Modelling Linkage Disequilibrium, And Identifying Recombination Hotspots Using SNP Data

Na Li* and Matthew Stephens†

July 25, 2003

*Department of Biostatistics, University of Washington, Seattle, WA 98195

†Department of Statistics, University of Washington, Seattle, WA 98195

Running Head: Linkage Disequilibrium and Recombination

Key Words: Linkage Disequilibrium, Population Genetics, Recombination Rate, Recombination Hotspot, Coalescent, Approximate Likelihood

Corresponding Author:

Matthew Stephens

Department of Statistics
University of Washington
Box 354322
Seattle, WA 98195
(206) 543-4302 (ph.)
(206) 685-7419 (fax)
stephens@stat.washington.edu

Abstract

We introduce a new statistical model for patterns of Linkage Disequilibrium (LD) among multiple SNPs in a population sample. The model overcomes limitations of existing approaches to understanding, summarizing, and interpreting LD by (i) relating patterns of LD directly to the underlying recombination process; (ii) considering all loci simultaneously, rather than pairwise; (iii) avoiding the assumption that LD necessarily has a "block-like" structure; and (iv) being computationally tractable for huge genomic regions (up to complete chromosomes). We examine in detail one natural application of the model: estimation of underlying recombination rates from population data. Using simulation, we show that in the case where recombination is assumed constant across the region of interest, recombination rate estimates based on our model are competitive with the very best of current available methods. More importantly, we demonstrate, on real and simulated data, the potential of the model to help identify and quantify fine-scale variation in recombination rate from population data. We also outline how the model could be useful in other contexts, such as in the development of more efficient haplotype-based methods for LD mapping.

1 Introduction

Linkage disequilibrium (LD) is the non-independence, at a population level, of the alleles carried at different positions in the genome. The patterns of LD observed in natural populations are the result of a complex interplay between genetic factors, and the population's demographic history. In particular, recombination plays a key role in shaping patterns of LD in a population. When a recombination occurs between two loci, it tends to reduce the dependence between the alleles carried at those loci, and thus reduce LD. Although recombination events in a single meiosis are relatively rare over small regions, the large total number of meioses that occur each generation in a population have a substantial cumulative effect on patterns of LD, and so molecular data from population samples contain valuable information on fine-scale variations in recombination rate.

Despite the undoubted importance of understanding patterns of LD across the genome, most obviously because of its potential impact on the design and analysis of studies to map disease genes in humans, most current methods for interpreting and analyzing patterns of LD suffer from at least one of the following limitations:

1. They are based on computing some measure of LD defined only for *pairs* of sites, rather than considering all sites simultaneously.

2. They assume a "block-like" structure for patterns of LD, which may not be appropriate at all loci.

3. They do not directly relate patterns of LD to biological mechanisms of interest, such as the underlying recombination rate.

As an example of the limitations of current methods, consider Figure 1, which shows a graphical display of pairwise LD measures for six simulated data sets, simulated under various models for heterogeneity in the underlying recombination rate. The reader is invited to speculate on what the underlying models are in each case — the answer appears in the caption to Figure 8. In each of the six figures one could identify by eye, or by some quantitative criteria (e.g. DALY *et al.* 2001, OLIVIER *et al.* 2001, WANG *et al.* 2002), "blocks" of sites, such that LD tends to be high among markers within a block. In some cases there might also be little LD between markers in different "blocks", which might be interpreted as evidence for variation in local recombination rates: low recombination rates within the blocks, and higher rates between the blocks. Indeed, JEFFREYS *et al.* (2001) have shown, using sperm-typing, that in the class II region of MHC, variations in local recombination rate are indeed responsible for block-like patterns of LD. However, without this type of experimental confirmation, which is currently technically challenging and time consuming, it is difficult to distinguish between blocks that arise due to recombination rate heterogeneity, and blocks that arise due to chance, perhaps through chance clustering of recombination events in the ancestry of the particular sample being considered (WANG *et al.* 2002). The ability to distinguish between these cases would of course be interesting from a basic science standpoint — for example, in helping to identify sequence characteristics associated with recombination hotspots. In addition,

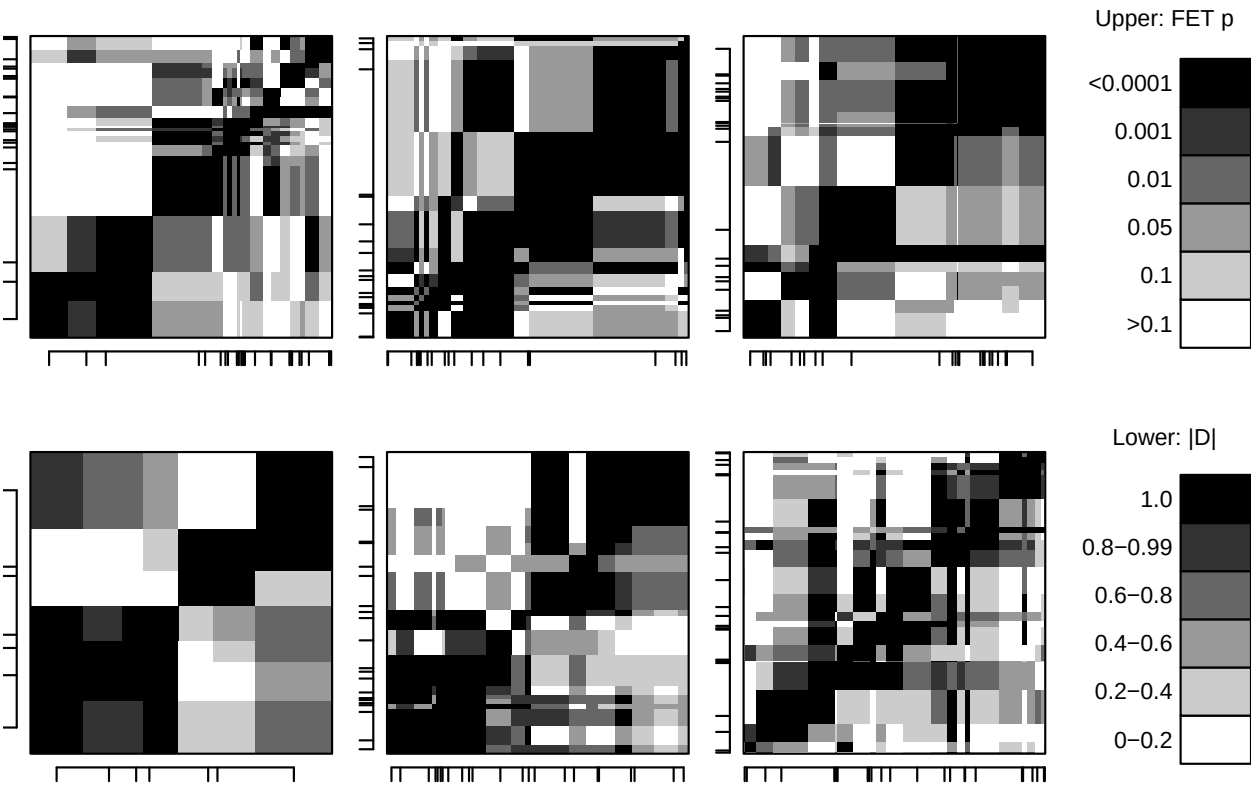

Figure 1: Plots of LD measurement, $|D'|$, (lower right triangle) and p-value for Fisher's exact test (upper right triangle) for every pair of sites with minor allele frequency > 0.15, in data sets simulated under varying assumptions about variation in the local recombination rate. Details of the models used to simulate each data set appear in the caption to Figure 8, which is based on the same six data sets.

it would have important implications for the design and analysis of LD mapping studies. For example, it would help in predicting patterns of variation at sites that have not been genotyped (perhaps sites influencing susceptibility to a disease), and it would provide some indication of whether block structures observed in one sample are likely to be replicated in other samples — a crucial requirement for being able to select representative "tag" SNPs (JOHNSON *et al.* 2001) based on LD patterns observed in some reference sample.

In this paper we introduce a statistical model for LD that overcomes the limitations of existing approaches by relating genetic variation in a population sample to the underlying recombination rate. We examine in detail one natural application of the model: estimation of underlying recombination rates from population data. Using simulation, we show that in the case where recombination is assumed constant across the region of interest, recombination rate estimates based on our model

are competitive with the very best of current available methods. More importantly, we demonstrate, on real and simulated data, the potential of the model to help identify and quantify fine-scale variation in recombination rate (including "recombination hotspots") from population data.

Although we focus here on estimating recombination rates, we view the model as being useful more broadly, in interpreting and analyzing patterns of LD across multiple loci. In particular, as we outline in our discussion, the model could be helpful in the development of more efficient haplotype-based methods for LD mapping, along the lines of, for example, MCPEEK and STRAHS (1999), MORRIS *et al.* (2000), and LIU *et al.* (2001).

2 Models

2.1 Background

The most successful current approaches to constructing statistical models relating genetic variation to the underlying recombination rate (and to other genetic and demographic factors) are based on the coalescent (KINGMAN 1982), and its generalization to include recombination (HUDSON 1983). Although these approaches are based on rather simplistic assumptions about the demographic history of the population from which individuals were sampled, and about the evolutionary processes acting on the genetic region being studied, they have nonetheless proven useful in a variety of applications. In particular, they provide a helpful simulation tool (e.g. software described in HUDSON 2002), allowing more realistic data to be generated under various assumptions about underlying biology and demography, and hence aid exploration of what patterns of LD might be expected under different scenarios (KRUGLYAK 1999; PRITCHARD and PRZEWORSKI 2001).

Despite the ease with which coalescent models can be *simulated* from, using these models for *inference* remains extremely challenging. For example, consider the problem of estimating the underlying recombination rate in a region, using data from a random population sample. It follows from coalescent theory that population samples contain information on the value of the product of the recombination rate c, and the effective (diploid) population size N, but not on c and N separately. It has therefore become standard to attempt to estimate the compound parameter $\rho = 4Nc$, and several methods have been proposed. Some (e.g. GRIFFITHS and MARJORAM 1996, NIELSEN 2000, KUHNER *et al.* 2000, FEARNHEAD and DONNELLY 2001) try to make use of the full molecular data available. However, although such methods have been applied successfully to small regions and non-recombining parts of the genome (HARDING *et al.* 1997; HAMMER *et al.* 1998; NIELSEN 2000; KUHNER *et al.* 2000; FEARNHEAD and DONNELLY 2001), for even moderate-sized autosomal regions (e.g. a few kilobases in humans) they become computationally impractical (FEARNHEAD and DONNELLY 2001). Other methods, many of which are considered by WALL (2000), make use of only summaries of the data, substantially reducing computational requirements at the expense of some loss in efficiency.

More recently, HUDSON (2001) and FEARNHEAD and DONNELLY (2002) proposed "composite likelihood" methods for estimating ρ over moderate to large genomic regions. Hudson's

method is based on multiplying together likelihoods for every pair of sites genotyped, where these pairwise likelihoods are computed via simulation, assuming an "infinite-sites" mutation model (i.e. no repeat mutation). This method has been modified by McVean *et al.* (2002) to allow for repeat mutation. Fearnhead and Donnelly's method is based on dividing data on a large region into smaller regions, and multiplying likelihoods obtained for each smaller region. These methods, together with the best of the summary-statistic-based methods of Wall (2000), appear to be the most accurate of existing methods for estimating recombination rates from patterns of LD over moderate to large genomic regions. None of these methods, as currently implemented, allows explicitly for variation in recombination rate along the region under study.

2.2 A New Model

Here we describe a new model for LD, which enjoys many of the advantages of coalescent-based methods (e.g. it directly relates LD patterns to the underlying recombination rate) while remaining computationally tractable for huge genomic regions, up to entire chromosomes. Our model relates the distribution of sampled haplotypes to the underlying recombination rate, by exploiting the identity

$$\Pr(h_1,\ldots,h_n|\rho) = \Pr(h_1|\rho)\Pr(h_2|h_1;\rho)\ldots\Pr(h_n|h_1,\ldots,h_{n-1};\rho), \tag{1}$$

where $h_1,\ldots,h_n$ denote the n sampled haplotypes, and ρ denotes the recombination parameter (which may be a vector of parameters if the recombination rate is allowed to vary along the region). This identity expresses the unknown probability distribution on the left as a product of conditional distributions on the right. For simplicity we will often use the notation π to denote these conditional distributions. While the conditional distributions are not computationally tractable for models of interest, they are amenable to approximation, as we describe below. Our strategy is to substitute an approximation for these conditional distributions ($\hat{\pi}$ say) into the right hand side of (1), to obtain an approximation to the distribution of the haplotypes h given ρ:

$$\Pr(h_1,\ldots,h_n|\rho) \approx \hat{\pi}(h_1|\rho)\hat{\pi}(h_2|h_1;\rho)\ldots\hat{\pi}(h_n|h_1,\ldots,h_{n-1};\rho). \tag{2}$$

We refer to this model as a "Product of Approximate Conditionals" (PAC) model, and to the corresponding likelihood as a PAC likelihood, which we denote L_{PAC}. Explicitly

$$L_{\mathrm{PAC}}(\rho) = \hat{\pi}(h_1|\rho)\hat{\pi}(h_2|h_1;\rho)\ldots\hat{\pi}(h_n|h_1,\ldots,h_{n-1};\rho). \tag{3}$$

Similarly, we will refer to the value of ρ that maximizes L_{PAC} as a *maximum PAC likelihood estimate* for ρ, and denote it by $\hat{\rho}_{\mathrm{PAC}}$.

The utility of the model (3) will naturally depend on the use of an appropriate approximation for the conditional distribution π. This approximation should be designed to answer the following question: if, at a particular locus, in a random sample of k chromosomes from a population, we observe genetic types $h_1,\ldots,h_k$, what is the conditional distribution of the type of the next sampled chromosome, $\Pr(h_{k+1}|h_1,\ldots,h_k)$? We are aware of three forms for π in the literature, each

of which attempts to answer this question under different assumptions for the genetic model underlying the loci being studied. The first and best-known comes from the *Ewens sampling formula* (EWENS 1972). This arises from considering a neutral locus in a randomly-mating population, evolving with constant (diploid) size N and mutation rate μ per generation, and assuming an "infinite alleles" mutation model, where each mutation creates a novel (previously unseen) haplotype. Under these idealized conditions, if we let $\theta = 4N\mu$, then with probability $k/(k+\theta)$ the $k+1$st haplotype is an exact copy of one of the first k haplotypes chosen at random, otherwise it is a novel haplotype. Although the assumptions underlying this formula will never hold in practice, it does capture the following properties that we would expect to hold more generally:

(i) the next haplotype is more likely to match a haplotype that has already been observed many times than one that has been observed less frequently.

