Am. J. Hum. Genet. 74:40–49, 2004

Population-Genetic Basis of Haplotype Blocks in the 5q31 Region

Eric C. Anderson and Montgomery Slatkin

Department of Integrative Biology, University of California at Berkeley, Berkeley

We investigated patterns of nucleotide variation in the 5q31 region identified by Daly et al. as containing haplotype blocks, to determine whether the blocklike pattern requires the assumption of hotspots in recombination. Using extensive simulations that generate data matched to the Daly et al. data set in (*a*) the method of ascertainment of single-nucleotide polymorphisms, (*b*) the heterozygosity of ascertained markers, (*c*) the number of block boundaries, and (*d*) the diversity of haplotypes within blocks, we show that the patterns found in the Daly et al. data are not consistent with the assumption of uniform recombination in a population of constant size but are consistent either with the presence of hotspots in a population of constant size or with the absence of hotspots if there was a period of rapid population growth. We further show that estimates of local recombination rate can distinguish between population growth and hotspots as the primary cause of a blocklike pattern. Estimates of local recombination rates for the Daly et al. data do not indicate the presence of recombination hotspots.

Introduction

Many regions in the human genome are organized into haplotype blocks of 5–100 kb, within which most variation is accounted for by two to four haplotypes (Daly et al. 2001; Jeffreys et al. 2001; Patil et al. 2001; Gabriel et al. 2002). The extent of linkage disequilibrium (LD) between sites within blocks is greater than that between nearby blocks (Stumpf and Goldstein 2003). The discovery of haplotype blocks has important implications for the prospects of mapping complex inherited diseases. In genomic regions containing blocks, relatively few SNPs within each block may characterize patterns of genetic variation sufficiently that, in association studies, many fewer SNPs will have to be typed. The recently launched HapMap project (Couzin 2002) will survey patterns of haplotype variation in the human genome.

The discovery of haplotype blocks creates a challenge for theoreticians to account for their presence. One important question is whether haplotype blocks reflect underlying heterogeneity in recombination rate. Jeffreys et al. (2001) identified several recombination hotspots in the human leukocyte antigen region. Single-sperm typing showed that the per-nucleotide rates of recombination within each hotspot are up to several hundred times larger than the genomewide average. Furthermore, Jeffreys et al. (2001) showed that boundaries of haplotype blocks in that region correspond to hotspots. These results do not necessarily imply that there is a similar correspondence in other genomic regions. Wang et al. (2002), Phillips et al. (2003), and Zhang et al. (2003) show that haplotype blocks can be generated in simulations of genetic drift in the absence of hotspots, although Phillips et al. (2003) and Zhang et al. (2003) conclude that the presence of long haplotype blocks suggests the existence of regions of unusually low recombination rate ("coldspots"). Wall and Pritchard (2003), on the other hand, found that the fraction of the Gabriel et al. (2002) data set covered by haplotype blocks is too small to be accounted for by genetic drift alone and instead requires the assumption of hotspots.

In this article, we focus attention on the haplotype-block structure of one genomic region, 5q31, studied by Daly et al. (2001). We ask whether patterns of variation within and between haplotype blocks in that region require the assumption of hotspots. We show that the diversity of haplotypes in each block is affected both by the demographic history of a population and by the presence of hotspots. We conclude first that the data of Daly et al. (2001) are not consistent with the assumption of constant population size and uniform recombination rates. Although haplotype blocks can be generated in simulations under those assumptions, the within-block levels of haplotype diversity found in those simulated data sets tend to be too low. In a population of constant size, hotspots can result in low within-block haplotype diversity, but low-diversity haplotype blocks can also be generated in the absence of hotspots if a population has undergone rapid growth. To distinguish between these two explanations, we implement Hudson's (2001) method for estimating local recombination rates.

Properties of Haplotype Blocks in the 5q31 Data Set

The Daly et al. (2001) data were among the first to document a blocklike pattern of LD in high-density SNP

Received September 2, 2003; accepted for publication October 13, 2003; electronically published December 17, 2003.

Address for correspondence and reprints: Dr. Montgomery Slatkin, Department of Integrative Biology, University of California at Berkeley, Berkeley, CA 94720-3140. E-mail: slatkin@socrates.berkeley.edu

0002-9297/2004/7401-0005$15.00

markers, and they still comprise the largest sample (in number of chromosomes typed) of a densely genotyped region in the human genome. Our analysis of the 5q31 data (Whitehead/MIT Center for Genome Research) focuses initially on the number of block boundaries and on the diversity of haplotypes within each block.

The first step is to delineate the blocks in the 5q31 chromosome region. The presence of blocks of LD in the 5q31 data would be difficult to dispute, since blocks have been detected in the data by several different methods (Wang et al. 2002; Anderson and Novembre 2003; Koivisto et al. 2003). For the analysis in the present article, we use the program MDBlocks (Anderson and Novembre 2003), which partitions a genomic region into blocks by use of the minimum-description-length (MDL) criterion. Methods that employ the MDL criterion select among models based on information theoretic criteria and can be thought of as penalized log-likelihood methods (Rissanen 1989). The statistical model underlying the algorithm in MDBlocks explicitly takes account of both haplotype diversity within blocks and association between haplotypes at adjacent blocks. We chose this method because it was shown elsewhere to outperform other methods in locating block boundaries at regions with large drops in LD, because it can be conveniently applied to large numbers of simulated data sets, and because the block boundaries it finds in the 5q31 data are similar to those discerned by the several methods used by Daly (2001) (fig. 1*a* and 1*b*).

