Am. J. Hum. Genet. 77:781–794, 2005

A Novel Method with Improved Power To Detect Recombination Hotspots from Polymorphism Data Reveals Multiple Hotspots in Human Genes

Paul Fearnhead and Nick G. C. Smith*

Department of Mathematics and Statistics, Lancaster University, Lancaster, United Kingdom

We introduce a new method for detection of recombination hotspots from population genetic data. This method is based on (*a*) defining an (approximate) penalized likelihood for how recombination rate varies with physical position and (*b*) maximizing this penalized likelihood over possible sets of recombination hotspots. Simulation results suggest that this is a more powerful method for detection of hotspots than are existing methods. We apply the method to data from 89 genes sequenced in African American and European American populations. We find many genes with multiple hotspots, and some hotspots show evidence of being population-specific. Our results suggest that hotspots are randomly positioned within genes and could be as frequent as one per 30 kb.

Introduction

Recombination is a fundamental evolutionary process that shapes patterns of sequence variation by breaking up allelic associations. Recombination is intertwined with the other fundamental molecular evolutionary processes of mutation and selection, since recombination itself may be mutagenic and because recombination aids selection by reducing Hill-Robertson inference. In addition, understanding recombination rates is crucial for practical applications in human genetics, most notably in the use of association mapping of complex traits. However, recombination rates per generation are small, which means that pedigree studies can reveal recombination rate variation only at the megabase scale. The recent development of sperm typing allows the consideration of very large numbers of male meioses, to reveal fine-scale variation in recombination rates, with resolution limited only by the density of polymorphic markers. Sperm typing has revealed the existence of narrow recombination-rate hotspots of ~1–2 kb in size. But sperm typing is laborious and is not applicable to genomewide studies, so increasing attention has been paid to the inference of recombination-rate hotspots through use of population-genetics models, to infer recombination rates from population data. Such population-genetics approaches yield recombination-rate estimates that differ from those of sperm typing and pedigree studies in a number of important respects. Population-genetics methods estimate a population-scaled compound parameter rather than the raw recombination rates given by the pedigree and sperm-typing methods. The time scale over which recombination rates are inferred with population-genetics methods is also different from the other methods, since those methods consider the history of polymorphism in the sample rather than simply measure present-day recombination rates (sperm typing) or recombination rates in recent generations (pedigree studies). Finally, population-genetics methods estimate a sex-averaged recombination rate (at least for autosomal data), in contrast to sperm typing, which measures only male recombination rates, and pedigree studies, which can measure both male and female recombination rates.

Recent evidence from both sperm data (Jeffreys et al. 2001, 2005) and population data (Crawford et al. 2004; McVean et al. 2004) show that there is large local variation in the recombination rate across the human genome. A simple qualitative description of this is that there are relatively large regions (on the order of 10–100 kb) of the genome that have a small "background" recombination rate, and these regions are separated by recombination "hotspots." Recombination hotspots are generally 1–2 kb in width (Jeffreys et al. 2001, 2005) and have a recombination rate that is $\geq$1 orders of magnitude larger than the background rate.

Currently, there is little understanding about the biological factors that produce hotspots. Jeffreys and Neumann (2002) show a hotspot that is controlled by the nucleotide at a single polymorphic site. Comparisons of hotspots in humans and chimps (Ptak et al. 2005; Winckler et al. 2005) show that recombination hotspots are not conserved between humans and chimps. Thus, hotspots appear to evolve over time scales substantially shorter than the 10–12 million years of evolution that

Received May 10, 2005; accepted for publication August 25, 2005; electronically published September 16, 2005.

Address for correspondence and reprints: Dr. Paul Fearnhead, Department of Mathematics and Statistics, Fylde College, Lancaster University, Lancaster, LA1 4YF, United Kingdom. E-mail: p.fearnhead@lancaster.ac.uk

* Deceased.

0002-9297/2005/7705-0009$15.00

separate humans and chimps. Whether the rate of recombination-rate evolution is sufficiently quick to cause different human populations to have different hotspots is an open question, although large recombination-rate differences among individuals have been reported.

Detection of hotspots is important both for the design and analysis of association studies aimed at finding genetic factors of diseases and for the correct interpretation of patterns of diversity in population data. Furthermore, detection of a large number of recombination hotspots from the extensive human population data that is currently being generated will produce substantial data that can be used to address questions relating to the biology and evolution of these hotspots.

Currently, there are three methods of detection of hotspots from population data (Li and Stephens 2003; Fearnhead et al. 2004; McVean et al. 2004). A recent comparison of these three methods on a 206-kb region of chromosome 1 for which sperm typing had been separately used to detect hotspots (Jeffreys et al. 2005) suggests that the method of Fearnhead et al. (2004) is the most powerful. Of eight hotspots found by sperm typing, the method of Fearnhead et al. (2004) found seven, with one false-positive result. By comparison, the method of McVean et al. (2004) found four hotspots, with zero false-positive results, and that of Li and Stephens (2003) found five, with three false-positive results. Here, we present an extension of the method of Fearnhead et al. (2004) that, in simulation studies, produces an increase in power (for comparable false-positive rates, the new method found 65% of hotspots, as compared with just 53% for the old method [see the "Results" section]). We have applied our new method to polymorphism data for African American (AA) and European American (EA) samples from 89 genes, and we have found strong evidence of multiple recombination hotspots in single genes and some indication of potential differences in hotspots between the two populations.

Material and Methods

Data

We used human sequence data generated by the SeattleSNPs Program for Genomic Applications. The SeattleSNPs database provides polymorphism data that are based on resequencing of a large number of candidate genes thought likely to be involved in genetic disease and so in no way represents a random sampling of the genome. Sequencing spanned the transcribed regions. We chose 89 genes (*AGTRAP, ALOX12, ALOX15, ALOX5AP, APOH, C3, CAT, CD36, CD9, CHUK, CKM, CRF, CSF2, CSF3R, CYP4A11, DCN, F10, F12, F13A1, F3, F5, F9, GP1BA, HABP2, ICAM1, IFNAR1, IFNGR1, IFNGR2, IGF2AS, IL10RB, IL11, IL11RA, IL12RB2, IL15RA, IL16, IL17, IL1R1, IL1R2, IL1RN, IL20, IL21R, IL26, IL2RA, IL2RB, IL4R, IL5RA, IL6, IL7R, IL9R, IRAK4, ITGA2, ITGA8, JAK3, MAP3K8, MMP3, NFKBIA, NOS3, PLAUR, PLG, PLTP, PON1, PON2, PPARA, PPARG, PROC, PTGS1, RIPK1, SELE, SELL, SELP, SERPINA5, SFTPA1, SFTPA2, SFTPB, STAT3, STAT6, TF, TFPI, THBS4, TIRAP, TNFAIP2, TNFRSF1A, TNFRSF1B, TRAF2, TRAF6, TRPV6, TTRAP, TYK2,* and *VCAM1* [see table 1 and the SeattleSNPs Web site for more information]) from an early version of the data set, discarding only those genes that appeared to have too few SNPs or to be too short (in physical distance) to provide sufficient power to infer hotspots.