(ii) the probability of seeing a novel haplotype decreases as k increases.

(iii) the probability of seeing a novel haplotype increases as θ increases.

However, for modern molecular data, and for sequence data and SNP data in particular, it fails to capture the two following properties:

(iv) if the next haplotype is not exactly the same as an existing (i.e. previously-seen) haplotype, it will tend to differ by a small number of mutations from an existing haplotype, rather than to be completely different from all existing haplotypes.

(v) due to recombination, the next haplotype will tend to look somewhat similar to existing haplotypes over contiguous genomic regions, the average physical length of these regions being larger in areas of the genome where the local rate of recombination is low.

STEPHENS and DONNELLY (2000) suggested a form for π that captures properties (i)-(iv) above. In their suggested form for π, the next haplotype differs by M mutations from a randomly-chosen existing haplotype, where M has a geometric distribution with $\Pr(M=0) = k/(k+\theta)$ (so that it reproduces the Ewens sampling formula in the special case of the infinite alleles mutation model). Thus the next haplotype is a (possibly imperfect) "copy" of a randomly-chosen existing haplotype. FEARNHEAD and DONNELLY (2001) (henceforth FD) extended this form for π to also capture property (v) above. In FD's approximation, the $k+1$st haplotype is made up of an imperfect mosaic of the first k haplotypes, with the size of the mosaic fragments being smaller for higher values of the recombination rate.

Here we use two new forms for π that also capture properties (i)-(v) above. The first, described in detail in Appendix A and illustrated in Figure 2, and which we denote π_A, is a simplification of FD's approximation that is easier to understand and slightly quicker to compute. (Dr. N. Patterson, personal communication, has independently suggested a similar simplification.) The second, which we describe in detail in Appendix B and denote π_B, is a slight modification of π_A, developed using empirical results from Section 3.1 to produce a likelihood L_{PAC} that gives more accurate

estimates of ρ. Where necessary, we denote the PAC likelihoods and maximum PAC likelihood estimates corresponding to π_A (respectively π_B) by $L_{\mathrm{PAC-A}}$ and $\hat{\rho}_{\mathrm{PAC-A}}$ (respectively $L_{\mathrm{PAC-B}}$ and $\hat{\rho}_{\mathrm{PAC-B}}$).

A key property of both π_A and π_B is that they are easy and fast to compute. Unlike the Ewens sampling formula, but like the approximations of STEPHENS and DONNELLY (2000) and FD, neither corresponds exactly to the actual conditional distribution under explicit assumptions about population demography and the evolutionary forces on the locus under consideration. Indeed, no closed-form expressions for π, based on such explicit assumptions, and capturing (iv) or (v), are known. However, the suggested forms for π were motivated by considering both the Ewens sampling formula, and the underlying genealogy (or, in the case with recombination, genealogies) relating a random sample of haplotypes from a neutrally-evolving, constant-sized panmictic population. As such, it may be helpful to view them as approximations to the (unknown) true conditional distribution under these assumptions. In particular, there are certain aspects of many real populations (e.g. population expansion, or population structure), and biological factors (e.g. gene conversion, selection) that these forms for π do not attempt to capture. For some applications this may not matter very much. For others it may be necessary to develop forms for π that do capture these aspects — a point we return to in the discussion.

An unwelcome feature of the PAC likelihoods corresponding to our choices of π — and indeed the forms for π from STEPHENS and DONNELLY (2000) and FD — is that they depend on the order in which the haplotypes are considered. In other words, although these likelihoods each correspond to a valid probability distribution on the haplotypes, these probability distributions do not enjoy the property of exchangeability that we would expect to be satisfied by the true (unknown) distribution. Practical experience, and theory in STEPHENS and DONNELLY (2000) (their Proposition 1, part d) suggests that this problem cannot be rectified by making a simple modification to π. Although in principle the dependence on ordering could be removed by averaging the PAC likelihood over all possible orderings of the haplotypes, in practice this would require a sum over $n!$ terms, which is infeasible even for rather small values of n. Instead, as a pragmatic alternative solution, we propose to average L_{PAC} over several random orders of the haplotypes. Unless otherwise stated, all results reported here were obtained by averaging over 20 random orders. In our experience, the performance of the method is not especially sensitive to the number of random orders used — results based on 100 random orders gave qualitatively similar results, and results based on a single random order were often not much worse (data not shown) . It is, however, important that when comparing likelihoods for different values of ρ, the same set of random orders should be used for each value of ρ.

3 Estimating constant recombination rate

In this section we consider estimating the recombination rate when it is assumed to be constant across the region of interest. More precisely, we assume that crossovers in a single meiosis occur

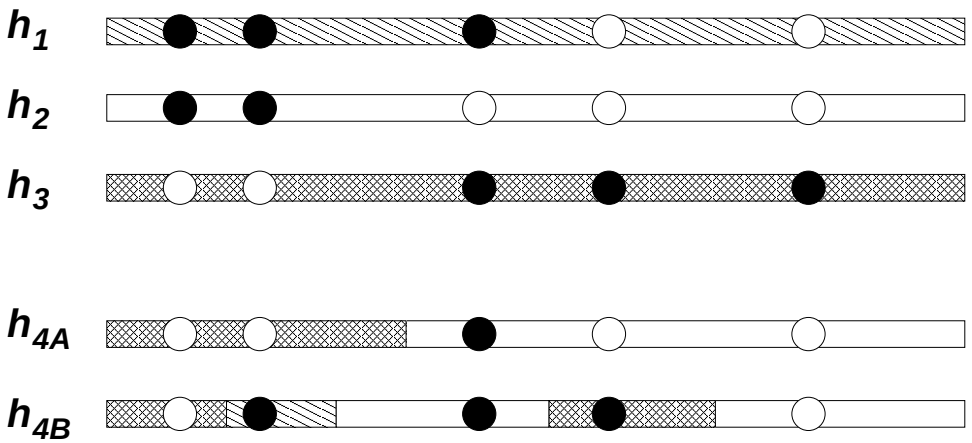

Figure 2: Illustration of how $\pi_A(h_{k+1}|h_1,\ldots,h_k)$ builds h_{k+1} as an imperfect mosaic of $h_1,\ldots,h_k$. The figure illustrates the case $k=3$, and shows two possible values (h_{4A} and h_{4B}) for h_4, given h_1,h_2,h_3. Each of the possible h_4s can be thought of as having been created by "copying" (imperfectly) parts of h_1,h_2 and h_3. The shading in each case shows which haplotype was "copied" at each position along the chromosome. Intuitively we think of h_4 as having recent shared ancestry with the haplotype that it copied in each segment. We assume that the copying process is Markov along the chromosome, with jumps (i.e. changes in the shading) occuring at rate ρ/k per physical distance. Thus the more frequent jumps in h_{4B} suggest a higher value of ρ than the less frequent jumps in h_{4A}. Note that for very large numbers of ρ the loci become independent, as they should. Each column of circles represents a SNP locus, with black and white representing the two alleles. The imperfect nature of the copying process is exemplified at the third locus, where each of h_{4A} and h_{4B} have the black allele, although they "copied" h_2, which has the white allele. In practice, of course, the shading is not observed, and so to compute the probability of observing a particular h_4 we must sum over all possible shadings. The Markov assumption allows us to do this efficiently using standard methods for Hidden Markov Models, as described in Appendix A.

as a Poisson process of constant rate c per unit (physical) distance, and consider estimating the scalar parameter $\rho = 4Nc$. We first use simulated data to examine the properties of the estimator $\hat{\rho}_{\mathrm{PAC-A}}$, corresponding to the conditional distribution π_A described in Appendix A, under what we will call the "standard coalescent model": constant-sized, panmictic population, with an infinite-sites mutation model. We show that, although quite accurate, $\hat{\rho}_{\mathrm{PAC-A}}$ exhibits a systematic bias. We use the empirical results to develop a modified conditional distribution π_B (described in detail in Appendix B), whose corresponding estimator $\hat{\rho}_{\mathrm{PAC-B}}$ exhibits considerably less bias, and is more accurate. We compare the performance of models based on both π_A and π_B with results from other methods.

3.1 Properties of the point estimate $\hat{\rho}_{\mathrm{PAC}}$

We used the program `mksample` (HUDSON 2002) to simulate data sets consisting of samples of SNP haplotypes from the standard coalescent model, for various values of

1. the number n of haplotypes in the sample.
2. the number S of markers typed.
3. the value of ρ (we measure physical distance so that the total physical length of each simulated haplotype equals 1.0. Thus the value of ρ is also the total value of ρ across the region.)

For each data set we found $\hat{\rho}_{\mathrm{PAC-A}}$ by numerically maximizing the PAC likelihood (using a golden bisection search method, PRESS *et al.* 1992), and compared it with the true value of ρ used to generate the data.

It seems natural to measure the error in estimates for ρ on a relative, rather than an absolute, scale. For example, WALL (2000) reported the frequency with which different methods for estimating ρ gave estimates within a factor of 2 of the true value, and both FD and HUDSON (2001) examine the distribution of the ratio $\hat{\rho}/\rho$ for their estimates $\hat{\rho}$, and the deviation of this ratio from the "optimal" value of 1. A problem with working with this ratio directly is that it tends to penalize over-estimation more heavily than under-estimation. For example, overestimating ρ by a factor of 10 gives a larger deviation from 1 than underestimating ρ by a factor of 10. To avoid this problem, we quantify the relative error of an estimate $\hat{\rho}$ for ρ by $Err(\rho,\hat{\rho}) = \log_{10}(\hat{\rho}/\rho)$. This gives, for example, an error 0 if $\hat{\rho} = \rho$, an error of 1 if $\hat{\rho}$ overestimates ρ by a factor of 10, and an error of -1 if $\hat{\rho}$ underestimates ρ by a factor of 10.

We note that $Err(\rho,\hat{\rho})$ can also be viewed as the error (on an absolute scale) in estimating $\log_{10}(\rho)$ by $\log_{10}(\hat{\rho})$. Thus, if the usual asymptotic theory for maximum likelihood estimation applies for estimation of $\log_{10}(\rho)$ in this setting (which, as discussed in FD, it may not) then for the actual MLE $\hat{\rho}_{\mathrm{MLE}}$ of ρ, $Err(\rho,\hat{\rho}_{\mathrm{MLE}})$ would be asymptotically normally distributed, centered on 0. Optimistically, we might therefore hope that for sufficiently-large data sets (large in terms of the number of haplotypes, the number of markers, or both) $Err(\rho,\hat{\rho}_{\mathrm{PAC-A}})$ might be approximately normally distributed, centered on 0. In our simulations, we found that for some combinations of

n, S and ρ this did indeed appear to be the case (e.g. Figure 3(b)), but that for other combinations, although the distribution often appeared close to normal, it was centered around some non-zero value (e.g. Figure 3(a),(c)), indicating a systematic tendency for $\hat{\rho}_{\mathrm{PAC-A}}$ to over- or under-estimate ρ. We will refer to the median of *Err* as the "bias" (of $\log(\hat{\rho}_{\mathrm{PAC-A}})$ in estimating $\log(\rho)$). Although bias is usually defined as a *mean* error, this is not particularly helpful here since the mean is often heavily influenced by a small number of very large values, and may even be infinite in some cases (see also FD). We therefore follow previous authors, including HUDSON (2001) and FD, in concentrating on the behavior of the median, rather than the mean, of the error.

Despite the biases evident in Figure 3(a) and (c), $\hat{\rho}_{\mathrm{PAC-A}}$ gives reasonably accurate estimates of ρ. For example, even in the right-hand panel (c) of Figure 3, which shows one of the most extreme biases we observed in our simulations, the bias corresponds to underestimating ρ by approximately a factor of 2, and $\hat{\rho}_{\mathrm{PAC-A}}$ is within a factor of 2 of the true value of ρ in 68% of cases. Although in many statistical applications estimates within a factor of 2 of the truth would not be considered particularly helpful or impressive, in this setting this kind of accuracy is often not easy to achieve (see for example WALL 2000).

We performed extensive simulations to better characterize the bias noted above, and found that although the bias depends on all 3 variables (n, S, and ρ), it is especially dependent on the average spacing between sites. More specifically, for fixed n and S we observed a striking linear relationship between the bias, and the log of the average marker spacing (Figure 4). This linear relationship was also apparent for data simulated under an assumption of population expansion (data not shown). The slope of the linear relationship is negative in each case, indicating a tendency for $\hat{\rho}_{\mathrm{PAC-A}}$ to overestimate ρ when the markers are very closely spaced and underestimate ρ when the markers are far apart. As the number of sampled haplotypes increases, both the slope and intercept of the line appear to get closer to 0 (Table 1). Based on these empirical results we can modify π_A to reduce the bias of the point estimates (see Appendix B for details). The improved performance of this modified conditional distribution, which we denote π_B, is illustrated in the next section.

Figure 4 also illustrates the effect of varying parameter values on the variability of point estimates. As might be expected, the variance of the error reduces with increased sample size, and increased number of sites, with the latter providing the more substantial decrease. For example, doubling the number of sites from 50 to 100 roughly halved the variance of the error in most cases, while doubling the number of individuals from 50 to 100 resulted in much smaller decreases. For a fixed sample size, and number of sites, the variance of the error decreases as the spacing between sites grows. This may be due to the fact that for larger spacings more recombination events occur, increasing the relative accuracy with which ρ can be estimated — although we would not expect this pattern to continue indefinitely as the marker spacing is increased beyond the range considered here.

