Chromosome-wide distribution of haplotype blocks and the role of recombination hot spots

M.S. Phillips[1], R. Lawrence[2], R. Sachidanandam[3], A.P. Morris[2], D.J. Balding[4], M.A. Donaldson[1], J.F. Studebaker[1], W.M. Ankener[1], S.V. Alfisi[1], F.-S. Kuo[1], A.L. Camisa[1], V. Pazorov[1], K.E. Scott[1], B.J. Carey[1], J. Faith[3], G. Katari[3], H.A. Bhatti[1], J.M. Cyr[1], V. Derohannessian[1], C. Elosua[1], A.M. Forman[1], N.M. Grecco[1], C.R. Hock[1], J.M. Kuebler[1], J.A. Lathrop[1], M.A. Mockler[1], E.P. Nachtman[1], S.L. Restine[1], S.A. Varde[1], M.J. Hozza[1], C.A. Gelfand[1], J. Broxholme[2], G.R. Abecasis[5], M.T. Boyce-Jacino[1] & L.R. Cardon[2]

Published online 18 February 2003; doi:10.1038/ng1100

Recent studies of human populations suggest that the genome consists of chromosome segments that are ancestrally conserved ('haplotype blocks'; refs. 1–3) and have discrete boundaries defined by recombination hot spots[4,5]. Using publicly available genetic markers[6], we have constructed a first-generation haplotype map of chromosome 19. As expected for this marker density[7], approximately one-third of the chromosome is encompassed within haplotype blocks. Evolutionary modeling of the data indicates that recombination hot spots are not required to explain most of the observed blocks, providing that marker ascertainment and the observed marker spacing are considered. In contrast, several long blocks are inconsistent with our evolutionary models, and different mechanisms could explain their origins.

The ability to identify genomic regions that are shared within and between human populations holds promise for the detection of predictors of common multifactorial disease and for the advancement of personalized medicines[8]. Accordingly, several studies have recently been carried out to catalog conserved regions[1,2,5,7,9], leading to the initiation of a 'Haplotype Map' follow-on to the human genome sequencing project[10]. By evaluating population frequencies of specific haplotypes, or individual chromosomes containing polymorphic sites that are transmitted from generation to generation, the project aims to identify segments that have remained intact over time[2]. It is anticipated, although it is unproven and hotly debated, that characterization of these haplotype segments will help in the discovery of the etiological basis of common disease[11,12].

Linkage disequilibrium (LD) refers to the non-random assortment of genetic variants in a population, as with polymorphic sites in haplotypes that are ancestrally conserved. Among recent studies of LD at various locations throughout the genome (reviewed in ref. 13), some of the most striking patterns emerged from direct assessments of site-specific recombination and showed a series of discrete tracts of low recombination bounded by recombination hot spots[5,14]. The rest of the genome may also follow this heterogeneous pattern of successive recombination hot and cold spots[1–4,9]. Regions of low recombination have been labeled haplotype blocks[1,2].