The SeattleSNPs data set contains information from two populations: 23 unrelated EAs from CEPH pedigrees and 24 AAs from the African-American Human Variation Panel. On average, genes in the SeattleSNPs data set are 20–25 kb long and contain 115 segregating sites, of which 100 are segregating in AAs, and 65 in EAs. For the analyses of recombination-rate variation, we required haplotypes that are not directly ascertained in the SeattleSNPs data set but that are instead inferred using PHASE v2.1 software (Stephens et al. 2001; Stephens and Donnelly 2003). Note that only sites with a minor-allele frequency >5% were used for haplotype reconstruction and, hence, for hotspot detection. Simulations indicate that the estimation of haplotypes by PHASE causes very little bias in recombination-rate estimation (Smith and Fearnhead 2005).

Approximate Likelihood Methods

Our approach to detection of hotspots is based on use of the approximate marginal-likelihood (AML) method of Fearnhead and Donnelly (2002) to obtain (approximate) likelihood curves for the recombination rate within short subregions of each data set. Each subregion consists of six (consecutive) SNPs; we calculated likelihood curves for each such subregion within our data set. The likelihood curves were calculated using the program sequenceLDsr, which is available at P.F.'s Web site. This approach uses importance sampling (see Stephens and Donnelly [2000] and Fearnhead and Donnelly [2001] for more details) and is based on (*a*) simulation of a set of possible genealogical histories for the data, (*b*) calculation of an (approximate) likelihood curve for the recom-

Table 1

Inferred Hotspots: Single Analysis of Each Population

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

bination rate for each history, and (*c*) combination of these curves, by use of a suitable weighted average, to produce the final likelihood curve for the subregion. In practice, we used 100,000 histories in our calculation (step *a* above) for each subregion, and the computational cost was 10–30 minutes for each subregion.

Note that we attempt to detect recombination hotspots, using recombination models that consider only crossing over; that is, gene conversion is not explicitly modeled. However, previous simulation studies have indicated that variation in gene-conversion rates should be revealed to some extent by methods that consider only crossing over (Smith and Fearnhead 2005).

This AML has been shown to give a very good approximation of the true likelihood curve and is 1–3 orders of magnitude quicker to compute (Fearnhead and Donnelly 2002). Although the likelihood curve is calculated under the assumption of a constant-sized, panmictic population, inferences that are based on it are reasonably robust (in terms of estimation of relative rates of recombination) to a variety of deviations from this model (see Smith and Fearnhead 2005). The likelihood curve is also calculated under the assumption of a constant recombination rate across the subregion. Smith and Fearnhead (2005) show that, if the recombination rate varies across the subregion, then the method estimates an average recombination rate across the subregion.

Detecting Hotspots: Single Population

Consider inferring hotspots within a single gene with S SNPs (or segregating sites). Let $l_i(\rho)$ denote the log AML curve for the ith subregion of the gene that extends from the ith to the $(i + 5)$th SNP, inclusive. (Our approach can be used for subregions that include either fewer or more than six SNPs, and the optimal width of subregion may depend on the SNP density; here, we follow the approach of Fearnhead et al. [2004]). We use $l_i(\rho)$ for $i = 1, \ldots, S - 5$ to fit a hotspot model of how the recombination rate varies with the position along the gene. In particular, we consider recombination surfaces $\rho(x)$, which gives the local recombination rate (per kb) at any position of the gene, of the form

$$\rho(x) = \begin{cases} \rho_1 & \text{for } s_1 \leqslant x \leqslant e_1 \text{ ,} \\ \vdots & \quad \vdots \\ \rho_h & \text{for } s_h \leqslant x \leqslant e_h \text{ ,} \\ \rho_b & \text{otherwise ,} \end{cases} \tag{1}$$

where ρ_b is the background recombination rate and h is the number of hotspots, with the ith hotspot having rate ρ_i and extending from position s_i to e_i. We assume that hotspots do not overlap or touch. See figure 1 for an example.

We define a penalized log likelihood for a recombination surface $\rho(x)$ by

$$\mathrm{Pl}[\rho(x)] = \sum_{i=1}^{S-5} l_i(\rho_i) - \lambda h \text{ ,} \tag{2}$$

where ρ_i is the average recombination rate across subregion i, as specified by the recombination surface $\rho(x)$ (see fig. 1), and λ is a positive constant included to penalize overfitting with hotspots. (Note that the

$$\sum_{i=1}^{S-5} l_i(\rho_i)$$

term is itself not a true log likelihood, since dependence between the data within different subregions is being ignored.)

We estimate the number and position of the hotspots by maximizing $Pl[\rho(x)]$ with respect to recombination surfaces of the form (eq. [1]). This gives a "maximum penalized-likelihood estimate" of the recombination surface from which the number and positions of the hotspots can be read. This approach is an extension of that of Fearnhead et al. (2004). The approach of Fearnhead et al. (2004) was to consider each subregion individually and to evaluate the evidence of a hotspot within that subregion by use of a likelihood-ratio test. The approach proposed here aims to use the information in all subregions that contain (part of) a hotspot to detect the hotspot.

Maximization of Penalized Likelihood

We maximize the penalized log likelihood (eq. [2]), using a recursive segmentation that iteratively adds hotspots to the recombination surface. Formally, we proceed as follows:

A. Estimate the background rate, $\rho_b^{(0)}$, and set $\rho^{(0)}(x)$ to be equal to $\rho_b^{(0)}$ for all x;
B. Given a current recombination surface $\rho^{(i)}(x)$, evaluate the change in penalized log likelihood obtained by adding different hotspots to this surface, subject to the condition that each new considered hotspot does not overlap or touch a hotspot in $\rho^{(i)}(x)$;
C. Add hotspots to $\rho^{(i)}(x)$ that increase the penalized log likelihood. The order in which hotspots are added is determined by the amount they increase the penalized log likelihood (largest increase first) and are subject to each new hotspot not overlapping or touching any hotspots that have already been added; and
D. If no new hotspots are added in step C, then the estimated recombination surface is $\rho^{(i)}(x)$. Otherwise, re-estimate the background rate, update the