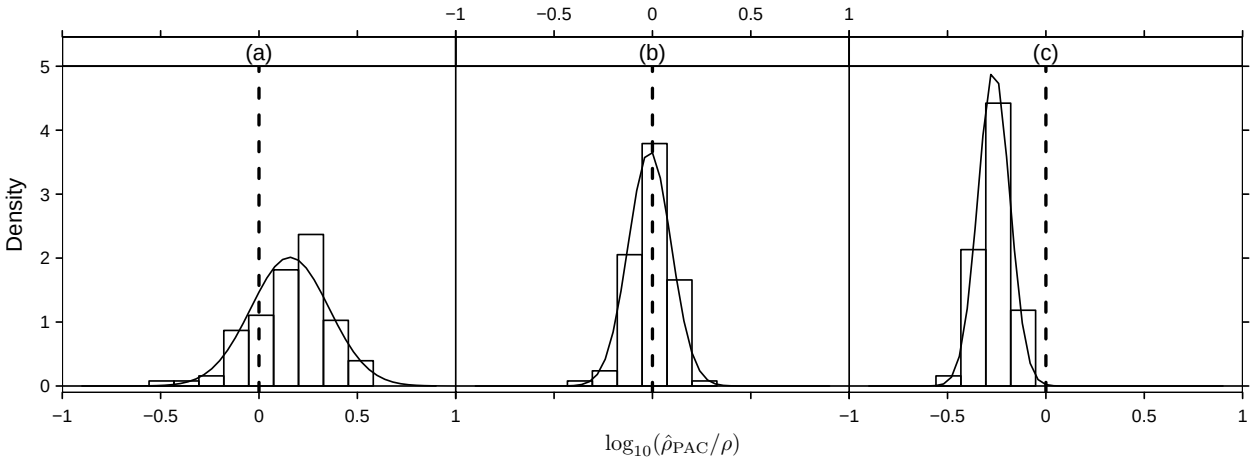

Figure 3: Histograms of the error $Err(\rho, \hat{\rho}_{\mathrm{PAC-A}}) = \log_{10}(\hat{\rho}_{\mathrm{PAC-A}}/\rho)$, each based on 100 data sets simulated from the standard coalescent model with $n = 50$ haplotypes and $S = 50$ segregating sites. The values of ρ are a) $\rho = 5$, b) $\rho = 25$, and c) $\rho = 500$. Superposed curves are normal densities with the same mean and standard deviation as the 100 values making up the histogram. These results, as well as those in Figure 4 and Table 1 are based on averaging the likelihoods over 10 random orders of the haplotypes.

	Intercept			Slope		
$n\backslash S$	20	50	100	20	50	100
20	-0.16	-0.12	-0.09	-0.18	-0.21	-0.26
50	-0.12	-0.07	-0.06	-0.16	-0.21	-0.24
100	-0.12	-0.06	-0.04	-0.09	-0.17	-0.21
200	-0.10	-0.05	-0.02	-0.06	-0.14	-0.17

Table 1: The intercepts and slopes of the linear relationship between $\log_{10}(\hat{\rho}_{\mathrm{PAC}}/\rho)$ and $\log_{10}(\mathrm{spacing})$ (see also Figure 4).

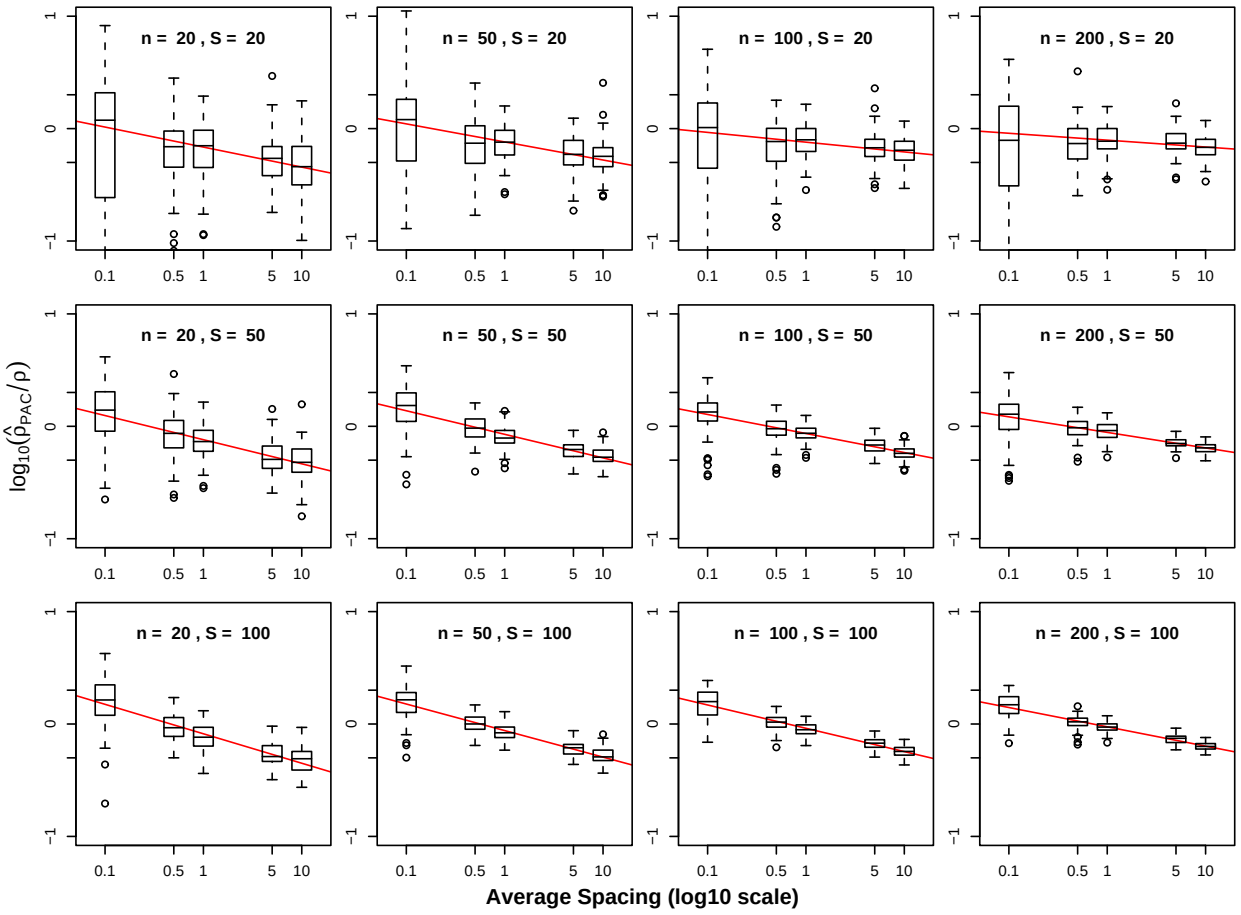

Figure 4: Box plots showing the relationship of the bias to the average marker spacing. For each combination of parameters, 100 data sets each were simulated under the standard coalescent model. The parameters involved are: the number of haplotypes in each sample $n = 20, 50, 100, 200$; the number of segregating sites $S = 20, 50, 100$; the average marker spacing $\rho/S = 0.1, 0.5, 1.0, 5.0$ and 10.0. In humans a marker spacing of $\rho/S = 0.5$ corresponds to roughly 1kb between markers. The unlabeled tick marks on the y-axis correspond to $\hat{\rho}_{\mathrm{PAC-A}} = \pm 2\rho$.

3.2 Comparison of point estimates with other methods

HUDSON (2001) introduced a composite likelihood method for estimating ρ, based on multiplying together the likelihood computed for every pair of SNPs. He compared the performance of this method with others in the literature (HUDSON and KAPLAN 1985; HUDSON 1987; HEY and WAKELEY 1997; WAKELEY 1997; WALL 2000), under the standard coalescent model, and found it to be as good as, or better than, the best of these. We compared the results reported by HUDSON (2001) for his maximum composite likelihood estimate, $\hat{\rho}_{\mathrm{CL}}$, with the results for $\hat{\rho}_{\mathrm{PAC-A}}$ and $\hat{\rho}_{\mathrm{PAC-B}}$ on data sets simulated under the same conditions (Figure 5). For data sets with small numbers of SNPs ($\leq\sim 12$) $\hat{\rho}_{\mathrm{CL}}$ provides the most accurate estimates of ρ, although all three methods struggle to produce reliable estimates. For larger numbers of SNPs both $\hat{\rho}_{\mathrm{PAC-A}}$ and $\hat{\rho}_{\mathrm{PAC-B}}$ tend, desirably, to exhibit less variability than $\hat{\rho}_{\mathrm{CL}}$. Further, $\hat{\rho}_{\mathrm{PAC-B}}$ exhibits little or none of the bias present in $\hat{\rho}_{\mathrm{PAC-A}}$, and provides the most accurate estimates of ρ.

The superior performance of the pairwise composite likelihood method for datasets with small numbers of SNPs is perhaps not surprising — indeed, for data sets with only 2 SNPs $\hat{\rho}_{\mathrm{CL}}$ is precisely the maximum likelihood estimate for ρ. However, we note that almost all of the improvement in accuracy comes from the increase in the 10th percentile of the estimator towards the true value, rather than a decrease in the 90th percentile. One possible explanation for this is that $\hat{\rho}_{\mathrm{CL}}$ uses a likelihood based on an infinite-sites mutation model (i.e. assumes no repeat mutation), and so is able essentially to rule out very small values for ρ if there is even one pair of sites at which all four gametes are present. (The effect of this may be compounded by the fact that $\hat{\rho}_{\mathrm{CL}}$ was found by maximizing over a grid of possible values, which forces all non-zero estimates of ρ to be above some threshold.) Our estimator does not make the infinite-sites assumption, and so will be more inclined to estimate very small values of ρ, possibly leading to occasional substantial underestimates. Since in real data it will typically be unclear whether or not the infinite-sites assumption holds, the advantage of $\hat{\rho}_{\mathrm{CL}}$ for even small numbers of sites is perhaps less clear-cut than it appears in Figure 5.

We used the same simulated data to examine the accuracy of estimates of ρ obtained by the methods described by KUHNER *et al.* (2000) and FD, both of which use computationally-intensive Monte-Carlo procedures to attempt to approximate the full coalescent likelihood. The computational complexity of these approaches increases with what might be called "the total value of ρ across the region", or "per-locus ρ", which we denote $\tilde{\rho}$ (more precisely, in our notation $\tilde{\rho} = \rho L$ where L is the physical length of the region). Results from FD suggest that even for small values of $\tilde{\rho}$ (< 3 say), the approximate likelihood curves obtained by these methods may be poor approximations to the actual likelihood curve, and so it seems unlikely that the curves will be accurate for larger $\tilde{\rho}$. However, point estimates based on these methods could still be accurate, if the maximum of the approximate likelihood curve occurs in about the right place. To investigate this possibility, we applied both methods, using approximately one day of CPU time per method per dataset (compared with roughly 30 seconds per dataset for $\hat{\rho}_{\mathrm{PAC-B}}$), to 10 of the data sets simulated with $\tilde{\rho} = 40$. Computational considerations make a more comprehensive simulation study inconvenient.

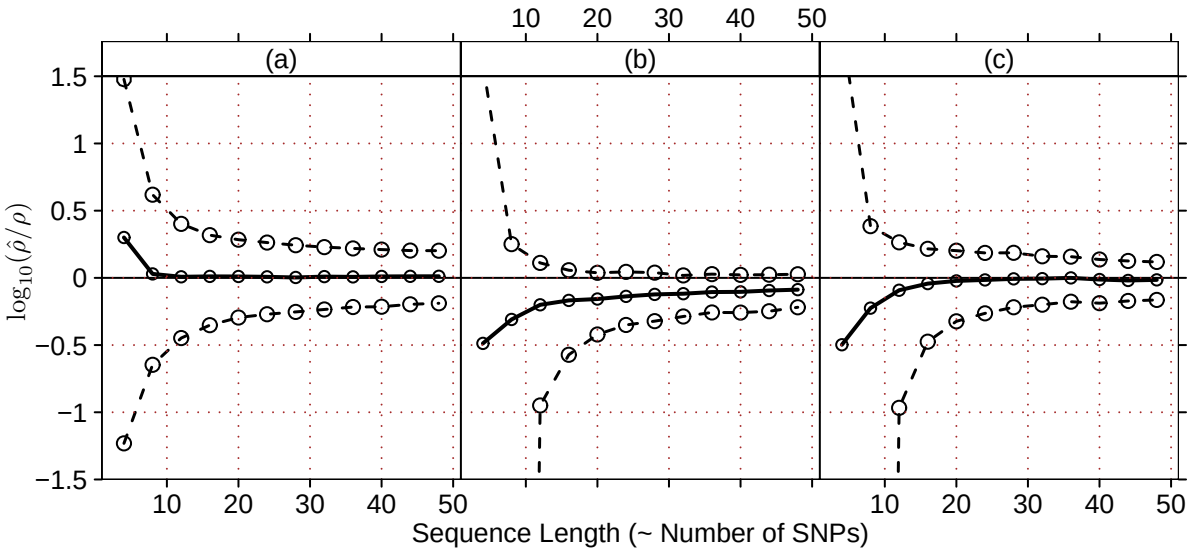

Figure 5: Comparison of $\hat{\rho}_{\mathrm{PAC-A}}$ and $\hat{\rho}_{\mathrm{PAC-B}}$ with Hudson's pairwise composite likelihood estimator $\hat{\rho}_{\mathrm{CL}}$ (HUDSON 2001), on data sets of $n = 50$ haplotypes simulated from the neutral infinite-sites model. The datasets were simulated with haplotypes of physical length $L = 4, 8, 12, 16, 20, 24, 28, 32, 36, 40, 44$ and 48 units, with $\rho = 1$ per unit physical length, and $\theta = 1/4$ per unit physical length. (With these parameters the expected number of SNPs in each dataset is approximately equal to the physical length of the haplotypes.) The results for $\hat{\rho}_{\mathrm{CL}}$ come from (HUDSON 2001), and were kindly provided by R. Hudson. The results for $\hat{\rho}_{\mathrm{PAC-A}}$ and $\hat{\rho}_{\mathrm{PAC-B}}$ are based on 1000 datasets we simulated for each set of parameters using the program `mksample` (HUDSON 2002). (We discarded the few simulated datasets that had only one SNP.) The panels are: (a) $\hat{\rho}_{\mathrm{CL}}$; (b) $\hat{\rho}_{\mathrm{PAC-A}}$; (c) $\hat{\rho}_{\mathrm{PAC-B}}$. In each panel, the solid line is the median of $Err(\hat{\rho}, \rho) = \log_{10}(\hat{\rho}/\rho)$ and the dashed lines are the 10% and 90% quantiles.

Each of the methods was run with θ fixed at the value used to simulate the data, giving them some advantage over how they could be used in practice. Nevertheless, neither method produced point estimates of $\tilde{\rho}$ as accurate as those from $\hat{\rho}_{\mathrm{PAC-B}}$ (Table 2). Of the two full likelihood schemes, the maximum of the likelihood curve obtained by `infs` was consistently closer to the true value of $\tilde{\rho}$ than the maximum of the likelihood curve obtained by `Recombine`. Indeed, the estimates obtained from `Recombine` were often close to an order of magnitude smaller than the true value of $\tilde{\rho}$, which raises a danger that when the method is applied to real data (for which the value of $\tilde{\rho}$ is of course not known) the user might be mislead into thinking that the value of $\tilde{\rho}$ is small enough for the method to produce reliable results. Results from longer runs of `infs` taking roughly five days of CPU time each, produced improved results, competitive with $\hat{\rho}_{\mathrm{PAC-B}}$ (data not shown).