Once blocks are delineated, we quantify within-block haplotype diversity. To estimate haplotype frequencies in the 5q31 chromosome region, we first imputed missing data—~20% of the data—according to the frequencies of SNPs within a sliding window (see appendix C in Anderson and Novembre [2003]). This method, though heuristic, has the advantages that it does not assume a blocklike structure and that it takes little time. For each of the 11 blocks found by MDBlocks, the cumulative frequency of the haplotypes (sorted in order of decreasing frequency) increases rapidly (fig. 2). For 10 of the blocks, the four most frequent haplotypes account for >75% of all the haplotypes; only in the block extending from marker 77 to 86 do the four most frequent haplotypes account for <75% of the haplotypes. For the purpose of characterizing haplotype diversity within blocks, we refer to a block in which the four most frequent haplotypes account for >75% as a "low-diversity block." Blocks with the four most common haplotypes comprising <75% of the haplotypes are termed "high-diversity blocks." In the 5q31 data, the pattern of 10 low-diversity blocks and 1 high-diversity block was observed in 200 different random imputations of the missing data. The 75% threshold value is arbitrary and was chosen for this study because it provides a convenient way to characterize the pattern of low, within-haplotype diversity noted by Daly et al. (2001). In studies in which the density of markers and ascertainment criteria are different, a different threshold value might be more appropriate.

Simulation Study

We undertook a series of coalescent simulations to discover the conditions under which the patterns observed in the 5q31 data can and cannot be obtained. Genealogies of 516 chromosomes were simulated using the pro-

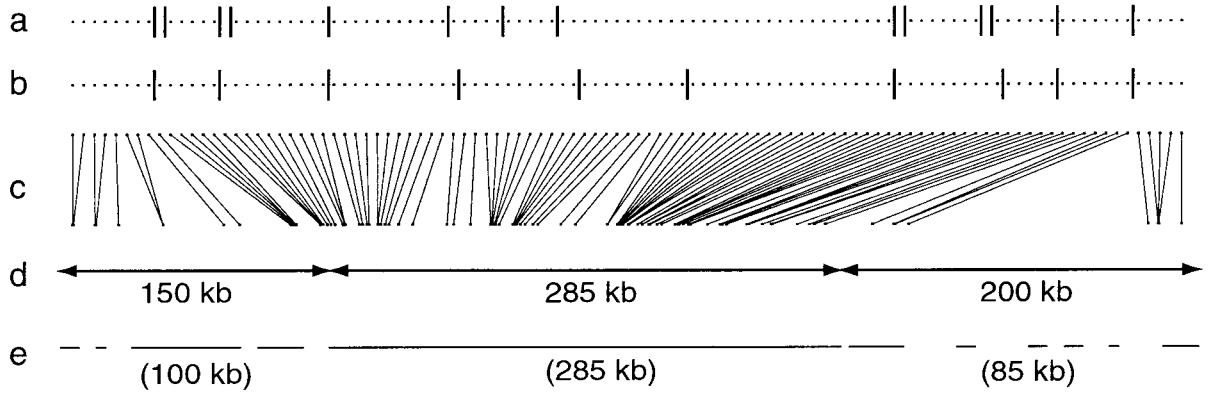

Figure 1 Blocks in 5q31. Blocks inferred by Daly et al. (2001) (*a*) and MDBlocks (*b*). Each dot represents 1 of the 103 SNPs. Vertical bars represent block boundaries. Some SNPs were not included by Daly et al. (2001) in any blocks. These are denoted by the absence of a dot between the adjacent block boundaries. The bottom row of dots (*c*) shows the physical map location of the markers in the 635-kb span of the chromosome within which Rioux et al. (2001) sequenced. The map is relative to panel d, in which arrows delimit the extent of the 285-kb "central core" region that was exhaustively resequenced in eight chromosomes by Rioux et al. (2001) and the 150- and 200-kb flanking regions that were only partially resequenced. *e,* Regions in which markers were ascertained during our simulations. The darkened portions of the discontinuous line were regions in which markers could be ascertained. This pattern approximates the resequencing tiles described in Rioux et al. (2001).

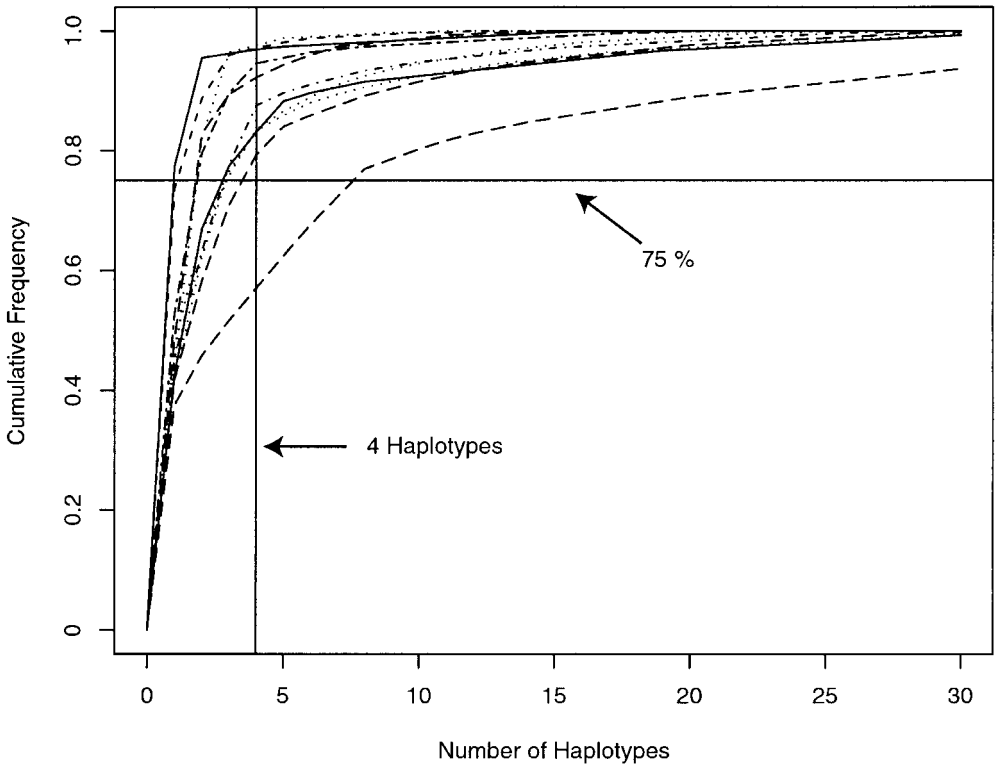

Figure 2 Haplotype diversity within blocks. Each line shows the cumulative frequency of haplotypes within one of the 11 blocks identified by MDBlocks in the 5q31 data. As is apparent, in every block except one, the four most frequent haplotypes account for >75% of the haplotypes observed.

gram "makesamples" (Hudson 2002), allowing for recombination at 500 different points in the simulated interval; "makesamples" was modified to allow for hotspots, as described by Anderson and Novembre (2003). Mutations were simulated on the genealogies by use of the infinite-sites model; that is, each SNP was assumed to arise by mutation only once. Four different scenarios were investigated: (*a*) constant population size, with uniform recombination, (*b*) constant population size, with recombination hotspots, (*c*) population growth, with no bottleneck and no hotspots, and (*d*) population growth, with a bottleneck and no hotspots. The parameter values for each scenario are given in table 1.