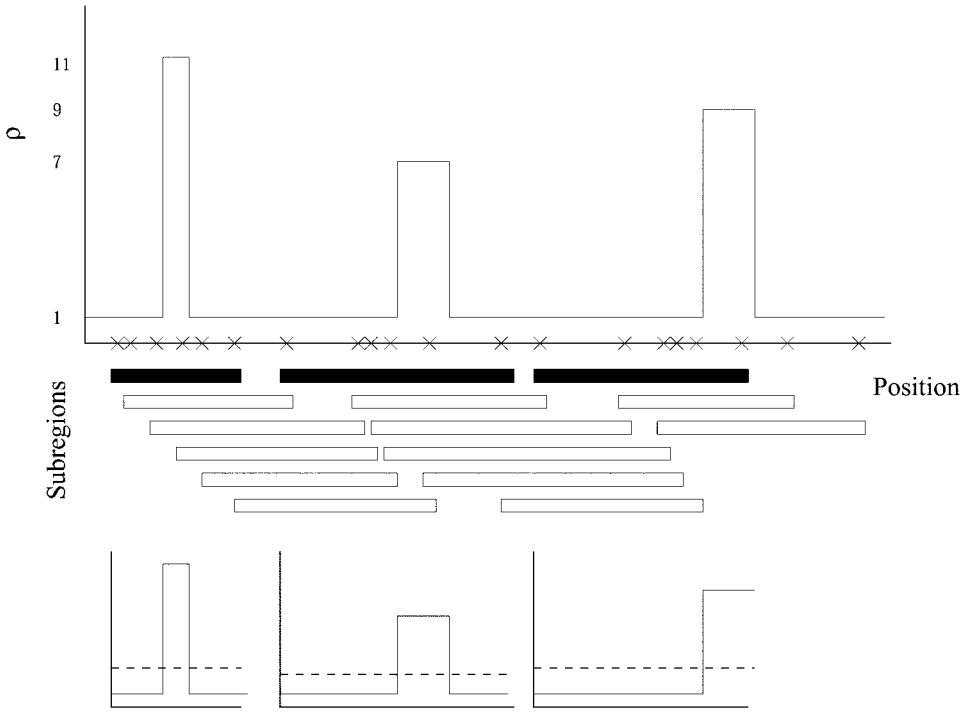

Figure 1 An example recombination surface of a gene. This surface consists of a background recombination rate of 1 and three hotspots with lengths of 1 kb, 2 kb, and 2 kb and with recombination rates of 11, 9, and 7, respectively. The position of 20 SNPs are marked by crosses, and these SNPs define 15 subregions denoted by rectangles below the *X*-axis. The penalized likelihood for the recombination surface is calculated from the AMLs for each of these subregions. The AMLs are calculated under an assumption of constant recombination rate across the subregion. The recombination surface restricted to the 1st, 7th, and 13th subregions (*blackened rectangles*) are shown at the bottom of the figure, together with the average recombination rate defined by the recombination surface (*dashed line*). For example, for subregion 1, the recombination surface is 4 kb with a recombination rate of 1 and 1 kb with a recombination rate of 11, which produces an average recombination rate of 3 kb. This average rate is the value of the recombination rate at which the AML for this subregion is evaluated (the ρ_1 in eq. [2]).

current recombination surface $\rho^{(i+1)}(x)$, and return to step B.

To estimate the background recombination rate at each iteration, we use the composite likelihood of Fearnhead and Donnelly (2002), after exclusion of any subregions that contain part of a hotspot. For estimation of the background rate in step A, we first use the method of Fearnhead et al. (2004) to detect subregions that include hotspots, and we omit those in the composite likelihood. When considering new hotspots in step B, we consider all hotspots specified by a grid of possible start positions and lengths for the hotspot. In the analyses here, the grid had start positions every 250 bp, and the hotspots lengths were either 250 bp to 3 kb or 1 kb to 3 kb, in steps of 250 bp. The maximization was performed using the function HotspotEstimate written in R.

This approach does not find the true maximum penalized-likelihood estimate. Sources of error include (*a*) consideration of only a grid of possible hotspots, (*b*) misestimation of the background recombination rate, and (*c*) the recursive procedure fitting a single hotspot to a cluster of hotspots. The errors due to source *a* are small compared with the uncertainty in the data about the position of hotspots. Any errors that are due to source *c* can be corrected by reanalysis of hotspots that are detected, to determine whether they are single hotspots or hotspot clusters. Errors due to source *b* are potentially more problematic; accurately estimating a background rate requires the exclusion of subregions that contain hotspots, but finding such subregions may depend on a reasonable estimate of the background rate. Therefore, if, in step A, we overestimate the recombination rate (e.g., because of too many hotspots within the gene), then that may preclude us from correctly detecting the hotspots that are there.

Specifying the Likelihood Penalty

The accuracy of our penalized-likelihood approach to finding hotspots depends on the choice of the penalty for hotspots λ. To determine a suitable choice for λ, we resorted to simulation studies (see below). We chose $\lambda = 16$, which gave an approximate false-positive rate

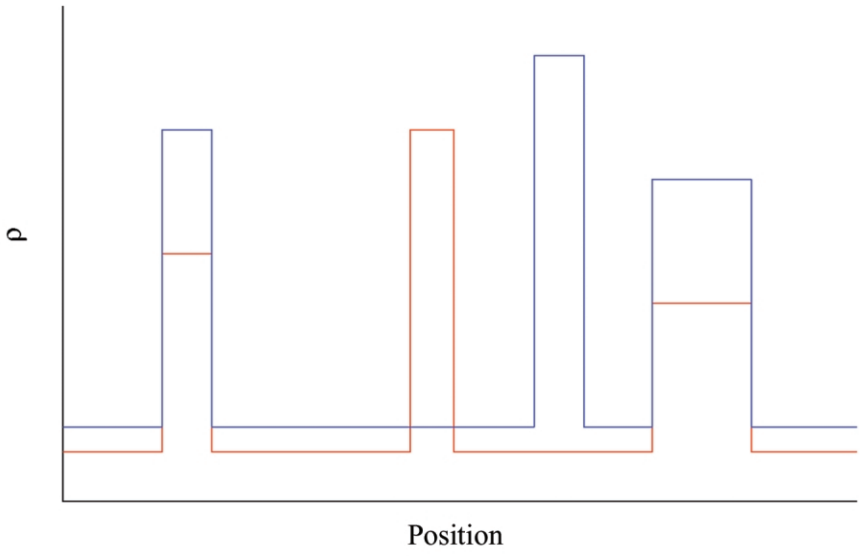

Figure 2 An example two-population recombination surface. The recombination surfaces for population 1 (*red*) and population 2 (*blue*) are shown for a hypothetical gene. In this example, the effective population size of population 2 is 1.5 times that of population 1; the two populations share two hotspots, and each has one population-specific hotspot.

of one hotspot per 60 25-kb genes (see the "Results" section for further details).

We also introduced an edge correction to this hotspot penalty. Hotspots at most positions within a gene will occur within at least five different subregions, whereas hotspots at the edge of the gene may occur within fewer subregions. To account for this and for the positive correlation between likelihood curves at nearby subregions, we reduce the hotspot penalty for such hotspots. The penalty incurred for a hotspot ranges from 8 (if it appears in just one subregion) to 16 (if it appears in five or more subregions); these different penalties were calculated from the simulation studies, to maintain the same false-positive rate at the edge of a gene as elsewhere.

Detecting Hotspots: Joint Analysis

We also considered an extension of this approach that jointly infers hotspots that are based on the data from two populations (in our application, EA and AA populations). Our aim is to jointly infer the recombination surfaces for both populations, $\rho^{(1)}(x)$ and $\rho^{(2)}(x)$. Our model for these recombination surfaces depends on γ (the ratio of effective population sizes between the two populations), ρ_b (the background recombination rate in population 1), and three types of hotspots: those in only population 1, those in only population 2, and those in both populations. We denote by "$h^{(1)}$," "$h^{(2)}$," and "h," respectively, the number of each type of hotspot. Our model for the recombination surface is

$$\rho^{(1)}(x) = \begin{cases} \rho_1^{(1)} & \text{for } s_1^{(1)} \leqslant x \leqslant e_1^{(1)}, \\ \vdots & \vdots \\ \rho_{h^{(1)}}^{(1)} & \text{for } s_{h^{(1)}}^{(1)} \leqslant x \leqslant e_{h^{(1)}}^{(1)}, \\ \rho_1 & \text{for } s_1 \leqslant x \leqslant e_1, \\ \vdots & \vdots \\ \rho_h & \text{for } s_h \leqslant x \leqslant e_h, \\ \rho_b & \text{otherwise}, \end{cases}$$

$$\rho^{(2)}(x) = \begin{cases} \rho_1^{(2)} & \text{for } s_1^{(2)} \leqslant x \leqslant e_1^{(2)}, \\ \vdots & \vdots \\ \rho_{h^{(2)}}^{(2)} & \text{for } s_{h^{(2)}}^{(2)} \leqslant x \leqslant e_{h^{(2)}}^{(2)}, \\ \gamma\rho_1 & \text{for } s_1 \leqslant x \leqslant e_1, \\ \vdots & \vdots \\ \gamma\rho_h & \text{for } s_h \leqslant x \leqslant e_h, \\ \gamma\rho_b & \text{otherwise}, \end{cases} \tag{3}$$

We again assume that hotspots do not overlap within each of the two populations. Our model is based on the recombination surfaces of the two populations differing only by a constant of proportionality (the ratio of effective population sizes), except at population-specific hotspots. We include the possibility of population-specific hotspots because of the biological evidence that hotspots evolve over time. For an example of our two-population recombination surface, see figure 2.