Data Set	`Recombine`	`infs`	$\hat{\rho}_{\mathrm{PAC-B}}$
1	13	26	57
2	9	21	24
3	24	27	30
4	10	29	26
5	10	27	34
6	5	33	42
7	4	25	45
8	8	23	38
9	7	53	50
10	12	27	29
Median	10	27	36
Median $\lvert Err(\hat{\rho},\rho)\rvert$	0.62	0.17	0.11

Table 2: Comparison of $\hat{\rho}_{\mathrm{PAC-B}}$ with estimates of $\tilde{\rho}$ from `Recombine` and `infs`, for 10 datasets simulated with $\tilde{\rho} = 40$, $\theta = 10$, $n = 50$. Both `infs` and `Recombine` were run with θ fixed at its true value. `infs` was run for 20,000 iterations with 5 driving values for $\tilde{\rho}$ (10, 30, 40, 50, 60). The ESS at the MLE is always less than 4, indicating that `infs` had very little confidence in its estimated likelihood curve (and the estimated 95% CIs failed to include the true $\tilde{\rho}$ in all but one case). `Recombine` was run with five short runs of 20000 iterations and one long run of 1 million iterations, using three heating temperatures, initializing the runs at the true value of $\tilde{\rho}$, with θ fixed as 10. The CPU time (on an 800MHz Pentium III processor) for data set 8 was about 30 hours for `infs` and `Recombine`, and 30 seconds for $\hat{\rho}_{\mathrm{PAC-B}}$.

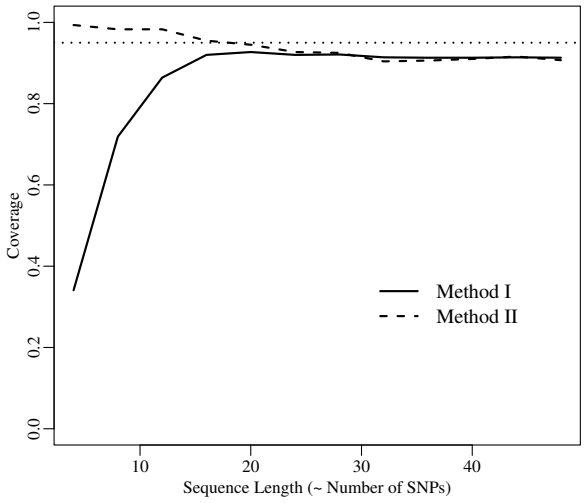

Figure 6: Empirical coverage properties of confidence intervals produced using two different methods described in text. Each number in the table is based on analysis of 1000 datasets, and shows the proportion of cases in which the CI contained the true value of ρ used to generate the data. The datasets used are the same as those used to produce Figure 5(c).

3.3 Properties of PAC likelihood curves

3.3.1 Construction of confidence intervals

We examined the coverage properties of confidence intervals (CIs) constructed from the PAC likelihood curve in two ways:

I: include all values of ρ for which $\log_e(L_{\mathrm{PAC-B}}(\rho))$ is within 2 of the maximum.

II: include all values of $\log_e(\rho)$ within $\pm 1.96\sigma$ of $\hat{\rho}_{\mathrm{PAC-B}}$, where σ is the square root of the inverse of minus the second derivative (found numerically) of the log of the PAC-B likelihood curve (as a function of $\log_e(\rho)$) evaluated at at $\rho = \hat{\rho}_{\mathrm{PAC-B}}$.

The rationale for looking at such CIs is that, under standard asymptotic theory for likelihood estimation, CIs constructed in this way using the *true* likelihood curve would include the true value of ρ approximately 95% of the time. (For I this follows from the asymptotic χ^2 distribution of the log likelihood ratio statistic; for II it follows from asymptotic normal distribution of the MLE).

Figure 6 shows the coverage properties for CIs produced using the two methods (i.e. the proportion of times that CIs formed using each method contained the true value of ρ), for the data sets used to obtain Figure 5(c). For moderate sequence length both methods produce CIs that are slightly anti-conservative, with coverage properties that approach ~ 0.91, compared with the expectation of ~ 0.95 under asymptotic theory. Based on these results we speculate that the curvature of the PAC-B likelihood curve does not deviate grossly from that of the true likelihood curve. We

note that the coverage properties are also closer to asymptotic expectations than those reported by (FEARNHEAD and DONNELLY 2002) for their composite likelihood using the same methods of CI construction.

3.3.2 Comparison with other methods

We compared the $L_{\mathrm{PAC-B}}$ likelihood curves with likelihood curves obtained from three other methods: the full-data coalescent method of FD (implemented in the computer program `infs`), and the pairwise composite-likelihood methods of HUDSON (2001) (implemented by one of us (NL), using tables available from R. Hudson's website) and MCVEAN *et al.* (2002) (implemented in the computer program `LDhat`). Figure 7 shows likelihood curves obtained using each method for the 20 data sets considered by WALL (2000), which were simulated under the standard coalescent model with $\rho = \theta = 3.0$ across a region of physical length 1, and were kindly supplied by J. D. Wall. (These likelihood curves are plotted with ρ on the *x*-axis, rather than $\log(\rho)$, because `infs` and `LDhat` output likelihood curves for evenly-spaced values of ρ.)

Interpreting the results of this comparison is slightly tricky. Unlike the other three methods we consider, the full-data coalescent method can, in principle, provides a fully accurate representation of the true likelihood curve. As such it is tempting to treat this as a "gold-standard" against which to compare the other methods. However, as mentioned previously, even for the rather small value of $\rho = 3$ used to generate these data, accurate approximation of the true likelihood curve may be computationally impractical. Indeed, the estimated effective sample sizes (ESSs) obtained for these data sets, shown above each panel in the figure, suggest that we should not place much confidence in the accuracy of many of the curves. Our attempts to obtain more accurate likelihood curves by performing longer runs for some of the data sets (numbers 15 and 16) actually produced *smaller* estimated ESSs, suggesting that the effective sample sizes quoted for the other data sets are optimistic (see FD for further discussion of this problem). A further complication in comparing the methods is that both our method, and that of MCVEAN *et al.* (2002), allow (implicitly, and explicitly, respectively) for the possibility of multiple mutations, and thus the likelihoods from these methods are in some sense not directly comparable with those from the other two methods. Finally, we note that the methods deal in different ways with the unknown mutation parameter θ: the likelihood curves shown from `infs` are profile likelihoods for ρ at the true value of θ; Hudson's method and our method avoid explicitly estimating θ; `LDhat` estimates θ using an analogue of Watterson's estimate, but allowing for multiple mutations.

Notwithstanding these issues, we attempt to draw some general conclusions:

1. In general, the likelihood curves produced by the 4 methods seem to agree rather more closely than might have been expected. (Compare, for example, the variability here with the variability observed for different runs of a single method in FD). However, the closeness of the agreement between the methods differs appreciably across datasets. Dataset 12 consists of only 4 sites, three of which are singletons, and so the differences in the curves for this

dataset seem not to be particularly interesting. We were unable to discern a systematic reason for the larger differences among methods observed in some of the other datasets (e.g., 16).

2. The two pairwise composite-likelihood methods tend to produce likelihood curves that are slightly more peaked than the other two methods. This might be expected since, as pointed out by McVean *et al.* (2002), pairwise composite-likelihood curves are typically more peaked than the true likelihood curve because they treat each pair of sites as independent, when in fact many pairs are highly dependent.

3. The method implemented in `LDhat`, which allows for multiple mutations, tends to achieve its maximum at larger values of ρ than Hudson's method, which does not allow for multiple mutations. This is surprising; indeed, the opposite might have been expected, since multiple mutations could be used in place of recombination events to explain certain patterns of LD. One possible explanation is that the run lengths we used for computing the likelihood in `LDhat` might be insufficient (we used the default values).

4. Different orderings of the haplotypes can give PAC likelihood curves that differ appreciably from one another. In addition, the maximum of the likelihood curve based on the average over several orderings tends to be towards the left end of the distribution of maxima obtained from different orderings. This is because, although not shown on the figure, the curves with maxima at smaller values of ρ tend to be larger (in absolute value) than those with maxima at larger values of ρ (presumably because they correspond to orderings of the haplotypes that, in some sense, require fewer recombination events to explain them), and thus contribute more to the average. Although this dependence on ordering is bothersome, in simulation studies (results not shown) we have found that the variability in the position of the maxima of the PAC likelihood over different orderings of the haplotypes is typically small compared with the uncertainty in estimation of ρ.

4 Variable recombination rate.

4.1 Models for variation in recombination rate

One of our main motivations for developing this model is to explore fine-scale variation in recombination rates. A simple (no interference) model for variation in recombination rates is that crossovers in a single meiosis occur as an inhomogeneous Poisson process, of rate $c(x)$ at position x. Here we consider two specific cases of this general model:

1. A simple single-hotspot model, where

$$c(x) = \begin{cases} \lambda \bar{c} & \text{for } a \le x \le b, \\ \bar{c} & \text{otherwise.} \end{cases} \tag{4}$$

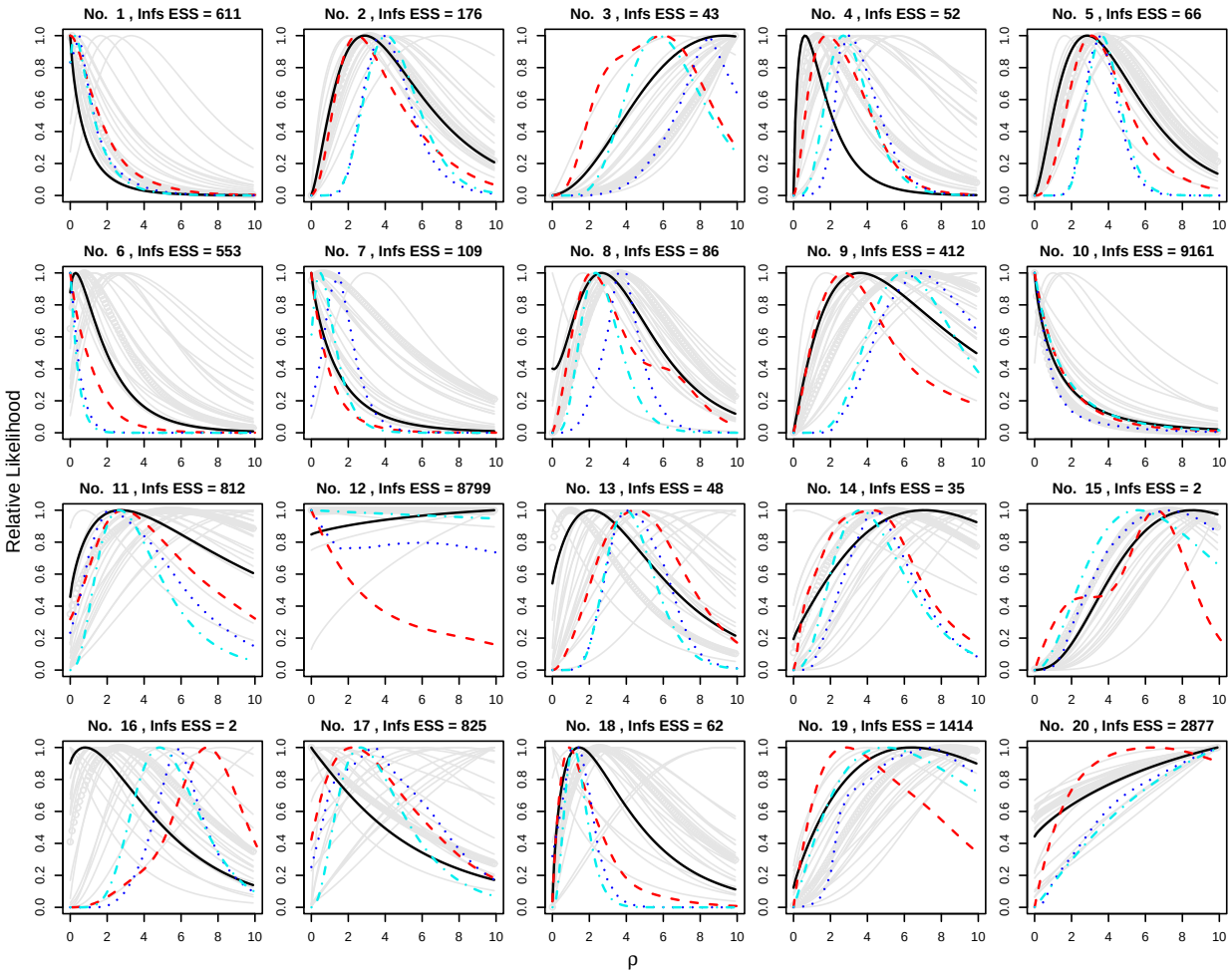

Figure 7: Comparison of the relative PAC likelihood curves, with coalescent-based and pairwise composite relative likelihood curves, for the first 20 data sets in WALL (2000). In each case the relative likelihood is obtained by normalizing each likelihood curve to have a maximum of 1. The light gray lines show 20 PAC likelihood curves, each from a different random order of the haplotypes, and the solid dark line is based on the PAC likelihood averaged over the 20 random orders. The other lines correspond to likelihood curves computed using the methods of: FD, implemented in the computer program `infs` (red dashed line); MCVEAN *et al.* (2002), implemented in `LDhat` (blue dotted line); HUDSON (2001), using the table generated by program `eh` written by Hudson (cyan dot-dashed line). The effective sample sizes (ESS) for `infs` at the MLE is given for each data set above the graph, and is a measure of the confidence `infs` has in its estimated likelihood curve (the larger the better). Results for `infs` for all data sets except numbers 15 and 16 were kindly provided to us by P. Fearnhead, and were obtained using between 50,000 and 5,000,000 iterations. Results for datasets 15 and 16, were obtained by ourselves using 10,000,000 iterations.

Here $\bar{c}$ represents the background rate of crossover, a and b represent the left and right ends of the hotspot region, and λ (> 1) quantifies the magnitude of the recombination hotspot. The PAC likelihood for this model is a function of four parameters: a, b, λ, and $\bar{\rho} = 4N\bar{c}$.

2. A more general model, where if x is a position between markers j and $j+1$ then

$$c(x) = \lambda_j \bar{c}. \tag{5}$$

Here $\bar{c}$ represents the background rate of crossover, and λ_j is a multiplier controlling how the crossover rate between markers j and $j+1$ deviates from the background rate. The PAC likelihood for this model is a function of the parameters $\lambda_1, \ldots \lambda_{S-1}$ (where S is the number of SNPs), and $\bar{\rho} = 4N\bar{c}$.