Simulating the 5q31 Data–Collection Process

Care was taken to design our simulations to mimic as closely as possible the 5q31 data–collection process. The simulated chromosomal segment corresponds to a 635-kb region. Mutations were ascertained for inclusion in the sample according to the procedure adopted by Rioux et al. (2001). First, only mutations falling within 470 kb of "sequenced" chromosome were included in the sample. These 470 kb were in a 285-kb contiguous central core, flanked by two incompletely sequenced regions. The positions of sequenced portions (see fig. 1*e*) follow roughly what was described in figure 1*f* of Rioux et al. (2001). Second, markers were ascertained for the sample only if they were polymorphic in 8 randomly selected chromosomes and were at a frequency >5% among all 516 chromosomes.

Approximately 13% of the data in the 5q31 data are missing because of genotyping failure, and an additional 7% are treated as missing because their haplotype phase could not be resolved. To control for possible artifacts caused by missing data, we generated missing data in our simulated samples. Twenty percent of the data in each simulated sample was randomly selected to be missing in a way that enforced a pattern of missingness similar to that observed in the 5q31 data. Specifically, our scheme for creating missing data was designed so that both the distribution of the number of missing sites per chromosome and the proportion of missing sites at a SNP, as a function of its heterozygosity, were similar to that observed in the 5q31 data. When computing haplotype diversity within simulated blocks, we used a data set with holes imputed as described above for the 5q31 data.

Matching Simulated Data to 5q31

For each scenario, we sought to adjust the parameters of the simulations so as to obtain simulated data that

Table 1

Summary of Simulations Performed

Scenario	θ[a]	ρ[b]	Hotspot Intensity[c] (%)	No. Simulated	No. of Data Sets in Selected Subset	No. of Data Sets with 10, 11, and 12 Blocks[d]	% HD[e]
CPS[f], UR[g]	63	200	...	80,000	2,000	183	.0
CPS[f], UR[g]	63	400	...	80,000	2,000	1007	.0
CPS[f], UR[g]	63	600	...	80,000	2,000	606	.0
CPS[f], 5 HS[h]	63	150	90	80,000	1,987	2	.0
CPS[f], 10 HS[h]	63	150	50	80,000	2,855	326	.6
CPS[f], 10 HS[h]	63	150	75	80,000	2,636	706	7.9
CPS[f], 10 HS[h]	63	150	90	80,000	2,307	752	34.3
CPS[f], 15 HS[h]	63	150	90	80,000	2,426	1051	8.8
PG[i], no BN[j]	62	15	...	50,000	389	269	27.1
PG[i], BN[j]	78	21	...	32,900	360	176	69.9

[a] The scaled population-mutation rate. For constant-sized populations, this is $4Nm$. For scenarios involving population growth, this was the parameter used in "makesamples" (Hudson 2001), under the assumption that time is scaled in units of the pregrowth population size.

[b] The scaled population-recombination rate, $4Nr$, for constant-sized population. Definition is as for θ under a scenario of population growth.

[c] Hotspot intensity measured in terms of the percentage of recombination events occurring at hotspots.

[d] The number of data sets from the selected subset that have 10, 11, or 12 blocks.

[e] Percentage of data sets having 10, 11, or 12 blocks that have one or fewer high-diversity blocks, as defined in text.

[f] CPS = constant population size.

[g] UR = uniform recombination (no hotspots).

[h] HS = hotspots. Number preceeding "HS" gives the number of equally spaced hotspots in the simulation.

[i] PG = population growth scenario, as described in text.

[j] BN = bottleneck, as described in text.

resembled the 5q31 data. The two parameters to be adjusted were θ, the scaled mutation rate, and ρ, the scaled population recombination rate. In all simulations, θ was adjusted so that the mean number of markers ascertained in a simulated data set was 103 ± 1. Values of θ over ranges that could plausibly give 103 ascertained SNPs produced similar results (data not shown). The recombination rate, ρ, was adjusted so that the number of blocks found in the simulated data had at least moderate probability of being close to the 11 blocks found in the 5q31 data. In general, increasing ρ increases the number of blocks found by MDBlocks. For the scenario involving a constant-sized population without hotspots, we used several values of ρ to cover the range of possible values. These values of ρ correspond well to the range of plausible values of ρ (158–474), given estimates of the per-nucleotide recombination rate in the region (Kong et al. 2002) and under the assumption of a population size of 10,000–20,000 individuals.

With each set of parameters, we simulated a large number (32,900–80,000) of data sets and then selected a subset of those, on the basis of number of markers and average marker heterozygosity, for further comparison to the 5q31 data. Data sets were selected to be in the subset if they had 101–105 markers (103 ± 2), if the average heterozygosity of the markers was .34–.36 (roughly within .01 of the value, .3517, observed in the 5q31 data), and if the frequency of markers having allele frequencies in the range (.05–.1) was <.185. Although such a screening procedure greatly restricted the number of data sets we focused on, the selected data sets were still typical of the data sets simulated. Under each scenario, the median number of markers in all simulated data sets was always close to 103, and the average heterozygosity of .3517 observed in the 5q31 data fell at least in the middle 60% of the simulated values. Thus, this selection scheme ensures that we base our comparisons on simulated data sets that are well matched to the 5q31 data yet are typical for the simulation scenarios used.

