PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B

Phil. Trans. R. Soc. B (2008) **363**, 3921–3930
doi:10.1098/rstb.2008.0172
Published online 7 October 2008

Bayesian inference of fine-scale recombination rates using population genomic data

Ying Wang[1,2] **and Bruce Rannala**[1,2,*]

[1]*Genome Center, and* [2]*Department of Evolution and Ecology, University of California Davis, One Shields Avenue, Davis, CA 95616, USA*

Recently, several statistical methods for estimating fine-scale recombination rates using population samples have been developed. However, currently available methods that can be applied to large-scale data are limited to approximated likelihoods. Here, we developed a full-likelihood Markov chain Monte Carlo method for estimating recombination rate under a Bayesian framework. Genealogies underlying a sampling of chromosomes are effectively modelled by using marginal individual single nucleotide polymorphism genealogies related through an ancestral recombination graph. The method is compared with two existing composite-likelihood methods using simulated data. Simulation studies show that our method performs well for different simulation scenarios. The method is applied to two human population genetic variation datasets that have been studied by sperm typing. Our results are consistent with the estimates from sperm crossover analysis.

Keywords: recombination rate; recombination hotspot; linkage disequilibrium; ancestral recombination graph; Bayesian inference

1. INTRODUCTION

Inferring how recombination rates vary across chromosomes is a fundamental problem in population genetics owing to its implications for evolutionary studies, disease association mapping and understanding the molecular basis of recombination. Pedigree-based estimates of recombination rate are typically at a coarse scale due to few informative meioses. Although sperm typing can provide fine-resolution estimates of recombination rate, the laborious nature of the experimental techniques limits its usefulness to comparisons on a small region (typically less than a megabase interval). Sperm typing is also restricted to estimates of male recombination. Recently, population-genetics-based methods have been applied to estimate sex-averaged recombination rates and to measure the changes in recombination rates over human genomes (reviewed in Hellenthal & Stephens 2006).

Although a number of population genetic models have been proposed based on the use of a coalescent process with recombination (Kingman 1982*a*,*b*; Hudson 1990), the inference methods that are currently widely used for inferring recombination rates are based on approximated likelihoods (typically a composite likelihood). Approximate-likelihood methods may provide consistent point estimates, but the likelihoods obtained do not have standard properties, and especially when the data contain limited information, the approximate-likelihood methods may not perform as well as methods that use the full likelihood. Full-likelihood methods have the advantage that they use all information contained in the data. However, existing full-likelihood methods (Griffiths & Marjoram 1996; Kuhner *et al.* 2000; Nielsen 2000; Fearnhead & Donnelly 2001) are too computationally expensive to be applied to large-scale genomic data (reviewed in Stumpf & McVean 2003).

The use of composite-likelihood methods in estimating ρ was first suggested by Hudson (2001), although similar ideas have previously been used in linkage disequilibrium (LD) mapping methods (reviewed by Rannala & Slatkin 2000). McVean *et al.* (2002) extended Hudson's (2001) composite-likelihood approach to allow recurrent mutation. Other approximate methods have also been developed. Li & Stephens (2003), for example, developed a method based on an approximation to the conditional likelihood, which they called the 'product of approximate conditionals (PAC)' likelihood. The two approximate-likelihood methods have recently been applied to study properties of recombination rate across human genomes (e.g. Myers *et al.* 2005; Graffelman *et al.* 2007).

Here, we develop a computationally tractable full-likelihood-based method in a Bayesian framework. In our method, the chromosomal intervals are treated as vectors of discrete points (corresponding to sampled marker sites), and the genealogy underlying a sample is effectively modelled by excluding non-ancestral recombination and lineages. We compare the performance of our method with the composite-likelihood method (implemented in `LDhat`) and the 'PAC' likelihood method (implemented in `PHASE`) using simulated data, and applied our method to two human population genetic variation datasets (Jeffreys *et al.* 2001, 2005).

*Author and address for correspondence: Genome Center, University of California Davis, One Shields Avenue, Davis, CA 95616, USA (brannala@ucdavis.edu).

One contribution of 17 to a Discussion Meeting Issue 'Statistical and computational challenges in molecular phylogenetics and evolution'.

2. BAYESIAN INFERENCE VIA SINGLE NUCLEOTIDE POLYMORPHISM GENEALOGIES

Coalescent theory provides a general framework for population genetic inference (Kingman 1982*a*,*b*; Hudson 1990). We let G denote the genealogy underlying a sample of single nucleotide polymorphism (SNP) haplotypes or genotypes (denoted by $\boldsymbol{X}$). Let $\boldsymbol{\Theta}$ be a vector of parameters, including $\theta = 4N_e\mu$ and $\rho = 4N_e c$, where N_e is the effective population size; μ is the site-specific mutation rate per generation; and c is the recombination rate per generation in cM Mb^{-1}. The posterior density of parameters,

$$f(\boldsymbol{\Theta}|\boldsymbol{X}) = \frac{1}{f(\boldsymbol{X})}\int f(\boldsymbol{X}|G,\boldsymbol{\Theta})f(G|\boldsymbol{\Theta})f(\boldsymbol{\Theta})\mathrm{d}G, \qquad (2.1)$$

can be estimated using Markov chain Monte Carlo (MCMC). However, developing inference methods based on the coalescent with recombination (ancestral recombination graph, ARG) is challenging because the problem is high dimensional (the probability of any one instance of the genealogy is small) and the dimension varies (due to the variable number of coalescent–recombination pairs). These factors can lead to inefficient parameter estimates because a large portion of the ARG is not related to the data and does not contribute to the likelihood calculations.

If $\boldsymbol{X}$ represents the data for one marker site, the genealogy underlying the sample is a coalescent tree (e.g. τ_0). With k linked markers ($k>1$), the genealogy can be considered an array of correlated coalescent trees (e.g. $\boldsymbol{\tau} = \{\tau_0, \ldots, \tau_{k-1}\}$), which are jointly sufficient for the likelihood calculation. The ARG provides an indirect way to sample genealogies for each of the markers and is the biologically appropriate prior for the genealogies under a model of the coalescent with recombination.

We develop a method for estimating ρ using population samples of SNPs based on the ARG in a Bayesian framework. Both constant and variable recombination models are considered. We let G_S denote the joint multilocus SNP genealogy (described in §3) underlying the sample. Given G_S, the genealogical trees ($\boldsymbol{\tau}$) for each marker position can be obtained (figure 1). The posterior distribution of $\boldsymbol{\rho}$ is

$$f(\boldsymbol{\rho}|\boldsymbol{X}) = \frac{1}{f(\boldsymbol{X})}\int f(\boldsymbol{X}|\boldsymbol{\tau}\in G_S,\theta)f(G_S|\boldsymbol{\rho})f(\boldsymbol{\rho})f(\theta)\mathrm{d}G_S\mathrm{d}\theta, \qquad (2.2)$$

which can be numerically evaluated by MCMC.

(a) *Prior on genealogy (SNP genealogy, G_S)*

The most general representation of the ARG (Griffiths & Marjoram 1996) models the positions of recombinations as events on an interval of the real line. The probability density of recombination events on this interval depends on the spatial distribution of local recombination rates, hotspots, etc. Each recombination event separates a chromosomal interval into two disjoint intervals (split at the recombination point) that subsequently undergo independent coalescence and recombination processes. To analyse SNP markers, it is efficient to reduce this description to exclude chromosomal segments that are not ancestral to the sample and for which markers are not informative.