For the simple single-hotspot model it is straightforward to obtain numerically the maximum PAC likelihood estimates for all 4 parameters simultaneously, although in the examples we consider we assume that a and b are known, and maximize the PAC likelihood in terms of λ and $\bar{\rho}$. The evidence for the presence of a hotspot can be summarized by the log likelihood ratio (LLR) for the null hypothesis of no hotspot, $H_0 : \lambda = 1$, versus the alternative $H_1 : \lambda > 1$. If $\bar{\rho}_0$ denotes the value of $\bar{\rho}$ that maximizes $L_{\mathrm{PAC-B}}$ under H_0, and $\bar{\rho}_1$ and λ_1 denote the values of $\bar{\rho}$ and λ that maximize $L_{\mathrm{PAC-B}}$ under H_1, then

$$\mathrm{LLR} = \log_e L_{\mathrm{PAC-B}}(\bar{\rho}_1, \lambda_1) / L_{\mathrm{PAC-B}}(\bar{\rho}_0, \lambda = 1), \tag{6}$$

and large values of LLR represent evidence for the existence of a hotspot. Under standard asymptotic theory, 2 times LLR would have (asymptotically) a chi-square distribution on 1 degree of freedom, and so rejecting H_0 if LLR > 1.92 would give a hypothesis test with a type I error rate of 0.05. Although it seems unlikely that standard asymptotic theory will apply here, we found that for data sets simulated under the null hypothesis, rejecting H_0 for LLR > 1.92 gave empirical type I error rates close to 0.05 (Table 3), which provides some guidance as to what might be considered a "large" value of LLR.

For the second, more general model, obtaining maximum PAC likelihood estimates for the parameters creates problems. First, the maximum likelihood estimates are not unique (indeed there are infinitely many of them), because multiplying all the λ_j by any constant, and dividing $\bar{\rho}$ by the same constant, gives exactly the same likelihood. (In technical terms, the parameters are said to be *unidentifiable*.) Second, even if the identifiability problem is solved (for example by first obtaining an estimate for $\bar{\rho}$ assuming that there is no hotspot, and then fixing this when estimating the other parameters) there is the practical problem that the likelihood curve for some λ_j will often be very flat, resulting in estimates for many λ_j being very close to (or equal to) either 0 or infinity. This seems undesirable: if the likelihood for a particular λ_j is very flat, this indicates that there is little information about the recombination rate in that marker interval, in which case it seems sensible to estimate that the recombination rate is close to the background rate (i.e. $\lambda_j \approx 1$), rather than (close to) infinitely bigger or smaller!

To solve both these problems, we assume a "prior" distribution for the λ_js: specifically that the λ_js are independent, and identically distributed, with $\log_{10}(\lambda_j) \sim \mathcal{N}(0, 0.5^2)$. This prior was chosen to allow occasional deviations from the background rate of recombination by a factor of 10 or more (with probability approximately 95%, λ_j lies in the range $0.1 - 10$). This choice of prior could be motivated from a Bayesian viewpoint as reflecting our prior beliefs about the λ_js, but it also has the more pragmatic justification that identifying variations of this kind of magnitude seems both interesting and, perhaps, attainable. We consider alternative prior specifications in the discussion.

In principle, given the prior distribution for the λ_js described above, we could also place a prior distribution on $\bar{\rho}$, and obtain an approximation to the posterior distribution of all parameters, using Markov chain Monte Carlo for example. Although this would be our preferred approach, for simplicity we avoid this here, and instead use the following *ad hoc* two-stage approach: first obtain point estimates for $\bar{\rho}$ and λ by jointly maximizing the product of $L_{\mathrm{PAC}}(\bar{\rho}, \lambda)$ and the prior density of λ; second obtain a "posterior distribution" for each λ_j, by fixing all other parameters at their estimated values, discretizing the prior on λ_j (truncated at $\lambda_j = \pm 3$), and computing the corresponding discretized posterior distribution as being proportional to the prior times the PAC likelihood. For data sets with large number of sites the first stage (optimization over $\bar{\rho}, \lambda$) can be very time-consuming, requiring large numbers of evaluations of the likelihood function. Further, it seems unlikely that the simple optimization method we used will reliably find the global maximum of the likelihood surface. Both these problems could be alleviated by exploiting the fact that the derivatives of the PAC likelihood can be computed efficiently, but we do not pursue this here.

4.2 Power to detect recombination hotspots, and robustness

In this section we assume that there is a single recombination hotspot (model 1 above), whose putative position is known, and examine the power of our model to detect the hotspot under various assumptions about the population demography, and SNP marker ascertainment. Although the assumptions made here (in particular, that there is a single hotspot with known putative position) are unrealistic, they provide a convenient framework within which to examine quantitatively the power of our approach, and how it is affected by population demographic history, and marker ascertainment schemes.

We consider the following scenarios:

1. constant-size randomly-mating population, all markers.
2. constant-size randomly-mating population, only markers at frequency > 0.1.
3. exponentially expanding population, with expansion starting $t = 500$ generations ago.
4. exponentially expanding population, with expansion starting $t = 5000$ generations ago.

5. haplotypes sampled from a structured population, consisting of two islands exchanging migrants at a rate of one per generation (scaled migration parameter $4Nm = 4$).

6. haplotypes sampled from only one of the islands in the structured population described above.

These last four models are the same as those considered in PRITCHARD and PRZEWORSKI (2001). In the two expanding-population scenarios, the population was assumed constant-sized until t generations ago, when it started to expand exponentially, continuing until the present. The current population size N_0 is set to be 10^5 and the population growth rate is chosen, as a function of t, to match the expected diversity in a population of constant size 10^4. (The necessary growth rates of $\alpha = 1960, 350$ for $t = 500, 5000$ were kindly provided by M. Przeworski.)

For each scenario we simulated data sets under the simple single-hotspot model described above, using `mksample`, and the postprocessing algorithm described in Appendix C. For the first two scenarios each data set was simulated to have about 50 segregating sites and 60 chromosomes, with $a = 0.4$, $b = 0.5$, $\bar{\rho} = 20$, and $\lambda = 10$ (these values were chosen to approximately match values for the TAP2 data from JEFFREYS *et al.* (2000) considered in Section 4.3.2). For the expanding population scenarios we set $\bar{\rho} = 4N_0c = 200$, and for the structured scenarios $\bar{\rho} = 20$ within each population. For each scenario we also simulated data sets under the same conditions, but with no hotspot (i.e. $\lambda = 1$).

We applied the likelihood-ratio test to each data set to test the null hypothesis $H_0 : \lambda = 1$, against the alternative $H_1 : \lambda > 1$ (Table 3). For the scenarios not involving population expansion, the test gave type I error rates of approximately 0.05 when applied to data without a hotspot, and a power of approximately 0.90 when applied to data simulated with a hotspot, although the test based on just the common SNPs had a slightly reduced power. The two scenarios involving population expansion gave either a substantial reduction in power, or an inflated type I error rate (which is in some sense equivalent to a reduction in power). This might be due to a reduction in the number of "common" SNPs under these scenarios, as common SNPs tend to be most informative for estimating recombination rates.

We also examined the robustness of estimates of λ under the various scenarios (Table 3). As noted by FD, the recombination rate $\bar{\rho} = 4Nc$ depends on how the effective population size N is defined. In contrast, the definition of the parameter λ does *not* depend on how the effective population size is defined, and so we might hope that estimation of λ will be robust to departures from the assumption of a constant-sized panmictic population. For the levels of population structure we used in our simulations this does indeed appear to be the case – in both cases estimates were more accurate than for the sample from a single random-mating population, perhaps because population structure makes recombinants easier to "spot". As might be expected, estimates based only on common SNPs were less accurate than those based on all SNPs. A drop in accuracy is also evident for the scenarios simulated under population expansion, probably again due to a reduction in the number of "common" SNPs under these scenarios. Some of the scenarios also resulted in an upward bias for estimates of λ, notably one of the expansion scenarios in which the median of the

	a) $\lambda = 10$			b) $\lambda = 1$		
	Power	Med $\hat{\lambda}$	Med $\lvert Err \rvert$	Type I Error	Med $\hat{\lambda}$	Med $\lvert Err \rvert$
All Sites	0.90	13.02	0.19	0.04	0.95	0.45
Common Sites ($f > 0.1$)	0.81	16.57	0.29	0.05	1.13	0.57
Island (mixed)	0.92	11.35	0.18	0.07	0.90	0.33
Island (single)	0.94	10.24	0.16	0.07	1.01	0.31
Expansion ($t = 500$)	0.94	24.29	0.40	0.13	1.41	0.56
Expansion ($t = 5000$)	0.53	11.66	0.36	0.07	1.23	0.70

Table 3: Performance of the simple single-hotspot model, for data sets simulated under various demographic scenarios, a) with a hotspot of magnitude $\lambda = 10$, and b) with no hotspot (i.e. $\lambda = 1$). In each case, the first column shows the proportion of "significant" LR tests (LLR > 1.92) for testing the null hypothesis of no hotspot, the second column shows the median estimate of λ, and the third column shows the median of $|Err(\hat{\lambda}, \lambda)|$. Each number in the table is based on results for 200 simulated data sets.

estimates was almost 2.5 times the true value.

4.3 Estimating recombination rates along a region

4.3.1 Simulated Data

We fitted the more general varying recombination rate model to the simulated data used to produce the pairwise LD plots in Figure 1; the results are shown in Figure 8. From the latter figure we might conclude, correctly, that the data sets corresponding to the top left and bottom middle panels had recombination hotspots somewhere in the region 0.4-0.6. We might also conclude that the other data sets had no hotspots, which would be correct except in the case of the bottom left figure, which was actually generated from data with a hotspot between 0.4 and 0.5. One possible reason that the hotspot shows up less well in this case is that there are fewer sites at high frequency (> 0.15) in this data set. Despite the fact that we might have been mislead in 1 case out of 6, we view Figure 8 as considerably more informative than Figure 1, from which we find it difficult to draw any conclusions.

4.3.2 TAP2 Data

JEFFREYS *et al.* (2000) used patterns of LD (measured by haplotype diversity) in a population sample to refine the location of a putative recombination hotspot in the human TAP2 gene, and provided a more detailed characterization of its properties through sperm typing. The population sample consists of 30 individuals from UK typed at 47 polymorphisms (45 SNPs, 2 insertion/deletions) across 9.7kb, with haplotypes determined by allele-specific PCR. Through analysis

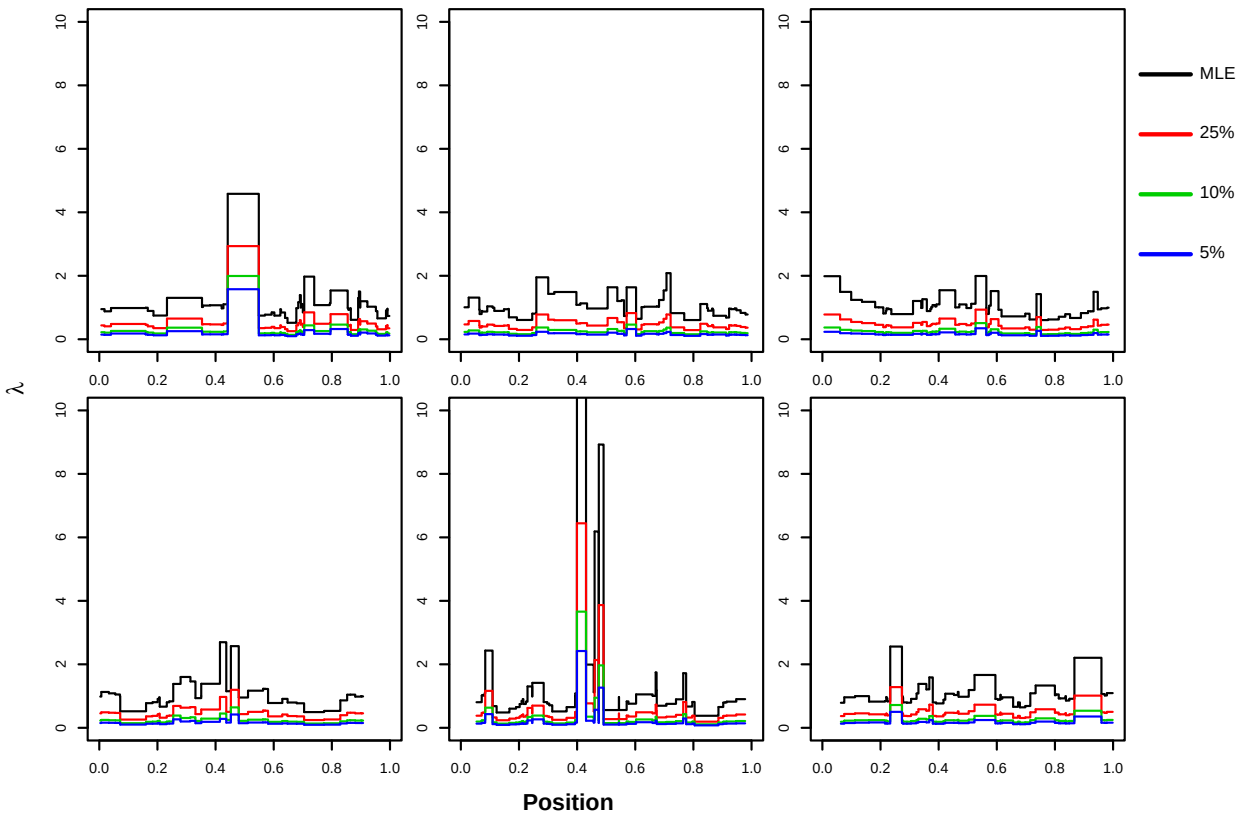

Figure 8: Estimates of variation from background recombination rate within each marker interval, for the same simulated data sets that were used to produce Figure 1. The panels top-left, bottom-left, and bottom-middle correspond to data sets simulated with a single hotspot of magnitude $\lambda = 10$ between positions 0.4 and 0.5. The other panels correspond to data simulated with constant recombination rate across the region.

of sperm crossover events JEFFREYS *et al.* (2000) identified a region of increased crossover intensity, located approximately in the interval from 4 to 5.2kb.

We fitted the simple single-hotspot model to the haplotype data (kindly provided in convenient electronic format by A. Jeffreys), assuming a hotspot between 4 and 5.2 kb, and obtained estimates of $\bar{\rho} = 1.3$ per kb, and $\lambda = 12$ (95% CI $[6, 21]$), with a LLR of 12, indicating strong evidence for the presence of the hotspot. Our estimate for the average magnitude of the hotspot being $\lambda = 12$ times the background rate agrees well with the sperm-typing results from JEFFREYS *et al.* (2000). In particular, JEFFREYS *et al.* (2000) (their Fig 4) observed 128 crossovers within the interval 4 to 5.2 kb, in 2.4 million progenitor molecules, giving an average rate of 4.4 cM/Mb, which is 11 times the approximate background rate (for males) they quote, of 0.4 cM/Mb (although they warn that this estimate of the background rate should be "treated with caution"). Since our estimate is based on a population sample, it is actually an estimate of the magnitude of the hotspot in the sex-average crossover rates. JEFFREYS *et al.* (2000) point out that the crossover rate in this region appears to be substantially higher in females than in males, and so the sex-average crossover rates are likely to be dominated by the female crossover process. Our results therefore suggest that the crossover rate within the 1.2 kb hotspot in female meioses is roughly an order of magnitude higher than the (female) background rate.