Blocks were found, using MDBlocks, in every data set of the selected subset. Subsequently, the within-block haplotype diversity of each selected data set having 10, 11, or 12 blocks was assessed, and the distribution of the number of high-diversity blocks in those simulated data sets was computed.

Simulation Results

Constant Population Size without Hotspots

We found we could generate haplotype blocks comparable in number and size to those found in the 5q31 data under the model of constant population size with no hotspots. We found it impossible, however, to reproduce the pattern of low within-block haplotype diversity. Using $\theta = 63$, we simulated 80,000 data sets for each of three different values of ρ (200, 400, and 600) spanning the range of ρ values that could plausibly give rise to 11

blocks in the data. For each value of ρ, 2,000 data sets in the selected subset were analyzed using MDBlocks. Of these 2,000, with $\rho = 200$, 183 (9%) data sets had 10, 11, or 12 blocks. For ρ of 400 and 600, the corresponding numbers were 1,007 (50%) and 606 (30%), respectively (fig. 3*a*–3*c*). None of these simulated data sets had zero or one high-diversity block. Figure 3*d*–3*f* shows the distribution of the number of high-diversity blocks found for the three values of ρ.

Constant Population Size with Recombination Hotspots

Simulations with recombination hotspots reproduced the haplotype diversity observed in the 5q31 data but only under specific assumptions about hotspot intensity and number. Simulations were performed with $\theta = 63$ and $\rho = 150$. In simulations of "strong" hotspots, parameters were adjusted so that 90% of all recombination events were expected to occur at hotspots. Under this assumption, each strong hotspot was slightly more intense than the highest intensity hotspots found by Jeffreys et al. (2001). We simulated 80,000 data sets with 5, 10, and 15 equally spaced strong hotspots. Selected subsets of 1,987, 2,307, and 2,426 data sets, respectively, were drawn from these simulations and were analyzed using MDBlocks. With five hotspots, only two (.1%) of the data sets in the selected subset had 10, 11, or 12 blocks. With 10 hotspots and 15 hotspots, 752 (33%) and 1,051 (43%) data sets had 10, 11, or 12 blocks. With 10 hotspots, 34% of the data sets matched to the 5q31 data sets had one or zero high-diversity block, and, with 15 hotspots, the proportion was 8%.

We also did simulations at $\rho = 150$ with 10 hotspots of "weak" and "intermediate" intensity, in which the expected proportion of recombinations occurring at hotspots was 50% and 75%, respectively; 80,000 data sets were simulated for each condition. For the weak hotspots, 2,855 data sets were included in the selected subset. We found that 326 of these selected data sets had 10, 11, or 12 blocks, and, of those, only two data sets had one high-diversity block; the rest had two or more. The intermediate hotspot simulations produced 2,636 data sets in the selected subset, of which 706 had 10, 11, or 12 blocks. Only 56 (8%) of those had fewer than two high-diversity blocks. The results are summarized in table 1 and figure 4.

Population Growth with and without a Bottleneck

We simulated 50,000 data sets, under a scenario of population growth. We assumed that before 5,000 gen-

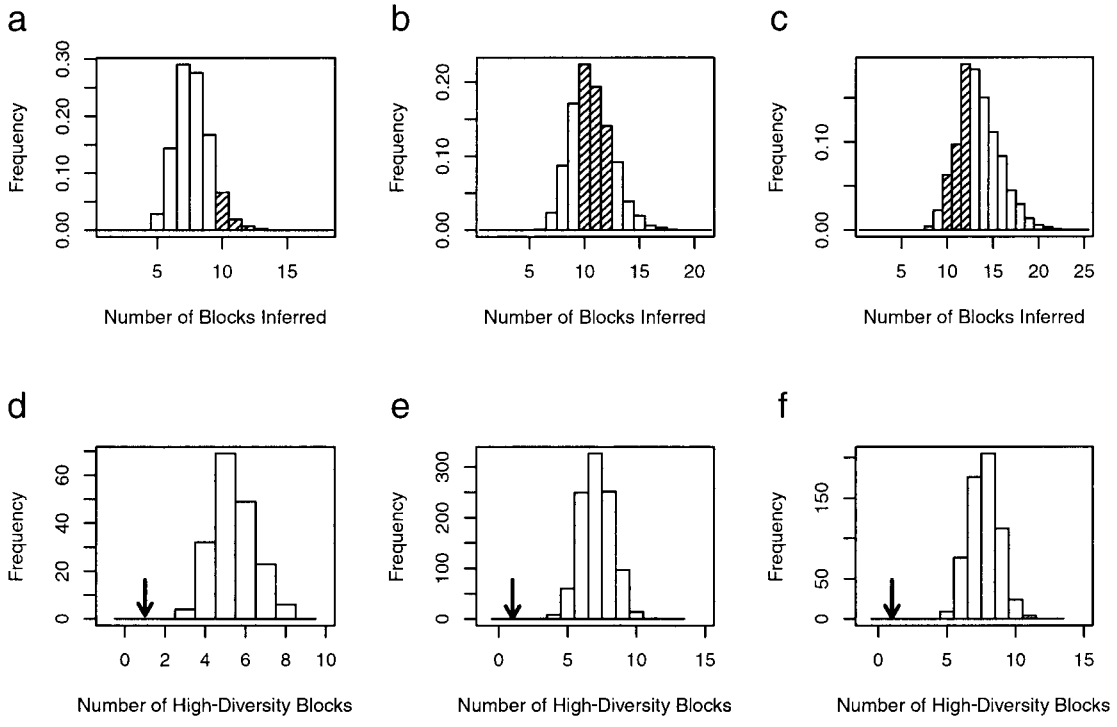

Figure 3 Results for data simulated under the neutral coalescent, with constant population size. *Top row,* The distribution of the number of blocks inferred by MDBlocks, with $\theta = 63$ and $\rho = 200$ (*a*), $\rho = 400$ (*b*), and $\rho = 600$ (*c*). Columns representing 10, 11, and 12 blocks are shaded. *Bottom row,* Number of high-diversity blocks found among 10, 11, or 12 blocks, with $\theta = 63$, $\rho = 200$ (*d*), $\rho = 400$ (*e*), and $\rho = 600$ (*f*). Dark arrows show the number of high-diversity blocks in the 5q31 data—well below the number observed in all the simulations.

