. equation (3.3)).

Figure 3. Profiles (left column) and corresponding cumulative distribution (right column) of the recombination intensity across the human *f5* gene in an African (top) and a European (bottom) population sample.

5. MINIMUM NUMBER OF RECOMBINATION EVENTS

To obtain the minimum number of recombination events for the following examples described in the paper, we have used the program RECMIN (Myers & Griffiths 2003) which provides a lower bound on the number of recombination events in the history of a sample on the basis of patterns of variation in the sample DNA. Suppose we have phased haplotype data with n segregating sites (sampled locations) $s_1, \dots, s_n$. RECMIN returns an upper triangular matrix B , whose elements B_{ij} are the minimum number of recombination events between sites s_i, s_j with $1 \leq i < j \leq n$. We treat the rows of the matrix B as independent observations. This is the strongest assumption we impose on the data and it can lead us to underestimate the posterior variance of our estimates. The data $N_{(s_i, s_j)}$ coincide with the elements of B , i.e. $N_{(s_i, s_j)} = B_{ij}$, $i < j$. Hence there are $M = n(n-1)/2$ data points.

Let $\mathbf{n}_i = (N_{(s_i, s_{i+1})}, \dots, N_{(s_i, s_n)})$ be the vector whose elements are the number of recombination events between s_i and s_j , with $i < j \leq n$. Then

$$\begin{aligned} p(\mathbf{n}_i) &= p(N_{(s_i, s_{i+1})}, \dots, N_{(s_i, s_n)}) \\ &= p(N_{(s_i, s_n)} | N_{(s_i, s_{i+1})}, \dots, N_{(s_i, s_{n-1})}) \\ &= p(N_{(s_i, s_{n-1})} | N_{(s_i, s_{i+1})}, \dots, N_{(s_i, s_{n-2})}) \dots p(N_{(s_i, s_{i+1})}) \end{aligned}$$

Rewriting each $N_{(s_i, s_{i+j})}$ as the sum of independent Poisson random variables

$$N_{(s_i, s_{i+j})} = \sum_{r=i}^{i+j-1} N_{(s_r, s_{r+1})},$$

and exploiting the *independent increments* property of the Poisson process, we obtain

$$p(\mathbf{n}_i) = \prod_{j=i}^{n-1} p(N_{(s_j, s_{j+1})}).$$

The likelihood for a pair of adjacent sites is

$$p(N_{(s_i, s_{i+1})} | (s_i, s_{i+1}) \in A_\ell, \lambda_\ell) = \text{Poisson}(\lambda_\ell(s_{i+1} - s_i)).$$

Assuming that the $\mathbf{n}_i$ s are independent leads to the following likelihood for all the data $\mathbf{n}_i$, $i = 1, \dots, n-1$,

$$\begin{aligned} \mathbf{n}_1, \dots, \mathbf{n}_{n-1} | k, \mathcal{T}_k, A_k &\sim \prod_{i=1}^{n-1} p(\mathbf{n}_i | k, \mathcal{T}_k, A_k) \\ &= \prod_{i=1}^{n-1} \prod_{j=i}^{n-1} p(N_{(s_j, s_{j+1})} | k, \mathcal{T}_k, A_k) \\ &= \prod_{i=1}^{n-1} \prod_{j=i}^{n-1} \text{Poisson} \left(\sum_{\ell=1}^{k+1} \lambda_\ell(s_{j+1} - s_j) I_{A_\ell}(s_j, s_{j+1}) \right) \end{aligned} \quad (5.1)$$

Figure 4. Profiles of the recombination intensity across the human *plau* (top) and *plaur* (bottom) genes in an African (left) and a European (right) population sample. Note the different scales on the vertical axis of the plots for the two genes.

where $A_k = \{\lambda_1, \dots, \lambda_k\}$ and

$$I_{A_k} = \begin{cases} 1 & \text{if } (s_j, s_{j+1}) \subseteq A_k \\ 0 & \text{otherwise} \end{cases}.$$

We can substitute equation (5.1) in equation (3.2) to obtain equivalent expressions for the posterior distribution of the λ_k s (equation (3.3)) and the marginal likelihood of the data (equation (3.4)).