JEFFREYS *et al.* (2000) found that if one assumes that the sex-average background recombination rate is equal to the male background recombination rate of ≈ 0.4 cM/Mb (a figure that we again emphasise they suggest should be treated with caution) then the observed patterns of LD in the population sample appear consistent with an effective population size of $N = 100,000$, which contrasts with the more commonly-quoted figure for humans of $N = 10,000$. Our estimate of $\bar{\rho} = 1.3$ per kb supports their analysis, as it corresponds to $N \approx 84,000$ if the background sex-average recombination rate is assumed to be 0.4cM/Mb. As pointed out by JEFFREYS *et al.* (2000), one possible explanation for this is differences between male and female recombination rates — in particular a sex-average background rate across the 9.7kb region of 3.4 cM/Mb would give $N \approx 10,000$. An alternative (or additional) explanation, suggested to us by M.Przeworski (personal communication), is that gene conversion events not detected by the sperm typing experiments could partially account for the unusually large estimated effective population size.

Figure 9 shows the estimates of λ_j and posterior quantiles obtained by fitting the more general model for recombination rate variation to the TAP2 haplotype data. The hotspot in the region 4-5.2 kb is fairly clear, with some suggestion that it may extend slightly further to the right than 5.2 kb. The peak of the recombination hot spot is estimated as being about 14 times the background rate. In the interval corresponding to this peak the posterior probability that λ_j is greater than 7 is 75% (compared with a prior probability of less than 5%). However, the large number of parameters estimated within this more general model results in generally poor precision for the estimate of each λ_j. In particular, confidence in the estimates is probably not sufficient to conclude that the three subpeaks present in the figure, within the hotspot, correspond to actual variation in the hotspot intensity.

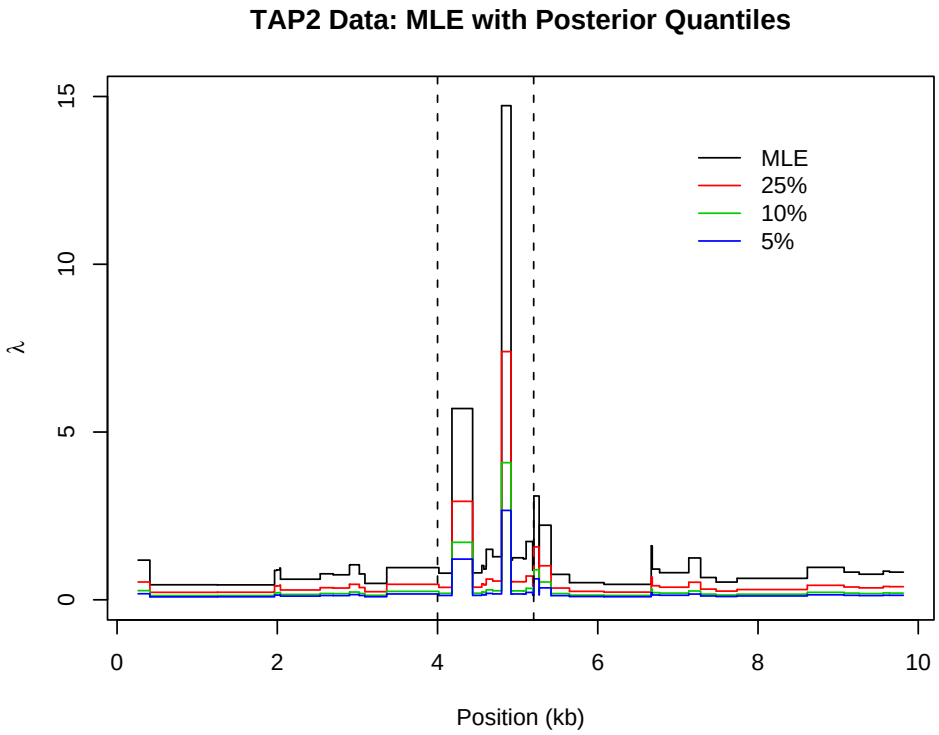

Figure 9: Results of fitting the more general model for varying recombination rate to the TAP2 data from JEFFREYS *et al.* (2000). The plot shows the estimated value, and 25, 10 and 5 percent posterior quantiles, for λ_j in each marker interval along the chromosome. The vertical lines indicate the approximate boundaries of the hotspot identified by JEFFREYS *et al.* (2000).

Population	$\bar{\rho}$ per kb	$\hat{\lambda}$	CI	LLR
Jackson	7.5	2.7	[1.3 , 4.8]	3
Rochester	0.74	12	[5.0 , 25]	8
Finland	0.15	104	[55 , 183]	22
Combined	7.0	6.5	[4.1 , 8.5]	23

Table 4: Results of fitting the simple single-hotspot model to the LPL data, to each subpopulation sample individually, and to the combined sample. The first two columns show estimated values for $\bar{\rho}$ and λ assuming a hotspot between 3 and 5 kb along the sequence. The column CI gives the range of values of λ whose log likelihood is within 1.92 of the log likelihood at $\hat{\lambda}$ (for $\bar{\rho}$ fixed at its estimated value). The column headed LLR gives the value of the log likelihood ratio for testing the null hypothesis of $\lambda = 1$.

4.3.3 Lipoprotein Lipase Data

The LPL data (NICKERSON *et al.* 1998; CLARK *et al.* 1998) consists of 9.7 kb of genomic DNA sequence from human lipoprotein lipase gene from 71 individuals from Jackson ($n = 24$), Rochester ($n = 23$) and North Karelia, Finland ($n = 24$). In the published data, the haplotypic phase for 69 sites was either determined by experiment or estimated by Clark's algorithm (CLARK 1990). Although the use of a statistical method to infer some of the phases means that there is some possibility that not all the published haplotypes are completely correct, the majority seem likely to be accurate, and in this analysis we assume them to be known without error. Based on patterns of LD, and on the results of phylogenetic-based methods that attempt to infer ancestral recombination events, TEMPLETON *et al.* (2000) suggested the existence of a putative recombination hot spot between $[2987, 4872]$.

Table 4 shows the results of fitting the simple single-hotspot model to the whole dataset, and to the data from each subpopulation individually, assuming a hotspot from 3 to 5 kb. Figure 10 shows the results for fitting the more general model for recombination rate variation. Overall, these results seem to support the existence of the putative hotspot, although there is considerable variation in the strength of the evidence (as measured by the LLR), and of the estimated magnitude of the hotspot, in different subpopulations. We note that the apparent magnitude of the hotspot in the Finnish population is smaller in Figure 10 than in Table 4, due the affect of the prior. There is also tentative evidence, mostly from the Jackson sample, for a smaller-magnitude hotspot between 8 and 9kb. Although no interval in that region produces a very large estimate for λ_i, the clustering together of three intervals with moderate λ_i provides stronger evidence than any one of these estimates taken separately.

Our results are consistent with those from FEARNHEAD and DONNELLY (2002), who found evidence for the $[2987, 4872]$ hotspot in the samples from Rochester and Finland, but not in those from Jackson. In addition both we and FEARNHEAD and DONNELLY (2002) found that the Rochester and Finland samples give much smaller estimates for $\bar{\rho}$ than the Jackson sample, prob-

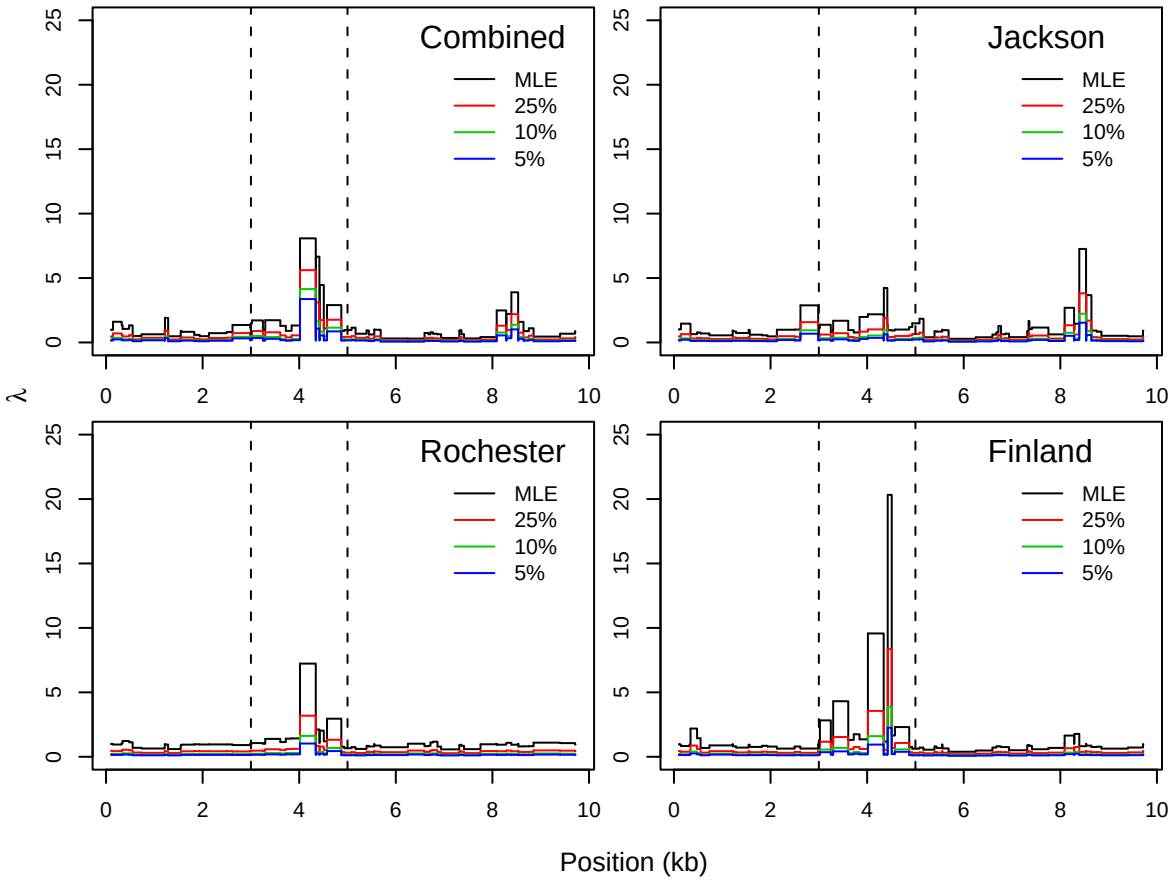

Figure 10: Results of fitting the more general model for varying recombination rate to the LPL haplotype data from (CLARK *et al.* 1998). The plot shows the estimated value, and 25, 10 and 5 percent posterior quantiles, for λ_j in each marker interval along the chromosome. The vertical dotted lines indicate the boundaries of the putative recombination hotspot identified by TEMPLETON *et al.* (2000)

ably reflecting smaller effective population sizes as a result of a recent bottleneck. One general advantage of the approach we take here, over the method of considering segments of the chromosome separately as do FEARNHEAD and DONNELLY (2002), is that it uses the patterns of LD not only between markers *within* the hotspot, but also between markers *either side* of the hotspot, to estimate the magnitude of the hotspot. This may explain why we detected a signal (albeit a modest one) for a $[2987, 4872]$ hotspot in the Jackson sample, where FEARNHEAD and DONNELLY (2002) did not.

The large differences among estimates for λ from the three separate population samples are surprising. To examine whether this might be simply due to poor precision for these estimates in one or more of the populations (due for example to too much or too little diversity), we constructed approximate 95% confidence intervals for λ using an analogue of method I in section 3.3.1 (column headed CI in Table 4). Although the coverage probabilities for these CIs are unlikely to be 95%, they give some indication of the curvature of the likelihood surface, and the fact that none of the three intervals overlaps with either of the others suggests that the hotspot intensity may indeed vary among the three populations. Additional simulation results (not shown) suggest that the larger effective population size of the Jackson population should actually *increase* power to detect the hotspot compared with the Finnish population, and so differences in effective population sizes do not appear to explain our results. An alternative explanation is the bias we observed for our estimates of λ under certain expansion scenarios (Table 3), which might partially explain the large estimate of λ in the Finnish population for example, although it is unclear whether this is enough to account for the fact that the estimated λ is almost 40 times greater than in the Jackson population. Biological mechanisms that could lead to different patterns of recombination rate heterogeneity in different populations are known to exist (e.g. JEFFREYS and NEUMANN 2002), and the kinds of method we introduce here should be helpful in determining how often this occurs in practice.

5 Discussion

In this paper we have introduced a new statistical model relating patterns of LD at multiple loci to the underlying recombination rate, and examined its effectiveness for inferring the underlying rate of recombination. Another potential application of our model is to methods for LD (association) mapping in "case-control" studies, where chromosomes have been collected and typed for both case and control individuals. Several authors, including MCPEEK and STRAHS (1999), MORRIS *et al.* (2000), and LIU *et al.* (2001) have developed methods to use genetic types at multiple loci to perform association mapping for case-control studies. These methods aim to improve on other common methods — which typically test small groups of markers, one group at a time, for association with a trait — by considering data at many SNP markers simultaneously. Although the methods differ in details, broadly speaking they all pursue a strategy of assuming that (subsets of) the case chromosomes share some region identical by descent about a causal mutation, and as a result will be more similar than would be expected by chance. The challenge then is to identify

regions where (subsets of) the case chromosomes are more similar than would be expected by chance. Models of LD play a key role here, because what would be expected "by chance" depends critically on the amount of LD among loci. In particular, correlations among loci will cause chromosomes to tend to be more similar by chance than if the loci were independent. MCPEEK and STRAHS (1999) use a first-order Markov chain to model LD, so that the probability of observing types $(x_1,\ldots,x_L)$ at L loci along a chromosome is $\Pr(x_1)\Pr(x_2|x_1)\Pr(x_3|x_2)\ldots\Pr(x_L|x_{L-1})$, where the conditional probabilities $\Pr(x_r|x_{r-1})$ are estimated using the control chromosomes. This model was also adopted by MORRIS *et al.* (2000). While the first-order Markov assumption is better than assuming that the loci are independent, and may suffice if there is little LD among markers, it seems not to be a good model for LD in general. In particular, it fails to capture the fact that markers may be in weak LD with neighboring markers, but in strong LD with more distant markers. Although MCPEEK and STRAHS (1999) mention that higher order Markov models might better model LD, such models seem unlikely to be helpful in practice because of the difficulty of estimating all the necessary parameters. The model we have introduced here provides a parsimonious method for modeling LD: even the more general model for varying recombination rates has fewer parameters than the first order Markov model used previously. Further, in these kinds of applications, where estimation of underlying recombination rates may be of only indirect interest, the usefulness of our model will depend only on whether $\Pr(h_1,\ldots,h_n \mid \rho)$ is a sensible distribution for $h_1,\ldots,h_n$ for *some* value of the parameters ρ, even if this ρ does not correspond precisely to the background recombination rate scaled by the effective population size. Under these circumstances our two approximations π_A and π_B should perform almost identically, and so π_A might be preferred on the grounds that it is simpler to understand and implement, and is more amenable to theoretical study.

