ARTICLE

A New Method for Detecting Human Recombination Hotspots and Its Applications to the HapMap ENCODE Data

Jun Li, Michael Q. Zhang, and Xuegong Zhang

Computational detection of recombination hotspots from population polymorphism data is important both for understanding the nature of recombination and for applications such as association studies. We propose a new method for this task based on a multiple-hotspot model and an (approximate) log-likelihood ratio test. A truncated, weighted pairwise log-likelihood is introduced and applied to the calculation of the log-likelihood ratio, and a forward-selection procedure is adopted to search for the optimal hotspot predictions. The method shows a relatively high power with a low false-positive rate in detecting multiple hotspots in simulation data and has a performance comparable to the best results of leading computational methods in experimental data for which recombination hotspots have been characterized by sperm-typing experiments. The method can be applied to both phased and unphased data directly, with a very fast computational speed. We applied the method to the 10 500-kb regions of the HapMap ENCODE data and found 172 hotspots among the three populations, with average hotspot width of 2.4 kb. By comparisons with the simulation data, we found some evidence that hotspots are not all identical across populations. The correlations between detected hotspots and several genomic characteristics were examined. In particular, we observed that DNaseI-hypersensitive sites are enriched in hotspots, suggesting the existence of human β hotspots similar to those found in yeast.

Meiotic recombination is one of the major sources of genetic diversity. It has been observed that the occurrence of meiotic recombination in the human genome (and some other genomes) is not uniform, but rather there are regions called "hotspots" (usually 1–2 kb in width) where the frequency of recombination is 10 to several thousand times higher than the average in the background, and almost all recombination events happen within them.[1–6] Recent studies have shown that hotspots are a ubiquitous feature of the human genome,[7,8] and recombination hotspots are also the main contributor of the block-like pattern of haplotypes.[9] Characterizing these hotspots is of critical importance for understanding molecular mechanisms of meiotic recombination and for designing better strategies in association studies of complex diseases.[10–16] Pedigree analysis can only specify recombination rate on a megabase scale, because of the small number of recombination events that can be observed within a few generations. The first fine-scale description of human recombination hotspots was achieved with the sperm-typing technique,[1–3] which types millions of sperm that contain hundreds of recombination events in the studied region (often ~10 kb). The resolution of sperm typing is very high, but it is costly and laborious, so it is not yet practical for application to long genomic segments, and it cannot provide any information about females. Up to now, <20 hotspots have been characterized by sperm-typing experiments, and genomewide fine-scale investigations in humans have largely relied on computational analysis of population polymorphism data.[7,8]

The problem of estimating a constant recombination rate from population polymorphism data has been intensively studied in recent years.[17,18] Among the many possible methods, likelihood-based methods are the most widely accepted. The basic idea is to search for a recombination rate that maximizes the likelihood of obtaining the observed phased (haplotype) or unphased (genotype) data from the population under the coalescent model.[19] Some methods use all information contained in the data to calculate the full likelihood, which is accurate but extremely expensive to compute.[20–22] Other methods use partial data to calculate approximate likelihoods.[23–26] These likelihoods can approximate the full likelihood well if the methods are designed properly.[27]

Since a constant recombination rate is rarely the case in the genome, detecting recombination hotspots is more challenging. Zhang et al.[28] proposed a nonparametric method based on haplotype-block partitioning, which is computationally effective but cannot give high-resolution prediction of hotspot locations. For more-precise predictions, three major parametric methods based on coalescent models have been developed. Their common basic idea is to compare approximate likelihoods under models with and without hotspot(s). The three methods use different approximations of the full likelihood. The LDhot method[7] uses a pairwise likelihood that is the product of two-locus likelihoods of all pairs of segregating sites. The Hotspotter method[26] defines another kind of likelihood, constructed by multiplying the approximate conditional likelihoods of each haplotype in a specific order. The

From the Bioinformatics Division, Tsinghua National Laboratory for Information Science and Technology and Department of Automation, Tsinghua University, Beijing (J.L.; M.Q.Z.; X.Z.); and Cold Spring Harbor Laboratory, Cold Spring Harbor, New York (M.Q.Z.)

Received May 9, 2006; accepted for publication July 25, 2006; electronically published August 30, 2006.

Address for correspondence and reprints: Dr. Xuegong Zhang, Bioinformatics Division, FIT 1-107, Tsinghua University, Beijing 100084, China. E-mail: zhangxg@tsinghua.edu.cn

Am. J. Hum. Genet. 2006;79:628–639. © 2006 by The American Society of Human Genetics. All rights reserved. 0002-9297/2006/7904-0005$15.00

method by Fearnhead et al.[29] and its improved version[30] divide a studied region into small subregions and calculate a composite likelihood by multiplying full likelihoods of all subregions.

These three parametric methods are differentiated by their ability to detect multiple hotspots in a genomic segment, the data types to which they can be applied, and the speed of calculation. Both LDhot and Hotspotter assume no more than one hotspot in the studied region, whereas the methods of Fearnhead et al. are able to detect multiple hotspots in a region. Hotspotter and the methods of Fearnhead et al. require phased data, so users need to do haplotype inference first, since almost all available polymorphism data are unphased (Hotspotter has been integrated into the PHASE[26,31–33] program to deal with unphased data). LDhot can be applied to both unphased and phased data directly. The computational costs of these three methods are also quite different. LDhot is very fast because of the use of the pairwise likelihood; therefore, it can be applied to the whole genome.[8] The methods of Fearnhead et al. are much slower, since they calculate the full likelihood in each of the subregions. Hence, it is very costly to apply them to genome-scale data. The speed of Hotspotter lies between those of LDhot and the methods of Fearnhead et al.

The three parametric methods had been compared on a 206-kb region on human chromosome 1 near the highly variable minisatellite MS32, where the fine-scale recombination-rate variation has been analyzed by sperm-typing experiments.[5] Of the eight hotspots detected by sperm-typing, LDhot detected four, with no false-positive result; Hotspotter detected five but gave three false-positive predictions; and the first method of Fearnhead et al. detected seven, with only one false-positive prediction, which shows the highest power for this data set.

The major limitations of the methods of Fearnhead et al. are their high computing cost and their inability to directly handle unphased data. To make the method more practical and flexible, we propose, in this article, a new method for hotspot detection. Our method uses a truncated, weighted pairwise log-likelihood (TWPLL) and can be applied to both phased and unphased data with a very fast computational speed. In simulation data, our method shows a high power to detect multiple hotspots, with a considerably low false-positive rate. In the two regions of the human genome where sperm-typing data have been reported, our method gets comparable or even better results than the best results obtained by all those other leading computational methods. We applied the method to the 10 human genome regions known as the HapMap ENCODE regions and identified 172 hotspots that exist in at least one of the three populations.

Nowadays, the mechanism of meiotic recombination is still poorly understood in higher eukaryotes.[34–36] We studied the molecular features of the predicted hotspots in the HapMap ENCODE regions and observed correlations of hotspots with some genomic features. In particular, we observed that DNaseI-hypersensitive sites (DHSSs) are enriched in hotspots. This is a strong sign that there are β hotspots in the human genome similar to those identified in yeast.

Material and Methods

Data

We applied our method to the HapMap ENCODE data as a practical application. The pilot phase of the ENCODE Project focuses on a specified 1% (~30 Mb) of the human genome, aiming to identify all functional elements in the regions. Some of these regions (known as the HapMap ENCODE regions) have been genotyped by HapMap Centers, and contain 10 genomic segments (500 kb each) from seven chromosomes. The data were genotyped in four populations: Utah residents with northern and western European ancestry (CEU), Han Chinese in Beijing (CHB), Japanese in Tokyo (JPT), and Yoruba in Ibadan, Nigeria (YRI), with diploid population sizes of 90, 45, 44, and 90, respectively. To balance sample size in our experiments, we combined the two East Asian populations (CHB and JPT) into one group and called it “ASI.” Almost all SNPs in these regions have been genotyped. Only those markers with minor-allele frequency (MAF) >0.05 were used to infer hotspots.