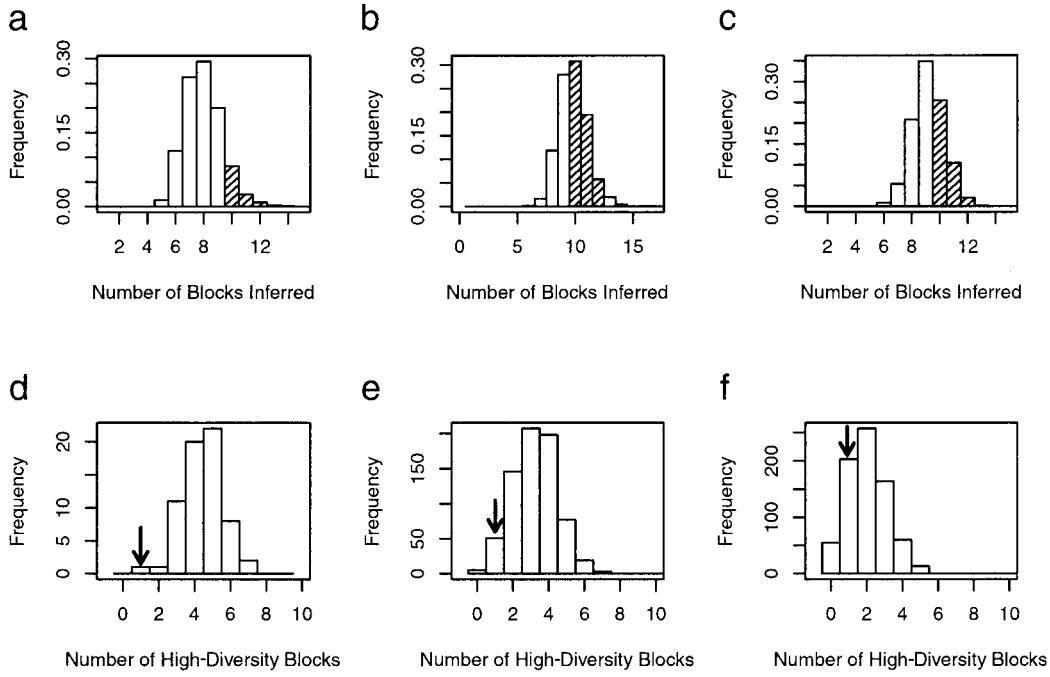

Figure 4 Results for data simulated under the neutral coalescent, with constant population size and 10 recombination hotspots. *Top row,* The distribution of the number of blocks inferred by MDBlocks, with $\rho = 150$ and weak (*a*), intermediate (*b*), and strong (*c*) hotspot intensity. *Bottom row,* Number of high-diversity blocks found among 10, 11, or 12 blocks with weak (*d*), intermediate (*e*), and strong (*f*) recombination. Dark arrows show the number of high-diversity blocks in the 5q31 data—not a likely outcome for any but the strong hotspots.

erations ago (100,000 years), the population contained 10,000 individuals. Beginning at 5,000 generations, the population grew exponentially to 10 million individuals. Using $\theta = 62$ as our scaled mutation rate resulted in an average of 102.4 ascertained markers. Of the simulated data sets, 389 data sets had the right SNP diversity to be included in the selected subset, and, of those, 269 (69%) had 10, 11, or 12 blocks. The distribution of the number of high-diversity blocks among these data sets is given in figure 5*a*. As can be seen, >27% of those data sets had one or zero high-diversity blocks.

The assumption that there was a bottleneck makes it more likely for a simulated data set to be comparable to the 5q31 data. The simulations of a bottleneck were identical to the simulations of exponential growth except that, at 102,000 years ago, the population declined instantaneously from 10,000 to 1,000 individuals and remained at that size until 100,000 years ago—when exponential growth to a present-day size of 10 million individuals began—implying a 90% reduction in population size for 100 generations. We simulated 32,900 data sets under these conditions. Because of the large variance in number of ascertained markers, only 360 data sets remained in the selected subset, and, of those, 176 had 10, 11, or 12 blocks. However, ~70% of those 176 data sets showed one or zero high-diversity blocks. The distribution of the number of high-diversity blocks is shown in figure 5*b*.

Estimating Recombination Rate across Block Boundaries

We used the composite likelihood method implemented in "maxhap" (Hudson 2001) to estimate recombination rates near block boundaries in the simulated data sets and in the 5q31 data. For every boundary separating blocks containing at least eight markers each (which we called a "testable boundary"), we estimated ρ using eight markers immediately to the left of the boundary, eight markers immediately to the right of the boundary, and eight markers straddling the boundary. A testable boundary was counted as showing evidence of a locally elevated recombination rate if $\hat{\rho}$ estimated from the eight markers straddling the boundary was at least four times as large as the average of the $\hat{\rho}$ values obtained from the markers on either side of the boundary. We refer to such testable boundaries as "potential hotspots."

The program "maxhap" was run with the default settings, allowing for gene conversion rates ≤ 10 times higher than recombination rates. Because the running time of "maxhap" increases rapidly with sample size (Hudson 2001), it was not feasible to analyze all 516

chromosomes of either the simulated or the actual data. Instead, we ran "maxhap" on subsamples of 100 chromosomes randomly selected from the 516 chromosomes in each data set. Although "maxhap" can accommodate missing data, the amount of missing data in the 5q31 data set increased the running time sufficiently to preclude the analysis of the data sets simulated with missing data. Therefore, in the 5q31 data and the simulated data sets, missing values were imputed, as described earlier. We found that results were similar for cases with no missing data, for cases with missing data handled by our imputation method, and for cases with missing data handled directly by "maxhap" (results not shown).