We define a penalized log likelihood for a recombination surface $\rho(x)$ by

$$\mathrm{Pl}[\rho^{(1)}(x),\rho^{(2)}(x)] = \sum_{j=1}^{2}\left\{\sum_{i=1}^{S-5} l_i^{(j)}(\rho_i^{(j)})\right\} - \lambda(h^{(1)} + h^{(2)} + h)\ , \qquad (4)$$

where $l_i^{(j)}(\cdot)$ is the log AML for subregion i in population j, and $\rho_i^{(j)}$ is the average recombination rate in subregion i of population j, as defined by $\rho^{(j)}(x)$. We assume the same penalty for each of the three possible hotspots, since each hotspot introduces the same number of parameters (three: size, start, and end of hotspot) regardless of its type.

We maximize equation (4) with respect to recombination surfaces of the form (3). We perform the maximization, using a recursive segmentation procedure similar to the one described above, implemented in R via a function HotspotEstimate2. We estimate the ratio of effective population sizes on the basis of the ratio in the estimates of average recombination rate in the two populations. We again used simulation to choose $\lambda = 18$, which gave an approximate false-error rate of one hotspot per 20 25-kb genes.

Coalescent Simulations

We simulated polymorphism data, using coalescent models for two different purposes: simulations without recombination hotspots were used to choose the likelihood penalty parameter for the penalized-likelihood method, and simulations with recombination hotspots were used to test the power of recombination-hotspot detection methods. For both classes of simulations, three sets of 100 data sets each were generated corresponding to three alternative demographic histories: a null history of constant population size and a panmictic population, a demographic history thought to match that of AAs, and a demographic history thought to match that of EAs. All simulations were for 50 samples of 25-kb sequences, chosen to match the SeattleSNPs data set. For simulation sets 2, 3, 5, and 6, we simulated genotype data by randomly combining the 50 haplotypes to produce 25 genotypes, and we used PHASE to infer haplotypes. The inferred haplotypes were then used in our analysis.

Simulation set 1: constant recombination rate and null demographic model.—These sequence data were simulated in two stages. First, the ms program (Hudson 2002) was used to simulate a tree file (consisting of a set of genealogies and branch lengths for different portions of the sequence) under the standard neutral coalescent, with a constant rate of crossing-over across the sequence. The recombination rate $\rho = 4N_e r$ was determined for each simulation by simulating r from the empirical distribution of autosomal crossing-over rate in the deCODE pedigree study (Kong et al. 2002) and by applying the standard assumption of $N_e = 10{,}000$. The average autosomal crossing-over rate in the deCODE pedigree study (excluding centromeres) is 1.2 cM/Mb, which means that the average ρ is 0.48 per kb. DNA sequence data was then simulated, according to a two-allele finite-sites mutation model, on the basis of the tree file, using the seq-gen program (Rambaut and Grassly 1997). The population-scaled mutation parameter θ was set to 0.9 per kb, similar to average nucleotide diversity in the pooled AA and EA SeattleSNPs data set, and mutation rates were modeled as constant among sites.

Simulation sets 2 and 3: constant recombination rates and AA and EA demographic histories.—Coalescent simulations invoking complex human demographic scenarios were simulated using the Cosi program of Schaffner et al. (in press), available at the Cosi Web site. The Cosi program is designed to simplify simulation of complex demographic scenarios, including asymmetric migration rates between subpopulations, admixture, populations splitting into subpopulations, and various population-size changes, including exponential growth and bottlenecks. Cosi allows for both crossing-over and gene conversion, as well as variation in rates of crossing-over (determined by a piecewise constant recombination surface), and assumes the infinite-sites model of mutation (with mutation positions converted into discrete base-pair positions). We used the detailed model of human demography, which was calibrated by Schaffner et al. (in press), using several large data sets of human sequence variation. The model of Schnaffner et al. (in press) simulates four populations that include an AA population and a European population. Unlike the African population in America, the European population in America has experienced very little recent admixture with other populations, so a European population is equivalent to an EA population. The mutation rate of 1.3×10^{-8} per base pair per generation, with no variation among sites, was chosen to give the correct average numbers of segregating sites in the AA and EA populations. The crossing-over rate, constant across the sequence for each simulation, was simulated from the empirical deCODE distribution, as described above. The same low level of gene conversion, 4.5×10^{-9} per bp per generation, was applied to each simulation (with a gene conversion tract length of 500 bp at all gene-conversion events). As a check of these model parameters, we compared our simulated data with the real SeattleSNPs data, with respect to two important summary statistics affected by demography, Tajima's D and the F_{ST} measure of genetic differentiation. For a SeattleSNPs collection of 201 genes, mean D was -0.53 for the AA population and $+0.20$ for the EA population. For simulation sets of 1,000 sequences, the mean D was -0.67

for the AA population and +0.03 for the EA population. For the SeattleSNPs genes, mean F_{ST} between the two populations was 9.9%, and we obtained a mean F_{ST} of 8.8% in our simulations (we used equation 3 in the work of Hudson et al. [1992] to estimate F_{ST}, with negative F_{ST} estimates adjusted to zero).

Simulation set 4: recombination hotspots and the null demographic model.—These sequence data were simulated as for simulation set 1, except for the addition of recombination hotspots. The background crossing-over rate was chosen using the deCODE distribution, as for simulation set 1, but, in addition, a single crossing-over hotspot was simulated. The hotspot rate was chosen with ρ distributed uniformly between 20 and 30 per kb (so, on average, 50 times higher than background). Hotspot width was distributed uniformly between 1 kb and 2 kb, and hotspot position in the sequence was chosen at random (more precisely, the start of the hotspot was distributed uniformly between 0 kb and 24 kb in the sequence). We simulated data, using the approach described in appendix C of Li and Stephens (2003), to convert the output of Hudson's ms program.

Since the deCODE distribution represents the average crossing-over rate for both background and hotspot regions, we could have chosen a different distribution, concentrated on smaller values, for the background rate. Precisely how to do this is unclear, and the approach we take should be considered conservative (in terms of estimating power), since we allow for larger background rates. Furthermore, our results (see below) suggest that the hotspot intensity (ratio of hotspot rate to background rate) had, at most, a small effect on the power of our approach.