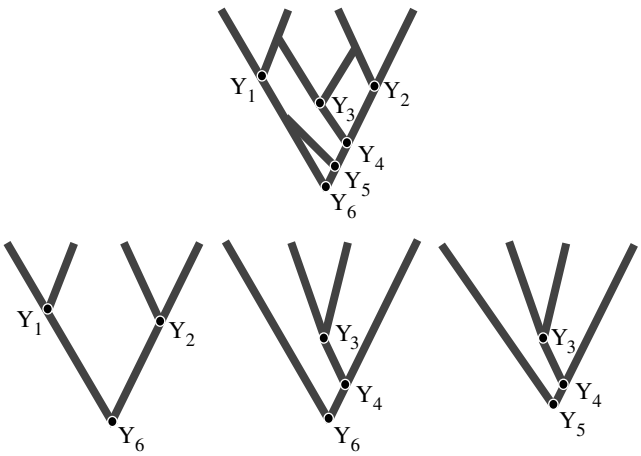

Figure 1. Illustration of SNP genealogy (G_S) and marker trees resolved from G_S.

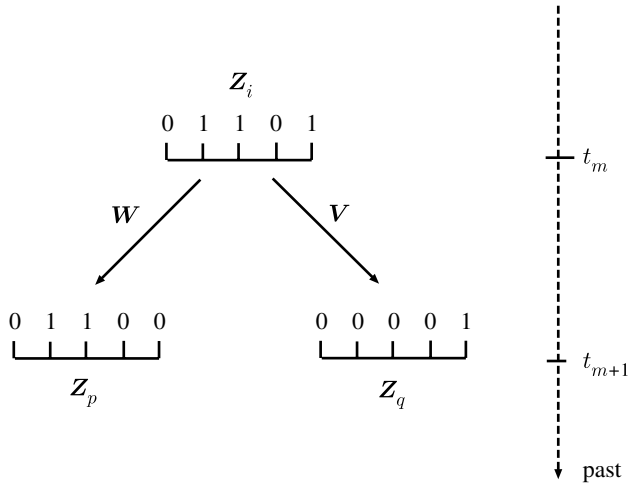

Figure 2. Illustration of the MA vector. In this example, a recombination occurred at time t_{m+1} and split $\boldsymbol{Z}_i$ to $\boldsymbol{Z}_p$ and $\boldsymbol{Z}_q$.

Let $\boldsymbol{r} = [r_0, \ldots, r_{k-1}]$ be a vector of the total recombination rate between adjacent marker sites for k markers on an interval. Note that $r_j = \rho_j \times d_j$, where d_j denotes the distance in Mb between markers j and $j+1$, and $\rho = \rho_0, \ldots, \rho_{k-1}$ is a vector of the ρ for each interval between markers, where $\rho_j = 4N_e c_j \times 0.01$ is the rate in units of per cent recombination per Mb on the interval between markers j and $j+1$ (in time units of $2N_e$ generations). Let $\boldsymbol{Z}_i = \{Z_{ij}\}$ be a Boolean vector, such that $Z_{ij}=1$ if the jth marker for the ith lineage is ancestral to the sampled chromosomes and 0 otherwise. We refer to this as the marker ancestry (MA) vector.

If a recombination occurs on the interval between markers j and $j+1$, the two resulting lineages p and q receive MA vectors determined by taking the Hadamard products,

$$\boldsymbol{Z}_p = \boldsymbol{W}\cdot\boldsymbol{Z}_i, \qquad (2.3)$$

$$\boldsymbol{Z}_q = \boldsymbol{V}\cdot\boldsymbol{Z}_i, \qquad (2.4)$$

where $\boldsymbol{W}$ and $\boldsymbol{V}$ are both vectors of dimension k with $W_l=0$ for all $l\le j$, $W_l=1$ for all $l>j$, $V_l=1$ for all $l\le j$ and $V_l=0$ for all $l>j$. An example is illustrated in figure 2. A coalescence event involves Boolean vector addition of elements in the two descendent

vectors ($\boldsymbol{Z}_p$ and $\boldsymbol{Z}_q$),

$$\boldsymbol{Z}_A = \boldsymbol{Z}_p + \boldsymbol{Z}_q, \tag{2.5}$$

where $\boldsymbol{Z}_A$ is the MA vector for the ancestral lineage (A).

The total recombination rate at any point in time is the sum of recombination rates across lineages. Note that if a lineage i contains only one non-zero element in $\boldsymbol{Z}_i$, the recombination rate becomes 0 (e.g. $\boldsymbol{Z}_q$ in figure 2). Consider the example illustrated in figure 2; the prior density of G_S at time t_{m+1} (in $2N_e$ generations) at which a recombination event occurred on lineage i between markers x and $x+1$, given m lineages at t_m, is

$$\exp\left\{-\left[\binom{m}{2} + \sum_{y=1}^{m}\sum_{j=L}^{R-1} r_j^{(y)}\right](t_{m+1}-t_m)\right\}\frac{r_x}{d_x}, \tag{2.6}$$

where L is the first marker at the left, with $Z_{il}>0$, and R is the last marker at the right, with $Z_{iI}>0$, for lineage i, and $r_j^{(y)}$ indicates the recombination rate between markers j and $j+1$ on lineage y, obtained by multiplying $\boldsymbol{r}^{(y)}$ by $\boldsymbol{Z}_y$. Equation (2.6) represents the joint probability of exponential waiting time and recombination breakpoint.

(b) *Prior on recombination rate*

Currently, the pattern of recombination hotspots from several genomic regions has been studied by sperm typing (Arnheim *et al.* 2007). As was discovered from these studies, the majority of recombination clusters in narrow regions known as recombination hotspots and recombination only occasionally occurs in non-hotspot regions. Since only recombination rates between markers contribute to the probability of the genealogy, a general model of variable recombination is assumed such that $\boldsymbol{\rho}_i$ between markers i and $i+1$, $0 \le i < k-1$, is independent with a common prior density,

$$f(\rho_i) = p_H f_H(\rho_i) + (1-p_H) f_{\bar{H}}(\rho_i), \tag{2.7}$$

where $f_H(\rho_i)$ and $f_{\bar{H}}(\rho_i)$ represent the prior densities of ρ within a recombination hotspot and background recombination rate (non-hotspot regions), respectively. It is assumed that $f_H(\rho_i) \sim \text{lognormal}(\mu_H, \sigma_H)$ and $f_{\bar{H}}(\rho_i) \sim \text{lognormal}(\mu_{\bar{H}}, \sigma_{\bar{H}})$, and $\mu_H = 10$, $\sigma_H = 1$, $\mu_{\bar{H}} = 5$ and $\sigma_{\bar{H}} = 1$. The 0.025 and 0.975 quantiles for $f_H(\rho_i)$ and $f_{\bar{H}}(\rho_i)$ are [3102.73,156 367.45] and [20.91,1053.60], respectively. Parameter p_H is integrated in the MCMC, since it is unknown for any given chromosomal intervals.

If assuming ρ is constant (e.g. across short intervals), since more data points (markers) contribute to the estimate of the parameter, a uniform distribution bounded at 0 is used as the prior on ρ. Similarly, the prior on θ is assumed to be uniform and bounded at 0. Although we try to use minimal prior information, the effect of different prior densities on ρ and θ needs to be further investigated.

By increasing the efficiency of computations on the ancestral graph using SNP genealogies to allow the use of a full-likelihood-based model, the accuracy of estimates of ρ can be improved by comparison with approximate likelihoods, since all of the information contained in the data is used for obtaining the estimates for full-likelihood methods. Prior independent knowledge of how recombination rate varies and how recombination hotspots are distributed across chromosomes obtained from other studies (e.g. by sperm typing; reviewed in Arnheim *et al.* 2007) could also be incorporated into the prior used in an analysis to make use of such information and obtain more refined estimates of recombination rate.

