Population genomic inference of recombination rates and hotspots

Ying Wang[1] and Bruce Rannala

Genome Center and Department of Evolution and Ecology, University of California, Davis, CA 95616

Edited by Jasper Rine, University of California, Berkeley, CA, and approved February 12, 2009 (received for review January 14, 2009)

As more human genomic data become available, fine-scale recombination rate variation can be inferred on a genome-wide scale. Current statistical methods to infer recombination rates that can be applied to moderate, or large, genomic regions are limited to approximated likelihoods. Here, we develop a Bayesian full-likelihood method using Markov Chain Monte Carlo (MCMC) to estimate background recombination rates and hotspots. The probability model is inspired by the observed patterns of recombination at several genomic regions analyzed in sperm-typing studies. Posterior probabilities and Bayes factors of recombination hotspots along chromosomes are inferred. For moderate-size genomic regions (e.g., with <100 SNPs), the full-likelihood method is used. Larger regions are split into subintervals (typically each having between 20 and 50 markers). The likelihood is approximated based on the genealogies for each subinterval. The background recombination rates, hotspots, and parameters are evaluated by using a parallel computing approach and assuming shared parameters across the subintervals. Simulation analyses show that our method can accurately estimate the variation in recombination rates across genomic regions. In particular, clusters of hotspots can be distinguished even though weaker hotspots are present. The method is applied to SNP data from the HLA region, the MS32, and chromosome 19.

ancestral recombination graph | Bayesian inference | linkage disequilibrium | Markov Chain Monte Carlo | recombination hotspot

Describing fine-scale recombination rate variation and the distribution of recombination hotspots across chromosomes are important goals in population genetics. Recombination is one of the fundamental evolutionary forces affecting patterns of polymorphism variation across genomes. The distribution of recombination rates and hotspots helps to reveal the molecular basis of meiotic cross-overs and is a crucial factor in association study designs because of its effect on the pattern of linkage disequilibrium in human genomes (1). Traditional linkage data from pedigrees typically do not provide estimates of recombination on a fine scale because of the limited number of meioses (2, 3). Sperm-typing analyses can experimentally provide estimates of recombination rates but are laborious and expensive, so only a few regions of the human genome have been studied. In addition, only recombination rates in males can be revealed from these studies (reviewed in ref. 4). Statistical inferences of recombination rates based on population genetic data represent a major approach to obtain an overall picture of fine-scale recombination rates and hotspot locations in the human genome, especially at present, with more population genomic data become available daily [e.g., data generated by the HapMap project (5), etc.].

Both sperm-typing analyses and statistical inferences based on population genetic data (typically SNPs) reveal similar patterns of fine-scale recombination rates across chromosomes. These studies suggest that recombination rates vary significantly over genomic regions, and most recombination tends to occur in particular regions of chromosomes (with interval sizes of $\approx$1–2 kb) known as recombination "hotspots," whereas in other "background" regions, many fewer recombinations occur (4, 6).

A number of statistical methods have been developed to estimate recombination rates by using population genetic data, including those using summary statistics, full likelihoods, and approximated likelihoods (reviewed in refs. 7 and 8). Full-likelihood methods use all information contained in the data and, in principle, should provide more accurate estimates. However, because full-likelihood methods based on the Ancestral Recombination Graph (ARG) (9, 10) involve integrating a large number of variable dimension genealogies, it has been challenging to develop efficient methods based on full likelihoods that are applicable to large-scale data (10–13). Several methods based on approximate likelihoods have therefore been developed (14–16) and applied to human genomic data (17–19). We recently developed a full-likelihood Bayesian Markov Chain Monte Carlo (MCMC) method for estimating fine-scale recombination rates (20). In our method, genealogies underlying a sampling of chromosomes are effectively modeled by using marginal individual SNP genealogies related through an ARG. Simulation studies showed that our full-likelihood method performed well under different simulation scenarios and can be applied to small-to-moderate-size chromosomal intervals (e.g., with $\leq$100 SNPs).