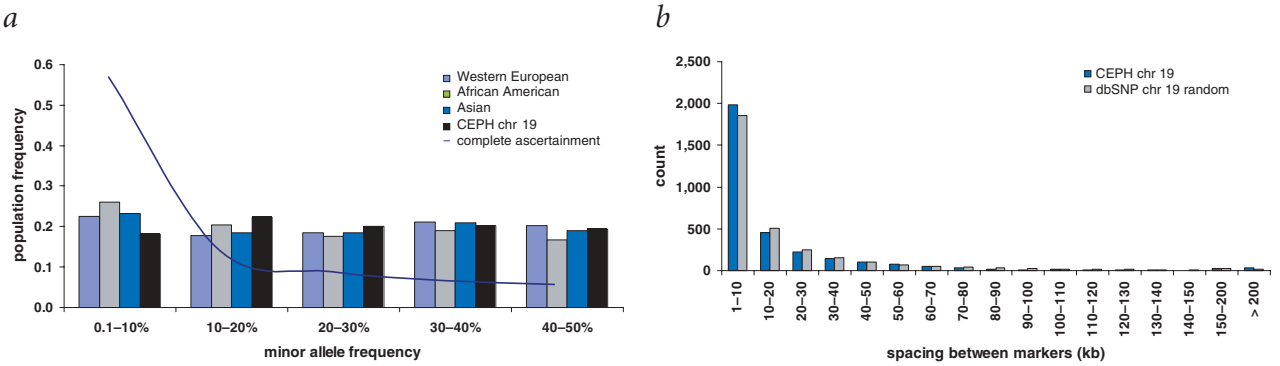

Fig. 1 Marker allele frequencies and physical spacing on chromosome 19. ***a***, The distributions of minor allele frequencies for the 3,297 CEPH markers in this study, compared with those of >15,000 randomly ascertained genomic markers that were genotyped on non-familial samples of different ancestry by The SNP Consortium. The uniform distribution of allele frequencies reflects the bias towards common alleles in public databases[16,30]. Had more samples been initially assessed for detection, the distribution would be expected to resemble that in the superimposed line (complete ascertainment), as shown empirically on chromosome 21 (ref. 9). This bias towards common alleles are inherent to all studies that use the current sample of publicly available SNPs. ***b***, The distribution of physical gaps between our final set of markers (mean gap, 17.65 kb; median gap, 5.5 kb), as compared with the mean spacing of 3,297 markers that were randomly selected from all chromosome 19 SNPs. The observed marker distribution is compared with the random distribution on chromosome 19, as summarized on the basis of 10,000 random draws of 3,297 marker positions from the complete set of 36,240 public domain markers (dbSNP database). The marker distribution and allele frequency profile of chromosome 19 SNPs seem to be generally consistent with the rest of the human genome.

[1]*Orchid Biosciences Inc., 303A College Road East, Princeton, New Jersey 08540, USA.* [2]*Wellcome Trust Centre for Human Genetics, University of Oxford, Roosevelt Drive, Oxford OX3 7BN, UK.* [3]*Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, USA.* [4]*Department of Epidemiology and Public Health, Imperial College Faculty of Medicine, Norfolk Place, London W2 1PG, UK.* [5]*University of Michigan, Ann Arbor, Michigan 48109, USA. Correspondence should be addressed to L.C. (e-mail: lon.cardon@well.ox.ac.uk).*

If any block is flanked by sites of recurrent recombination or if the population haplotypes reflect multiple copies of only a few ancestral recombination events, then the high LD within it will render many variant sites redundant and will help reduce genotyping requirements for large-scale association studies[15]. It is also conceivable, however, that block patterns arise from stochastic recombination and other forces of genetic variability[16], in which case the block boundaries may not be consistently demarcated within or between populations. In such cases, apparent blocks may depend on the specific history of the DNA samples in which they were detected and the markers used to define them. The utility of haplotype blocks for association studies then becomes less clear.

To examine block patterns, we constructed a first-generation haplotype map of chromosome 19, a relatively short human chromosome that was selected for its high gene density. (Coding bases make up > 5% of the chromosome, the highest in the human genome.) We genotyped 9,550 of