We performed the analysis described above on the 5q31 data and on simulated data sets that produced 10, 11, or 12 blocks with fewer than three high-diversity blocks. The results are summarized in figure 6. In 200 replicates of random imputations of the missing data in the 5q31 data, it was rare that any of the five testable boundaries were detected as potential hotspots (fig. 6 [*line c*]). Similar results were obtained from the simulations of a growing population with or without bottlenecks; the fraction of testable block boundaries inferred as potential hotspots in those data sets was most often zero and was always <50% (fig. 6 [*lines a* and *b*]). In simulations of strong hotspots in a population of constant population size, most of the testable block boundaries were detected as potential hotspots (fig. 6 [*line d*]).

We also considered hotspots of intermediate intensity. We show in figure 4*e* that intermediate hotspots resulted in a small proportion (8%) of simulated data sets containing only one high-diversity haplotype. We analyzed the 202 data sets that were simulated under the model with intermediate hotspots and that had zero, one, or two high-diversity blocks. In 91% of these data sets, ⩾20% of the testable boundaries were detected as potential hotspots. Although intermediate hotspots have only a small chance of producing low within-block haplotype diversity, many of the testable boundaries produced by those hotspots are likely to be detected as potential hotspots.

Discussion and Conclusions

Whether haplotype blocks in the human genome reflect the presence of recombination hotspots or the stochastic effects of genetic drift is important both for designing mapping studies and for understanding causes of genetic variation in humans. Hotspots in recombination exist and correspond to boundaries of haplotype blocks in some genomic regions (Jeffreys et al. 2001; Kauppi et al. 2003). However, it is far from clear that all or even most block boundaries correspond to hotspots.

In statistical terms, finding evidence for hotspots in genomic survey data can be regarded as testing a null hypothesis or a null model (no hotspots in a population of constant size). How a failure to reject the null model is interpreted depends on the power of the test. Wang et al. (2002), Phillips et al. (2003), and Zhang et al. (2003) found in their simulation studies that they could not reject the null model as an explanation for shorter haplotype blocks, although Phillips et al. (2003) and Zhang et al. (2003) concluded that very long blocks required the assumption of regions of unusually low recombination rate.

Wall and Pritchard (2003), in contrast, rejected the null model. They concurred with Wang et al. (2002), Phillips et al. (2003), and Zhang et al. (2003) that simulations of the null model can generate blocks comparable to those found in data. But they tested for an additional property of the data—the "coverage," which is the fraction of the genome in which haplotype blocks are detected—and found that simulations of the null model resulted in too little coverage. Only if they assumed a sufficient density of hotspots was the coverage comparable to that found in their analysis of the Gabriel et al. (2002) data set. Wall and Pritchard (2003) concluded that the pattern of haplotype blocks in the Gabriel et al. (2002) data set could be accounted for only if recombination hotspots were abundant in the genomic regions surveyed.

Our study is similar in spirit to that of Wall and Pritchard (2003) in that it tries to construct a more powerful test of the null model by examining additional features of the data and by finding the conditions under which those features can be reproduced in simulations. Our study differs from theirs in our focusing on a genomic region in which haplotype blocks are definitely present. Our initial simulation results of the null model are consistent with those of Wang et al. (2002), Phillips et al. (2003), Zhang et al. (2003), and Wall and Pritchard (2003). We found that the null model can readily generate haplotype blocks of sizes comparable to those found in the Daly et al. (2001) data set, provided that we model their ascertainment method. We then showed that within-block haplotype diversity allowed us to reject the null model for the Daly et al. (2001) data set. Under the null model, it does not appear possible to generate a sufficient number of low-diversity blocks. The extent of within-block haplotype diversity was consistent with either of two alternative models (strong hotspots in a population of constant size and no hotspots with recent population growth) but did not allow us to distinguish between them.

We then showed that an additional test that is based on estimates of local recombination rate does allow us to distinguish between the two alternatives; our results suggest that population growth rather than the presence of hotspots is the primary cause of the haplotype blocks found by Daly et al. (2001). Hotspots of intensity suf-

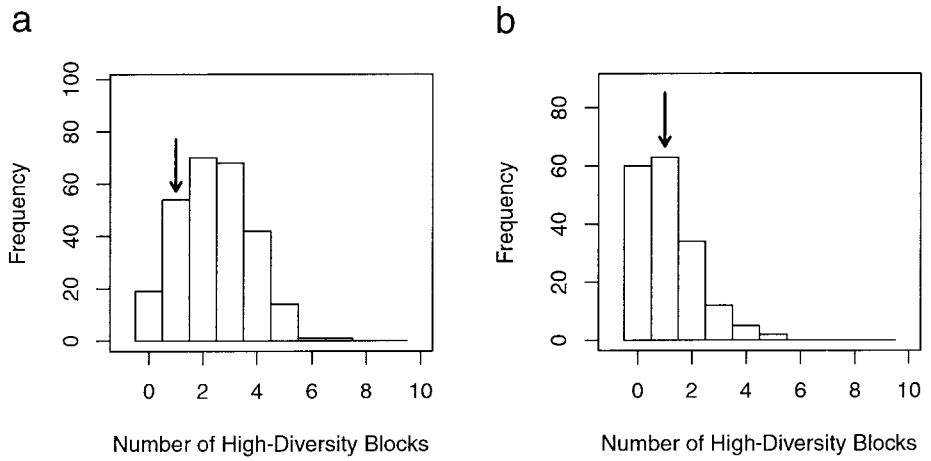

Figure 5 Number of high-diversity blocks found among 10, 11, or 12 blocks, in data simulated under the neutral coalescent with population growth and bottlenecks. *a,* Scenario of growth with no bottleneck: population is of constant size, 10,000 individuals until 100,000 years ago, at which point exponential growth commences. Present population size is 10 million individuals. $\theta = 62; \rho = 15$. Dark arrows show the number of high-diversity blocks in the 5q31 data. *b,* Scenario with a bottleneck: simulation parameters correspond to a population that was of constant size, 10,000 individuals, until 102,000 years ago, at which point there was a bottleneck of 1,000 individuals lasting for 2,000 years. After the bottleneck, the population size increases to 10 million individuals in 100,000 years. $\theta = 78$, $\rho = 21$.

ficient to create haplotype blocks in a population of constant size should be evident in estimates of local recombination rate, but they are not evident in the 5q31 region. Our results do not rule out the presence of recombination hotspots, but they suggest that hotspots, if they are present at all, are not present at most haplotype block boundaries.