After hotspot detection, some genomic features of the HapMap ENCODE regions were downloaded from the ENCODE Project at UCSC Web site to investigate their possible correlations with hotspot locations. These data include DNA sequences, RefSeq genes, CpG islands, repeats, and DHSSs. The total sequence length of the regions is 5 Mb, and we use the sequences to calculate the G+C content. There are 70 CpG islands in these regions, with an average length ~0.9 kb. Repeats were identified by use of the RepeatMasker software and the repeat libraries available, and they cover 45.6% of the 5-Mb region. Among the ~30 repeat families, 7 of them (Alu, L1, MIR, Simple_repeat, L2, Low_complexity, and MaLR) occur >500 times in the studied regions. A total of 56 RefSeq genes are found in the studied regions. These gene areas (counted from 1 kb upstream of 5′ sites of the first exons to 1 kb downstream of 3′ sites of the last exons) cover ~39% of the 5-Mb region. DHSSs are associated with all kinds of gene regulatory regions, including enhancers, silencers, promoters, insulators, and locus-control regions.[37] The available data are from four groups: (1) DHSSs identified by DNase-chip in the GM06990 lymphoblastoid cell line, (2) DHSSs identified by DNase-chip in the nonactivated CD4$^+$ T cells, (3) DHSSs identified by massively parallel signature sequencing (MPSS) in the nonactivated CD4$^+$ T cells, and (4) DHSSs identified by MPSS in the activated CD4$^+$ T cells. Among these four groups, 144, 143, 26, and 30 DHSSs were identified in the 5-Mb region, and their average length is ~0.29 kb. More detailed description about these genomic features can be found at the ENCODE Project at UCSC.

The Pairwise Log-Likelihood (PLL)

Suppose there are S segregating sites in the studied segment, and the recombination rate between site i and site j is ρ_{ij}. We use $L_{ij}(\rho_{ij})$ to denote the two-locus likelihood between site i and site j, defined as the probability of observing the sample configuration at these two sites in the data, given ρ_{ij}. This can be easily calculated according to the definition of recombination rate at the two sites.

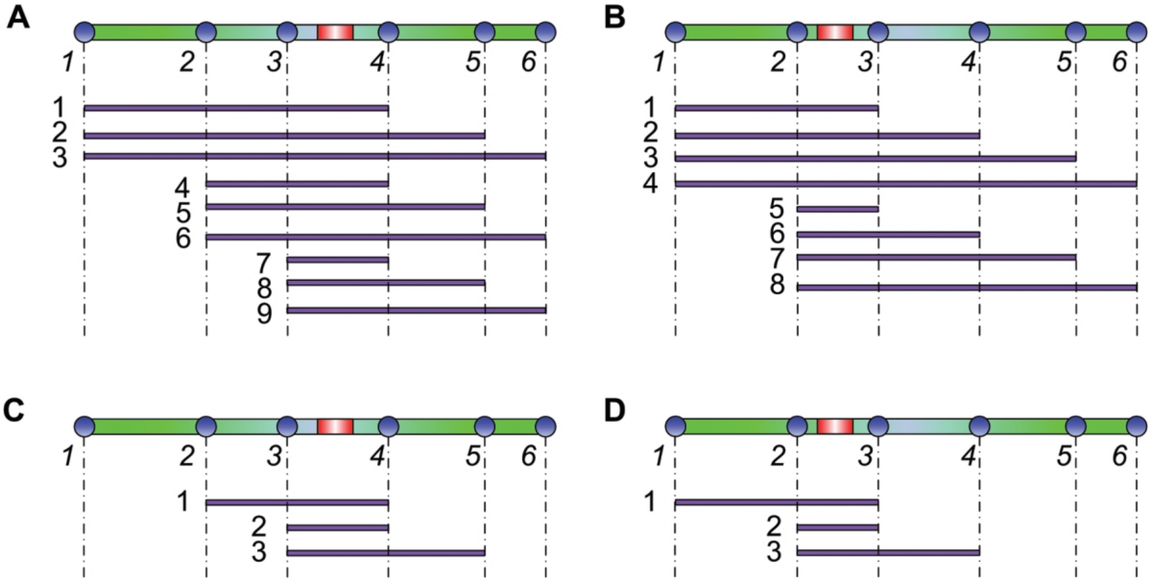

Figure 1. Subregions covering a hotspot in different log-likelihoods. In each panel, the example region (*green bar*) contains six segregating sites (*blue circles*), with a hotspot (*red bar*) located between a pair of sites. The purple lines indicate the subregions between pairs of sites that cover the hotspot in the log-likelihood. *A,* With PLL, the hotspot located between the third and fourth sites is covered by nine subregions. *B,* With PLL, the hotspot located between the second and third sites is covered by eight subregions. *C,* With TWPLL, the hotspot located between the third and fourth sites is covered by three subregions. *D,* With TWPLL, the hotspot located between the second and third sites is also covered by three subregions. The number of subregions covering a hotspot depends on the location of the hotspot in PLL, whereas it does not in TWPLL.

The pairwise likelihood is defined as the product of all pairs of sites in the segment,

$$L(\rho) = \prod_{j=i+1}^{S}\prod_{i=1}^{S-1} L_{ij}(\rho_{ij}) \ ,$$

and the PLL is defined as

$$l(\rho) = \log L(\rho) = \sum_{j=i+1}^{S}\sum_{i=1}^{S-1} \log L_{ij}(\rho_{ij}) \ .$$

There are several advantages of using this likelihood. Since likelihoods of all possible sample configurations at any two sites for a given sample size can be calculated beforehand and stored in a lookup table, calculation of PLL can be extremely fast. Moreover, PLL can be applied to both phased and unphased data directly, since two-locus likelihoods for unphased pairs can be inferred straightforwardly from those of phased pairs.[23] In addition, when two-locus likelihoods are calculated, it is convenient to use different recombination models, such as gene conversion,[25,38] or to calculate under a finite-site mutation model.[25] The recombination rate of a region can be estimated as the rate that maximizes the pairwise likelihood. Smith and Fearnhead[27] have shown that it is one of the most accurate methods for estimating a uniform recombination rate. When applied to a region with variable recombination rates, the estimated rate will be the average rate across the region.

Recombination hotspots can be detected using the likelihood by investigating whether a model with hotspot(s) can produce a higher likelihood of the data than can the uniform-rate model. This can be done by studying the log-likelihood ratio (LLR) of the two models, defined as the log ratio of the likelihood under the model with hotspots to that under the model without hotspots. However, there is a problem if we use PLL for this purpose. Indeed, the PLL is defined on likelihoods of $S(S-1)/2$ pairs of segregating sites. Suppose there is a hotspot located between the ith and $(i+1)$th segregating sites; then, this hotspot region will affect $i(S-i)$ terms in the PLL. A simple example with $S = 6$ is illustrated in figure 1*A* and 1*B*. Note that $i(S-i)$ depends on the location i, such that it first ascends and then descends with i. Hotspots at different locations in the studied segment will have unequal effects on the likelihood, and hotspots near the center of the segment are more likely to be detected. This will cause a loss of detection power and will also lead to bias in the discovery of hotspots.