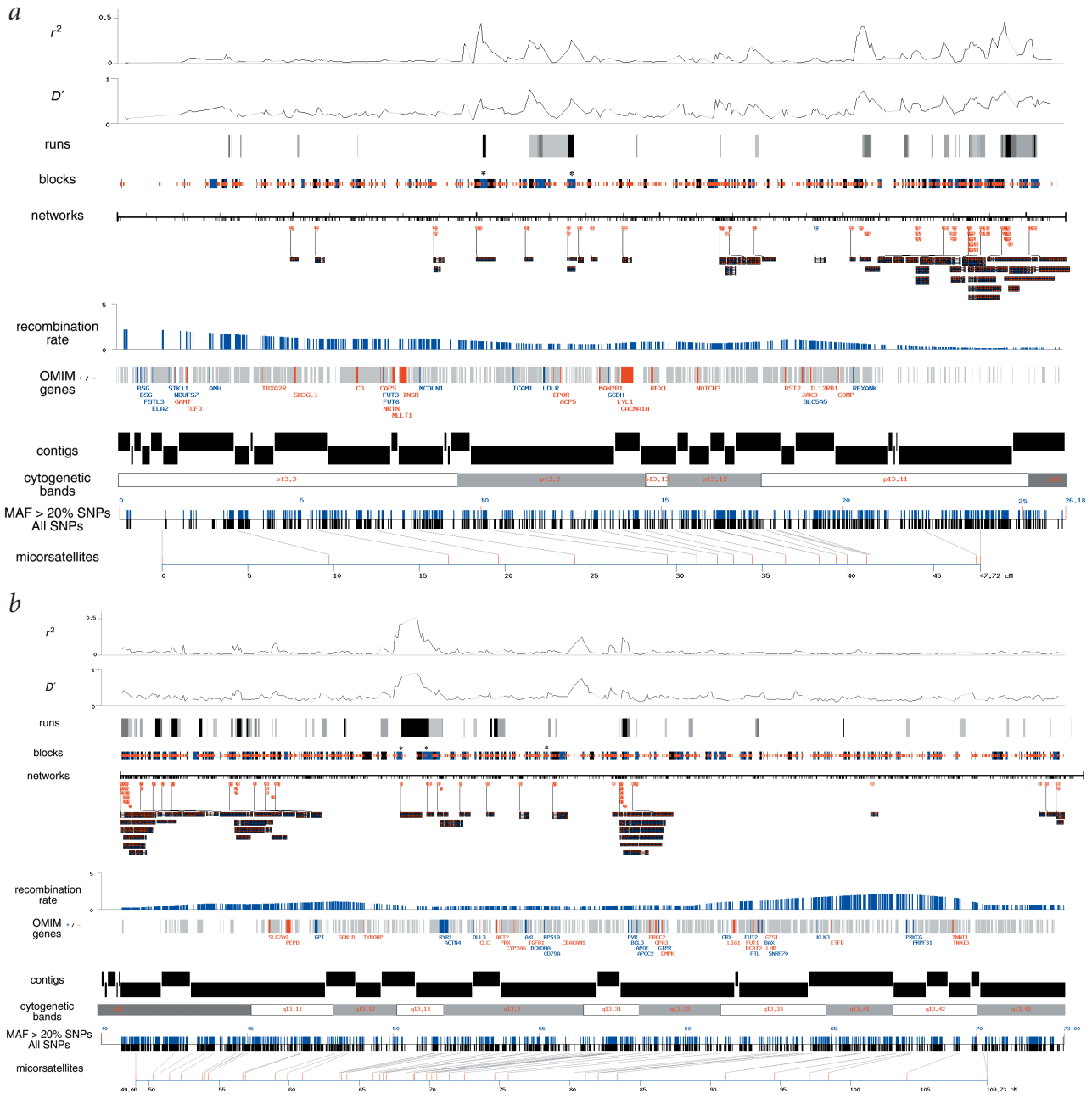

Fig. 2 Patterns of LD on chromosome 19. ***a***, The short (p) arm of the chromosome. ***b***, The long (q) arm of the chromosome. In each panel, the top 2 sections present sliding-window plots of r^2 and $|D'|$ coefficients for common allele markers (overlap of 500 kb; markers with minor allele frequency ≥ 0.20). The range of r^2 and $|D'|$ values is (0, 0.5) and (0, 1.0), respectively. The statistical significance of runs of excess LD[7] are shown below the sliding windows (gray bars, $P < 10^{-4}$; darker bars, greater significance). Stringent blocks used for evolutionary comparisons are shown, with each successive block offset in color to illustrate demarcation points. Asterisks indicate locations of the long blocks listed in Table 1. Beneath the stringent blocks are 'haplotype networks', another form of block with more permissive boundaries that allow non-contiguous markers to occupy the same block[7]. Haplotype networks of length 75 kb or greater are shown. Values above the networks indicate physical position along the chromosome (Mb). Recombination rates for each SNP (maximum 5 cM Mb^{-1}) are shown as histograms. A graphic depiction of all genes listed in Online Mendelian Inheritance in Man (OMIM) is provided. The orientation of each gene is indicated as +/–. Finally, the chromosome 19 sequence contigs for NCBI release build 29 are shown (successively offset to show breakpoints), followed by the chromosome 19 cytogenetic map.

roughly 36,000 publicly available single-nucleotide polymorphisms (SNPs) on Centre d'Etude du Polymorphisme Humain (CEPH) reference individuals[17]. Among genotyped SNPs, 3,297 were polymorphic, mapped unambiguously to the chromosome and met a series of stringent quality control procedures. The spacing distribution of these markers follows that of all available chromosome 19 markers, with many closely spaced markers and several long gaps (Fig. 1).

There are several regions of excessively high LD on this chromosome (Fig. 2), particularly near the centromere, which is consistent with reduced centromeric recombination[18]. There are additional regions at p13.2 and q13.12–13 in which the mean LD is much higher than background. Statistical assessment of LD runs[7] indicates that the tracts with high LD were significantly larger than the expected size for this chromosome ($P < 10^{-4}$; Fig. 2). As expected, the regions of highest LD were located in areas of low recombination, supporting earlier indications that rates from existing genetic maps are of immediate benefit for prediction of broad-scale LD[7].