Simulation sets 5 and 6: recombination hotspots and AA and EA demographic histories.—These sequence data were simulated as for simulation sets 2 and 3, except for the addition of recombination hotspots. The details of the hotspots were the same as for simulation set 4, except that the Cosi program requires recombination rates in terms of *r,* the recombination probability per meioses, and we chose r to be in the range of 5×10^{-4} to 7.5×10^{-4} per kb.

Polymorphism Frequencies

The SeattleSNPs data were examined for a signature of biased gene conversion (BGC) (Marais 2003) on polymorphism frequencies—that is, the higher frequency of AT→GC mutations compared with GC→AT mutations. The direction of mutation of SNPs (GC→AT or AT→GC) was inferred by parsimony with use of the chimpanzee outgroup information that is incorporated in the SeattleSNPs data sets. Coding SNPs were ignored to avoid the effects of selection, as were those SNPs possibly generated by CpG mutations, for which parsimony may be unreliable (for more details, see Webster and Smith [2004]).

Results

Method for Detecting Hotspots

We first performed a detailed simulation study, both to choose the value of the penalty in our penalized likelihood and to evaluate the power of our new method for detecting hotspots and its robustness to deviations from the simple model under which the AMLs are calculated. We compared our new approach with that of Fearnhead et al. (2004), an existing method that performed better than did LDhot (McVean et al. 2004) and Hotspotter (Li and Stephens 2003) at detecting hotspots in a 206-kb region of chromosome 1 (Jeffreys et al. 2005).

The method of Fearnhead et al. (2004) requires a threshold for a likelihood-ratio test to be set. We chose both this threshold and our likelihood penalty from analyses of ~300 simulated 25-kb data sets, none of which contained recombination hotspots. For each method, we chose two values for the respective constants, one that gave a false-error rate of approximately one hotspot per 100 genes and one that gave a false-error rate of approximately five hotspots per 100 genes. We then analyzed ~300 data simulated data sets, each of which contained a hotspot under both methods with each choice of constant. See the "Materials and Methods" section for full details of the simulations and table 2 for the results of the power and false-positive rates of these two methods. For a fair comparison with the likelihood-ratio method, we allowed hotspots of lengths between 250 bp and 3 kb in the penalized-likelihood approach (rather than use the knowledge we have about hotspot lengths to impose a minimum hotspot length of 1 kb, information that cannot be incorporated into the likelihood-ratio test). The power results in table 2 count a hotspot to be found if an estimated hotspot overlaps with the true hotspot, and an estimated hotspot is a false-positive result if it does not. The new penalized-likelihood approach is consistently more powerful, for similar false-positive rates, than the existing likelihood-ratio approach. The performance of the penalized-likelihood method is only slightly worse for the EA and AA data sets than for the null data sets.

We can further examine the accuracy of methods for estimation of the actual position of the hotspot. The median absolute error for estimating the edge of a hotspot is 400 bps for the EA data sets and 320 bps for the AA data sets, with 80% and 86%, respectively, of inferred hotspots containing the middle of the true hotspot. The increased accuracy for the AA data sets is due

Table 2

Power and False-Positive Rates for New and Existing Approaches for Hotspot Detection

Data Simulation and Demographic Scenario[a]	Power (False-Positive) Rates[b] for Penalized-Likelihood Approach When $\lambda = 13$	$\lambda = 16$	Likelihood-Ratio Approach When $c = 10$	$c = 12$
Simulations without recombination hotspots:				
Null	… (3)	… (0)	… (1)	… (0)
EA	… (3)	… (1)	… (3)	… (1)
AA	… (5)	… (1)	… (5)	… (2)
Simulations with recombination hotspots:				
Null	78 (3)	75 (2)	65 (2)	60 (2)
EA	72 (3)	63 (1)	71 (4)	56 (1)
AA	70 (7)	67 (4)	56 (5)	44 (2)
Average	73 (4)	65 (1.5)	62 (3.3)	53 (1.3)

Note.—For each method, we chose two values of a user-specified parameter: the penalty, λ, for the penalized-likelihood approach, and the threshold, c, for the likelihood-ratio approach. We allowed hotspot lengths between 250 bp and 3 kb for the penalized-likelihood approach. Each row relates to results from ~100 data sets.

[a] We considered three demographic scenarios: the null model assumes a panmictic constant population and is the model under which the likelihoods are calculated; EA and AA refer to data simulated jointly under demographic models that roughly match the patterns of diversity seen in EA and AA populations.

[b] Power is given as a percentage, and the false-positive rate as number of hotspots per 100 genes.

to the higher density of segregating sites in those data sets.

The AA and EA data sets simulated to produce the results in table 2 were each simulated jointly in pairs; this models the generation of data from a single gene that is sampled in two diverse populations. We tested our method for detecting hotspots, given data from two populations on these pairs of data sets. Again, we used the results of the analysis of simulated data sets that did not contain a hotspot to choose the penalty in the penalized likelihood, and we chose $\lambda = 18$, which gave a false-positive rate of approximately five hotspots per 100 genes. Since it is the approach we take for the real data, we fixed a minimum hotspot length of 1 kb in our analysis—although imposing this minimum has only a small effect on the results. Table 3 gives the results of the joint analysis, in terms of power and false-positive rates for each population individually and in terms of hotspots that were inferred jointly in both populations.

The joint analysis gives improved power for detecting hotspots, as compared with the analyses of single populations with similar false-positive rates. In particular, the method performs well for hotspots inferred in both populations. The joint analysis also gives slightly more accurate inference of the hotspot boundaries, with median absolute error of 320 bps and 95% of inferred hotspots containing the middle of the true hotspot. Although there is limited data, a simple analysis of the features of the data sets for which hotspots were detected suggests that the most important feature is the number of SNPs inside or near (within 2 kb) the hotspot (logistic regression $P < .01$); by comparison, the width or amount of recombination within the hotspot or background recombination rate had little effect (across the range of values that these varied for our simulation study).

Hotspots in SeattleSNPs Genes

We applied our method for inferring hotspots to 3 Mb of sequence data from 89 genes (see the "Materials and Methods" section). We allowed for hotspots of 1–3 kb in size, since sperm-typing results suggest that hotspots tend to be within this range. Imposition of a suitable

Table 3

Power and False-Positive Rates for the New Penalized-Likelihood Approach

Rate	Hotspots Inferred in Population(s) EA	AA	Both
Power (%)	86	85	81
False-Positive[a]	5	7	3

Note.—A penalty $\lambda = 18$ was used, and we allowed hotspots to have lengths of 1–3 kb.

[a] Number of hotspots per 100 genes.

Table 4

Inferred Hotspots in 89 Genes

No. of Hotspots in Gene (kb)	No. of Inferred Hotspots, by Population, as Determined by				
	Separate Analysis[a]		Joint Analysis[b]		
	EA	AA	EA	AA	Both
0	51	42	33	26	36
1	26	32	36	39	37
2	8	9	10	14	10
3	3	4	3	4	4
4	1	1	6	5	2
5	0	1	1	1	0
Total[c]	55	71	94	104	77

[a] The separate analyses used a likelihood penalty of 16.

[b] The joint analysis used a likelihood penalty of 18.

[c] Total number of hotspots inferred across the 89 genes.

minimum length for the hotspot should improve later inferences about the features of detected hotspots.