3. SIMULATION STUDIES

Smith & Fearnhead (2005) performed a simulation study to evaluate the performance of three existing composite-likelihood methods (Hudson 2001; Fearnhead & Donnelly 2002; Li & Stephens 2003) for estimating population recombination rate using sequence data. Here, we focused on conducting simulation studies to examine the statistical performance of our method (implemented in the program `InferRho`) and to compare our method with two existing widely used approximate-likelihood methods implemented in `LDhat` (McVean *et al.* 2004) and `PHASE` (Li & Stephens 2003; Crawford *et al.* 2004). The first two simulation studies were aimed at examining the performance of the methods by assuming a constant recombination rate across intervals, and examining the effect of different SNP ascertainment criteria, using either haplotypes (simulation study 1) or genotypes (simulation study 2). The third simulation study examined the performance of the methods when allowing for variable recombination rates across intervals. For all three simulation studies, the sample size was set to be 50 chromosomes, and it was assumed that $N_e = 10^4$ and $\mu = 10^{-8}$ per site per generation under a Jukes–Cantor model, both realistic values for humans. Three program packages were used for these comparisons: `InferRho` v. 1.0; `LDhat` v. 2.1; and `PHASE` v. 2.1.

(a) *Simulation study 1* ($\boldsymbol{S}_1$)

In the first simulation study, we generated 150 independent genealogies of 50 chromosomes and simulated complete sequences, each for an interval of length 45 kb. We further assumed a constant $c = 1.13$ cM Mb^{-1}, which is the average recombination rate across the human genome estimated by segregation analysis on pedigrees (Kong *et al.* 2002). In order to conduct a realistic simulation study, given a simulated genealogy and samples, the first 10 markers that satisfied the SNP ascertainment criterion were taken as the sample for estimating ρ. Three marker ascertainment strategies were considered, including no ascertainment (no restriction on minor allele frequency (MAF), $\boldsymbol{S}_{1_A}$), MAF ≥ 0.05 ($\boldsymbol{S}_{1_B}$) and MAF ≥ 0.1 ($\boldsymbol{S}_{1_C}$). The length of the chromosome interval spanned by polymorphic markers is a random variable that varies across simulations, from a minimum of 2.15 kb to a maximum of 39.5 kb in $\boldsymbol{S}_1$.

For analyses using `InferRho`, the mode of the posterior distribution was used as the point estimate and the highest posterior density region was used as the credible set. The number of iterations was set to be 10^7, and the parameters were sampled every 500 iterations using the last 5×10^6 iterations. For analyses using `LDhat`, the program `pairwise` in the `LDhat` v. 2.1 package was used. The parameter θ per site was set to

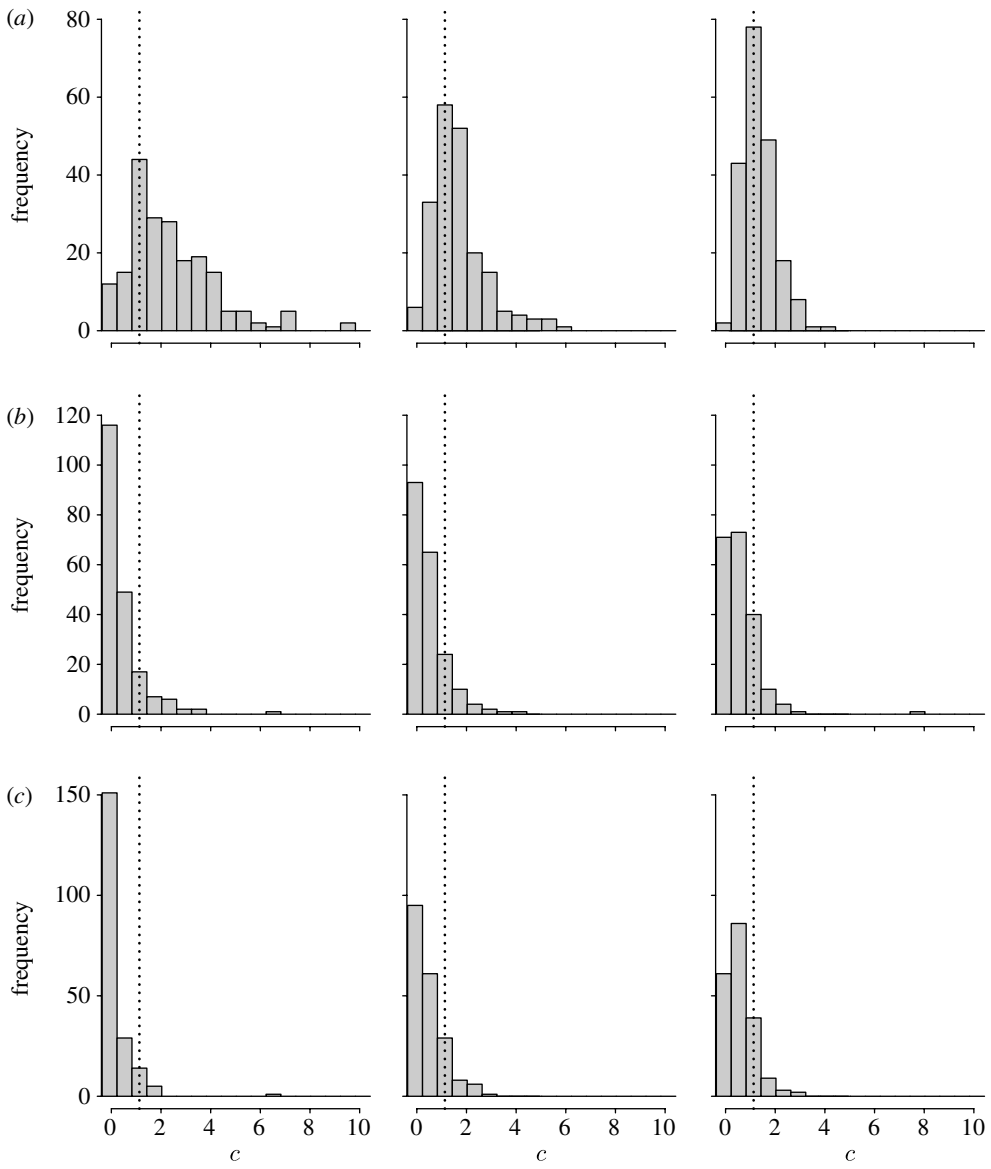

Figure 3. Distribution of estimated recombination rate ($\hat{c}$) generated by programs (*a*) `InferRho`, (*b*) `LDhat` and (*c*) `PHASE` using simulated dataset 1 ($\boldsymbol{S}_1$). The estimates are converted to cM Mb^{-1} for all three programs. The true c (=1.13 cM Mb^{-1}) used in the simulation is illustrated by a vertical dotted line.

the default value provided by the program (a finite-sites version of Watterson's estimate), max $4N_e r$ was set to be 100 and the number of points on the grid was set to be 201. For analyses using `PHASE`, the flag `-MR3` was used to indicate a constant recombination rate. The `-X10` and `-k999` options were also used. The first is recommended to increase the number of iterations of the final run by 10, and the second is for indicating that haplotype phases in the sample are known. The number of iterations, thinning intervals and burn-in were set to be 10 000, 100 and 10 000, respectively, which are 100 times larger than the default settings. As was recommended, the median of the results from file `*.out_recom` was taken as the point estimate. The distributions of the point estimates of ρ using datasets $\boldsymbol{S}_{1_A}$, $\boldsymbol{S}_{1_B}$, $\boldsymbol{S}_{1_C}$ obtained by use of the three programs are shown in figure 3.

(b) *Simulation study 2* ($\boldsymbol{S}_2$)

The second set of data (150 genealogies of 50 chromosomes) was simulated using the same simulation method and parameters as in $\boldsymbol{S}_1$, except that haplotypes were randomly paired to create multilocus genotypes of individuals. The datasets are labelled as $\boldsymbol{S}_{2_A}$, $\boldsymbol{S}_{2_B}$, $\boldsymbol{S}_{2_C}$, corresponding to the three ascertainment criteria described above. In our method, we integrate over haplotypes in the MCMC and haplotype phases are jointly estimated. The posterior distribution of haplotypes for each genotype in the sample is reported.