Haplotype blocks make up 32% (17.8 of 56.5 Mb) of the finished sequence (Fig. 2). To examine the dependence of location and length on the operational definition of block boundaries, we used two definitions to derive the blocks. The pattern of blocks is generally similar using both approaches. Blocks were dispersed throughout both chromosome arms and were clustered in the regions of high LD around the centromere. These patterns are unlikely to result from study-specific marker characteristics, as the uniform allele frequency distribution of our markers closely followed that of dbSNP markers throughout the genome (Fig. 1*a*) and the marker spacing distribution closely resembled that of all chromosome 19 markers (Fig. 1*b*).

To explore the influence of localized recombination patterns and SNP selection on haplotype blocks, we compared the overall distribution of blocks with that expected under uniform recombination. We modeled different features of our data under the assumption of a coalescent process with recombination for haplotype evolution[19,20]. We first simulated the chromosome under assumptions of a constant recombination rate (using the chromosome 19 median, 1.63 cM Mb^{-1}), fixed marker spacing (median 1 SNP per 5.5 kb) and allele frequency profile (minor allele frequency >1%). This model provided a poor fit to the data, predicting too few short blocks and too many blocks of moderate length (Fig. 3*a*). We then attempted to improve the fit by more precisely modeling the features of our markers. First, we incorporated a model of ascertainment that approximately replicated the distribution of observed allele frequencies. This actually worsened the fit (Fig. 3*b*). Second, we relaxed the assumption of fixed marker spacing and modeled the observed distribution of gaps instead. This also provided a poor fit to the data (Fig. 3*c*). Joint consideration of empirical marker spacing and uniform allele frequencies, however, provided a good approximation to the observed block distribution (Fig. 3*d*). No recombination hot spots, population bottlenecks or selective forces were required to explain the observed distribution of blocks. Analyses incorporating the observed distribution of chromosome 19 recombination rates[21] produced similar results (data not shown).