Several methods for detecting recombination hotspots have also been developed that search for regions of accelerated recombination rates by comparison with surrounding regions or with overall background rates (16, 18, 21, 22). Auton and McVean (24) recently incorporated a model of hotspots and background recombination into the LDhat package to simultaneously estimate fine-scale recombination rates and detect recombination hotspots. Our method differs from Auton and McVean's composite likelihood method in several ways. Most importantly, for moderate-size genomic regions (e.g., $\leq$100 SNPs), the posterior probability of recombination rates is obtained by a full-likelihood method. If genomic regions are larger, they are divided into n-marker subregions (typically, an appropriate choice for n is between 20 and 50). The likelihood is then calculated conditional on the genealogies for each subregion, and parameters are evaluated jointly across all subregions.

Here, we present a model of recombination rates and hotspots whose design is based on the observed distribution of recombination hotspots at several genomic regions obtained from sperm-typing studies (4, 6). The background rates between SNPs are assumed to be independently distributed following a Γ-distribution. Piecewise estimators of recombination rate change have been developed that accommodate recombination hotspots (18, 23). Auton and McVean (24) presented a model in which

Author contributions: Y.W. and B.R. designed research; Y.W. performed research; B.R. contributed new reagents/analytic tools; Y.W. analyzed data; and Y.W. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence should be addressed. E-mail: ygwang@ucdavis.edu.

This article contains supporting information online at www.pnas.org/cgi/content/full/0900418106/DCSupplemental.

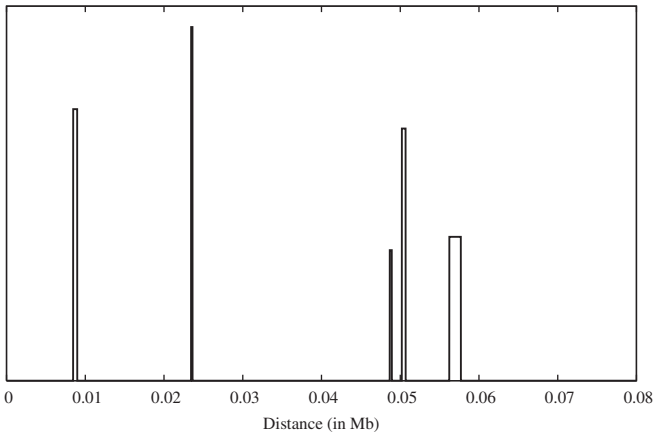

Fig. 1. An example of recombination hotspots simulated by using the 2-parameter Markov process model.

recombination hotspots are uniformly scattered over the region being analyzed.

In our hotspot model, a 2-parameter Markov process is used to describe the distribution of the intervals between hotspots and the duration of hotspots. Both duration of hotspots and the distances between hotspots are exponentially distributed. One feature of the distribution of recombination hotspots that has been revealed by sperm typing analyses is that hotspots are sometimes clustered. The advantage of using an exponential distribution to describe the distances between hotspots is that the mode of the distribution is 0, and the variance is large. Thus, scenarios in which hotspots are far apart and in which they are clustered can be accommodated. An example of a pattern of recombination hotspots generated by using the model is given in Fig. 1. The recombination rates across chromosomal regions are a combination of the 2 independent processes (the summation of the rates from the background rate and the hotspots rates).

A reversible jump MCMC scheme is used to estimate background recombination rate and hotspots by using a population sample of SNPs from regions of the genome (25). The position and intensity of hotspots, background recombination rates, and other parameters are sampled from the Markov chains. The chromosomal intervals are divided into bins (e.g., with size 100 bp) for estimating the posterior probability and the Bayes factor (BF) that each interval contains a hotspot and the average intensity of the hotspot in each interval. To identify hotspots, a hotspot is defined by 2 BF thresholds: HT_1 and HT_2. If BF (hotspot at location i) $> HT_1$, the local mode is used to estimate the modal BF of the hotspot, and the hotspot extends until BF (hotspot at location j) $< HT_2$. The hotspot is then inferred to be on the interval (i, j). Parameter HT_1 represents the criterion for detecting hotspots. Larger HT_1 implies higher confidence that the identified hotspots are true. Parameter HT_2 determines the boundaries, and thus the width, of an estimated hotspot conditional on HT_1. The power and type I error rates can be adjusted by modifying values of HT_1 and HT_2.