Another model for LD across multiple sites, introduced by DALY *et al.* (2001), is based on the empirical observation that in some regions of the genome LD exhibits a "block-like" structure. DALY *et al.* (2001) model each observed haplotype as a mosaic of "ancestral haplotypes", with the transition rates among these ancestral states (representing the "historical recombination frequency" between each pair of consecutive markers) being estimated by maximum likelihood. The ancestral haplotypes are identified by an initial scan for regions of low haplotype diversity, although in principle they could instead be treated as parameters in the model. DALY *et al.* (2001) used this model to produce a summary of patterns of LD that illustrates the haplotype structure in their data more clearly, and in more detail, than would plots of pairwise LD measures. However, it is currently unclear to what extent this model might be helpful for applications involving statistical inference, or prediction, particularly in regions where patterns of LD are less "block-like".

There are several challenges that might arise in applying our method to real data that we have ignored here. In particular, we have assumed in our examples that haplotypes are known, and that there are no missing genotypes or genotyping errors. A new version of the software package PHASE (STEPHENS *et al.* 2001) is under development, which will deal with these problems by incorporating the PAC likelihood into a Markov chain Monte Carlo (MCMC) algorithm to jointly estimate the recombination rate parameters, haplotypes, missing genotypes, and potential locations of genotyping errors. This algorithm also produces a method for estimating haplotypes that

takes account of decay of LD along chromosomes. Preliminary results for simulated data suggest that these ideas result in slightly more accurate haplotype estimates than the method described in STEPHENS *et al.* (2001).

There are also biological aspects of real data that we have not accounted for here, including for example gene conversion, whose affect on patterns of LD in humans has been the subject of considerable recent interest (see FRISSE *et al.* 2001 for example). The effect that the presence of gene conversion will have on our method will vary, depending on how the tract length — about which little is known in humans — compares with the marker density. Gene conversion events with very small tract-lengths compared to the marker density will only rarely involve a typed marker, and so will tend to have a small effect on our method unless such events are extremely common. Conversely, gene conversion events with longer tract lengths — comparable to the typical distance between markers — will often affect one or more markers, and will tend to look like double crossover events to our method. The presence of gene conversion with this kind of tract length will thus elevate our estimates of recombination rate, perhaps substantially, and regions with elevated rates of such gene conversion may appear as recombination hotspots in our method. In principle the PAC model could be extended to account explicitly for gene conversion, by suitable modification of the conditional distribution π. A concrete suggestion for how to achieve this would be to augment the space of the hidden Markov Model for the mosaic process (described in detail in appendix A) to include both the current and previous "copied" chromosome, and then to modify the Markov jump process to make jumps back to the previously-copied chromosome more likely than jumps to other chromosomes. However, this would greatly increase the computational expense of the model, making it unappealing in practice. A more attractive possibility would be to settle for modeling only those gene conversion events that affect a single marker (which, depending on tract length and marker density, may be the vast majority of gene conversion events affecting patterns of LD). This would require only a simple modification of the conditional distribution (it could be handled similarly to the way that mutations are currently handled), with essentially no increase in the computation required.

Another aspect of real data that we have not accounted for explicitly is population structure. Our simulation results in Table 3 suggest that for the purposes of identifying recombination hotspots our method is robust to a certain amount of population structure. Nevertheless, modeling population structure explicitly might prove helpful in some settings. For example, it could be used to extend methods for detecting population structure from unlinked markers (e.g. PRITCHARD *et al.* 2000) to allow them to be applied to sets of tightly-linked markers. Again, a natural approach is to modify the conditional distribution π to account explicitly for population structure. One suggestion is to modify the copying process in the $k+1$st chromosome (see Appendix A) so that, rather than being equally likely to copy all r existing chromosomes, it is more likely to copy chromosomes from the same population than chromosomes from a different population. This would effectively model population structure by increasing the probability of seeing similar chromosomes in the same population, compared with in different populations. We are currently investigating the effectiveness of a similar idea for LD mapping in case-control studies: treating cases and controls

as separate populations, and examining whether there appears to be evidence in some regions for the case chromosomes to be more similar to other case chromosomes than to control chromosomes.

While we have concentrated here on models for biallelic loci, the ideas we have introduced could also be used to model LD among multi-allelic loci such as micro-satellites. There is a natural analogue of π_A for loci with K alleles (see also the conditional distribution for K-allele loci suggested in FD), and this could form a starting point for further investigation.

To deal with the problem that the PAC likelihood depends on the order in which the haplotypes are considered, we have chosen to average the likelihood over several random orders. One possible alternative would have been to use the pseudo-likelihood (BESAG 1974) based on our conditional distribution:

$$L_{\text{pseudo}}(\rho) = \prod_{k=1}^{n} \pi(h_k | H_{-k}), \tag{7}$$

where H_{-k} denotes the set of all haplotypes excluding h_k. The pseudo-likelihood, by definition, does not depend on the ordering of the haplotypes. This idea is more along the lines of the way that these conditional distributions are used in STEPHENS *et al.* (2001). However, in preliminary studies we found that this pseudo-likelihood performed poorly for estimating ρ. Our intuitive explanation for this is that the pseudo-likelihood in effect contains only information on the recombination that is occurring in the tips of the trees, and not on the structure of the tree as a whole. (Interestingly, under our approximation the first two haplotypes contain no information on ρ, so in some sense the information on ρ is coming from intermediate haplotypes.) Nonetheless, it is possible that the pseudo-likelihood may prove useful in settings where estimating ρ is not of direct interest.

We have introduced here two models for variation in recombination rate: a simple single-hotspot model, and a more general model that allows recombination rates to vary along the chromosome. Each of these models has weaknesses. The simple single-hotspot model makes some unrealistic assumptions: the background recombination is unlikely to be constant, and neither is the recombination rate within the hotspot; furthermore, there could be more than one hotspot. The more general model makes few assumptions, and allows more flexible investigation of patterns of recombination rate variation along a region. However, this extra flexibility comes at the expense of the introduction of extra parameters, which can result in a reduction in the precision with which parameters can be estimated. When using the model as a general model for LD, rather than for parameter estimation as we have concentrated on here, the precision of parameter estimates may be unimportant, and the few assumptions make by the more general model make it particularly attractive in this situation. When estimation of recombination rates is the main goal, the more general model may be viewed as most suited to exploratory data analysis, identifying plausible positions for hotspots, whose magnitudes might be estimated by a more parsimonious model. In this situation it might prove fruitful to consider modifying the more general model by putting a more informative prior on the λ_js. In particular a prior in which the λ_js are correlated along the chromosome (e.g. an autoregressive prior) would reduce the variance of parameter estimates, at the expense of assuming that changes in recombination rate occur more or less smoothly along the chromosome (which may or may not be the case).

In assessing our model as a method for estimating recombination rates from sequence data over moderate genomic regions, perhaps the most natural comparisons to make are with the composite likelihood methods of HUDSON (2001) and FEARNHEAD and DONNELLY (2002). (While some other methods based on summaries of the data might be competitive with these approaches when the recombination rate is assumed constant, they seem likely to suffer from loss of information when fitting models with more parameters, such as either of our models for recombination rate variation.) Of the two composite likelihood methods, although both are tractable for estimating ρ over large genomic regions, only Hudson's method is comparable with our own in terms of computational expense: our method and Hudson's method typically takes seconds, or less, to compute a likelihood, while Fearnhead and Donnelly's method can take hours per likelihood computation. Although the time and effort expended in collecting these kinds of data make it not unreasonable to wait hours or days for the results of analysis, the extra computational burden may make Fearnhead and Donnelly's method difficult to extend to more general settings involving missing genotype data, genotyping error, and/or unknown haplotypic phase for example. The approach of splitting the sequence data into contiguous segments also has the disadvantage noted earlier, of estimating recombination rates in a region only on the basis of sites within the region, and not sites either side of the region, resulting in potential loss of information. Our limited comparisons with Hudson's method suggest that it performs similarly to our method for estimating the recombination rate when it is assumed constant across the region. In principle Hudson's method could also be applied to fit models of varying recombination rate along the sequence, and the existence of more than one method to fit such models would be welcome. Both approaches seem to offer considerable advantages over other available methods for modelling LD and inferring patterns of recombination rate heterogeneity.

6 Software

The methods discussed in this paper have been implemented in a C++ package, `Hotspotter`, which is freely available for download at `http://stat.washington.edu/stephens/hotspotter.html`.

7 Acknowledgments

We thank Peter Donnelly, Paul Fearnhead, Gil McVean, Jonathan Pritchard, and Molly Przeworski for helpful conversations; and Paul Fearnhead, Richard Hudson, and Mary Kuhner for helpful advice and/or results from software. Two anonymous referees gave helpful comments on the submitted version. This work was supported by a grant from the University of Washington, and NIH Grant number 1RO1HG/LM02585-01.

Appendices

A The conditional distribution π_A

Here we give a formal description of π_A. We also provide some additional motivation for the form of this approximation, and describe briefly some of the variations on this form that we have also experimented with.

A.1 Formal description of π_A

Let $h_1,\ldots,h_n$ denote the n sampled haplotypes typed at S biallelic loci (SNPs). Typically $h_1,\ldots,h_n$ would come from a sample of n haploid individuals, or $n/2$ diploid individuals. We assume that the distribution of the first haplotype is independent of ρ (e.g. all 2^S possible haplotypes are equally likely, so $\pi_A(h_1) = 1/2^S$). Consider now the conditional distribution of h_{k+1} given $h_1,\ldots,h_k$, for $k \geq 1$. Recall (Figure 2) that h_{k+1} is an imperfect mosaic of $h_1,\ldots,h_k$. That is, for $k \geq 1$, at each SNP, h_{k+1} is a (possibly imperfect) copy of one of $h_1,\ldots,h_k$ at that position. Let X_j denote which haplotype h_{k+1} copies at site j (so $X_j \in \{1,2,\ldots,k\}$). For example, for haplotype h_{4A} in Figure 2, $(X_1,X_2,X_3,X_4,X_5) = (3,3,2,2,2)$. To mimic the effects of recombination, we model the X_j as a Markov chain on $\{1,\ldots,k\}$, with $\Pr(X_1 = x) = 1/k$ ($x \in \{1,\ldots,k\}$), and

$$\Pr(X_{j+1} = x' | X_j = x) = \begin{cases} \exp(-\rho_j d_j/k) + (1-\exp(-\rho_j d_j/k))(1/k) & \text{if } x' = x; \\ (1-\exp(-\rho_j d_j/k))(1/k) & \text{otherwise,} \end{cases} \tag{A1}$$

where d_j is the physical distance between markers j and $j+1$ (assumed known); and $\rho_j = 4Nc_j$, where N is the effective (diploid) population size, and c_j is the average rate of crossover per unit physical distance, per meiosis, between sites j and $j+1$ (so that $c_j d_j$ is the genetic distance between sites j and $j+1$). This transition matrix captures the idea that, if sites j and $j+1$ are a small genetic distance apart (i.e. $c_j d_j$ is small) then they are highly likely to "copy" the same chromosome (i.e. $X_{j+1} = X_j$). We note the following special cases used in this paper:

1. Constant recombination rate (Section 3): $c_j = \bar{c}$ for all j.
2. Simple single-hotspot model (Section 4.2): $c_j = \bar{c}$ if markers j and $j+1$ are both outside the hotspot, and $c_j = \lambda\bar{c}$ if markers j and $j+1$ are both inside the hotspot. (For brevity we omit details of the more tedious, though straightforward, case where one marker is in the hotspot, and the other outside the hotspot.)
3. General variable recombination rate model (Section 4.3): $c_j = \lambda_j\bar{c}$.

To mimic the effects of mutation the copying process may be imperfect: with probability $k/(k+\tilde{\theta})$ the copy is exact, while with probability $\tilde{\theta}/(k+\tilde{\theta})$ a "mutation" will be applied to the copied

haplotype. Specifically, if $h_{i,j}$ denotes the allele (0 or 1) at site j in haplotype i, then given the copying process $X_1,\ldots,X_S$ the alleles $h_{k+1,1}$, $h_{k+1,2}$, …, $h_{k+1,S}$ are independent, with

$$\Pr(h_{k+1,j}=a|X_j=x,h_1,\ldots,h_k)=\begin{cases}k/(k+\tilde{\theta})+(1/2)\times\tilde{\theta}/(k+\tilde{\theta}), & h_{x,j}=a\\ (1/2)\times\tilde{\theta}/(k+\tilde{\theta}), & h_{x,j}\neq a.\end{cases} \tag{A2}$$

(The factor of (1/2) appears in both cases, so that as $\tilde{\theta}\longrightarrow\infty$ both alleles become equally likely.)

We fix the value of $\tilde{\theta}$ to be

$$\tilde{\theta}=\left(\sum_{m=1}^{n-1}\frac{1}{m}\right)^{-1}, \tag{A3}$$

where n is the total number of sampled haplotypes. (See the section on motivation below for more discussion.)