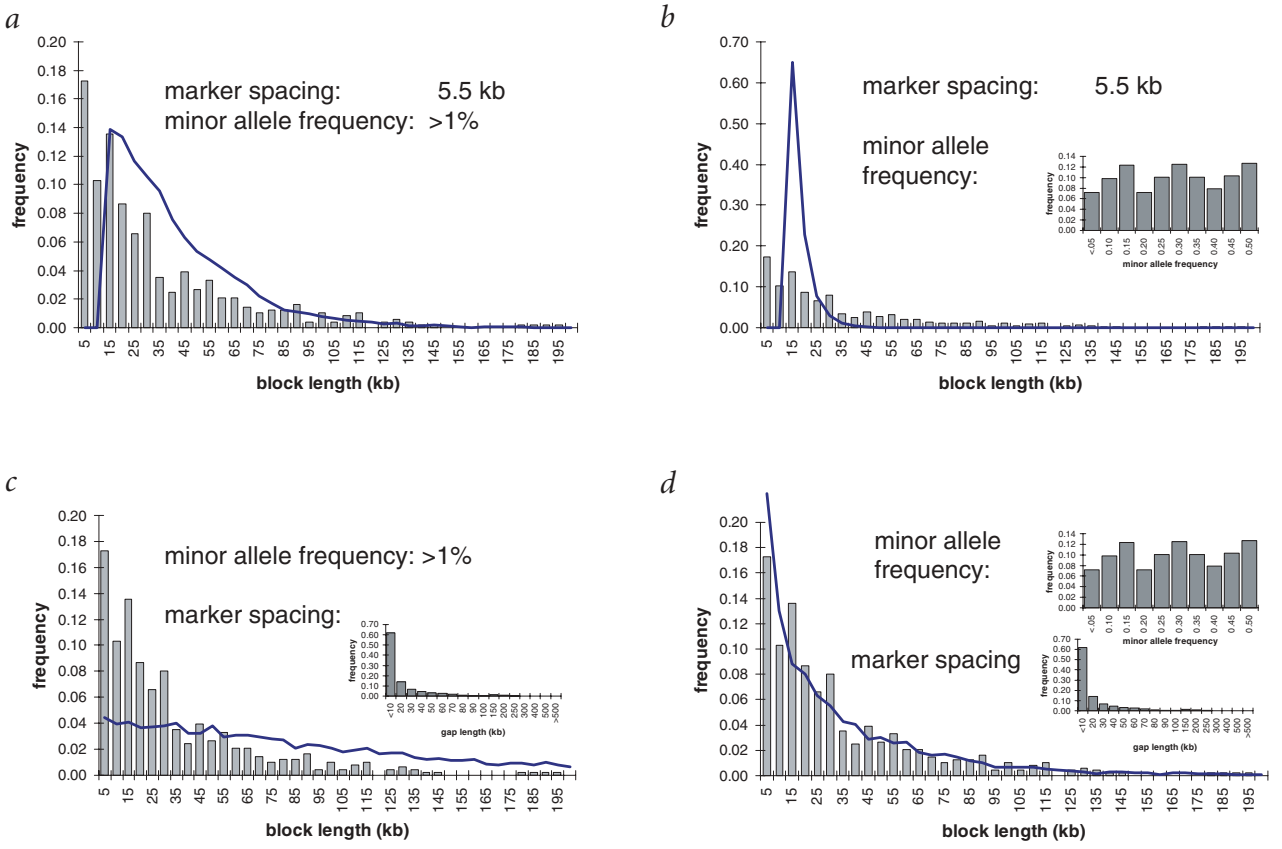

Fig. 3 Comparisons of observed block distributions (bars) and evolutionary model expectations (lines) under uniform recombination. ***a***, Comparisons made under the standard coalescent model of random SNP ascertainment, with a marker spacing of 1 marker per 5.5 kb (the median of this study) and uniform recombination of 1.63 cM Mb^{-1} (chromosome 19 median). ***b***, Modeling the minor allele frequencies according to the empirical uniform distribution, with the spacing and recombination as in (*a*). ***c***, Modeling the observed marker spacing distribution by coalescent SNP selection on the basis of that of the 3,297 markers studied, with recombination and allele frequencies as in (*a*). ***d***, Jointly modeling the allele frequencies and marker spacing under uniform recombination. Modeling uniform recombination rates other than the observed 1.63 cM Mb^{-1} did not have any detectable influence on the findings.

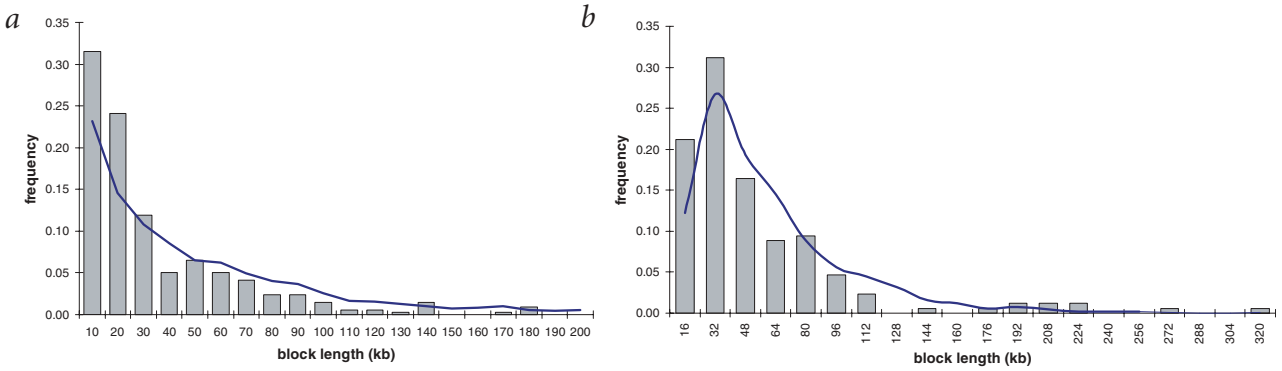

Fig. 4 Distribution of block lengths in different marker sets. ***a***, The observed (bars) and predicted (line) distributions of all chromosome 19 markers that had minor allele frequencies ≥ 0.10 (2,694 markers). ***b***, Observed (bars) and predicted (line) results from an independent study of chromosome 22 markers[7].

Coalescent modeling of the subset of markers having common alleles also yielded a close approximation to the empirical distribution (Fig. 4*a*). In addition, re-analysis of the haplotype map data for chromosome 22 (ref. 7), which came from a different laboratory and a different marker ascertainment scheme, yielded a similarly close approximation under uniform recombination and selective neutrality (Fig. 4*b*). The dependence of block-length predictions on marker spacing and allele frequencies indicates that apparent blocks can result from incomplete coverage of the chromosome genealogy. Genotyping additional markers may uncover additional ancestral recombination events, breaking up larger blocks and refining their boundaries.