Results

We examined the performance of our method by applying it to both simulated data and human population genetic data. Population genetic datasets spanning the HLA (26) and MS32 (27) regions that have been previously studied by sperm cross-over analysis were analyzed by our method and compared with previous results. The method was also applied to a SNP dataset across human chromosome 19 sampled from the African-American population (28).

Table 1. Summary of the statistical performance of the IR program in the simulation study

Parameter	Statistical criteria	Dataset S_1	Dataset S_2
Hotspot	False positive	5 of 100	4 of 100
	Power	0.92	0.74
	Average width	1.83	1.58
	Average intensity	12.70	9.56
	MSE of intensity	781.02	28.86
Background rate	Average	0.094	0.127
	MSE	0.005	0.014
	Coverage	0.95	0.92
	Average with of 95% CS	0.177	0.242

Hotspots thresholds $HT_1 = 5$ and $HT_2 = 2.5$ were used.

Simulation Studies. To evaluate the statistical performance of the method, we used the msHOT program (29, 30) to simulate 3 sets of data. Common parameters used in all simulations include a sample size of 50 chromosomes, a population size (N_e) equal to 10^4, a mutation rate per site per generation (μ) equal to 10^{-8}, a background recombination rate of 0.15 cM/Mb ($\rho = 0.06$/kb, given $N_e = 10^4$), and a chromosomal interval of size 30 kb. Only sites with minor allele frequencies (MAF) ≥0.05 were retained and used in the analyses.

We first examined the performance of the method by considering hotspots with 2 different intensities: $\rho = 40$/kb (for dataset $\mathbf{S}_1$) and $\rho = 10$/kb (for dataset $\mathbf{S}_2$), representing a relatively strong and a weaker recombination hotspot. For all 3 datasets, the location of the hotspots is assumed to be the same (at a position between 15 and 16.5 kb from the left of the interval). The average [minimum, maximum] number of SNPs for $\mathbf{S}_1$ and $\mathbf{S}_2$ are 34.58 [14, 73] and 34.22 [12, 64], respectively.

The BF of recombination hotspot locations and other parameters of the model, obtained by using the program IR, were reported. Different hotspot threshold values, HT_1 and HT_2, were considered to examine how the false-positive rate, power, and the estimated hotspot intensity and width change. Here, a hotspot is counted as correct when the estimated hotspots overlapped with a true hotspot; otherwise, it is counted as incorrect. The false-positive rate is the number of incorrect hotspots over the number of intervals examined, and the power is the percentage of successfully identified hotspots over the number of true hotspots.

As expected, both the power and the false-positive rates increase as HT_1 decreases [see supporting information (SI) Fig. S1 *A* and *B*]. The width of the estimated hotspots is determined by HT_2 given HT_1. In general, HT_2 can be assumed to be 2.5, and the estimated widths are approximately consistent for different HT_1 values (Fig. S1*C*). The results assuming $HT_1 = 5$ and $HT_2 = 2.5$ are summarized in Table 1. When a hotspot is strong, the estimated intensity tends to be underestimated, because the genealogical trees are independent due to the large number of recombinations, as was pointed out by other authors (24).

The second simulation study aimed to examine the ability of the method to identify clustered hotspots. A sperm-typing analysis of the HLA region (15) revealed 2 clusters of hotspots: DNA1-3 and DMB1-2. The distance between hotspot centers is 4.01 kb for DNA1 and DNA2, 7.97 kb for DNA2 and DNA3, and 3.25 kb for hotspots DMB1 and DMB2. In our simulation study, we assumed the centers of the 2 recombination hotspots are 5 kb apart, with one hotspot locating between 11.75 and 13.25 and the other between 16.75 and 18.25 kb. As was revealed by the sperm-typing analysis, weaker hotspots can exist with stronger hotspots within clusters. In the simulation study, we assumed the

Table 2. Comparison of the performances in the simulation study using the full likelihood and the approximated likelihood

Likelihood	Dataset	False positive	Power	Hotspot		
				Average width	Average intensity	MSE of intensity
Full	S_1	0 of 31	1	1.77	11.42	852.62
	S_2	0 of 30	0.83	1.56	8.11	14.07
	S_3*	1 of 16	(0.56, 0.94)	(1.72, 1.53)	(5.18, 10.73)	(5.77, 390.28)
Approximate	S_1	0 of 31	1	1.56	13.48	726.74
	S_2	2 of 30	0.77	1.40	10.31	24.44
	S_3*	0 of 16	(0.44, 1)	(1.79, 1.68)	(10.73, 12.57)	(52.13, 325.60)