Two parallel chains were run for the program `InferRho`, and the number of iterations, thinning interval and burn-in parameters were 3×10^6, 500 and 3×10^6, respectively. For `LDhat`, the program `complete` was used for generating the likelihood lookup table and `pairwise` was used for obtaining estimates of ρ for each sample. Because it is computationally expensive to obtain the likelihood table, the number of points on the grid was set to be 101, which is the default value, and other parameters are the same as in $\boldsymbol{S}_1$. For `PHASE`, the `-k999` flag was removed but the remaining settings are the same as in $\boldsymbol{S}_1$. The distributions of the point estimates obtained using the three programs are shown in figure 4. The

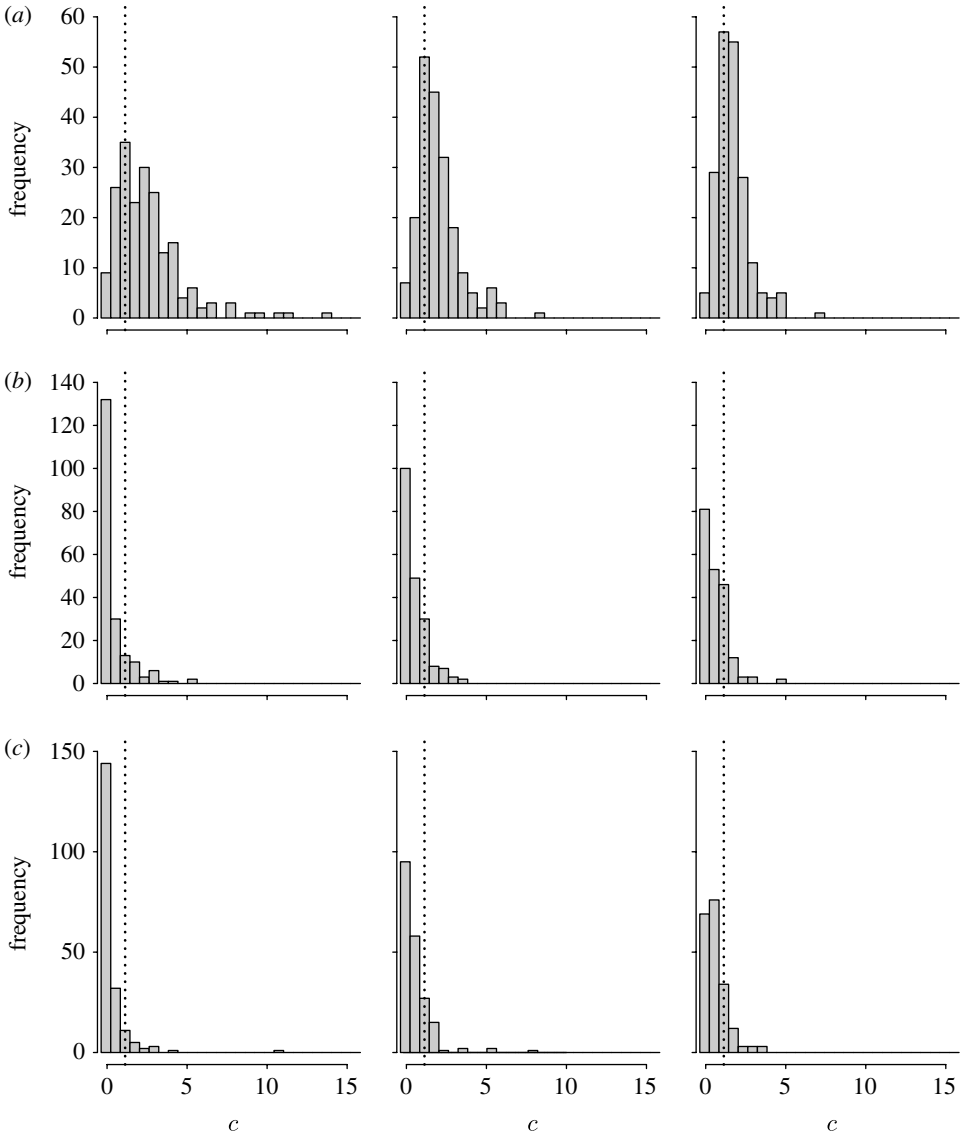

Figure 4. Distribution of estimated recombination rate ($\hat{c}$) generated by programs (*a*) `InferRho`, (*b*) `LDhat` and (*c*) `PHASE` using simulated dataset 2 ($\boldsymbol{S}_2$). The true c ($=1.13$ cM Mb^{-1}) used in the simulation is illustrated by a vertical dotted line.

average accuracy of haplotype phasing (the percentage of accurately phased genotypes in a dataset) across all three datasets obtained from `InferRho` and `PHASE` was 0.939 and 0.967, respectively.

(c) *Simulation study 3* ($\boldsymbol{S}_3$)

The third set of simulations assumes that one recombination hotspot exists near the centre of a 30 kb chromosomal interval, located at a position between 14 and 15.5 kb from the left of the interval. The strength of the hotspot is 25 cM Mb^{-1} and the background recombination rate is 0.5 cM Mb^{-1}. In total, 100 genealogies were simulated and polymorphic sites with MAF$\geq$0.05 were used for the analysis. In this simulation study, all of the polymorphic markers spanning the region were used, so the number of markers in the sampled haplotypes varied (e.g. from 16 to 64 in $\boldsymbol{S}_3$).

For analyses using `InferRho`, the number of chains, burn-in and number of iterations are the same as in $\boldsymbol{S}_2$. The options `-MR0`, `-X10` and `-k999` were used for `PHASE`, and other iteration parameters are the same as in $\boldsymbol{S}_2$. For analyses using `LDhat`, `complete` was used to generate the likelihood lookup table, and `interval` was used to estimate variable recombination rates. Default values for the iteration and penalty parameters `-its 10000000 -bpen 5 -samp 2000` were set for the program `interval`, and other parameters are the same as in $\boldsymbol{S}_2$. The program `stat` in the `LDhat` package was used for summarizing the results with option `-burn 50` (corresponding to 10^5 iterations) to obtain the point estimate and confidence interval of ρ.

Table 1. Summary of estimated c ($\hat{c}$) obtained by the three programs from simulated study 3 ($\boldsymbol{S}_3$). Results are based on 100 datasets and a total of 3394 estimates.

program	bias	mean square error	coverage	average width of 95% credible set
`InferRho`	−0.425	21.822	0.935	2.173
`LDhat`	0.787	102.192	0.331	2.669
`PHASE`	−0.885	18.436	0.961	4.390

The results from all three programs are summarized in table 1. Because the values of ρ_i vary between each