A.2 Computation

Computing $\pi_A(h_{k+1}|h_1,\ldots,h_k)$ requires a sum over all possible values of the X_j, which can be done efficiently using the forward part of the Forward–Backward algorithm for hidden Markov Models (HMMs) (e.g. RABINER 1989). Specifically, let $h_{k+1,\leq j}$ denote the types of the first j sites of haplotype h_{k+1}, and let $\alpha_j(x)=\Pr(h_{k+1,\leq j},X_j=x)$. Then $\alpha_1(x)$ can be computed directly for $x=1,\ldots k$, and $\alpha_2(x),\ldots,\alpha_S(x)$ can be computed recursively using

$$\alpha_{j+1}(x)=\gamma_{j+1}(x)\sum_{x'=1}^{k}\alpha_j(x')\Pr(X_{j+1}=x|X_j=x') \tag{A4}$$

$$=\gamma_{j+1}(x)\left(p_j\alpha_j(x)+(1-p_j)\frac{1}{k}\sum_{x'=1}^{k}\alpha_j(x')\right), \tag{A5}$$

where $\gamma_{j+1}(x)=\Pr(h_{k+1,j+1}|X_{j+1}=x,h_1,\ldots,h_k)$ is given in (A2), and $p_j=\exp(-\rho_j d_j/k)$. The value of $\pi_A(h_{k+1}|h_1,\ldots,h_k)$ can then be computed using

$$\pi_A(h_{k+1}|h_1,\ldots,h_k)=\sum_{x=1}^{k}\alpha_S(x). \tag{A6}$$

The second term in the parenthesis of (A5) does not depend on x and needs to be computed only once for each j (as noted in FD). Thus the computational complexity of $\pi_A(h_{k+1}|h_1,\ldots,h_k)$ increases linearly in the number of SNPs and linearly in k. As a result the computation of $L_{\mathrm{PAC-A}}$ is linear in the number of SNPs, and quadratic in the number of chromosomes in the sample.

A.3 Motivation and variations

Although it seems intuitively sensible that the transition matrix in (A1) should have the property that the rate at which jumps occur in the copying process should increase with ρ, and decrease with the number of previous sampled chromosomes k, it is perhaps not so obvious why we chose the rate ρ/k. Indeed, the empirical results in Figure 4 suggest that this rate is not quite ideal, and Appendix B describes a corrected rate, based on these empirical results. However, we can get some idea of why ρ/n is a sensible starting point for the rate parameter, from the following informal argument. Assume that $h_1,\ldots,h_{k+1}$ are a random sample of haplotypes from a neutrally-evolving, randomly-mating, constant-sized population, and consider the unknown genealogical tree relating $h_1,\ldots,h_{k+1}$ at a single site. It follows from the Ewens sampling formula that in this tree, the probability that h_{k+1} is separated by at least one mutation from each of $h_1,\ldots,h_k$ (unconditional on the actual values of $h_1,\ldots,h_k$) is $\theta/(k+\theta)$, where $\theta = 4N\mu$, and μ is the probability of mutation per meiosis at that site. Similarly, if we consider marking on the tree recombination events that occur between this site and the next site, the probability that there will be at least one such event separating h_{k+1} from each of $h_1,\ldots,h_k$ is $\rho/(k+\rho)$, where $\rho = 4Nc$ and c is the probability of recombination between two adjacent sites per meiosis. Since ρ is small $\rho/(k+\rho) \approx \rho/k$, giving the rate that we used. (We emphasize that this is not intended to be a formal argument, and that in particular that it is unclear how our mosaic process relates formally to the genealogical tree relating the haplotypes. It is merely intended to provide additional motivation for the use of this rate, and perhaps to stimulate research into a more formal connection.)

The reason for our choice of $\tilde{\theta}$ is that $\tilde{\theta} = \sum_{m=1}^{n-1} \frac{1}{m}$ is the expected number of mutation events at a single site on the genealogical tree relating a random sample of n chromosomes, so (A3) gives *a priori* an expected number of mutation events at each site of 1 (although it does not force the number of mutations to be exactly 1, and so our method should be somewhat robust to the presence of multiple mutations at some sites).

We performed simulation experiments along the lines of those used to produce Figure 4 to see whether variations on the conditional distribution π_A described above might eliminate the bias we observed for π_A. In particular, we tried: using values for $\tilde{\theta}$ that were up to 4 times bigger or smaller than that in (A3); estimating $\tilde{\theta}$ from the data; replacing the transition probability in (A1) with a transition probability of $\rho d_i/(k+\rho d_i)$, as in FD; and making use of the more complex mutation mechanism (involving Gaussian quadrature) used in STEPHENS and DONNELLY (2000) and in FD. Although these different variations gave different quantitative results, they all produced similar qualitative patterns, and in particular the bias we observed for π_A remained in every variation that we tried. We therefore resorted to the empirical correction described in Appendix B below.

B π_B: a bias-corrected version of π_A.

To correct the bias observed in the results for $\hat{\rho}_{\mathrm{PAC-A}}$, we modified the transition matrix in (A1) by replacing ρ_j by $\delta_j\rho_j$, where $\delta_j = \exp(a + b\log_{10}\rho_j)$. The intercept a and slope b are interpolated

based on the number of haplotypes n and segregating sites S in the data (Table 1) using tensor product interpolation with natural cubic splines first in the direction of varying n, and then in the direction of varying S (UEBERHUBER 1997).

C Simulating data with a recombination hotspot.

We use the following algorithm to postprocess the output from `mksample` (HUDSON 2002) to simulate data under the simple single-hotspot model for recombination variation. Suppose we would like to simulate a sample with approximately S segregating sites. The background recombination rate is ρ. A hotspot of width $w = (b-a)$ lies between positions a and b, with recombination rate $\lambda\rho$ where $\lambda > 1$. We follow these steps:

1. Simulate samples with $S' = (1 + w(\lambda - 1))S$ segregating sites and constant recombination rate $\rho' = (1 + w(\lambda - 1))\rho$.

2. Multiply the position of each site by a factor of $1 + w(\lambda - 1)$ so that the total length of the haplotypes is $1 + w(\lambda - 1)$ instead of 1 (and the background recombination rate is ρ).

3. For sites within a and $a + w\lambda$, randomly delete them with probability $1 - 1/\lambda$.

4. For the remaining sites within a and $a + w\lambda$, shrink the distance of adjacent sites by a factor of λ.

5. Shift the positions of the sites to the right of the hotspot (subtract $w(\lambda - 1)$) so that the total length is again 1.

Shrinking the distance between sites in the hotspot produces the effect of elevated recombination rate. Deleting some sites keeps the mutation rate constant over the region.

LITERATURE CITED

BESAG, J., 1974 Spatial interaction and the statistical analysis of lattice systems. Journal of the Royal Statistical Society, series B **36:** 192–236.

CLARK, A. G., 1990 Inference of haplotypes from PCR-amplified samples of diploid populations. Molecular Biology and Evolution *7*(2)**:** 111–122.

CLARK, A. G., K. M. WEISS, D. A. NICKERSON, S. L. TAYLOR, A. BUCHANAN, J. STENGAARD, V. SALOMAA, E. VARTIANEN, M. PEROLA, E. BOERWINKLE, and C. F. SING, 1998 Haplotype structure and population genetics inferences from nucleotide-sequence variation in human lipoprotein lipase. American Journal of Human Genetics **63:** 595–612.

DALY, M. J., J. D. RIOUX, S. F. SCHAFFNER, T. J. HUDSON, and E. S. LANDER, 2001 High-resolution haplotype structure in the human genome. Nature Genetics **29:** 229—232.

EWENS, W. J., 1972 The sampling theory of selectively neutral alleles. Theoretical Population Biology **3:** 87–112.

FEARNHEAD, P. and P. DONNELLY, 2002 Approximate likelihood methods for estimating local recombination rates. Journal of the Royal Statistical Society, Series B **64:** 657–680.

FEARNHEAD, P. N. and P. DONNELLY, 2001 Estimating recombination rates from population genetic data. Genetics **159:** 1299–1318.

FRISSE, L., R. R. HUDSON, A. BARTOSZEWICZ, J. D. WALL, J. DONFACK, and A. D. RIENZO, 2001 Gene Conversion and Different Population Histories May Explain the Contrast between Polymorphism and Linkage Disequilibrium Levels. American Journal of Human Genetics **69:** 831–843.

GRIFFITHS, R. C. and P. MARJORAM, 1996 Ancestral inference from samples of DNA sequences with recombination. Journal of Computational Biology **3:** 479–502.

HAMMER, M. C., T. KARAFET, A. RASANAYAGAM, E. T. WOOD, T. K. ALTHEIDE, T. JENKINS, R. C. GRIFFITHS, A. R. TEMPLETON, and S. L. ZEGURA, 1998 Out of Africa and Back Again: Nested Cladistic Analysis of Human Y Chromosome Variation. Molecular Biology and Evolution *15*(4)**:** 427–441.

HARDING, R. M., S. M. FULLERTON, R. C. GRIFFITHS, J. BOND, M. J. COX, J. A. SCHNEIDER, D. S. MOULIN, and J. B. CLEGG, 1997 Archaic African and Asian Lineages in the Genetic Ancestry of Modern Humans. American Journal of Human Genetics **60:** 772–789.

HEY, J. and J. WAKELEY, 1997 A coalescent estimator of the population recombination rate. Genetics **145:** 833–846.

HUDSON, R. R., 1983 Properties of a neutral allele model with intragenic recombination. Theoretical Population Biology **23:** 183–201.

HUDSON, R. R., 1987 Estimating the recombination parameter of a finite population model without selection. Genetical Research. **50:** 245–250.

HUDSON, R. R., 2001 Two-locus sampling distribution and their application. Genetics **159:** 1805–1817.

HUDSON, R. R., 2002 Generating samples under a Wright–Fisher neutral model of genetic variation. Bioinformatics **18:** 337–338.

HUDSON, R. R. and N. L. KAPLAN, 1985 Statistical properties of the number of recombination events in the history of a sample of DNA sequences. Genetics **111:** 147–164.

JEFFREYS, A. J., L. KAUPPI, and R. NEUMANN, 2001 Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. Nature Genetics **29:** 217–222.

JEFFREYS, A. J. and R. NEUMANN, 2002 Reciprocal crossover asymmetry and meiotic drive in a human recombination hot spot. Nature Genetics **31:** 267–271.

JEFFREYS, A. J., A. RITCHIE, and R. NEUMANN, 2000 High resolution analysis of haplotype diversity and meiotic crossover in the human TAP2 recombination hotspot. Human Molecular Genetics **9:** 725–733.

JOHNSON, G. C. L., L. ESPOSITO, B. J. B. A. N. SMITH, J. HEWARD, G. D. GENOVA, H. UEDA, H. J. CORDELL, I. A. EAVES, F. DUDBRIDGE, R. C. J. TWELLS, F. PAYNE, W. HUGHES, S. NUTLAND, H. STEVENS, P. CARR, E. TUOMILEHTO-WOLF, J. TUOMILEHTO, S. C. L. GOUGH, D. G. CLAYTON, and J. A. TODD, 2001 Haplotype tagging for the identification of common disease genes. Nature Genetics **29:** 233–235.

KINGMAN, J. F. C., 1982 The coalescent. Stochastic Processes and their Applications **13:** 235–248.

KRUGLYAK, L., 1999 Prospects for whole-genome linkage disequilibrium mapping of common disease genes. Nature Genetics **22:** 139–144.

KUHNER, M. K., J. YAMATO, and J. FELSENSTEIN, 2000 Maximum likelihood estimation of recombination rates from population data. Genetics **156:** 1393–1401.

LIU, J. S., C. SABATTI, J. TENG, B. J. KEATS, and N. RISCH, 2001 Bayesian analysis of haplotypes for linkage disequilibrium mapping. Genome Research **11:** 1716–1724.

MCPEEK, M. S. and A. STRAHS, 1999 Assessment of linkage disequilibrium by the decay of haplotype sharing, with application to fine-scale genetic mapping. American Journal of Human Genetics **65:** 858–875.

MCVEAN, G., P. AWADALLA, and P. FEARNHEAD, 2002 A coalescent-based method for detecting and estimating recombination from gene sequences. Genetics **160:** 1231–1241.

MORRIS, A. P., J. C. WHITTAKER, and D. J. BALDING, 2000 Bayesian fine-scale mapping of disease loci, by hidden Markov models. American Journal of Human Genetics **67:** 155–169.

NICKERSON, D. A., S. L. TAYLOR, K. M. WEISS, A. G. CLARK, R. G. HUTCHINSON, J. STENGAARD, V. SALOMAA, E. VARTIAINEN, E. BOERWINKLE, and C. F. SING, 1998 DNA sequence diversity in a 9.7-kb region of the human lipoprotein lipase gene. Nature Genetics **19:** 233–240.

NIELSEN, R., 2000 Estimation of Population Parameters and Recombination rates from Single Nucleotide Polymorphisms. Genetics **154:** 931–942.

OLIVIER, M., V. I. BUSTOS, M. R. LEVY, G. A. SMICK, I. MORENO, J. M. BUSHARD, A. A. ALMENDRAS, K. SHEPPARD, D. L. ZIERTEN, A. AGGARWAL, C. S. CARLSON, B. D. FOSTER, N. VO, L. KELLY, X. LIU, and D. R. COX, 2001 Complex high-resolution linkage disequilibrium and haplotype patterns of single-nucleotide polymorphisms in 2.5 Mb of sequence on human chromosome 21. Genomics **78:** 64–72.

PRESS, W. H., S. A. TEUKOLSKY, W. T. VETTERLING, and B. P. FLANNERY, 1992 *Numerical recipes in C : the art of scientific computing*. New York: Cambridge University Press.

PRITCHARD, J. K. and M. PRZEWORSKI, 2001 Linkage disequilibrium in humans: models and data. American Journal of Human Genetics **69:** 1–14.

PRITCHARD, J. K., M. STEPHENS, and P. DONNELLY, 2000 Inference of population structure using multilocus genotype data. Genetics *155*(2)**:** 945–959.

RABINER, L. R., 1989 A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE **77:** 257–286.

STEPHENS, M. and P. DONNELLY, 2000 Inference in molecular population genetics. Journal of The Royal Statistical Society, Series B **62:** 605–655.

STEPHENS, M., N. J. SMITH, and P. DONNELLY, 2001 A new statistical method for haplotype reconstruction from population data. American Journal of Human Genetics **68:** 978–989.

TEMPLETON, A. R., A. G. CLARK, K. M. WEISS, D. A. NICKERSON, E. BOERWINKLE, and C. F. SING, 2000 Recombination and Mutation at the LPL Locus. American Journal of Human Genetics **66:** 69–83.

UEBERHUBER, C. W., 1997 *Numeric Computation: Methods, Software, and Analysis*, Volume I. Berlin Heidelberg: Springer-Verlag.

WAKELEY, J., 1997 Using the variance of pairwise differences to estimate the recombination rate. Genetical Research. **69:** 45–48.

WALL, J. D., 2000 A Comparison of Estimators of the Population Recombination Rate. Molecular Biology and Evolution **17:** 156–163.

WANG, N., J. M. AKEY, K. ZHANG, R. CHAKRABORTY, and L. JIN, 2002 Distribution of recombination crossovers and the origin of haplotype blocks: the interplay of population history, recombination, and mutation. American Journal of Human Genetics **71:** 1227–1234.