A summary of the results is given in table 4, and details of the position of the inferred hotspots are given in tables 1 and 5. The results differ across the two analyses, with the joint analysis inferring more hotspots, although this is consistent with the higher power and slightly higher false-positive rate observed in the simulation study. There is also a difference in the number of hotspots inferred in the AA sample compared with the EA sample. One possible explanation of the increased power in the AA population is the increased SNP density in that population—although we did not observe any difference in power in our simulation study, where we had the same difference in SNP density across the two populations.

We are able to compare our results with those of Crawford et al. (2004), who analyzed data from 74 genes, 39 of which are also in our study. Their analysis was based on a Bayesian approach that compared the hypotheses of no hotspot within a gene with that of one hotspot within a gene, and they analyzed each population separately. If we compare their results with those of our separate analyses, we have a slightly higher proportion (69% as compared with 62%) of genes for which a hotspot was inferred in at least one population. Of the 39 genes in common, the two analyses agree on the presence of hotspots in at least one population, or absence from both, in 28 cases. Of the remaining 11, we inferred hotspots not found by Crawford et al. (2004) in 7 genes. It is encouraging that, for all the 20 genes for which both approaches inferred a hotspot, the position of the hotspot was consistent for the two methods. (We define results as consistent if one of the hotspots inferred in the gene by our approach overlapped with the position of the hotspot inferred by Crawford et al. 2004).

Our results suggest hotspots occur more frequently than did the results of Crawford et al. (2004), whose estimate of a hotspot frequency of one every 63 kb was biased by their assumption that genes had, at most, a single hotspot. A simple estimate of the frequency of hotspots from our analysis of the AA population suggests hotspots either every ~30 kb (on the basis of the results from the joint analysis) or every ~40 kb (on the basis of results from the separate analysis). Our higher-density estimates are roughly consistent with the sperm-typing results from the major histocompatibility complex (*MHC*) gene (Jeffreys et al. 2001), in which six hotspots are found in ~200 kb, and from a 206-kb region of chromosome 1 that contains eight hotspots (Jeffreys et al. 2005).

Our results provide information on the clustering of hotspots within genes, which is unavailable from the analysis of Crawford et al. (2004). There is no evidence from our results against a model in which hotspots are placed uniformly at random along the genome (see fig. 3). The most extreme examples of clustering of hotspots were for *HABP2, PLAUR,* and *C3,* in which 5, 4, and 4 hotspots were inferred (from the joint analysis) in 39-kb, 24-kb, and 45-kb sequences, respectively.

Hotspot Position within Genes

Although our data suggest that the position of hotspots within the genome (or at least our biased sample of the genome) is random, there remains the issue of what factors might determine where hotspots are located within the genes in our data set. To investigate factors affecting hotspot position within genes, we used our largest set of hotspots, the AA hotspots identified in the joint analysis.

We first considered the position of hotspots within the genes. We calculated the position of the center of each hotspot as a proportion of the entire gene sequence, and we found the mean position to be close to the center of the gene (mean proportional position 0.45). We also tested whether the proportion of hotspot sequence in exons differed from the proportion of background sequence in exons (considering only those genes in which hotspots were found). We found a slightly higher propor-

Table 5

Inferred Hotspots: Joint Analysis

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

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

Figure 3 Quantile-quantile plot for residuals of a Poisson model for the number of hotspots per gene. The legend is available in its entirety in the online edition of *The American Journal of Human Genetics.*

tion of hotspot sequence to be exonic (5.4%) compared with background sequence (4.2%), although simulations indicated that the observed proportion of exonic sequence in hotspots was not significantly greater than for randomly positioned hotspots distributed across genes, according to the Poisson distribution with mean equal to the observed density of joint AA hotspots ($P = .34$).

We also searched for primary sequence motif differences between hotspot sequences and background sequences. We first considered testing for motifs that were more common overall in hotspots than in background sequences, similar to the approach of Crawford et al. 2004, but we found this test identified motifs that were found to be highly abundant in just a few hotspots—in particular, motifs repeated in rare large microsatellites. Instead, we looked for motifs found in a high proportion of hotspots, testing for significance by generating random sets of hotspots (uniform density within and among genes). We first tested the complete set of all $4^8 = 65{,}536$ octamers for their presence in the 104 hotspots. Four octamers—AAAAAAAA, CAGCCTGG, GGAGGCTG, and TCCCAGCA—were found in the highest proportion (45 of 104) of hotspots, so no single octamer is a powerful predictor of hotspots. To find octamers found more commonly in hotspots than would be expected by chance, we assessed significance for the 185 octamers that were present in >30% of hotspots. We used this relatively low cutoff point because the proportion of hotspots containing a motif is a function of both enrichment in hotspots and also overall frequency in the genome. One of the octamers, ACAGAGCA, had a P value of <1 in 100,000; a further octamer, TCCCAGCA, had a P value <1 in 10,000. However, no octamers had significant P values after a Bonferroni correction.

Comparison of Populations

The results from the joint analysis of the data in the two populations (see table 4) show that a substantial number of hotspots are inferred only in one of the two populations. This is in marked contrast to the results from the simulation study (see table 3). There are various explanations for this difference; one possibility is that the recombination landscape is actually different for the two distinct populations and that there are hotspots that exist solely in one of the two populations. There is evidence to show that recombination hotspots evolve over time (see the Introduction), although there is currently little idea as to whether this evolution has resulted in different hotspots in different populations and, if so, to what extent there are population-specific hotspots.

To quantify the evidence of population-specific hotspots within the 89 genes that we analyzed, we used a simple likelihood-ratio test (similar to that used by Crawford et al. [2004]). For each population-specific hotspot that was inferred by our joint analysis, we first chose a single region that overlapped with the hotspot. We centered each region around the center of the inferred hotspot and chose the width of that region to be the smallest of 1 kb, 2 kb, 3 kb, 4 kb, or 5 kb, such that there were at least six SNPs within the region for each of the two populations. For some hotspots, even the 5-kb regions did not contain six SNPs in each population; in these cases, we decided to use 5-kb regions. Note that these regions are not necessarily any of the subregions used when we originally inferred the hotspots. We chose regions in this way with the aim of getting small regions, so that the majority (ideally all) of each region would be contained within the area of the true (although potentially population-specific) hotspot, while ensuring enough SNPs within each population to have reasonable power for inferring the recombination rates within each hotspot.

For each hotspot, we then calculated the log-likelihood-ratio statistic for different recombination rates (relative to the different effective population sizes) in the two populations:

$$LR = 2\log\left[l_1(\hat{\rho}_1) + l_2(\hat{\rho}_2) - l_1(\hat{\rho}) - l_2(\gamma\hat{\rho})\right],$$

where $l_1(\cdot)$ and $l_2(\cdot)$ are the log AMLs for the AA and EA populations, γ is the estimate of the ratio of the average recombination rates in the two populations, $\hat{\rho}_i$ is the maximum-likelihood estimate (MLE) for the recombination rate in population i, and $\hat{\rho}$ is the MLE for the recombination rate (defined for the AA population), with the assumption that the rate differs only by a factor of γ across the two populations.

We converted the likelihood-ratio statistic values into approximate P values, using a χ^2_1 assumption for the likelihood-ratio statistic. Simulations suggests that the likelihood-ratio statistic does at least *roughly* have a χ^2_1 distribution and that the resulting P values are likely to be conservative (because of the positive dependence of the likelihood curves in the two populations).