6. APPLICATIONS

The power of statistical approaches can be assessed by testing against (i) simulated data and (ii) experimentally well-characterized data. As a general test of the new method's usefulness and an illustration of its power, we apply it first to the data of Jeffreys *et al.* Here the recombination hotspot intensity was fine-mapped in sperm-typing experiments. In figure 2 we show the data of Jeffreys *et al.* (2001) on a log-scale and the inferred recombination intensity using the new method. Apart from one notable exception (at approximately 75 000 bp) the positions of regions with high recombination intensity coincide well with the experimentally mapped hotspot positions. The reasons for this disagreement are somewhat puzzling but we note that we ran two other recombination rate estimation procedures (Li & Stephens 2003; McVean

et al. 2004) which also indicate increased recombination rates in this area. Therefore, we believe that the new method performs adequately, and quite remarkably given its simplicity and low computational cost.

While LD patterns are currently the main area of application of recombination rate estimators there has also been considerable interest in the evolutionary consequences of, and reasons for recombination. We have previously looked at a set of genes to determine the extent to which recombination relates to the organization of genes. The new approach allows us to determine hotspot positions with a resolution determined in principle by the single nucleotide polymorphism (SNP) density. In figure 3 we show recombination intensity profiles and cumulative distributions for the *f5* gene in an African and a European population sample, respectively (de Silva *et al.* 2004). We find good agreement between the positions of inferred hotspots, although the relative intensity varies considerably. The cumulative recombination intensity plots show how the inferred genetic distance relates to physical distance. As it is already apparent from the other two plots, the most pronounced vertical jumps in the two curves correspond well between the two populations. Also, and entirely as expected, the genetic distance corresponding to the *f5* gene in the African population sample is almost twice as large as the genetic distance in the European sample. This reflects the fact that the

number of observable recombination events in the data captures properties of the demographic process.

In addition to recombination hotspots, the recombination intensity is also able to estimate regions where the recombination intensity varies on a much more modest scale. In figure 4 we show the inferred profiles for two genes, *plau* and its receptor *plaur*, again in an African and a European population sample. The maximum intensity is markedly reduced compared to the *f5* gene in figure 4. We find, however, that there is still some evidence for variation in the underlying recombination rate. In the shorter *plau* gene we notice that the region of increased recombination intensity extends further in the African sample; the maximum absolute values of the recombination rates in the two populations also confirm that there is more evidence for historical recombination events in the African populations. Furthermore, our estimates are obtained under the assumption that the proportion of detected recombination events in the two populations is the same. This complicates the comparison of rates in different populations.

For the *plaur* gene recombination appears much more localized than for *plau*. The positions of the inferred hotspots agree very well between the two populations, even if their intensities differ. Interestingly, here the maximum is slightly higher for the European sample than for the African sample. We note that for the three cases discussed here, the inferred profiles of the recombination intensity agree very well with results obtained using composite likelihood approaches. The computational cost is however, greatly reduced.

7. DISCUSSION

We have introduced a new method for detecting variation in the recombination rate. The particular attraction of our method lies in the fact that it treats recombination rate (or our equivalent recombination intensity) as a physical rate and therefore uses the concepts introduced by Wiuf & Hein (1999) in an inferential framework. Because of the conjugacy of the prior, calculations are also computationally cheap and even for large regions calculations typically last only for a few minutes.

The data we use, and the number of observed recombination events, does, of course, depend on the population history. At the moment, however, estimating good lower bounds for the minimum number of recombination events requires haplotypes. Inferring reliable numbers from genotype data is difficult but we can average over estimates obtained from many haplotype reconstructions.