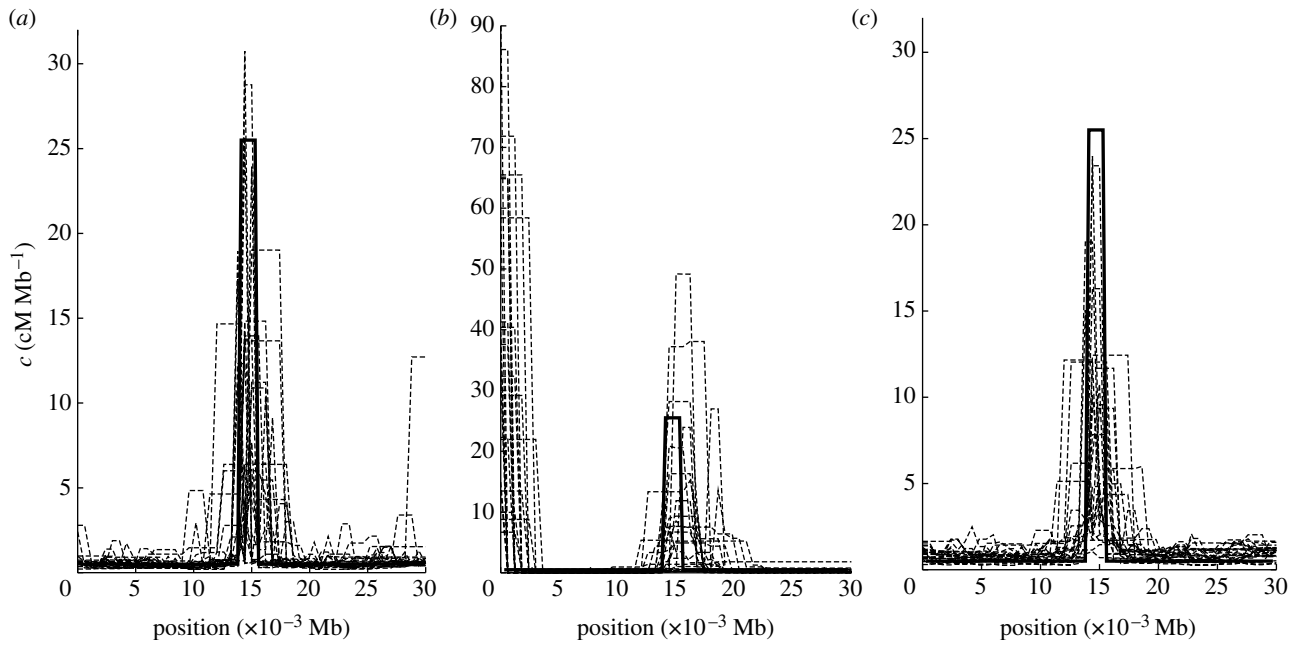

Figure 5. The plot of estimated recombination rate ($\hat{c}$) obtained from programs (*a*) `InferRho`, (*b*) `LDhat` and (*c*) `PHASE` using the first 25 samples in simulated dataset 3 ($\boldsymbol{S}_3$). The true c used in the simulation across the intervals is illustrated by a solid line, which is equal to 25.5 cM Mb^{-1} if the site is between 0.014 and 0.0155 Mb, and is equal to 0.5 cM Mb^{-1} otherwise.

marker, bias was estimated as

$$\mathrm{BIAS}(\hat{\rho}) = \frac{1}{\sum_{i=1}^{n} k^{(i)} - 1} \sum_{i=1}^{n} \sum_{j=0}^{k^{(i)}-2} \hat{\rho}_j^{(i)} - \rho_{j_T}^{(i)}, \tag{3.1}$$

where n denotes the number of datasets ($=100$ in this study); $k^{(i)}$ denotes the number of SNPs; and $\rho_{j_T}^{(i)}$ denotes the true ρ between markers j and $j+1$ in dataset i. Mean square error, coverage and the average width of 95 per cent credible set are calculated similarly. The estimates of ρ_j across the interval from the first 25 samples are plotted in figure 5.

4. ANALYSIS OF HUMAN LEUCOCYTE ANTIGEN AND MS32 REGIONS

We applied our method to two datasets from the human leucocyte antigen (HLA) and MS32 regions that have been previously studied by sperm typing (Jeffreys *et al.* 2001, 2005). The HLA dataset consists of 274 SNPs distributed across 0.216 Mb, sampled from 50 unrelated individuals. Six hotspots were revealed in the sperm-typing study (Jeffreys *et al.* 2001) and the data have been previously analysed using a composite-likelihood method (McVean *et al.* 2004). To reduce the computation time, the region is divided into sub-regions (each with 20 markers), although it is feasible to analyse the entire region simultaneously. The variable recombination model described above was used and the results obtained from `InferRho` are shown in figure 6. The estimates of the centres of the hotspots discovered by sperm typing are also indicated. Recombination rates obtained using population genetic data include both female and male rates, sex averaged over many generations, and only the population size-scaled recombination rate can be obtained (population size may vary across region of genomes due to the effects of selection, demographic events, etc.); we therefore would not expect the inferred intensities of hotspots from sperm-typing and population genetic inferences to agree completely. Nonetheless, the inferred locations of the hotspots obtained from sperm typing versus our method are very close. This is consistent with the findings from a recent study suggesting that recombination hotspot locations are similar between males and females (Coop *et al.* 2008).

Jeffreys *et al.* (2005) investigated recombination rates in the MS32 and surrounding region by both sperm-typing and coalescent analysis of genotypes (recombination rate estimated using `LDhat` and `PHASE`). The MS32 dataset consists of 206 SNPs sampled from 80 individuals and distributed across 0.206 Mb. For our analysis using `InferRho`, the region was again divided into sub-regions, each with 20 markers, and the results are shown in figure 7. The estimates are in general consistent with those obtained from sperm crossover analysis. In particular, hotspots MS32, MSTM1a, MSTM1b and MSTM2 discovered by sperm typing, which are only weakly evident from a `LDhat` and `PHASE` analysis (fig. 1*b* of Jeffreys *et al.* 2005), are very intense when these data are analysed by our method.

5. DISCUSSION

Here, we present a new method for estimating recombination rate using a population sample of either haplotypes or genotypes. The genealogy of the sampled chromosomes is represented by a SNP genealogy, which is composed only of lineages carrying sites ancestral to the sample. The chromosome intervals in the SNP genealogies are represented by MA vectors. By using the SNP genealogies as a prior on the genealogy relating the sampled haplotypes, full-likelihood estimation of recombination becomes feasible. If little information is available about the recombination rate for the sampled interval, an uninformative prior on

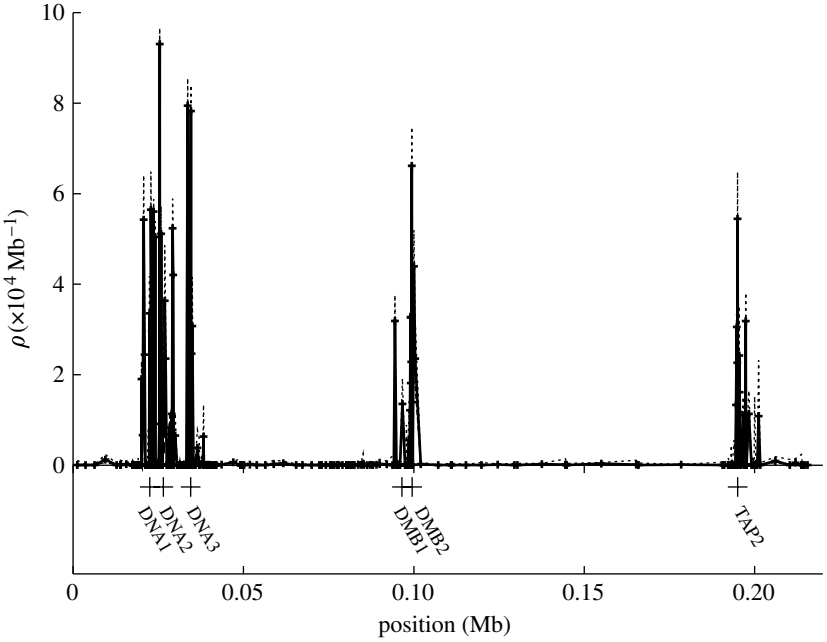

Figure 6. Estimated recombination rate generated by `InferRho` using the HLA dataset. The solid line shows the point estimate (the mode of the posterior distribution) and the dashed lines indicate the low and high bounds of 95% credible set. The plus signs indicate the positions of the centre of a recombination hotspot obtained from the sperm-typing study (Jeffreys *et al.* 2001).