The TWPLL

The idea of a weighted PLL that was introduced by Fearnhead[39] makes it possible to define a likelihood that is unrelated to the location of hotspots. It was originally defined as

$$l_w(\rho) = \sum_{k=1}^{S-i}\sum_{i=1}^{S-1} w_k \log L_{i,i+k}(\rho_{i,i+k}) \ , \qquad (1)$$

where $w_k \geqslant 0$, $k = 1,\dots,S-1$, are a set of weights decreasing in k. If we assign $w_k = 1$ for $k = 1,\dots,S-1$, weighted PLL will degenerate into PLL. Weighted PLL was suggested for the estimation of uniform recombination rates.[39] We adopt the idea for estimation of recombination models with hotspots. In our method, we define the weights in equation (1) as

$$w_k = \begin{cases} \omega_k & \text{if } k \leqslant N \\ 0 & \text{otherwise} \end{cases}$$

and call it the "TWPLL," where N is the number of segregating sites in the effective neighborhood region (usually $N \ll S$). By defining the distance of two segregating sites as the number of segregating sites between them, only the pairwise likelihoods of pairs of segregating sites with distance no more than $N-1$ are considered in calculation of the pairwise likelihood. In our experiments, we use $N=7$. The ω_k should decrease with k, but, currently, the optimal choice of ω_k is still unclear.[39] We set $\omega_k = 1/k$, $k=1,\ldots,N$, according to experiments on simulation data, with special attention paid to the balance between precision in detecting hotspot boundaries and sensitivity to noise. With this truncated, weighted pairwise likelihood, for any $N \leqslant i \leqslant S - N + 1$, a hotspot located between any site i and the site $i+1$ will be considered $N(N+1)/2$ times in equation (1). Since $N(N+1)/2$ is unrelated to i, it is equally possible to detect hotspots located at different positions. Figure 1*C* and 1*D* illustrates the effect of the TWPLL in the simplified example. Only hotspots at the boundaries of the region ($i < N$ or $i > S - N + 1$) are not evenly covered. They usually compose a very small proportion of the region, because $N \ll S$, and we will introduce a compensation for the boundaries when searching for the solutions (see "The Searching Strategy" section).

The Recombination-Rate Model

The models used by LDhot[7] and Hotspotter[26] are all restricted, with at most one hotspot in a data segment. We use a recombination-rate model that allows for multiple hotspots in each studied segment as did the one Fearnhead et al. used.[30] The recombination rate as a function of the location x in a region is called the "recombination surface," denoted $\rho(x)$. A multiple hotspot model has the form

$$\rho(x) = \begin{cases} \rho_1 & \text{for } s_1 \leqslant x \leqslant e_1 \\ \vdots & \qquad \vdots \\ \rho_h & \text{for } s_h \leqslant x \leqslant e_h \\ \rho_b & \text{otherwise} \end{cases},$$

where h is the number of hotspots in the region and the kth hotspot extends from position s_k to e_k. Any hotspot does not overlap with or touch the other hotspots. The recombination rates for hotspots are $\rho_1,\ldots,\rho_h$, respectively, and the background rate of the region is ρ_b.

The Searching Strategy

After defining TWPLL, we need to search for the recombination surface that maximizes the likelihood of the data. We adopted the standard forward-selection procedure to search for the solution. The procedure starts with a model of no hotspot in the region and adds hotspots one by one if adding them increases the likelihood by no less than an LLR threshold T, which is decided by simulation. The searching procedures are as follows:

Step 1. Assume no hotspot in the region and estimate an average recombination rate as the initial $\rho(x)$ by maximizing the likelihood.

Step 2. Consider all potential hotspot positions, given the current recombination surface $\rho(x)$. For each potential hotspot, use the following steps to find the best-fit model:

(*a*) Reestimate the background recombination rate that maximizes the likelihood after exclusion of the potential hotspot under current consideration and all hotspots that are already accepted. The recombination surface under this reestimated background rate and accepted hotspots is denoted as $\rho'(x)$.

(*b*) Assume the current potential hotspot is a real hotspot. Estimate its intensity (recombination rate of the hotspot) that maximizes the likelihood of the whole region. Add this hotspot to the surface $\rho'(x)$ to get a new surface, $\rho''(x)$.

(*c*) Calculate LLR, which is the likelihood under $\rho''(x)$ subtracted by the likelihood under $\rho'(x)$. If the potential hotspot is at one of the boundaries of the studied region, its LLR is amplified by a factor depending on the number of subregions that cover the hotspot, to compensate for its insufficient representation in the likelihood. This set of factors is decided with simulation experiments.

Step 3. After checking all potential hotspots in step 2, find the one that gives the highest LLR. If this LLR is $\geqslant T$, accept this potential hotspot, refresh the recombination surface, update the set of all potential hotspots, and go to step 2. Otherwise, stop the searching procedures.

In the above procedures, the set of all potential hotspots is collected in the following way. From the beginning of the region, consider every 200-bp position as a possible starting position of a hotspot, if the position is not at a hotspot already detected. From each of these starting positions, we generate a set of potential hotspots with lengths varying from 800 bp to 2.4 kb, with a 200-bp step length. If any potential hotspot thus generated overlaps with or touches any of the hotspots that are already accepted in the prediction, we remove this potential hotspot. This setting considers all possible hotspots of lengths from 800 bp to 2.4 kb, at a resolution of 200 bp.

When we estimate the background rate in step 2(*a*), some regions are excluded in advance, to avoid estimation of background rates that are too high. For this, we first slide a window of four adjacent SNPs along the whole region and estimate the average recombination rate in each window. If the rate is >10 times the genome average, the sites in this window will not be used for estimating the background rate. A similar strategy was also used by Fearnhead et al.[30]

It should be noted that the above procedure may not reach the global optimum solution, since it is a greedy forward-selection method. However, since the TWPLL only considers two-locus likelihoods between pairs of segregating sites with distance <*N*, hotspots will be independent with respect to the TWPLL if they are apart from each other by more than this distance. Therefore, if the distances between hotspots are >*N*, the greedy searching method can reach the global optimum. When two hotspots are very close to each other—for example, when there are only one or two sites between them—our method will tend to detect them as one larger hotspot. Considering that the average density of SNPs (with MAF >0.05) in the human genome is denser than 1 per kb and that the estimated average density of recombination hotspots is ~1 hotspot per 50 kb,[8] global optimum can be reached with this forward-searching strategy in most situations.

The whole method is implemented in a package named "HotspotFisher" that is written in C++ and works on different operation systems. The software is available at Jun Li's Web site.

Coalescent Simulations

Simulation data based on coalescent models are used for tuning some parameters and for assessing the performance of the method. We used the Cosi program[40] to simulate polymorphism data. Cosi is conceptually similar to Richard Hudson's widely used program,[41] but it has the extra benefit of allowing variable recombination rates—users can set multiple hotspots with different densities and at arbitrary locations. Moreover, Cosi calibrates population genetic models with genomewide data and provides users with four detailed human demographic histories that take into consideration events like population splits, admixture, changes in size, bottlenecks, and migration. These four populations include a European population, an Asian population, an African population, and an African American population. We used the first three, since they correspond to the CEU, ASI, and YRI populations in the HapMap ENCODE data.