Hotspots thresholds $HT_1 = 5$ and $HT_2 = 2.5$ were used.
*For 2 hotspots cases, except false positive, all values are given for the first and the second hotspots.

hotspot on the left was weak with $\rho = 6$ (H_1) and the other hotspot had a moderate intensity with $\rho = 30$ (H_2). Fifty replicate datasets were simulated (dataset S_3), and the average [minimum, maximum] number of SNPs was 35.36 [15, 66]. If assuming $HT_1 = 5$ and $HT_2 = 2.5$, the false-positive rate was 2/50 and the average power for both hotspots was 0.67 (0.5 for H_1 and 0.84 for H_2). Of 50 intervals, there were 21 intervals for which both hotspots were identified, 4 intervals for which only H_1 was identified, 21 intervals for which only H_2 was identified, and 4 intervals for which no hotspots were identified. In all cases, none of the estimated hotspots span both H_1 and $H2$, indicating that the method can discriminate between single hotspots and clusters.

The third simulation study examined the performance of the method when the approximated-likelihood method was used by splitting larger intervals into n-SNP subintervals. Of the 250 simulated samples from the above 2 simulation studies, samples that contain $\geq$40 SNPs were used in the third simulation study (77 in total). The intervals were broken into 2 subintervals with an approximately equal number of SNPs for each interval. The likelihood was approximated by multiplying likelihoods from subintervals with parameters shared across the entire interval. Results are listed in Table 2. The estimates are comparable between the 2 methods, even though for 14 of 77 intervals, the true hotspot locations were split across 2 subintervals.

Analysis of HLA and MS32 Regions. We applied our method to 2 datasets from the HLA and MS32 regions that have been previously studied by sperm typing (26, 27). 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 (26). The MS32 dataset consists of 206 SNPs sampled from 80 individuals and distributed across 0.206 Mb. Both regions have been previously analyzed by using coalescent methods for the analysis of genotypes (18, 24, 27).

For our analysis, the region was divided into subregions, each with 20 markers. The posterior distribution of recombination hotspots and background rates was inferred across the entire region. Only the locations of recombination hotspots were compared. The intensities of hotspots predicted by using the 2 approaches were not expected to be the same, because the hotspot intensities inferred by using population genetic data are a product of $\rho = 4N_e c$ and N_e likely varies across chromosomal regions due to selection. Moreover, population genetic rates are average rates over females and males. The BF of recombination hotspots across the 2 regions is shown in Fig. 2.

The hotspot locations estimated by using our method are, in general, consistent with those obtained from sperm cross-over analysis. The hotspots that were independently discovered by sperm-typing analysis also had high BFs in our population genetic analyses. Only hotspot NID3 in the MS32 region showed a lower BF.

Analysis of Human Chromosome 19. We applied our method to a human variation dataset for chromosome 19 (28). The dataset consists of 23 African-American individuals. In total, there are 18,406 SNPs on chromosome 19 from the sample. The whole