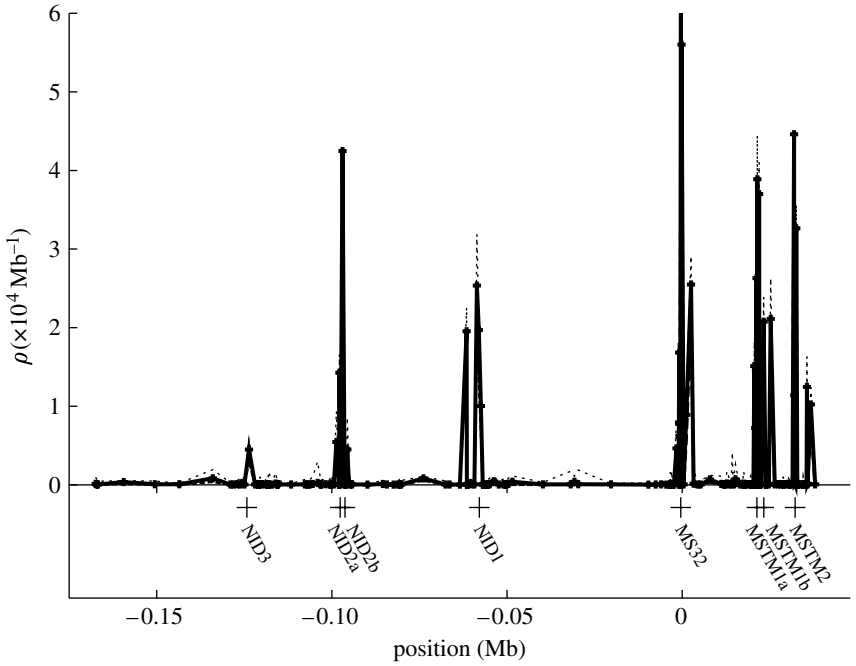

Figure 7. Estimated recombination rate generated by `InferRho` using the MS32 dataset. The solid line shows the point estimate (the mode of the posterior distribution) and the dashed lines indicate the low and high bounds of 95% credible set. The plus signs indicate the positions of the centre of a recombination hotspot obtained from the sperm-typing study (Jeffreys *et al.* 2005).

recombination rate can be used in the analysis. On the other hand, information on recombination rate obtained from other independent studies can be incorporated into the analysis through a more informative prior to obtain refined estimates of recombination rate, or possibly to address questions that cannot be obtained from either study alone (e.g. estimating female recombination rate by combining estimates of male recombination rate obtained from sperm typing and estimates of the sex-averaged recombination rate obtained by population genetic methods). The posterior distribution of recombination rates is approximated by a reversible-jump MCMC (RJMCMC). In the Metropolis–Hastings (MH) algorithm, proposed changes include modifying the SNP genealogy by changing a local topology or by adding (or removing) a pair of recombination and coalescent nodes, modifying ancestral alleles, modifying haplotypes (if the phase of the data is unknown), modifying alleles at sites with missing alleles in the sample and modifying the parameters θ and ρ. The method is implemented in the package `InferRho`.

Three simulation studies were conducted to evaluate our method and to compare its performance with two other approximate-likelihood methods, including a composite-likelihood method and a PAC likelihood method, implemented in packages `LDhat` and `PHASE`, respectively. The first (assuming haplotypes in the sample) and the second (assuming genotypes in the sample) simulation studies are aimed to examine the performance of three methods in the cases that the size of a chromosomal interval is relatively small and the recombination rate is constant across the interval. In this case, information in the data is limited due to the small interval size (or the small number of polymorphic sites), but all of the sites contribute to the estimate of ρ since a constant rate model is assumed. Three marker ascertainment criteria were considered. All three methods performed better as the number of informative markers increased (figures 3 and 4). As expected, the variance of the estimates of recombination rate obtained using genotypic data is larger than those obtained using haplotypes. The mode of the empirical distribution is centred on the true parameter value for our new method while the other methods have a mode near zero.

The third simulation study examined the performance of the methods when a recombination hotspot exists on an interval. The simulation scenario was chosen according to the recombination hotspot and background rate patterns obtained from previous sperm-typing studies. Because the number of markers on the haplotypes for each sample varies, samples with fewer polymorphic markers may contain less information about the recombination rate and hotspot. In general, more markers were more likely to yield accurate estimates of the location of a recombination hotspot. Table 1 shows the summary of the estimates of recombination rate across the region using all 100 simulated samples, and figure 5 shows the plots of recombination rate using the first 25 simulated samples obtained by use of the three methods. Our method has the smallest credible set with a coverage (0.935) close to the nominal value (0.95). In terms of MSE, `InferRho` and `PHASE` were very similar and both had MSE much lower than `LDhat`. The improved statistical performance of `InferRho` may be due to the fact that it uses the full likelihood and can therefore extract greater information from the data.

The fine-scale distribution of recombination rates in the HLA and MS32 regions has been studied previously by sperm typing. Although the sperm-typing studies only provide recombination estimates from males, the distribution of recombination hotspots obtained using population samples via our method is mostly consistent with those estimated by sperm typing. However, the intensity of the recombination hotspots obtained by the two different approaches is not completely in agreement. Reasons for this might be the differences in female and male recombination rates *per se*. Since only population size-scaled recombination rate ($\rho = 4N_e c \times 0.01$) can be obtained from population genetic approaches, the difference in the strengths of recombination hotspots might also be due to the variation in N_e across the region owing to natural selection, migration, etc. This could lead to a larger or smaller ρ for population genetic inferences versus sperm typing even though c is identical in both cases.

For the simulation studies presented in this paper, a large number of iterations and multiple Metropolis-coupled MCMC (MCMCMC) chains were used to ensure convergence and to avoid the necessity of checking each run individually. The computational time needed depends on the data and the number of chains used. The use of MCMCMC notably improved the mixing of the chains. For the results presented in this paper, the data were analysed multiple times, and the results were highly consistent. The convergence was also analysed using the Bayesian output analysis (BOA) package of R (Smith 2005). As a general guide, 2×10^6 burn-in and 2×10^6 iterations can be used initially, although convergence should be examined using a post-chain analysis package such as BOA. Because our program allows the final state of the chain to be saved, more iterations can be added to previous runs if needed to ensure convergence. The program `InferRho` is available from http://rannala.org.

This research was supported by NIH grant HG01988 to B.R.

APPENDIX A

(a) *MH algorithm*

A MH algorithm was used to evaluate the posterior distribution of parameters described in equation (2.2). The MH algorithm has two steps: (i) the 'proposal' step in which potential new parameter values are simulated from the proposal density $q(\boldsymbol{\Theta}'|\boldsymbol{\Theta})$, and (ii) the 'acceptance' step in which the proposed values are accepted with probability α or rejected with probability $1-\alpha$. If accepted, $\boldsymbol{\Theta}'$ becomes the current state in the chain, otherwise the chain remains at $\boldsymbol{\Theta}$ and $\boldsymbol{\Theta}'$ is discarded. The acceptance probability is

$$\alpha = \min\left\{1, \frac{L(\boldsymbol{\Theta}'|\boldsymbol{X})}{L(\boldsymbol{\Theta}|\boldsymbol{X})} \times \frac{q(\boldsymbol{\Theta}|\boldsymbol{\Theta}')}{q(\boldsymbol{\Theta}'|\boldsymbol{\Theta})} \times \frac{f(\boldsymbol{\Theta}')}{f(\boldsymbol{\Theta})}\right\}.$$

(b) *Modifying the SNP genealogy* ($G_S \rightarrow G'_S$)

There are three possible types of proposals for changes to G_S: (i) changing the local topology of the graph, and either (ii) adding, or (iii) removing, a pair of recombination and coalescent events.