In most published work on hotspot inference, simulation data were designed such that there is only one hotspot in one data segment.[7,26,29,30] In the study by Zhang et al.,[28] multiple hotspots in single segments were simulated, but their locations were fixed and equally spaced, and hotspot widths and intensities were also fixed. To make the simulation data more like the real situation, we simulated long genome regions (200 kb) with multiple hotspots at random locations, with variable widths and intensities. We use our method to estimate the locations of these variable hotspots. The detailed model is as follows:

(1) The length of each simulated data region is 200 kb. The expected average recombination rate in the simulated segments is set as 1.2 cM/Mb, the same as the human genome average.[42]

(2) On each simulated region, a proportion p of recombination events are expected to happen within hotspots. For convenience, we call this proportion the "hotspot quotient" (HQ). Sperm-typing analysis showed that >90% of recombination events in the human genome occur within hotspots (HQ > 90%), and the background recombination rate can be as low as 0.04 cM/Mb.[3,5] We use two HQ values (90% and 70%) in the simulation, to study the performance of the proposed method under different conditions. These two settings give background recombination rates of 0.12 cM/Mb and 0.36 cM/Mb. The model with HQ = 90% is consistent with the results from sperm-typing experiments and appears to be consistent with our results for the 10 HapMap ENCODE regions. The model with HQ = 70% is at the lower end of what is observed for the human genome.[7,8]

(3) The spacing between hotspot centers fits an exponential distribution with the mean of 50 kb, as the suggested average across the whole genome.[8] We also restrict the spacing between two hotspot centers to be not less than 2 kb.

(4) The width of hotspots follows a uniform distribution of 1–2 kb. This is in accordance with existing observations of hotspots.[3,5]

(5) The accumulated intensity (defined as the product of the intensity and the width of the hotspot) of each hotspot follows a gamma distribution, with gamma equal to 3 and the mean determined by parameters given in (1), (2), and (3) above. This distribution is chosen arbitrarily, because there is little knowledge about the true distribution of hotspot densities. The hotspot intensity of each hotspot is calculated from its accumulated intensity and width, and, if the resulting intensity is <10 times the background, this hotspot is discarded from the model and replaced by a new one.

We simulated six data sets. Each data set consists of 100 groups of data for estimating the false-positive rate and power, and each group consists of 90 diploid samples, so that the sample size is the same as for the HapMap ENCODE data. Every 90 diploid samples were obtained by combining 180 haplotypes randomly. Data sets 1, 2, and 3 all have HQ = 90% and are generated with the European, Asian, and African demographic histories, respectively. Data sets 4, 5, and 6 have HQ = 70% and are also generated with the three population histories. In the calibrated model of Schaffer et al.,[40] gene-conversion rate is set at 4.5×10^{-9} per bp per generation, with a tract length of 500 bp for all gene-conversion events. This is also what we used in our simulation. Cosi assumes an infinite-sites model of mutation, and mutation positions are converted into discrete base-pair positions. A constant mutation rate of 1.0×10^{-8} per bp per generation was chosen in our models, so that the average density of SNPs in our simulation data is the same as that in the HapMap ENCODE data.

Results

Hotspot Detection in Simulation Data

We applied our method to each group of the simulation data. The same lookup table was used for the pairwise likelihood in all the experiments. The original table for 192 haplotypes was downloaded from the LDhat version 2.0 (a package for recombination-rate analysis[7]) Web site, and we used the lkgen function in LDhat version 2.0 to convert it to a table for 180 haplotypes. The false-positive rate and power of the hotspot detection were assessed with the simulation experiments. If a detected hotspot overlaps with a hotspot built in the model, we regard it as a true-positive prediction; otherwise, we regard it as a false-positive prediction. Here, we define the false-positive rate as the expected number of false-positive results per Mb and define the power as the proportion of hotspots in the models that are detected by the algorithm.

The LLR threshold T was first estimated on the basis of a given false-positive rate with the data under HQ = 90%. It was observed that the false-positive rates are very similar under the two HQ values at the same T values. This is an important property, as it indicates that the same T can be used regardless of the background rate. Finally, $T = 26$ was chosen for all the experiments, which limits the expected false-positives in a 200-kb region to be no more than 0.08, or, equivalently, the false-positive rate is no more than 0.4 per Mb. The results for the simulation data with $T = 26$ are shown in table 1. With this setting, in the total 5 Mb of HapMap ENCODE data, the expected number of false-positive predictions in each population will be no more than 2.

Table 1 also shows the power reached with different groups of simulation data. It can be seen that almost the same power was reached in different populations, indicating that the method is not sensitive to the population

Table 1. Hotspot Prediction Performance on Simulation Data ($T = 26$)

	HQ = 90%				HQ = 70%			
Population	CEU	ASI	YRI	Total[a]	CEU	ASI	YRI	Total[a]
No. of false-positive results[b]	8	5	2	13	3	8	2	13
Power (%)[c]	69	66	66	87	38	37	35	58
Average position offset (bp)[d]	360	376	309	…	351	345	258	…
Center coverage (%)[e]	94	96	98	…	96	98	100	…

[a] The number of hotspots detected in at least one of the populations.

[b] The total number of false-positive predictions in all 100 segments (200 kb each).

[c] The percentage of true hotspots in the models that are correctly detected.

[d] Average offset from the predicted start and end sites to the real start and end sites.

[e] Percentage of predicted hotspots that cover centers of corresponding true hotspots in the models.

history. With the model of HQ = 90%, the average power for the three populations is as high as 0.67, but, with HQ = 70%, the power decreases to ~0.37. This indicates that HQ is a major factor that affects the prediction power; the higher the HQ is, the higher the power is. On the human genome, HQ is estimated to be around or more than 90%, according to sperm-typing experiments[3,5] as well as our calculation for the HapMap ENCODE data described below. Some hotspots in the simulation data are not detected in all populations. If we combine the hotspots detected in the three populations together, the power is even higher (0.87 for HQ = 90% and 0.58 for HQ = 70%), whereas the false-positive rate still maintains a low level of 0.65 per Mb, or 3.25 false-positive predictions in a 5-Mb region.

We further examined the accuracy of the hotspot locations that we detected. As shown in table 1, the mean offsets from the predicted start and end locations to the corresponding precise locations in the simulation models are ~310–380 bp when HQ = 90% and 250–350 bp when HQ = 70%. More than 94% of predicted hotspots cover the center of the corresponding hotspots in the models.

Finally, we examined whether our method is sensitive to some hotspot properties, such as hotspot intensities, hotspot widths, SNP densities in hotspots, and SNP MAFs in hotspots. Spearman's rank correlation coefficients were calculated between each of them and the detection of hotspots (table 2). Hotspot intensities, SNP densities in hotspots, and SNP MAFs in hotspots were weakly but significantly correlated with their detection, suggesting that stronger hotspots with denser SNPs and higher-MAF SNPs inside are easier to detect. Hotspot widths are uncorrelated with the detection of hotspots.

Hotspot Detection for Experimentally Verified Human Hotspots

To date, there are only two human genome regions for which multiple recombination hotspots have been characterized by sperm-typing experiments. One is a 216-kb segment of the class II region of the major histocompatibility complex (MHC) on chromosome 6, where six hotspots were found by experiments[3]; the other is a 206-kb segment on chromosome 1 near the highly variable minisatellite MS32, where eight hotspots were reported.[5] We used these two data sets to validate our method. The first data set contains 247 SNP sites ($\mathrm{MAF} > 0.05$) of 50 diploid samples, and the second data set contains 191 SNP sites ($\mathrm{MAF} > 0.05$) of 80 diploid samples. We used the diploid data directly. All parameters in our method were set to be the same as in the simulation experiments.