We found strongest evidence of a population-specific hotspot in the gene *TRPV6* ($P = .007$), although this P value is far from significant when the multiple testing is accounted for (we tested 44 inferred population-specific hotspots, but these were preferentially chosen from a total of 121 inferred hotspots; thus, we should correct for an

equivalent of 121 hypothesis tests). One further hotspot (the putative EA-specific hotspot in *HABP2*) had a P value <.01.

We studied the putative population-specific hotspot in *TRPV6* in more detail. The signal for a population specific hotspot is due to a breakdown of linkage disequilibrium (LD) across the region in the AA population, whereas there is substantial LD across the whole gene in the EA population (four distinct haplotypes from 50 SNPs that segregate in the EA population). However, the four distinct haplotypes found in the EA population consist of three haplotypes that differ only at two SNPs, and the fourth haplotype (which is at frequency 1 in the EA population) appears to be a recent migrant from the AA population. Thus, even the signal here appears to be caused by a combination of a lack of diversity in the EA population and this recent migrant, rather than strong evidence of a population-specific hotspot. (If we remove the migrant haplotype, the remaining haplotypes have almost no information about the amount of recombination in the EA population.)

Other Features of Hotspots

We also looked at features of SNPs and G+C content within the inferred hotspots. Throughout, we used the largest set of hotspots, those detected for the AA population from the joint analysis, and SNPs and G+C content from that population. To see the importance of inferring recombination rates at the kilobase scale—rather than the megabase scale—we also considered the features within the 10% of genes with highest recombination rates, as defined by the deCODE recombination map.

We compared the patterns of G+C content and SNP density for the background region of the AA data (G + C = 44.2%; SNP density of one per 310 bp), for the hotspot regions (G + C = 46.6%; SNP density of one per 190 bp), and for the nine genes with high deCODE recombination rates (G + C = 49.1%; SNP density of one per 260 bp). Both the large-scale and fine-scale high-recombination regions show higher levels of G+C and higher SNP densities. The difference in both G+C content and SNP density is significant ($P < .001$) for differences between the hotspot and background regions. These results are consistent with known correlations between G+C content and recombination rates and with the possible mutagenic effect of recombination (Hellmann et al. 2003).

One explanation for the correlation between G+C content and recombination rates is BGC: the bias in repair of hetero-mismatches in heteroduplex DNA formed by gene conversion that favors G+C over A+T (see Marais [2003] for a review). To look for evidence of BGC, we calculated the mean polymorphism frequencies of AA GC→AT and AT→GC SNPs within and outside recombination hotspots (see the "Materials and Methods" section).

We see that the signature of BGC, with higher mean frequency of AT→GC polymorphisms relative to GC→AT polymorphisms, is stronger within recombination hotspots than for background regions outside hotspots (see table 6), although the signal is small and not statistically significant. By comparison, a similar analysis of the SeattleSNPs data set found no discernible effect on polymorphism frequencies that were due to large-scale regional recombination-rate variation, as measured by the deCODE pedigree data (Webster and Smith 2004). When we partitioned our AA SNPs by deCODE recombination rates, we found no difference in polymorphism frequencies (see table 6).

The proportion of mutations that change GC content and are GC→AT is similar for the complete data and the hotspot regions (58% and 56%, respectively) but significantly smaller than for regions with high deCODE recombination rates (63%; $P = .01$ for difference in proportion to that of the complete data).

Discussion

We have presented a new method for detecting recombination hotspots from population-genetics data. Our simulation study shows that this method has greater power than what, on the basis of the results in Jeffreys et al. (2005), is currently the best method, that of Fearnhead et al. (2004). Both that approach and our new method are based on use of the AML for subregions of the genomic region of interest. It is intuitive that, for this approach to work well, the subregions must be informative (i.e., contain a sufficient number of SNPs) and be of a size similar to the width of recombination hotspots. (If subregions were much wider than hotspots, the signal from the hotspot would be weaker, since the main part of the subregion analyzed would be outside the hotspot.) This intuition is backed by our simulation results, which suggested that the main factor affecting our method's power for detecting hotspots is the number of SNPs within and near the hotspot. Thus, these approaches are well suited for analyzing the SeattleSNPs data considered in this work as well as the population data described by Jeffreys et al. (2005): in both cases, the density of SNPs is large. It may be that the methods of Li and Stephens (2003) and McVean et al. (2004) will be comparatively more powerful for data for which the SNP density is much smaller.

The main focus of our work has been on detecting recombination hotspots and not on estimating background or hotspot recombination rates. Results of the work of Smith and Fearnhead (2005) suggest that the pairwise likelihood method of McVean et al. (2002) and the approximate likelihood method of Li and Stephens

Table 6

Frequencies of Polymorphisms for Different Subsets of AA SNPs

	Findings by Mutation Direction					
	GC→AT		AT→GC			
Subset of AA SNPs	Frequency[b] (%)	No. of SNPs	Frequency[b] (%)	No. of SNPs	Average (%)	D[a] (%)
All SNPs	18.1	3,432	20.3	2,464	19.2	2.2
SNPs within hotspots	20.5	360	23.5	278	22.0	3.0
Background SNPs[c]	17.8	3,072	19.9	2,186	18.9	2.1
deCODE[d]	21.2	382	21.4	226	21.3	.2

[a] The difference in mean frequencies.
[b] Mean frequencies of polymorphisms that differ in their effects on GC content.
[c] Without hotspots.
[d] SNPs in genes with high deCODE regional recombination rates.

(2003) are more accurate for estimation of recombination rates than the AML that is the basis of our method. As a result, we would suggest that, for estimating recombination rates, one should use the pairwise likelihood approach of McVean et al. (2002), conditional on the hotspot positions found by our method.

The SeattleSNPs data that we analyzed consisted of population data from two distinct populations. The penalized-likelihood approach we propose can be easily extended to deal with data from two populations. This extension is "approximate," since it ignores the positive dependence that would be expected between the data from the two populations. However, in simulations under models that capture the main features of the data we analyze, we found that this joint analysis performed best. Whether such an approach will work more generally is unclear. The joint analysis is potentially less robust than the separate analysis, since there are extra features of the data, particularly the dependences between the two populations, that will affect the choice of penalization factor and would need to be roughly similar in the simulated and real data.

Our method is dependent on specifying a likelihood penalty λ, an estimate of the background-recombination rate $\hat{\rho}_b$, and, for the joint analysis, an estimate of the ratio of recombination rates between the two populations, γ. An important question is the degree to which results are robust to variations in these. The effect of varying λ is to change the amount of evidence required before a region is determined to be a hotspot. The choice of λ is a simple trade-off between power and false-positive rate; guidelines for choosing λ can be obtained from our simulation study. Although we considered only three possible demographic scenarios (for the separate analysis), the similarity in false-positive rates across these different scenarios gives us some confidence that these results will be roughly correct for a large variety of demographic scenarios.

The results of the joint analysis appear to be reasonably robust to variations in the choice of γ. For example, in our simulation study, varying γ from 1.5 to 2.3 had only a small effect on the number of both joint and population-specific hotspots inferred (on the order of one or two hotspots for each category). For the real data, an earlier analysis with $\gamma = 1.5$ (which was based on an estimate of the ratio of effective population size from polymorphism data), rather than $\gamma = 3.75$, did produce some noticeable differences in the results. In particular, whereas the total number of hotspots in the AA population was almost unaffected, the number of joint hotspots and the total number of hotspots in the EA population were reduced by an order of 10–20.