(i) *Local topology rearrangements*

To propose a change to the local topology, an internal node (Y_i) is randomly chosen. If Y_i is a recombination node, there will be two ancestors (denoted by $Y_{i_{A1}}$ and $Y_{i_{A2}}$) connecting with Y_i. With equal probability, one of the two ancestors is chosen to be the node that remains attached to Y_i (e.g. $Y_{i_{A1}}$ is chosen and its time is denoted by t_1). A new waiting time for Y_i is proposed, t', that is uniformly distributed between 0 and t_1. The new branch location for the node is chosen uniformly among all 'eligible' branches extant at time t'. The acceptance probability for the proposed change is

$$\min\left[1, \frac{f(\boldsymbol{X}, \boldsymbol{Y}|G'_S, \theta)}{f(\boldsymbol{X}, \boldsymbol{Y}|G_S, \theta)} \times \frac{n'_{br}}{n_{br}} \times \frac{f(G'_S|\boldsymbol{\rho})}{f(G_S|\boldsymbol{\rho})}\right],$$

where the first term of the product is the likelihood ratio of the proposed versus the current ancestral

graph; the second term is the proposal ratio of the move, where n_{br} and n'_{br} represent the number of eligible branches at times t and t', respectively; and the third term is a ratio of the prior probabilities of the proposed, G'_{S}, versus current, G_{S}, ancestral graphs given the current values of the model parameters, $\boldsymbol{\rho}$. Note that the eligible branches are determined by both the waiting time of the node and the MA vector for the branches (e.g. the MA vector must have more than one non-zero element, etc.). The MA vector is then updated according to the new topology.

If, on the other hand, Y_i is a coalescent node, it will connect with two descendants (denoted by $Y_{i_{\mathrm{D1}}}$ and $Y_{i_{\mathrm{D2}}}$). One of the two descendants is chosen to be the branch that remains attached to Y_i (e.g. $Y_{i_{\mathrm{D1}}}$ is chosen and its time is denoted by t_2). The new waiting time for Y_i is $t' = t_2 + \delta_{t'}$, where $\delta_{t'}$ is exponentially distributed with rate λ and the branch on which the node is placed is again chosen uniformly from among the eligible branches, n'_{br}, at time t'. The acceptance probability is

$$\min\left[1, \frac{f(\boldsymbol{X},\boldsymbol{Y}|G'_{\mathrm{S}},\theta)}{f(\boldsymbol{X},\boldsymbol{Y}|G_{\mathrm{S}},\theta)} \times \frac{\exp\{-\lambda(t-t_2)\}}{\exp\{-\lambda\delta_{t'}\}} \times \frac{n'_{\mathrm{br}}}{n_{\mathrm{br}}} \times \frac{f(G'_{\mathrm{S}}|\boldsymbol{\rho})}{f(G_{\mathrm{S}}|\boldsymbol{\rho})}\right].$$

(ii) *Adding (or removing) recombination and coalescence events*

Because the number of parameters changes with the addition or removal of a recombination or coalescent event, a RJMCMC scheme is used (Green 1995). When adding a recombination and coalescence node to the genealogy, three additional variables, u, v and w, are simulated, with $\{u,v\} \sim \mathrm{Uniform}(0, t_{\mathrm{H}})$, where t_{H} denotes the height of the genealogy. We let $t'_{\mathrm{r}} = \min\{u,v\}$, $t'_{\mathrm{c}} = \max\{u,v\}$, where t'_{r} and t'_{c} denote the waiting times of the new recombination and coalescent nodes, respectively. Given the ages of the new nodes, their positions on the graph are then randomly chosen among all eligible branches for each node. The two new nodes are inserted into their chosen positions and a new branch is added connecting them. Variable w is chosen uniformly on the interval of length $\sum_{j=L+1}^{R} d_j$, where L and R define the ancestral segment at the recombination node. The recombination breakpoint, denoted by s', is set equal to w. The acceptance probability is

$$\min\left[1, \frac{f(\boldsymbol{X},\boldsymbol{Y}|G'_{\mathrm{S}},\theta)}{f(\boldsymbol{X},\boldsymbol{Y}|G_{\mathrm{S}},\theta)} \times \frac{f(G_{\mathrm{S}}|G'_{\mathrm{S}})}{f(G'_{\mathrm{S}}|G_{\mathrm{S}})} \times \frac{f(G'_{\mathrm{S}}|\boldsymbol{\rho})}{f(G_{\mathrm{S}}|\boldsymbol{\rho})} \times \left|\frac{\partial(G'_{\mathrm{S}})}{\partial(G_{\mathrm{S}},u,v,w)}\right|\right],$$

where

$$\frac{f(G_{\mathrm{S}}|G'_{\mathrm{S}})}{f(G'_{\mathrm{S}}|G_{\mathrm{S}})} = \frac{T_{\mathrm{H}}^2 n'_{\mathrm{br}_1} n'_{\mathrm{br}_2} (l_2 - l_1)}{2n_{\mathrm{pair}}},$$

and the Jacobian term is 1 for the above described moves. Note that n_{pair} denotes all candidate recombination–coalescence pairs that can be deleted if removing a recombination and a coalescent.

(c) ***Modifying ancestral states, haplotype phases, missing data and parameters*** ($\boldsymbol{Y} \rightarrow \boldsymbol{Y}'$ ***and*** $\boldsymbol{\Theta} \rightarrow \boldsymbol{\Theta}'$)

For SNP haplotypes or genotypes, the ancestral states are proposed based on a discrete uniform distribution. At any site, the ancestral allele can change to any state with probability 0.25. First, a marker site is chosen uniformly among all sites (e.g. s_i), and the genealogical tree (τ_i) for marker s_i is obtained from G_{S}. A new set of ancestral states on τ_i is generated by either proposing new states or keeping the current states based on a pre-specified frequency parameter, which is used to adjust the proportion of accepted moves. The move is accepted with probability

$$\min\left[1, \frac{f(\boldsymbol{X}^{\{i\}}, \boldsymbol{Y}^{\{i\}'}|\tau_i,\theta)}{f(\boldsymbol{X}^{\{i\}}, \boldsymbol{Y}^{\{i\}}|\tau_i,\theta)}\right].$$

Missing data if they exist are proposed similarly. If genotypic data are applied, haplotypes are integrated by switching alleles of genotypes in the chains.

New parameters ρ and θ are proposed based on a sliding window with a reflecting boundary at 0 (e.g. Rannala & Yang 2003). The proposed parameters are accepted with probabilities

$$\min\left[1, \frac{f(G_{\mathrm{S}}|\boldsymbol{\rho}')}{f(G_{\mathrm{S}}|\boldsymbol{\rho})} \times \frac{f(\boldsymbol{\rho}')}{f(\boldsymbol{\rho})}\right]$$

and

$$\min\left[1, \frac{f(\boldsymbol{X},\boldsymbol{Y}|G_{\mathrm{S}},\theta')}{f(\boldsymbol{X},\boldsymbol{Y}|G_{\mathrm{S}},\theta)} \times \frac{f(\theta')}{f(\theta)}\right].$$

(d) ***Checking the MCMC algorithm***

The accuracy of the MH algorithm was evaluated by examining the stationary distribution when the chain is run with no data (e.g. with a constant likelihood ratio). By using a constant ρ, the distribution of the number of recombinations and the height of the SNP genealogy, generated by the MCMC algorithm can be compared with those obtained by straightforward Monte Carlo simulation. The marginal distribution of coalescent trees can also be checked against analytic expectations.