In the first region, our method detected seven hotspots, which included all six true hotspots and an additional one at ~6.5 kb downstream from the 3′ end of the *TAP2* hotspot (fig. 2*A*). The original sperm-typing data were uninformative about recombination in the 3′ end of the *TAP2* hotspot (the area indicated by a question mark in fig. 2*A*), and it was conjectured by Jeffreys et al.[3] that the *TAP2* hotspot might be part of a cluster. Fearnhead et al.[29] and Zhang et al.[28] also applied their methods to this data set. Fearnhead et al. found eight hotspots, including all six true hotspots, a hotspot downstream from the *TAP2* hotspot, and an extra hotspot not supported by the sperm-typing experiment.[29] Zhang et al. predicted four putative hotspot regions in this data, covering all the known hotspots, and their result also suggested a hotspot downstream from the *TAP2* hotspot.[28]

In the second region, our method correctly detected six true hotspots with no false-positive predictions (fig. 2*B*). Only two hotspots (*NID*2b and MSTM1a) were missed in the detection. The *NID*2b hotspot lies almost entirely within a region of intense marker association, so it is expected that coalescent-based methods would not detect it.[5] The other hotspot we missed, MSTM1a, was reported to be historically weak and a candidate for a young hotspot.[43] It lies very close to a historically strong hotspot, MSTM1b—their centers are only 2.0 kb apart.[5] As mentioned in our introduction, the methods of Fearnhead et al. showed highest power on this data set among the currently available methods.[5] It found the six hotspots we detected plus the MSTM1a hotspot, but it made a false-positive prediction between MSTM1b and MSTM2.

In total, our method detected 12 of 14 true hotspots, with zero false-positive results, in these two regions (not considering the putative hotspot in the 3′ region of the *TAP2* hotspot). The average offset from the predicted start and end locations to the corresponding true locations decided by the experiments is 409 bp. All predicted hotspots

Table 2. Spearman's Rank Correlation Coefficients (SCCs) between the Detection of Hotspots and Their Properties

	HQ = 90%			HQ = 70%		
Population	CEU	ASI	YRI	CEU	ASI	YRI
SCC with hotspot intensity	.35[a]	.34[a]	.34[a]	.30[a]	.29[a]	.34[a]
SCC with hotspot width	.06	−.01	.05	−.01	−.00	−.01
SCC with SNP density in hotspots	.30[a]	.27[a]	.40[a]	.42[a]	.45[a]	.50[a]
SCC with SNP MAF in hotspots	.35[a]	.35[a]	.41[a]	.32[a]	.37[a]	.40[a]

[a] Tested significant ($P < .05$). In calculating the coefficients, we assigned a 1 if a hotspot was detected, and a 0 if it was not.

cover the centers of the true hotspots. The performance is consistent with that for the simulation data with HQ = 90%.

Hotspot Detection in the HapMap ENCODE Regions

We applied our method to the HapMap ENCODE regions with the same set of parameters as in the simulation and validation experiments, and we used unphased data directly. Within the 10 500-kb regions, we detected 88, 110, and 87 hotspots in the CEU, ASI, and YRI populations, respectively. This gives us a total of 172 hotspots (or hotspot clusters, defined as sets of hotspots that overlap across populations) that occur in at least one population. The hotspot positions are listed in table 3. The widths of detected hotspots (or hotspot clusters) range from 0.8 kb to 9.8 kb, with average of 2.4 kb, covering ~8.14% of the studied regions. We downloaded hotspots estimated by LDhot from the ENCODE Project at UCSC, which are also a combination of predictions in the three populations. LDhot reported 95 hotspots (or hotspot clusters) with widths ranging from 2.75 to 16.25 kb, with an average of 4.9 kb. This suggests that the hotspots we identified are at a finer scale, and one hotspot (or hotspot cluster) identified by LDhot may contain several hotspots that we found. The overlapping of the two sets of predicted hotspots (or hotspot clusters) is 75 of the 95 predictions by the LDhot hotspots and 82 of the 172 predictions by our method.

According to the simulation results, our method is not sensitive to population histories, so the possible discrepancy between the real histories and the histories estimated by Schaffner et al.[40] will not lower the power significantly. The power of our method is mainly determined by the background recombination rate. We compared the estimated background rate (after detecting hotspots) in the simulation data and in the HapMap ENCODE data. From the results shown in table 4, it can be observed that the average background rates of the HapMap ENCODE regions are similar to those of simulation data with HQ = 90% and are much lower than those of the simulation data with HQ = 70%. This suggests that the power of our method for the HapMap ENCODE data is comparable to the power (67%) in the simulation data with HQ = 90%. In addition, the estimated background recombination rates of the two regions by sperm-typing experiments (all from the CEU population) are 0.087 and 0.120, which are similar to those of the ENCODE regions. The high power (~86%) achieved for those data also suggests a high power in the ENCODE regions.

Some of the hotspots are not discovered in all populations. Figure 3 shows the numbers of hotspots detected in one, two, and all three populations in the HapMap ENCODE data and those detected in the three populations in simulation data with HQ = 90%. We observe that, compared with the simulation (in which the three populations have exactly the same hotspots), there are more hotspots in the HapMap ENCODE data that are found only in one population, and there are fewer hotspots that are found in all three populations. This discrepancy is significant by the χ^2 test ($P < 1 \times 10^{-5}$), showing a systematic difference between the simulation data and HapMap ENCODE data. We checked known factors that may affect the power of the method, SNP density and SNP MAF, and found no evidence that they cause this difference. There has been a long discussion about whether recombination-rate variations are the same across human populations.[26,30,44,45] If we assume that the recombination rate model and other assumptions underlying the simulation are appropriate for the ENCODE data, the fact that significantly more population-specific hotspots are observed in the real data might be viewed as evidence that the presence of hotspots is not identical in the three populations on the basis of the current data. However, some other possibilities, such as inconsistent intensities of the hotspots in the three populations, may also explain the observed low consensus between the populations.

Correlation between Hotspot Positions and Genomic Features

Many sequence and gene-related features have been reported to be significantly correlated with hotspot positions at different scales, from several megabases to as fine as ~5 kb.[8,42,46,47] On the basis of the 172 hotspots we predicted with the HapMap ENCODE data, we investigated the possible correlation of hotspot positions with some major sequence factors and gene annotations. This was done by comparing the distributions of the major factors inside and outside the predicted hotspots, and a significant difference in the distributions may indicate correlation of the factor with the hotspots. The significance was tested with random permutations. First, the occurrence (for discrete features) or average (for continuous features) of a feature in the detected hotspots was counted. Then, we did permutation by randomly relocating the "hotspots" (without changing their widths) within the whole 5-Mb region, keeping in mind that they do not overlap or touch, and counted the occurrence or average of the feature in the permuted "hotspots." This procedure was done 10,000 times to get the null distribution of the occurrence or average of the feature in the permuted hotspots. The occurrence or average in the true detected hotspots was compared with this null distribution to calculate the *P* value of observing the occurrence or average solely by chance. If the true occurrence or average was significantly larger or smaller than that in the permuted hotspots, we inferred that the feature is enriched or depleted in hotspots; otherwise, a correlation was not observed.

The features and the test results are listed in table 5. It can be seen that high G+C content was enriched in hotspot regions, consistent with previous reports that hotspots have a weak positive correlation with the G+C content.[8,28,42,46,47] Significant correlation with the number of CpG islands was not observed. The relation with repeats on the genome was studied by calculating the length of repeat elements located in hotspots normalized by the width of the hotspots. We observed that repeats are significantly depleted in the predicted hotspots when all types of repeats are taken as a whole. This observation is roughly consistent with that in yeast, where Ty elements (the main family of large dispersed natural repeats) tend to have very low recombination rates.[48] We also studied each family of repeat elements separately by counting the number of repeats that overlap with hotspots. Of the seven most-frequent repeat families that each occur >500 times in the whole 5-Mb region, we observed significant enrichment of Low_complexity, L2, and MIR in hotspots; significant depletion of L1 in hotspots; and no significant

Table 3. Hotspots Detected in the HapMap ENCODE Region

The table is available in its entirety in the online edition of *The American Journal of Human Genetics*.