Our conclusion is not necessarily incompatible with that of Wall and Pritchard (2003). We examined a genomic region that is completely covered by haplotype blocks. It is an open question as to whether recent population growth would result in levels of coverage comparable to that found in much larger genomic regions containing no hotspots. But if population growth is important for the generation of blocks in one region, it must be allowed for in regions containing hotspots as well, and its consequences must be investigated.

Our results provide some insight into how genetic drift can create low-diversity haplotype blocks in the absence of hotspots. The existence of distinct blocks containing SNPs in perfect or nearly perfect LD implies that little recombination has occurred among chromosomes carrying the different haplotypes. The existence of only a few haplotypes that are distinguished by several SNPs implies that those haplotypes represent the descendents of distinct nonrecombining lineages that are quite old. Taken together, these two characteristics suggest a genealogy of each haplotype block (as illustrated in fig. 7), which can be described as "a palmetto genealogy," because it is similar in form to some (but by no means all) real palmettos.

Recent population growth can result in a palmetto genealogy. The terminal branches attached to each internal branch represent a starlike genealogy of the kind created by rapid growth (Slatkin and Hudson 1991). Furthermore, rapid growth would result in terminal

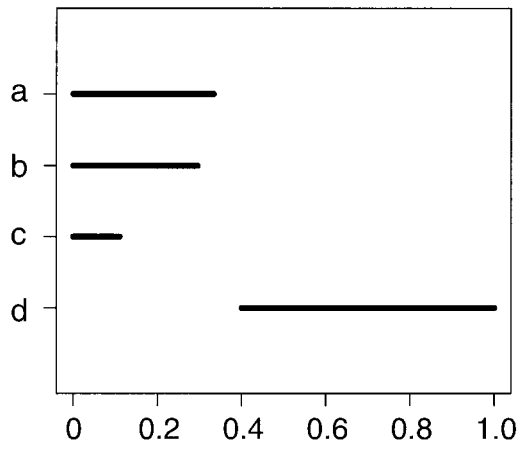

Figure 6 Proportion of testable block boundaries identified as potential hotspots in the Daly et al. (2001) data and under the three simulation models described in the text. *a,* Data simulated under scenario of population growth with no bottleneck. *b,* Data simulated under scenario of population growth with a bottleneck. *c,* Results of 200 imputations of the missing data in the 5q31 data. *d,* Data simulated under the model of constant population size with 10 strong hotspots. Lines a–c extend from the 0th percentile to the 95th percentile of the proportion of testable block boundaries identified as potential hotspots. Line d extends from the 5th percentile to the 100th percentile of the proportion of testable block boundaries identified as potential hotspots.

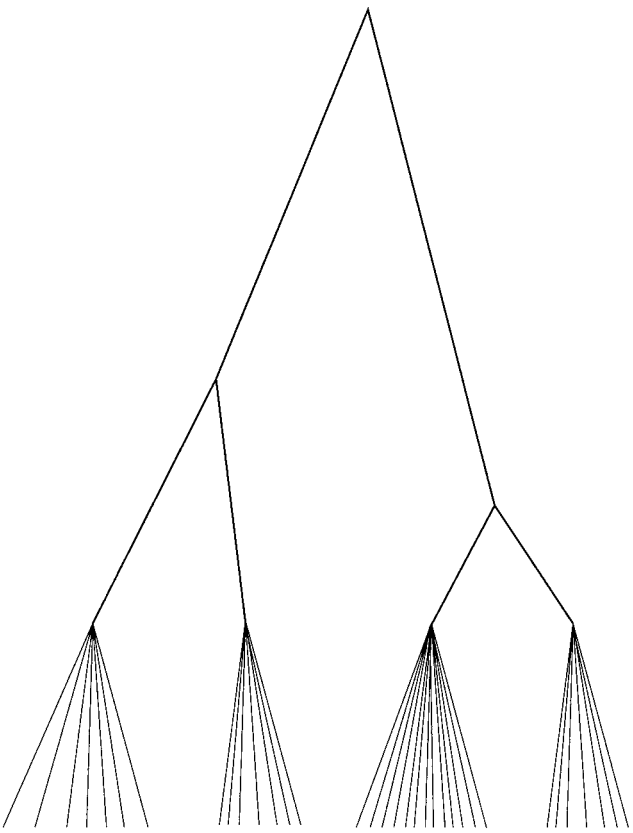

Figure 7 A palmetto genealogy representing the ancestry of a single haplotype block. The terminal branches connected to each internal branch represent lineages arising after the onset of population growth. The internal branches represent very old lineages on which mutations accumulated that distinguish different haplotypes within a block. Recombination events among the terminal branches would not affect the overall block structure but would create chimeric haplotypes of the kind found in the Daly et al. (2001) data set.

branches that are roughly the same length, because that length represents the time since the onset of rapid population growth. Adjacent haplotype blocks would differ somewhat because of recombination events among internal branches. Our characterization, in terms of a palmetto genealogy, is consistent with previous descriptions of haplotype blocks representing ancient or early recombination events (Daly et al. 2001; Goldstein 2001; Patil et al. 2001; Phillips et al. 2003). By representing the genealogy as we do in figure 7, we are emphasizing that the terminal branches have diverged recently from each long internal branch.

Demographic histories other than the two we simulated may also lead to a palmetto genealogy and haplotype blocks of the type analyzed here, although we did not explore that possibility. Also, natural selection acting in or near the 5q31 chromosome region may have helped create haplotype blocks.

Recent population growth does not necessarily lead to strong LD. Slatkin (1994) showed that there would be less LD in a rapidly growing population than in a population of constant size. The difference between those simulations and the ones described here is that, here, differences among haplotypes are established before population growth commences, whereas, in Slatkin's (1994) simulations, all variation was assumed to have arisen by mutation after rapid growth began.