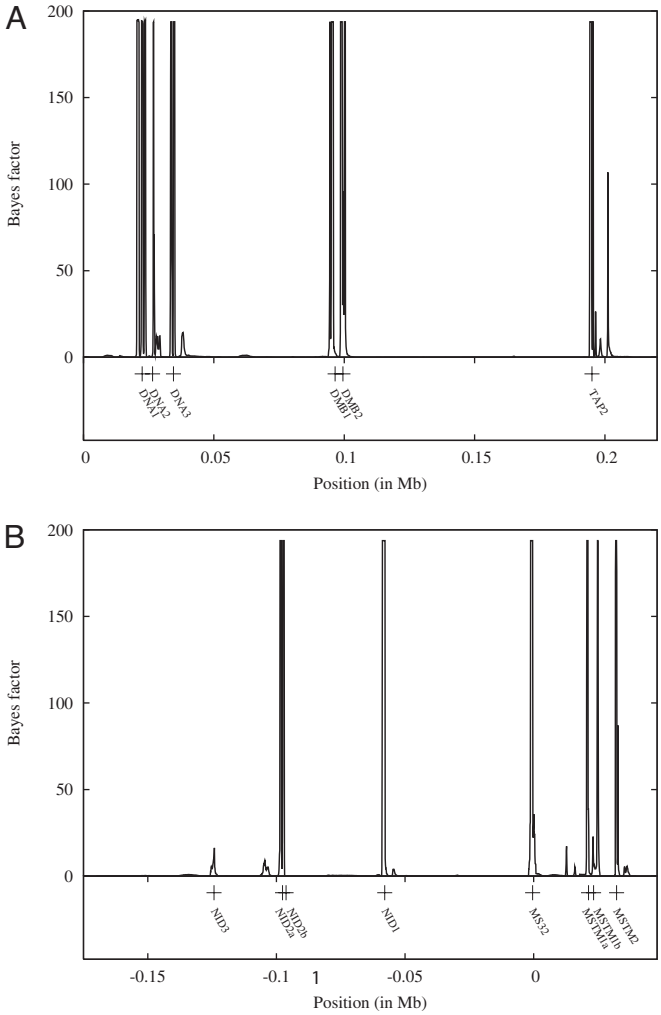

Fig. 2. Bayes factor of a hotspot inferred by the IR program as a function of location for HLA (*A*) region and MS32 (*B*) region. Location of the center of a hotspot previously inferred by sperm typing is indicated by ''+'' on the horizontal axis.

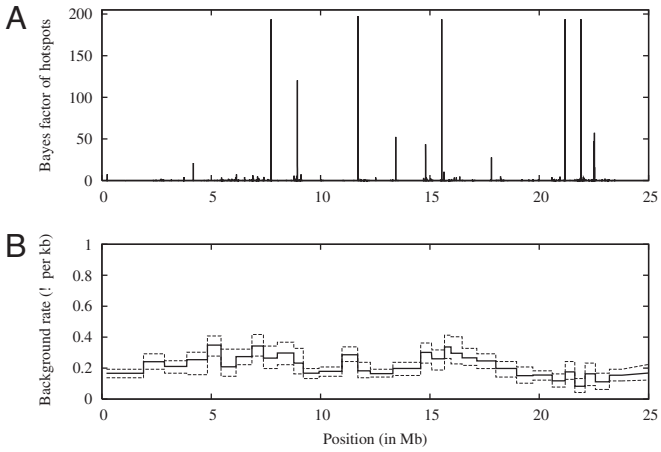

Fig. 3. Bayes factor of a hotspot (*A*) and the posterior mean (solid lines) and 95% credible intervals (dotted lines) of the expected background recombination rates (*B*) as a function of location across chromosome 19 p-arm estimated by IR in 23 African Americans.

chromosome was divided into 92 intervals that were analyzed separately. There are 200 markers on each interval for the first 91 intervals and 206 markers for the last interval. Each interval was split into subintervals with 20 markers (26 markers for the last segment of the last interval) in a parallel computation assuming shared parameters across the subintervals. Hotspots and background recombination rates were estimated for each interval. Figs. 3 and 4 show the BFs of recombination hotspots and the expected background recombination rates across chromosome 19 estimated by using our method. There is strong evidence for recombination hotspots in at least 10 locations on each chromosome arm. Comparing the hotspots locations with those inferred previously from the HapMap data, the majority appear to overlap (Table S1).

Discussion

In this article, we present a model of background recombination rates and hotspots to describe the changes of fine-scale recombination rates over genomes. It is an extension of our recently developed full-likelihood method for estimating recombination rates by using population genetic data. Full-likelihood methods use all information in the data and should provide more accurate estimates and have higher power to detect recombination hotspots, but the disadvantage of such methods is that they are computationally intensive. Our full-likelihood method efficiently models the ancestral genealogy of a sample by using marker ancestry vectors to avoid modeling nonancestral lineages, which not only add a computational burden but can cause convergence problems as well. Currently, the full-likelihood method can be applied to moderate-size chromosomal intervals. For larger intervals or whole genomes, an approximation to the full-likelihood is used that divides an entire region, or chromosome, into subregions with the likelihood approximated by using the products of likelihoods for genealogies on each subregion but with model parameters shared across the entire region.