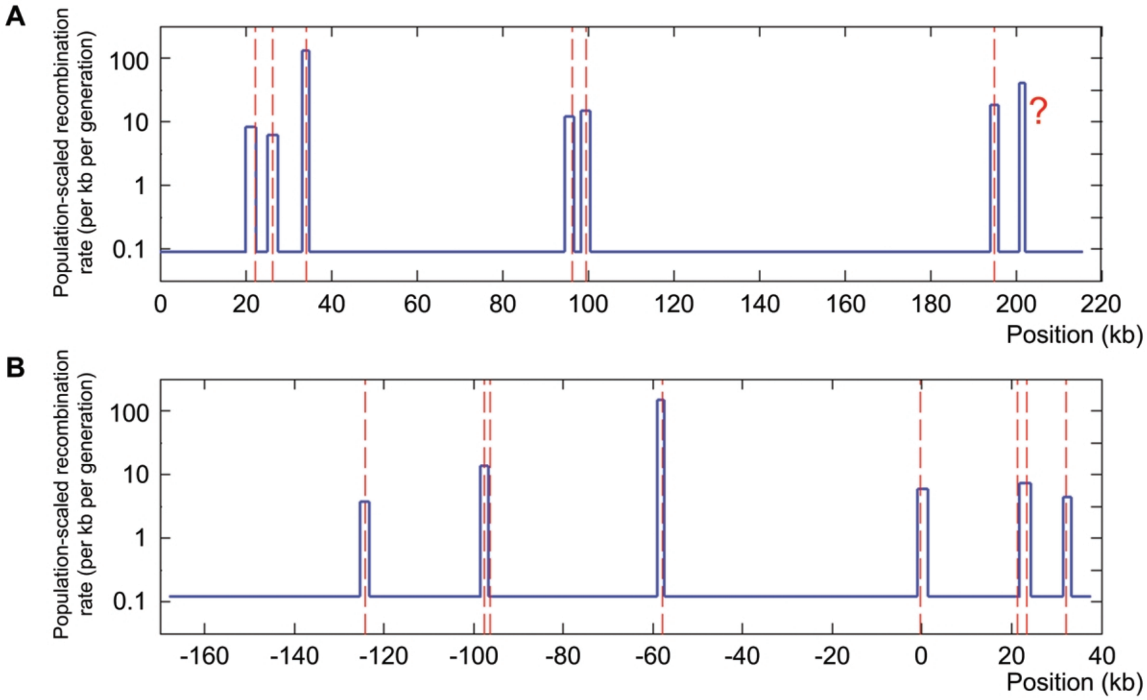

Figure 2. Hotspot detection in the two genomic regions where sperm-typing data are available. The blue lines are the recombination surface we estimated, and peaks in the line are recombination hotspots detected. The centers of true hotspots are shown by red dashed lines. *A,* The 216-kb segment of the class II region of the MHC. From left to right, the true hotspots are *DNA*1, *DNA*2, *DNA*3, *DMB*1, *DMB*2, and *TAP2,* and the question mark (?) indicates the hotspot that was not observed in sperm-typing experiments but that was conjectured by Jeffreys et al.[3] and predicted computationally.[28,29] These hotspots were all detected with our method, in the following order *TAP2* (LLR = 141.4), *DNA*3 (LLR = 137.2), *DMB*2 (LLR = 99.5), *DMB*1 (LLR = 67.6), *DNA*2 (LLR = 43.6), *DNA*1 (LLR = 30.9), and "?" (LLR = 29.3). *B,* The 206-kb segment on chromosome 1. From left to right, the true hotspots are *NID*3, *NID*2a, *NID*2b, *NID*1, MS32, MSTM1a, MSTM1b, and MSTM2.[5] We detected six of them, in the following order: *NID*1 (LLR = 77.0), *NID*2a (LLR = 75.3), MSTM2 (LLR = 58.4), MS32 (LLR = 41.5), MSTM1b (LLR = 36.4), and *NID*3 (LLR = 30.1). In both groups of data, every hotspot we detected contains the center of its corresponding true hotspot.

Table 4. Estimated Background Recombination Rate for the Simulation Data and the HapMap ENCODE Data

	Simulation Data					
	HQ = 90%		HQ = 70%		ENCODE Data	
Population	Mean	SEM	Mean	SEM	Mean	SEM
CEU	.078	.030	.202	.055	.059	.029
ASI	.070	.024	.177	.051	.079	.043
YRI	.133	.043	.362	.074	.165	.064

correlation with Alu, MaLR, and Simple_repeat. These observations are consistent with those of Myers et al.,[8] except that they did not observe significant relation with Low_complexity repeats. We also studied the correlation of the detected hotspots with gene annotations. Among the factors we studied, we observed that hotspots tend to avoid gene regions (from 1 kb upstream of the first exons to 1 kb downstream of the last exons). Of the 172 predicted hotspots, 56 are located at ±1 kb from annotated RefSeq genes. Among them, seven hotspots overlap with the ±1-kb areas around annotated transcription start sites. In yeast, there is a category of α hotspots that occur in promoter regions and that are related to certain transcription factor–binding sites. In humans, α hotspots have been reported in a small-scale study[28] but have not been found in other studies.[8,30] The observation in the current study does not show the correlation of hotspots with promoters but shows a few examples of hotspots in promoter regions.

The Existence of Human β Hotspots

It is known that in yeast open chromatin structure is necessary for the formation of double-stranded breaks (DSBs), which initiate meiotic crossover events.[35,36] To investigate whether such a relationship persists in humans, we calculated the correlation between hotspot positions and DHSSs, which are strong signals for open chromatin structure. Among 144, 143, 26, and 30 DHSSs in the four groups of DHSS data, 26, 24, 4, and 4 overlap with the detected hotspots, respectively. The lengths of DHSSs are also very short (~0.29 kb on average). From the results shown in table 5, we observed that the DHSSs from the first two data sets (the one identified by DNase-chip in the GM06990 line and the one identified by DNase-chip in the $CD4^+$ T cells) are significantly enriched in the detected hotspots. The other two data sets are rather small, and correlations with DHSSs therein are not significant.

In yeast, the hotspots that require open chromatin structure, which usually show DNaseI hypersensitivity, have been termed "β hotspots."[36] The significant correlation of the detected hotspots with DHSSs we observed suggests the existence of similar β hotspots in humans. If we take all DHSSs in the four groups of data together, they overlap with 26 hotspots among the 172, which indicates that ~15% of the hotspots in humans could be of the β type. It is interesting to note that, among the seven hotspots that are located at promoter regions, three also overlap with DHSSs where *cis*-regulatory elements are known to be abundant.

Discussion

In this article, we have presented a new method for detection of recombination hotspots, its validation with simulation and experimentally verified data, and its application to the HapMap ENCODE data. We introduced a TWPLL in the method and adopted models that allow multiple hotspots in a region. Simulation experiments, as well as validation with the two human genome regions that have available sperm-typing data, show that the method is comparable to the best methods, with regard to the detection power and false-positive rate. In addition, the proposed method is computationally fast and can work on both phased and unphased data.