(e) ***Likelihood***

Since the number of ancestral events in the sampled ancestral graph varies, and some of the ancestral events do not contribute to the likelihood calculation, the coalescent trees ($\boldsymbol{\tau}$) for each marker are first obtained based on the sampled ancestral graph (figure 1). The F81 mutation model (Felsenstein 1981) is assumed, although more general models can be easily incorporated. Given the genealogical tree τ_i for a marker site i, and conditional on one or more mutations having occurred, the likelihood is calculated by

$$\frac{\prod_{l\in\tau_i} \mathrm{I}\left(\boldsymbol{D}_l^{\{i\}} = \boldsymbol{D}_{l_A}^{\{i\}}\right)\left[(1-\mathrm{e}^{-(\theta t_l/2)})\pi_{\boldsymbol{D}_l^{\{i\}}} + \mathrm{e}^{-(\theta t_l/2)}\right] + \mathrm{I}\left(\boldsymbol{D}_l^{\{i\}} \neq \boldsymbol{D}_{l_A}^{\{i\}}\right)\left[(1-\mathrm{e}^{-\theta t_l/2})\pi_{\boldsymbol{D}_l^{\{i\}}}\right]}{1-\exp\left(-\frac{\theta T_i}{2}\right)},$$

where l indicates a branch in τ_i with length t_i, connecting $\boldsymbol{D}_l^{\{i\}}$ and $\boldsymbol{D}_{l_A}^{\{i\}}$, with $\boldsymbol{D}_l^{\{i\}}$ being the ancestral SNP of $\boldsymbol{D}_l^{\{i\}}$, and $\boldsymbol{D}^{\{i\}} = \{\boldsymbol{X}^{\{i\}}, \boldsymbol{Y}^{\{i\}}\}$. The likelihood is calculated across all branches in τ_i. Parameter T_i represents the total branch lengths in τ_i.

(f) *Metropolis-coupled MCMC (or (MC)^3)*

MCMCMC is a technique to improve the mixing of chains in MCMC methods (Geyer 1991), and has been used in Bayesian phylogenetic inferences (Altekar *et al.* 2004). The technique is particularly useful when the stationary distribution of interest has multiple modes by allowing large-step moves. This technique is also very helpful for improving mixing in our application since larger steps are frequently rejected due to the property of SNP genealogies, such that small changes in the MA vector on one node can result in large subsequent changes in the ancestry. Assuming m chains are run in parallel, for each chain, the MH algorithm is similar to the above method except that the acceptance, or rejection, probability is affected by the 'heating' parameter (β_i for chain i), and $\beta_1 = 1$ (Altekar *et al.* 2004). The swap of any two chains i and j is accepted with probability

$$\min\left[1, \left(\frac{f(\boldsymbol{\Theta}_i|\boldsymbol{X})}{f(\boldsymbol{\Theta}_j|\boldsymbol{X})}\right)^{\beta_j}\left(\frac{f(\boldsymbol{\Theta}_j|\boldsymbol{X})}{f(\boldsymbol{\Theta}_i|\boldsymbol{X})}\right)^{\beta_i}\right].$$

REFERENCES

Altekar, G., Dwarkadas, S., Huelsenbeck, J. P. & Ronquist, F. 2004 Parallel metropolis coupled Markov chain Monte Carlo for Bayesian phylogenetic inference. *Bioinformatics* **20**, 407–415. (doi:10.1093/bioinformatics/btg427)

Arnheim, N., Calabrese, P. & Tiemann-Boege, I. 2007 Mammalian meiotic recombination hot spots. *Annu. Rev. Genet.* **41**, 369–399. (doi:10.1146/annurev.genet.41.110306.130301)

Coop, G., Wen, X. Q., Ober, C., Pritchard, J. K. & Przeworski, M. 2008 High-resolution mapping of crossovers reveals extensive variation in fine-scale recombination patterns among humans. *Science* **319**, 1395–1398. (doi:10.1126/science.1151851)

Crawford, D. C., Bhangale, T., Li, N., Hellenthal, G., Rieder, M. J., Nickerson, D. A. & Stephens, M. 2004 Evidence for substantial fine-scale variation in recombination rates across the human genome. *Nat. Genet.* **36**, 700–706. (doi:10.1038/ng1376)

Fearnhead, P. & Donnelly, P. 2001 Estimating recombination rates from population genetic data. *Genetics* **159**, 1299–1318.

Fearnhead, P. & Donnelly, P. 2002 Approximate likelihood methods for estimating local recombination rates. *J. R. Stat. Soc. Ser. B-Stat. Methodol.* **64**(Pt 4), 657–680. (doi:10.1111/1467-9868.00355)

Felsenstein, J. 1981 Evolutionary trees from DNA-sequences—a maximum-likelihood approach. *J. Mol. Evol.* **17**, 368–376. (doi:10.1007/BF01734359)

Geyer, C. J. 1991 Markov chain Monte Carlo maximum likelihood. In *Computing Science and Statistics: Proc. 23rd Symp. on the Interface* (ed. E. M. Keramides), pp. 156–163. Fairfax Station, VA: Interface Foundation.

Graffelman, J., Balding, D. J., Gonzalez-Neira, A. & Bertranpetit, J. 2007 Variation in estimated recombination rates across human populations. *Human Genet.* **122**, 301–310. (doi:10.1007/s00439-007-0391-6)

Green, P. J. 1995 Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82**, 711–732. (doi:10.1093/biomet/82.4.711)

Griffiths, R. C. & Marjoram, P. 1996 Ancestral inference from samples of DNA sequences with recombination. *J. Comput. Biol.* **3**, 479–502.

Hellenthal, G. & Stephens, M. 2006 Insights into recombination from population genetic variation. *Curr. Opin. Genet. Dev.* **16**, 565–572. (doi:10.1016/j.gde.2006.10.001)

Hudson, R. R. 1990 Gene genealogies and the coalescent process. *Oxf. Surv. Evol. Biol.* **7**, 1–44.

Hudson, R. R. 2001 Two-locus sampling distributions and their application. *Genetics* **159**, 1805–1817.

Jeffreys, A. J., Kauppi, L. & Neumann, R. 2001 Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. *Nat. Genet.* **29**, 217–222. (doi:10.1038/ng1001-217)

Jeffreys, A. J., Neumann, R., Panayi, M., Myers, S. & Donnelly, P. 2005 Human recombination hot spots hidden in regions of strong marker association. *Nat. Genet.* **37**, 601–606. (doi:10.1038/ng1565)

Kingman, J. 1982*a* The coalescent. *Stoch. Process. Appl.* **13**, 235–248. (doi:10.1016/0304-4149(82)90011-4)

Kingman, J. 1982*b* On the genealogy of large populations. *J. Appl. Probab.* **19A**, 27–43. (doi:10.2307/3213548)

Kong, A. *et al.* 2002 A high-resolution recombination map of the human genome. *Nat. Genet.* **31**, 241–247.

Kuhner, M. K., Yamato, J. & Felsenstein, J. 2000 Maximum likelihood estimation of recombination rates from population data. *Genetics* **156**, 1393–1401.

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

McVean, G., 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. (doi:10.1126/science.1092500)

Myers, S., Bottolo, L., Freeman, C., McVean, G. & Donnelly, P. 2005 A fine-scale map of recombination rates and hotspots across the human genome. *Science* **310**, 321–324. (doi:10.1126/science.1117196)

Nielsen, R. 2000 Estimation of population parameters and recombination rates from single nucleotide polymorphisms. *Genetics* **154**, 931–942.

Rannala, B. & Slatkin, M. 2000 Methods for multipoint disease mapping using linkage disequilibrium. *Genet. Epidemiol.* **19**, S71–S77. (doi:10.1002/1098-2272(2000)19:1+<::AID-GEPI11>3.0.CO;2-D)

Rannala, B. & Yang, Z. H. 2003 Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. *Genetics* **164**, 1645–1656.

Smith, B. J. 2005 *Bayesian output analysis program (BOA)*, v. 1.1.5. See http://www.public-health.uiowa.edu/boa.

Smith, N. G. C. & Fearnhead, P. 2005 A comparison of three estimators of the population-scaled recombination rate: accuracy and robustness. *Genetics* **171**, 2051–2062. (doi:10.1534/genetics.104.036293)

Stumpf, M. P. H. & McVean, G. A. T. 2003 Estimating recombination rates from population-genetic data. *Nat. Rev. Genet.* **4**, 959–968. (doi:10.1038/nrg1227)