REFERENCES

- Consonni, G. & Veronese, P. 1995 A Bayesian method for combining results from several binomial experiments. *J. Am. Stat. Assoc.* **90**, 935–944.
- Denison, D. G. T. & Holmes, C. C. 2001 Bayesian partitioning for estimating disease risk. *Biometrics* **57**, 143–149.
- Denison, D. G. T., Mallick, B. K. & Smith, A. F. M. 1998 A Bayesian CART algorithm. *Biometrika* **85**, 363–377.
- Denison, D. G. T., Holmes, C. C., Mallick, B. K. & Smith, A. F. M. 2002 *Bayesian methods for nonlinear classification and regression*. New York: Wiley.
- de Silva, E., Kelley, L. & Stumpf, M. 2004 The extent and importance of intragenic recombination. *Hum. Genomics* **1**, 410–420.
- Ewens, W. 2004 *Mathematical population genetics*. New York: Springer.
- Fearnhead, P. & Donnelly, P. 2002 Approximate likelihood methods for estimating local recombination rates (with discussion). *J. R. Stat. Soc. Ser. B* **64**, 657–680.
- Gilks, W. R., Richardson, S. & Spiegelhalter, D. J. (eds) 1996 *Markov chain Monte Carlo in practice*. New York: Chapman & Hall. ISBN 0-412-05551-1
- Green, P. J. 1995 Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82**, 711–732.
- Green, P. J. & Richardson, S. 2002 Hidden Markov models and disease mapping. *J. Am. Stat. Assoc.* **97**, 1055–1070.
- Griffiths, R. & Marjoram, P. 1996 Ancestral inference from samples of DNA sequences with recombination. *J. Comput. Biol.* **3**, 479–502.
- Hein, J., Schierup, M. & Wiuf, C. 2004 *Gene genealogies, variation and evolution*. Oxford, UK: Oxford University Press.
- Hudson, R. R. 2001 Two-locus sampling distributions and their application. *Genetics* **159**, 1805–1817.
- Hudson, R. R. & Kaplan, N. 1985 Statistical properties of the number of recombination events in the history of a sample of DNA sequences. *Genetics* **111**, 147–164.
- Jeffreys, A., Kauppi, L. & Neumann, R. 2001 Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. *Nature Genet.* **29**, 217–222.
- Kass, R. & Raftery, A. E. 1995 Bayes factors. *J. Am. Stat. Assoc.* **90**, 773–795.
- Li, N. & Stephens, M. 2003 Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics* **165**, 2213–2233.
- McVean, G. A. T., Awadalla, P. & Fearnhead, P. 2002 A coalescent-based method for detecting and estimating recombination from gene sequences. *Genetics* **160**, 1231–1241.
- McVean, G. A. T., Myers, S. R., Hunt, S., Deloukas, P., Bentley, D. R. & Donnelly, P. 2004 The fine-scale structure of recombination rate variation in the human genome. *Science* **304**, 581–584.
- Myers, S. R. & Griffiths, R. C. 2003 Bounds on the minimum number of recombinations events in a sample history. *Genetics* **163**, 375–394.
- Nordborg, M. & Tavaré, S. 2002 Linkage disequilibrium: what history has to tell us. *Trends Genet.* **18**, 83–90.
- Posada, D. & Wiuf, C. 2003 Simulating haplotype blocks in the human genome. *Bioinformatics* **19**, 289–290.
- Pritchard, J. K. & Przeworski, M. 2001 Linkage disequilibrium in humans: models and data. *Am. J. Hum. Genet.* **69**, 1–14.
- Robert, C. P. & Casella, G. 1999 *Monte Carlo statistical methods*. New York: Springer.
- Stumpf, M. 2004 Haplotype diversity and SNP frequency dependence in the description of genetic variation. *Eur. J. Hum. Genet.* **12**, 469–477.
- Stumpf, M. & Goldstein, D. 2003 Demography, recombination hotspot intensity, and the block structure of linkage disequilibrium. *Curr. Biol.* **13**, 1–8.
- Stumpf, M. & McVean, G. 2003 Estimating recombination rates from population genetic data. *Nature Rev. Genet.* **4**, 959–968.
- Wiuf, C. & Hein, J. 1999 Recombination as a point process along sequences. *Theor. Popul. Biol.* **55**, 248–259.
- Wiuf, C. & Posada, D. 2003 A coalescent model of recombination hotspots. *Genetics* **164**, 407–417.