The results of our analyses suggest that it is possible to accurately infer recombination hotspot locations and intensities across chromosomes. In particular, our method can accurately distinguish clustered hotspots, even though weaker hotspots may be present. By choosing different criteria for identifying hotspots, the power and the false-positive rate changes accordingly. In general, it should be profitable to study those hotspots with high BFs at the molecular level; the false-positive rates are extremely low in these cases. Other parameters in the recombination-rate model might also be interesting, such as how the expected distances between hotspots and the background recombination rates change over the genome. In addition, variables and parameters of the ARG model may be of interest and can be estimated by sampling from the Markov chains. Such parameters include the posterior probabilities of genealogies at each SNP site and the distribution of recombination breakpoints over genomic regions.

The program InferRho (IR) and the simulated data in both msHOT and IR formats can be obtained from http://rannala.org, or by contacting Y.W.

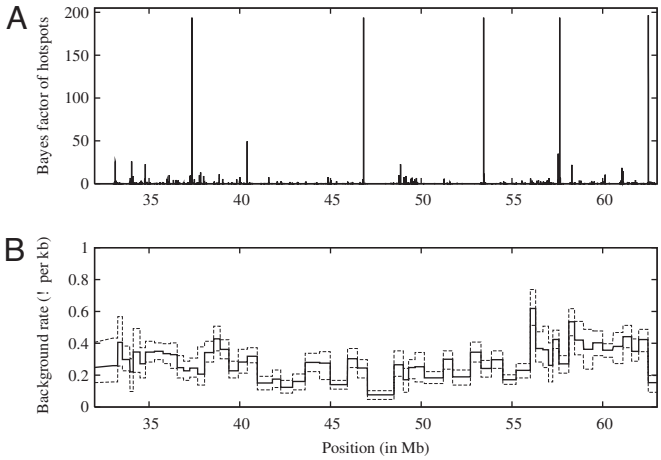

Fig. 4. Bayes factor of a hotspot (*A*) and the posterior mean (solid lines) and 95% credible intervals (dotted lines) of the expected background recombination rates (*B*) as a function of location across chromosome 19 q-arm estimated by IR in 23 African Americans.

Materials and Methods

Bayesian Inference of Fine-Scale Recombination Rates. Let Θ be a vector of parameters, including $\theta = 4N_e\mu$ and $\rho = 4N_ec$, 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. Given G_S, the genealogical trees ($\vec{\tau}$) for each marker position are then obtained. The posterior distribution of $\vec{\rho}$,

$$f(\vec{\rho}|\mathbf{X}) = \frac{1}{f(\mathbf{X})}\int f(\mathbf{X}|\vec{\tau}\in G_S,\ \theta)f(G_S|\vec{\rho})f(\vec{\rho})f(\theta)\mathrm{d}G_S\mathrm{d}\theta, \qquad [1]$$

is numerically evaluated by MCMC. 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 are unknown), modifying alleles at sites with missing alleles in the sample, and modifying the parameters θ and ρ.

Background Recombination Rates. The background recombination rates between SNPs are assumed to follow a Γ-distribution with shape parameter $a_{\rho*}$ and scale parameter $s_{\rho*}$. In the analyses, $s_{\rho*}$ is fixed, and $a_{\rho*}$ is estimated in the MCMC.

Recombination Hotspots. It is assumed that the distribution of recombination hotspots along chromosomes follows a Markov process. Hotspots arise with instantaneous rate λ_1 and revert with instantaneous rate λ_2. The waiting distance until the occurrence of a hotspot is therefore exponentially distributed with parameter λ_1, and the waiting distance until the loss of a hotspot is exponentially distributed with parameter λ_2. The values of $1/\lambda_1$ and $1/\lambda_2$ represent the average distance between hotspots and the average duration of a hotspot, respectively.