The method for obtaining $\hat{\rho}_b$ is important, particularly for small genes (or genes with only a small amount of background sequence). Our method—using the likelihood ratio test of Fearnhead et al. (2004) to extract regions that are likely to be hotspots and then using a composite likelihood based on likelihoods from all other subregions to estimate the background rate—is simple and appears to work reasonably well (on the basis of the simulation-study results). Even so, the choice of threshold used for detecting hotspot regions can have a noticeable effect on the results that are obtained; for some genes there can be relatively few subregions that are not classified as hotspots, which can lead to potentially large uncertainty in the estimate of $\hat{\rho}_b$. Better approaches for obtaining $\hat{\rho}_b$ may lead to important improvements in the accuracy of this penalized-likelihood approach.

One advantage of the joint analysis is that it allows us to directly compare the presence/absence of hotspots in the two populations. For the SeattleSNPs data, we found that a third of the hotspots were inferred to be present in only one of the two populations—much more than is expected on the basis of the simulation study. A simple analysis of these hotspots through use of a likelihood-ratio statistic does not give any evidence in favor of the presence of different recombination hotspots in the two populations. This may be due to lack of power,

or there may just be smaller differences in the recombination landscape in the two populations, such as a different ratio of hotspot to background rate, that are causing us to detect a large number of population-specific hotspots.

One important question is whether we would expect to have any power to detect population-specific hotspots, if they exist, given that differences in the recombination rates in the two populations are likely to be recent and that the populations have shared ancestry. One reason to be hopeful about having some power to detect population-specific hotspots is that it is the presence (or absence) of the most-recent recombination events that has most effect on inferences about recombination rates, and it is the recent recombination process that will be most different and most close to independent between the two populations.

Some of the SeattleSNPs data we analyzed has been analyzed by Crawford et al. (2004). An important difference between these two analyses is that our method for detecting hotspots allows us to find multiple hotspots (if they exist) within a gene rather than just answering the question of whether there is at least one hotspot within the gene. Our results show that many genes have multiple hotspots, and, accounting for this, we estimate that the frequency of hotspots is greater (one per 30–40 kb) than that suggested by Crawford et al. (2004) (one per 60 kb). Analysis of the position of the hotspots shows no evidence against the simple hypothesis that hotspots occur randomly.

To what extent the recombination hotspots that have been detected correspond to increased rates of gene conversion and/or increased rates of crossing-over is unclear. Whereas the AML is based on a model that assumes that all recombination events consist of a single recombination break point within the subregion being analyzed (pure crossing-over), the results of Smith and Fearnhead (2005) show that the AML actually infers some (weighted) sum of the crossing-over and gene-conversion rates. Thus, the recombination hotspots indicate a general increase in the rate of gene conversion and/or crossing-over, and our method gives no information about the relative rate of these two processes.

A qualitative look at the population data suggests that some of the hotspots may actually correspond to an increase only in the gene-conversion rate. For example, the hotspot detected at position 19,849–20,849 in *DCN* in the AA population corresponds to a region in which five neighboring SNPs show little LD (and, e.g., there are nine distinct haplotypes defined by the data at these five SNPs). However, if we consider the 16 SNPs that comprise the 8 SNPs immediately adjacent to either side of this hotspot, then these are in complete LD (as defined by D') and are consistent with no recombination. In fact, the LD is so strong that these 16 SNPs define only seven distinct haplotypes. This amount of LD is surprising if there is indeed a crossing-over hotspot in the middle of them, and a reasonable conclusion is that this hotspot corresponds solely to an increase in gene conversion (with short tract lengths).

Acknowledgments

This work is dedicated to the memory of Nick Smith. This work was supported by Engineering and Physical Sciences Research Council grant GR/S18786/01. We thank Andrea Jorgensen, for work during her M.Sc. project that motivated this; and two anonymous reviewers, for thoughtful and helpful comments. We particularly thank one of the reviewers for the idea of how to estimate γ.

Web Resources

The URLs for data presented herein are as follows:

Cosi, http://www.broad.mit.edu/personal/sfs/cosi/
P.F.'s Web site, http://www.maths.lancs.ac.uk/~fearnhea/
SeattleSNPs, http://pga.gs.washington.edu/

References

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

Fearnhead P, Donnelly P (2001) Estimating recombination rates from population genetic data. Genetics 159:1299–1318

Fearnhead P, Donnelly P (2002) Approximate likelihood methods for estimating local recombination rates (with discussion) J Roy Stat Soc B 64:657–680

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

Hellmann I, Ebersberger I, Ptak SE, Pääbo S, Przeworski M (2003) A neutral explanation for the correlation of diversity with recombination rates in humans. Am J Hum Genet 72: 1527–1535

Hudson RR (2002) Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics 18:337–338

Hudson RR, Slatkin M, Maddison WP (1992) Estimation of levels of gene flow from DNA-sequence data. Genetics 132: 583–589

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

Jeffreys AJ, Neumann R (2002) Reciprocal crossover asymmetry and meiotic drive in a human recombination hotspot. Nat Genet 31:267–271

Jeffreys AJ, Neumann R, Panayi M, Myers S, Donnelly P (2005) Human recombination hotspots hidden within regions of strong marker association. Nat Genet 37:601–606

Kong A, Gubdjartsson DF, Sainz J, Jonsdottir GM, Gudjonson

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

Li N, Stephens M (2003) Modelling LD, and identifying recombination hotspots from SNP data. Genetics 165:2213–2233

Marais G (2003) Biased gene conversion: implications for genome and sex evolution. Trends Genet 19:330–338

McVean GAT, Awadalla P, Fearnhead P (2002) A coalescent method for detecting recombination from gene sequences. Genetics 160:1231–1241

McVean GAT, 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

Ptak SE, Hinds DA, Koehler K, Nickel B, Patil N, Ballinger DG, Przeworski M, Frazer KA, Pääbo S (2005) Fine-scale recombination patterns differ between chimpanzees and humans. Nat Genet 37:429–434

Rambaut A, Grassly NC (1997) Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. Comput Appl Biosci 13:235–238

Schaffner SF, Foo C, Gabriel S, Reich D, Daly MJ, Altshuler D. Calibrating a coalescent simulation of human genome sequence variation. Genome Res (in press)

Smith NGC, Fearnhead P (2005) A comparison of three estimators of the population-scaled recombination rate: accuracy and robustness. Genetics (electronically published June 14, 2005; accessed September 7, 2005)

Stephens M, Donnelly P (2000) Inference in molecular population genetics (with discussion). J Roy Stat Soc B 62:605–635

——— (2003) A comparison of Bayesian methods for haplotype reconstruction from population genotype data. Am J Hum Genet 73:1162–1169

Stephens M, Smith NJ, Donnelly P (2001) A new statistical method for haplotype reconstruction from population data. Am J Hum Genet 68:978–989

Webster MT, Smith NGC (2004) Fixation biases affecting human SNPs. Trends Genet 20:122–126

Winckler W, Myers SR, Richter DJ, Onofrio RC, McDonald GJ, Bontrop RE, McVean G. AT, Gabriel SB, Reich D, Donnelly P, Altshuler D (2005) Comparison of fine-scale recombination rates in humans and chimpanzees. Science 308:107–111