The precision of our method in locating hotspots can be affected by the SNP density, which is high in our study. When the SNP density is low, hotspot locations cannot be determined as precisely, and their lengths should not be limited to 2.4 kb. This can be tackled by adjusting the detected hotspots in the following way. Suppose a hotspot

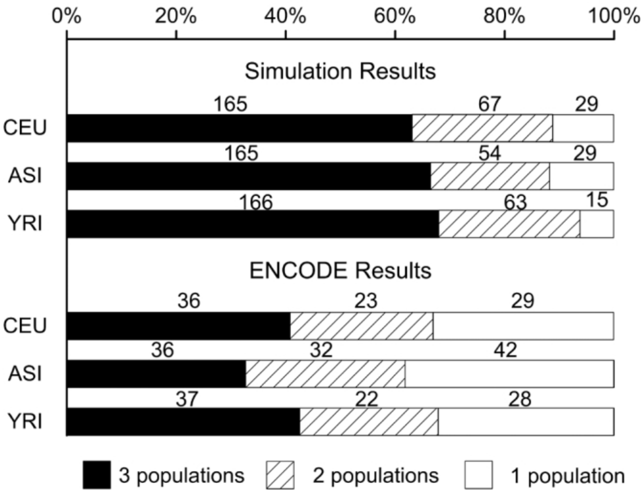

Figure 3. Numbers of hotspots detected in one, two, or all three populations in the simulation study and in the HapMap ENCODE data. The lengths of bars show the percentage of the hotspots in all detected hotspots, and the numbers on the bars are the numbers of hotspots following the corresponding categories. (The numbers of hotspots detected in all three populations are not the same for all populations, since there might be two hotspots in one populations overlapping with single hotspots in other populations.) It can be observed that, in the HapMap ENCODE data, we detected significantly smaller proportion of hotspots shared by all three populations but a larger proportion of hotspots that are detected only in one population, compared with those detected in the simulation data.

Table 5. Correlation between Detected Hotspot Positions and Genomic Features in the HapMap ENCODE Regions

Feature	Mean Occurrence or Average Value: In Putative Hotspots	Mean Occurrence or Average Value: In Random Hotspots	Enriched or Depleted in Putative Hotspots	*P*
Basic sequence features:				
G+C content	.432	.400	Enriched	<.0001
No. of CpG islands	10	7.78	…	.2587
Genomic repeats[a]:				
All families of repeats	.396	.456	Depleted	.0014
Alu	143	156.5	…	.2175
L1	117	149.7	Depleted	.0143
MIR	118	90.1	Enriched	.0061
Simple_repeat	78	63.9	…	.0671
L2	90	63.9	Enriched	.0041
Low_complexity	77	57.7	Enriched	.0180
MaLR	56	50.2	…	.2532
RefSeq genes and related features:				
Gene regions[b]	.2790	.3913	Depleted	.0009
Exonic bases	.0244	.0228	…	.3863
UTRs (5′ + 3′)	.0089	.0127	…	.3288
DHSSs:				
Sites in DNase GM069 Chip	26	13.1	Enriched	.0031
Sites in DNase CD4 Chip	24	13.0	Enriched	.0185
Sites in DNase CD4 MPSS	4	2.5	…	.2445
Sites in DNase CD4-act MPSS	4	2.7	…	.2889

[a] Only families of repeats that occur >500 times in the studied regions are listed here. All other repeat families are not tested significantly related with the hotspots.

[b] Gene regions are calculated as 1 kb upstream of the first exon to 1 kb downstream from the last exon.

(⩽2.4 kb) is detected whose starting location is between SNPs i and $i+1$ and whose ending location is between SNPs j and $j+1$; we adjust its location to be from the position of SNP i to that of SNP $j+1$. We did simulation with ~0.7 common SNPs per kb, similar to the SNP density of phase II of the main HapMap project. After the above adjustment to the hotspot boundaries, the power is ~59% for each population, with the same false-positive rate discussed above. The average position offset is ~1.5 kb, and >99% of detected hotspots cover the centers of true hotspots in each population. Therefore, the method can surely be applied to phase II HapMap data. When SNP density is too low—for example, 0.2 common SNPs per kb—our method is not recommended.

Another issue is the choice of N and ω_k in TWPLL. Simulations show that our method is not sensitive to either of them. For the data we used, $N=5$, 7, or 9 results in almost the same power, and, for $N=7$, $\omega_k=1/k$ and $\omega_k=1-(k-1)/N$ perform comparably. So, the choice of N and ω_k is not so critical within a certain range, and we suggest that $N=7$ and $\omega_k=1/k$ is generally a good choice for most data sets.

When N and ω_k are fixed, the key parameter to be decided is the threshold T, representing the trade-off between power and false-positive rate. Simulation results show that T can be affected by the SNP density and sample size. A lower T should be chosen for lower SNP density or smaller sample size, to give the same false-positive rate. For example, $T=19$ is appropriate when the SNP density decreases to the level of the phase II HapMap project, and $T=23$ is proper if the diploid sample size is 50. Because of the speed of HotspotFisher, it is straightforward to calibrate the choice of T for a specific set of real data by applying HotspotFisher with different values of T to data simulated with features that match the real data.

A single background rate in each region is assumed in our model. In practice, the background rates may vary across the chromosome, so a long chromosome segment should be divided into smaller pieces, to detect hotspots in each piece. Analysis of very small regions would lead to imprecise estimates (large variance), whereas analysis of regions that are too big may lead to poor estimates due to biases from the assumption of a constant background rate. We suggest 100–500 kb to be a good range of choices, and 200 kb may be chosen as the default.

In all our simulations, we applied the proposed method to unphased data directly. An alternative strategy is to detect hotspots on the basis of the haplotypes inferred from genotypes by use of software such as PHASE.[31,32] Since adopting such an additional step often increases the computational cost substantially and since our method can directly handle unphased data, we did not use this strategy in our study. However, it is recommended that others use haplotype data if the data are highly reliable; otherwise, the use of genotypes directly is reliable and convenient.

From another perspective, because the method can work on unphased data efficiently, it can also be incorporated into some haplotype-inference methods. Many current methods for inferring haplotypes from genotypes assume no recombination or minimum recombination events. Users may use the proposed method to detect recombination hotspots first and then use those haplotype-inference methods to infer haplotypes between each pair of adjacent hotspots. This strategy would increase the accuracy of haplotype inference, especially when applied to long genomic regions.

Applying the proposed method to the HapMap ENCODE data, we identified 172 putative hotspots in the 10 500-kb regions. We observed that hotspots are not completely identical across the three populations. Since there are many factors that can affect the prediction in the populations, the observation may indicate the existence of population-specific hotspots and/or that the intensity of the same hotspots in different populations is different, but more data and further experiments are needed to draw a conclusion on this point.

Evidence is accumulating that meiotic crossovers in humans and in yeast may share similar mechanisms—for example, similar short lengths of hotspots and similar correlation with G+C content.[34–36] In yeast, hotspots share no particular sequence features.[34–36] In humans, a recent report has shown that the presence or absence of at least some hotspots is not controlled by the sequence or polymorphisms.[43] An important determinant of the β-type hotspots in yeast is the open chromatin structure, and our results show that a significant portion of human hotspots may share a similar mechanism.

Acknowledgments

This work is partially supported by National Basic Research Program of China grant 2004CB518605 and National Natural Science Foundation of China grants 60234020 and 60575014. M.Q.Z. is partially supported by Changjiang Chair Professor Program. We also thank THPCC for their help during this work. Thanks to the anonymous reviewers for their constructive suggestions.