Three variables are associated with each hotspot (H), denoted by X_1, X_2, and Z, and represent the starting location, the ending location, and the strength of the hotspot. Variable Z is assumed to be log-normally distributed with parameters μ_Z and σ_Z. Considering s hotspots across a chromosomal region,

the distribution of the *i*th recombination hotspot given the adjacent hotspot on its left is

$$f(\mathrm{H}_i|\mathrm{H}_{i-1},\lambda_1,\ \lambda_2,\ \mu_Z,\ \sigma_Z) = f(\mathrm{H}_i \to X_1|\mathrm{H}_{i-1} \to X_2,\ \lambda_1)$$
$$f(\mathrm{H}_i \to X_2|\mathrm{H}_i \to X_1,\ \lambda_2) f(\mathrm{H}_i \to Z|\mu_Z,\sigma_Z), \quad [2]$$

where $i = \{1, \ldots, s-1\}$. If $i = 0$, replace $\mathrm{H}_{i-1} \to X_2$ with the location of the first site of the interval in the above equation. If $i = s-1$, and $\mathrm{H}_i \to X_2$ is equal to the right bound of the interval (or the last marker on the interval), then $f(\mathrm{H}_i \to X_2|\mathrm{H}_i \to X_1,\ \lambda_2) = \exp[-\lambda_2(\mathrm{H}_i \to X_2 - \mathrm{H}_i \to X_1)]$ to represent the fact that the end of a hotspot exceeds the right bound of the chromosomal interval. The joint density of s hotspots ($\vec{\mathrm{H}}$) on the chromosomal interval is given by the product of the above equation over s hotspots multiplied by the density on parameters λ_1, λ_2, μ_Z, and σ_Z.

Given $\vec{\rho}^*$ and $\vec{\mathrm{H}}$ across the chromosomal interval, the probability distribution of the SNP genealogy can be obtained. The posterior distribution described in Eq. **1** becomes

$$f(\vec{\rho}^*,\ \vec{\mathrm{H}},\ \lambda_1,\ a_{\rho^*}|\mathbf{X}) = \frac{1}{f(\mathbf{X})} \int f(\mathbf{X}|\vec{\tau} \in G_S,\theta) f(G_S|\vec{\rho}^*,\ \vec{\mathrm{H}})$$
$$f(\vec{\rho}^*|a_{\rho^*},\ s_{\rho^*}) f(a_{\rho^*}) f(\vec{\mathrm{H}}|\lambda_1,\ \lambda_2,\ \mu_Z,\ \sigma_Z) f(\lambda_1) f(\theta)\mathrm{d}G_S\mathrm{d}\theta. \quad [3]$$

Note that 4 parameters in the model are fixed to avoid parameter identifiability issues as well as to incorporate information from other independent studies. Parameter λ_2 is fixed to be 1,000, which corresponds to 0.001 Mb as the average width of hotspots. Other parameters are chosen to have a less informative prior, such that $\mu_Z = 9$ and $\sigma_Z = 1.5$, so the 95% interval for the strength of a hotspot is [428.40, 153,268.41] per Mb. Because background recombination rates are intended to describe low rates with relatively small variance, parameter s_{ρ^*} is fixed to be 50. For example, if $a_{\rho^*} = 2$, the 95% interval of the prior distribution on background rate is [12.11, 278.58] per Mb. If the size of a chromosomal interval is large, the interval is divided into K subintervals, denoted by $\mathbf{X} = \{\mathbf{X}_i\}$, to improve computational efficiency. The SNP genealogies underlying subintervals are treated as independent but allowing parameters to be jointly estimated across an entire interval. The independence assumption can be relaxed if needed, and the sizes of subintervals can be adjusted depending on the available computing resources. In this case, the posterior distribution is approximated by

$$f(\vec{\rho}^*,\ \vec{\mathrm{H}},\ \lambda_1,\ \mu_{\rho^*}|\mathbf{X}) \approx \frac{1}{f(\mathbf{X})} \int \prod_{i=1}^{K}$$
$$f(\mathbf{X_i}|\vec{\tau}_i \in G_{S_i},\ \theta) f(G_{S_i}|\vec{\rho}^*,\ \vec{\mathrm{H}}) f(\vec{\rho}^*|a_{\rho^*},\ s_{\rho^*}) f(a_{\rho^*})$$
$$f(\vec{\mathrm{H}}|\lambda_1,\ \lambda_2,\ \mu_Z,\ \sigma_Z) f(\lambda_1) f(\theta)\mathrm{d}\vec{G}_S\mathrm{d}\vec{\rho}^*\mathrm{d}\vec{\mathrm{H}}\mathrm{d}\theta. \quad [4]$$

For additional information see *SI Text*, Figs. S2 and S3, and Table S2.