If the blocklike pattern of the 5q31 data of Daly et al. (2001) is not caused by hotspots and instead arises because the underlying genealogy can be represented by a palmetto genealogy, then we would expect to find evidence of recombination among the terminal branches. Two kinds of evidence of this recombination could be obtained. First, recombination among terminal branches that descended from different internal branches should create rare chimeric haplotypes made up of abutting parts of two common haplotypes. Processes that affect single SNPs—gene conversion, recurrent mutation, and genotyping error—would not create chimeric haplotypes. We found several such chimeric haplotypes (results not shown) in the Daly et al. (2001) data set. Second, recombination among terminal branches should induce a decrease in LD with increased distance between pairs of sites within blocks. The number of long blocks in the 5q31 data is too small for us to detect a significant decrease in either D' or r^2 with the numbers of nucleotides separating pairs of markers. In the much larger data set of Gabriel et al. (2002), however, their figure 2*c* shows that both D' and r^2 decrease with increased distance within blocks. Phillips et al. (2003) found similar patterns in their data.

We conclude that recent population growth can result in haplotype blocks similar to those found in the 5q31 region examined by Daly et al. (2001) and, furthermore, that estimates of local recombination rates in this region do not support the presence of recombination hotspots. Although we have focused on the 5q31 region, other studies have emphasized the low within-block haplotype diversity. In fact, it is the low within-block diversity that makes haplotype blocks so important for mapping studies. Demographic history, as well as heterogeneity in local recombination rates, must be accounted for in attempts to understand the origin of haplotype blocks.

Acknowledgments

This research has been supported, in part, by National Institutes of Health grant GM40282 (to M.S.). We thank H. Chen, J. Felsenstein, J. Novembre, and M. P. H. Stumpf for helpful discussions of the topic of this paper.

Electronic-Database Information

URLs for data presented herein are as follows:

makesamples, http://home.uchicago.edu/~rhudson1/source/mksamples.html (for the program for generating samples under neutral models)

maxhap, http://home.uchicago.edu/~rhudson1/source/maxhap.html (for the program for estimating ρ ($4Nr$), a crossing-over parameter, and $f(= g/r)$, a gene conversion parameter, through use of a maximum composite likelihood method)

MDBlocks, http://ib.berkeley.edu/labs/slatkin/eriq/software/mdb_web/index.htm (for the haplotype block–partitioning program)

Whitehead/MIT Center for Genome Research, http://www-genome.wi.mit.edu/humgen/IBD5/index.html (for 5q31 data)

References

Anderson EC, Novembre J (2003) Finding haplotype block boundaries by using the minimum-description-length principle. Am J Hum Genet 73:336–354

Couzin J (2002) Human genome: HapMap launched with pledges of $100 million. Science 298:941–942

Daly MJ, Rioux JD, Schaffner SF, Hudson TJ, Lander ES (2001) High-resolution haplotype structure in the human genome. Nat Genet 29:229–232

Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M, Liu-Cordero SN, Rotimi C, Adeyemo A, Cooper R, Ward R, Lander ES, Daly MJ, Altshuler D (2002) The structure of haplotype blocks in the human genome. Science 296:2225–2229

Goldstein DB (2001) Islands of linkage disequilibrium. Nat Genet 29:109–111

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

——— (2001) Two-locus sampling distributions and their application. Genetics 159:1805–1817

Jeffreys AJ, Kauppi L, Neumann R (2001) Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. Nat Genet 29:217–222

Kauppi L, Sajantila A, Jeffreys AJ (2003) Recombination hotspots rather than population history dominate linkage disequilibrium in the MHC class I region. Hum Mol Genet 12: 33–40

Koivisto M, Perola M, Varilo T, Hennah W, Ekelund J, Lukk M, Peltonen L, Ukkonen E, Mannila H (2003) An MDL method for finding haplotype blocks and for estimating the strength of haplotype block boundaries. Pac Symp Biocomput 8:502–513

Kong A, Gudbjartsson DF, Sainz J, Jonsdottir GM, Gudjonsson SA, Richardsson B, Sigurdardottir S, Barnard J, Hallbeck B, Masson G, Shlien A, Palsson ST, Frigge ML, Thorgeirsson TE, Gulcher JR, Stefansson K (2002) A high-resolution recombination map of the human genome. Nat Genet 31:241–247

Patil N, Berno AJ, Hinds DA, Barrett WA, Doshi JM, Hacker CR, Kautzer CR, Lee DH, Marjoribanks C, McDonough DP, Nguyen BTN, Norris MC, Sheehan JB, Shen N, Stern D, Stokowski RP, Thomas DJ, Trulson MO, Vyas KR, Frazer KA, Fodor SPA, Cox DR (2001) Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. Science 294:1719–1723

Phillips MS, Lawrence R, Sachidanandam R, Morris AP, Balding DJ, Donaldson MA, Studebaker JF, et al (2003) Chromosome-wide distribution of haplotype blocks and the role of recombination hot spots. Nat Genet 33:382–387

Rioux JD, Daly MJ, Silverberg MS, Lindblad K, Steinhart H, Cohen Z, Delmonte T, et al (2001) Genetic variation in the 5q31 cytokine gene cluster confers susceptibility to Crohn disease. Nat Genet 29:223–228

Rissanen J (1989) Stochastic complexity in statistical inquiry. Vol 15. World Scientific, London

Slatkin M (1994) Linkage disequilibrium in growing and stable populations. Genetics 137:331–336

Slatkin M, Hudson RR (1991) Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. Genetics 129:555–562

Stumpf MPH, Goldstein DB (2003) Demography, recombination hotspot intensity, and the block structure of linkage disequilibrium. Curr Biol 13:1–8

Wall JD, Pritchard JK (2003) Assessing the performance of the haplotype block model of linkage disequilibrium. Am J Hum Genet 73:502–515

Wang N, Akey JM, Zhang K, Chakraborty R, Jin L (2002) Distribution of recombination crossovers and the origin of haplotype blocks: the interplay of population history, recombination, and mutation. Am J Hum Genet 71:1227–1234

Zhang K, Akey JM, Wang N, Xiong M, Chakraborty R, Jin L (2003) Randomly distributed crossovers may generate block-like patterns of linkage disequilibrium: an act of genetic drift. Hum Genet 113:51–59