Web Resources

The URLs for data presented herein are as follows:

Cosi, http://www.broad.mit.edu/personal/sfs/cosi/
HapMap ENCODE, http://www.hapmap.org/downloads/encode1.html.en
Jun Li’s Web site, http://bioinfo.au.tsinghua.edu.cn/member/~lijun/
LDhat version 2.0, http://www.stats.ox.ac.uk/~mcvean/LDhat/
ENCODE Project at UCSC, http://genome.ucsc.edu/ENCODE/

References

1. Jeffreys AJ, Murray J, Neumann R (1998) High-resolution mapping of crossovers in human sperm defines a minisatellite-associated recombination hotspot. Mol Cell 2:267–273
2. Jeffreys AJ, Ritchie A, Neumann R (2000) High resolution analysis of haplotype diversity and meiotic crossover in the human *TAP2* recombination hotspot. Hum Mol Genet 9:725–733
3. 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
4. Cullen M, Perfetto SP, Klitz W, Nelson G, Carrington M (2002) High-resolution patterns of meiotic recombination across the human major histocompatibility complex. Am J Hum Genet 71:759–776
5. 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
6. Holloway K, Lawson VE, Jeffreys AJ (2006) Allelic recombination and *de novo* deletions in sperm in the human β-globin gene region. Hum Mol Genet 15:1099–1111
7. McVean GA, Myers SR, Hunt S, Deloukas P, Bentley DR, Donnelly P (2004) The fine-scale structure of recombination rate variation in the human genome. Science 304:581–584
8. 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
9. The International HapMap Consortium (2005) A haplotype map of the human genome. Nature 437:1299–1320
10. Kruglyak L (1999) Prospects for whole-genome linkage disequilibrium mapping of common disease genes. Nat Genet 22:139–144
11. Jorde LB (2000) Linkage disequilibrium and the search for complex disease genes. Genome Res 10:1435–1444
12. Ott J (2000) Predicting the range of linkage disequilibrium. Proc Natl Acad Sci USA 97:2–3
13. Pritchard JK, Przeworski M (2001) Linkage disequilibrium in humans: models and data. Am J Hum Genet 69:1–14
14. Reich DE, Cargill M, Bolk S, Ireland J, Sabeti PC, Richter DJ, Lavery T, Kouyoumjian R, Farhadian SF, Ward R, Lander ES (2001) Linkage disequilibrium in the human genome. Nature 411:199–204
15. Zhang K, Calabrese P, Nordborg M, Sun F (2002) Haplotype block structure and its applications to association studies: power and study designs. Am J Hum Genet 71:1386–1394
16. Zhang K, Qin ZS, Liu JS, Chen T, Waterman MS, Sun F (2004) Haplotype block partitioning and tag SNP selection using genotype data and their applications to association studies. Genome Res 14:908–916
17. Wall JD (2000) A comparison of estimators of the population recombination rate. Mol Biol Evol 17:156–163
18. Stumpf MP, McVean GA (2003) Estimating recombination rates from population-genetic data. Nat Rev Genet 4:959–968
19. Kingman JFC (1982) The coalescent. Stoch Proc Appl 13:235–248
20. Griffiths RC, Marjoram P (1996) Ancestral inference from samples of DNA sequences with recombination. J Comput Biol 3:479–502
21. Kuhner MK, Yamato J, Felsenstein J (2000) Maximum likelihood estimation of recombination rates from population data. Genetics 156:1393–1401
22. Fearnhead P, Donnelly P (2001) Estimating recombination rates from population genetic data. Genetics 159:1299–1318
23. Hudson RR (2001) Two-locus sampling distributions and their application. Genetics 159:1805–1817
24. Fearnhead P, Donnelly P (2002) Approximate likelihood methods for estimating local recombination rates (with discussion). J R Statist Soc B 64:657–680

25. McVean G, Awadalla P, Fearnhead P (2002) A coalescent-based method for detecting and estimating recombination from gene sequences. Genetics 160:1231–1241
26. Li N, Stephens M (2003) Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. Genetics 165:2213–2233
27. Smith NG, Fearnhead P (2005) A comparison of three estimators of the population-scaled recombination rate: accuracy and robustness. Genetics 171:2051–2062
28. Zhang J, Li F, Li J, Zhang MQ, Zhang X (2004) Evidence and characteristics of putative human α recombination hotspots. Hum Mol Genet 13:2823–2828
29. Fearnhead P, Harding RM, Schneider JA, Myers S, Donnelly P (2004) Application of coalescent methods to reveal fine-scale rate variation and recombination hotspots. Genetics 167:2067–2081
30. Fearnhead P, Smith NG (2005) A novel method with improved power to detect recombination hotspots from polymorphism data reveals multiple hotspots in human genes. Am J Hum Genet 77:781–794
31. Stephens M, Donnelly P (2003) A comparison of Bayesian methods for haplotype reconstruction from population genotype data. Am J Hum Genet 73:1162–1169
32. Stephens M, Smith NJ, Donnelly P (2001) A new statistical method for haplotype reconstruction from population data. Am J Hum Genet 68:978–989
33. Crawford DC, Bhangale T, Li N, Hellenthal G, Rieder MJ, Nickerson DA, Stephens M (2004) Evidence for substantial fine-scale variation in recombination rates across the human genome. Nat Genet 36:700–706
34. Nishant KT, Rao MR (2006) Molecular features of meiotic recombination hot spots. Bioessays 28:45–56
35. Kauppi L, Jeffreys AJ, Keeney S (2004) Where the crossovers are: recombination distributions in mammals. Nat Rev Genet 5:413–424
36. Petes TD (2001) Meiotic recombination hot spots and cold spots. Nat Rev Genet 2:360–369
37. Crawford GE, Holt IE, Whittle J, Webb BD, Tai D, Davis S, Margulies EH, Chen Y, Bernat JA, Ginsburg D, Zhou D, Luo S, Vasicek TJ, Daly MJ, Wolfsberg TG, Collins FS (2006) Genome-wide mapping of DNase hypersensitive sites using massively parallel signature sequencing (MPSS). Genome Res 16:123–131
38. Frisse L, Hudson RR, Bartoszewicz A, Wall JD, Donfack J, Di Rienzo A (2001) Gene conversion and different population histories may explain the contrast between polymorphism and linkage disequilibrium levels. Am J Hum Genet 69:831–843
39. Fearnhead P (2003) Consistency of estimators of the population-scaled recombination rate. Theor Popul Biol 64:67–79
40. Schaffner SF, Foo C, Gabriel S, Reich D, Daly MJ, Altshuler D (2005) Calibrating a coalescent simulation of human genome sequence variation. Genome Res 15:1576–1583
41. Hudson RR (2002) Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics 18:337–338
42. Kong A, Gudbjartsson DF, Sainz J, Jonsdottir GM, Gudjonsson SA, Richardsson B, Sigurdardottir S, Barnard J, Hallbeck B, Masson G, Shlien A, Palsson ST, Frigge ML, Thorgeirsson TE, Gulcher JR, Stefansson K (2002) A high-resolution recombination map of the human genome. Nat Genet 31:241–247
43. Neumann R, Jeffreys AJ (2006) Polymorphism in the activity of human crossover hotspots independent of local DNA sequence variation. Hum Mol Genet 15:1401–1411
44. Ptak SE, Hinds DA, Koehler K, Nickel B, Patil N, Ballinger DG, Przeworski M, Frazer KA, Paabo S (2005) Fine-scale recombination patterns differ between chimpanzees and humans. Nat Genet 37:429–434
45. Winckler W, Myers SR, Richter DJ, Onofrio RC, McDonald GJ, Bontrop RE, McVean GA, Gabriel SB, Reich D, Donnelly P, Altshuler D (2005) Comparison of fine-scale recombination rates in humans and chimpanzees. Science 308:107–111
46. Jensen-Seaman MI, Furey TS, Payseur BA, Lu Y, Roskin KM, Chen CF, Thomas MA, Haussler D, Jacob HJ (2004) Comparative recombination rates in the rat, mouse, and human genomes. Genome Res 14:528–538
47. Smith AV, Thomas DJ, Munro HM, Abecasis GR (2005) Sequence features in regions of weak and strong linkage disequilibrium. Genome Res 15:1519–1534
48. Koren A, Ben-Aroya S, Kupiec M (2002) Control of meiotic recombination initiation: a role for the environment? Curr Genet 42:129–139