ACKNOWLEDGMENTS. We thank Matthew Stephens for suggestions regarding the prior distribution on background recombination rates and 2 anonymous reviewers for their helpful comments. This work was supported by National Institutes of Health Grant HG01988 (to B.R.).

1. Pritchard JK, Przeworski M (2001) Linkage disequilibrium in humans: Models and data. *Am J Hum Genet* 69:1–14.
2. Kong A, et al. (2002) A high-resolution recombination map of the human genome. *Nat Genet* 31:241–247.
3. Coop G, et al. (2008) High-resolution mapping of crossovers reveals extensive variation in fine-scale recombination patterns among humans. *Science* 319:1395–1398.
4. Arnheim N, Calabrese P, Tiemann-Boege I (2007) Mammalian meiotic recombination hot spots. *Annu Rev Genet* 41:369–399.
5. Altshuler D, et al. (2005) A haplotype map of the human genome. *Nature* 437:1299–1320.
6. Buard J, de Massy B (2007) Playing hide and seek with mammalian meiotic crossover hotspots. *Trends Genet* 23:301–309.
7. Stumpf M, McVean G (2003) Estimating recombination rates from population-genetic data. *Nat Rev Genet* 4:959–968.
8. Hellenthal G, Stephens M (2006) Insights into recombination from population genetic variation. *Curr Opin Genet Dev* 16:565–572.
9. Hudson RR (1990) Gene genealogies and the coalescent process. *Oxford Surv Evol Biol* 7:1–44.
10. Griffiths RC, Marjoram P (1996) Ancestral inference from samples of DNA sequences with recombination. *J Comput Biol* 3:479–502.
11. Kuhner MK, Yamato J, Felsenstein J (2000) Maximum likelihood estimation of recombination rates from population data. *Genetics* 156:1393–1401.
12. Fearnhead P, Donnelly P (2001) Estimating recombination rates from population genetic data. *Genetics* 159:1299–1318.
13. Nielsen R (2001) Estimation of population parameters and recombination rates from single nucleotide polymorphisms. *Genetics* 154:931–942.
14. Hudson RR (2001) Two-locus sampling distributions and their application. *Genetics* 159:1805–1817.
15. McVean G, Awadalla P, Fearnhead P (2002) A coalescent-based method for detecting and estimating recombination from gene sequences. *Genetics* 160:1231–1241.
16. Li N, Stephens M (2003) Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics* 165:2213–2233.
17. Crawford DC, et al. (2004) Evidence for substantial fine-scale variation in recombination rates across the human genome. *Nat Genet* 36:700–706.
18. McVean G, et al. (2004) The fine-scale structure of recombination rate variation in the human genome. *Science* 304:581–584.
19. 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.
20. Wang Y, Rannala B (2008) Bayesian inference of fine-scale recombination rates using population genomic data. *Phil Trans R Soc B* 363:3921–3930.
21. Li J, Zhang M, Zhang X (2006) A new method for detecting human recombination hotspots and its applications to the HapMap ENCODE data. *Am J Hum Genet* 79:628–639.
22. Fearnhead P (2006) SequenceLDhot: Detecting recombination hotspots. *Bioinformatics* 22:3061–3066.
23. De Lorio M, de Silva E, Stumpf M (2005) Recombination hotspots as a point process. *Phil Trans R Soc B* 360:1597–1603.
24. Auton A, McVean G (2007) Recombination rate estimation in the presence of hotspots. *Genome Res* 17:1219–1227.
25. Green PJ (1995) Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* 82:711–732.
26. Jeffreys AJ, Kauppi L, Neumann R (2001) Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. *Nat Genet* 29:217–222.
27. Jeffreys AJ, 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.
28. Hinds DA, et al. (2005) Whole-genome patterns of common DNA variation in three human populations. *Science* 307:1072–1079.
29. Hellenthal G, Stephens M (2007) msHOT: Modifying Hudson's ms simulator to incorporate crossover and gene conversion hotspots. *Bioinformatic* 23:520–521.
30. Hudson RR (2002) Generating samples under a Wright–Fisher neutral model. *Bioinformatics* 18:337–338.