Although block patterns are strongly correlated with the characteristics of the markers used to define them, density and frequencies do not predict the tail of the distribution. Several of the longest blocks emerging from our marker panel were under-represented in our simulations (Table 1). Although the true length of these blocks may be less than that observed (Figs. 1*b* and 3*c*), the long LD patterns previously noted on chromosome 22 (ref. 7) are also in the tail of the uniform recombination distribution. These large regions may reflect natural selection, recombination cold spots or characteristics of population demography[13].

Apart from the unusually long blocks, it is possible that our marker spacing is too sparse for elucidation of small hot spots and that some of the 'stochastic blocks' are interspersed with blocks flanked by genuine recombination hot spots. Presumably, at some higher marker density, genotyping further markers will not provide additional information about the local genealogy and will have little effect on block boundaries. To explore this possibility, we used the joint allele frequency and marker density features of our best-fitting model (Fig. 3*d*) and extrapolated the expected block lengths at different marker densities (Table 2). Long blocks are expected with sparse maps, whereas shorter blocks emerge from dense maps that are genotyped on the same data. These results indicate that a high marker density for genotyping, or an alternative strategy such as direct measurement of recombination events, may be required to distinguish blocks arising by means of recombination hot spots from those appearing by other forces, including chance. Direct comparison of these simulations with empirical studies is difficult because our predictions depend on the effective size and demographic history of the study population. Nevertheless, the results are not inconsistent with previous studies in several different populations, except when the density approaches about 1 marker per kb (Table 2).

The present study indicates that haplotype blocks can arise by several different mechanisms, including (i) heterogeneous recombination, which separates strongly conserved DNA segments[5,22]; (ii) natural selection, in the form of selective sweeps or background selection, which can create long-range LD[23]; (iii) population bottlenecks, which generate extended regions of LD[24], a principle that is widely used for mapping genes associated with a disease in population isolates; (iv) mating between populations with different allele frequencies (population admixture)[25], which can yield LD excesses that masquerade as ancestrally conserved segments; and (v) marker spacing and allele frequencies, which can result in incomplete coverage of the genealogy, as indicated here.

The origin of each individual block may be important for association studies. Because haplotype blocks can arise from several causes, simply identifying them does not ensure either their conservation within or between populations or their utility for mapping genes associated with a disease. Efficiencies for association mapping may be obtained from blocks that are truly flanked by recombination hot spots or by blocks originating from genetic drift, if the relevant recombination events occurred before the expansion of the human population. (That is, blocks that are induced by drift may be informative, but only for the oldest SNPs.) Other types of blocks will require closer scrutiny to be useful. Apparent blocks that arose from population bottlenecks, selective sweeps or recent mutations with drift may be specific to the population studied, and blocks that emerged from stochastic recombination and mutation events offer no clear advantages without further genotyping.

Table 1 • Unusual haplotype blocks on chromosome 19

Physical position (Mb)[a]	Genetic position (band)[b]	Length (kb)	Recombination rate[b]	Markers (*n*)	Probability[c]
51.027	62 (q13.13)	337.7	0.68	5	0.002
55.468	68.9 (q13.2)	222.7	0.70	8	0.018
10.149	28 (p13.2)	215.2	1.68	7	0.021
12.188	32 (p13.2)	180.9	1.68	6	0.032
50.137	62 (q13.13)	178.1	0.67	10	0.038

[a]Physical positions refer to the leftmost location of each block, taken from NCBI Golden Path release 29; genetic positions were obtained using Ensembl. [b]Taken from the flanking microsatellite markers on the most recent genetic map[21]. [c]The number of times a block at least as long as that observed arose in all simulations of the best fitting model.

Our results indicate that a high marker density, cross-validated in several populations, is required to classify blocks and assess their ultimate utility (Table 2). For such assessment, comparing average properties across populations is not sufficient, as mean levels can be similar because of the marker selection strategy rather than the common evolutionary histories. Instead, comparisons of the location, marker composition and boundary overlap for individual blocks in each population of interest are required to assist the design of disease association studies.

Methods

SNP selection. All chromosome 19 SNPs in the The SNP Consortium database[6] were initially mapped to build 27 of the Golden Path. This permitted the ascertainment of flanking DNA sequences for all SNPs that mapped to one position on chromosome 19. We then evaluated selected SNPs for their ability to undergo successful primer design using an automated software program, AutoPrimer. We selected a panel of 9,048 SNPs with unique map positions and primer designs. After preliminary SNP selection and analysis, we selected additional chromosome 19 SNPs from the dbSNP database to maximize coverage of the SNPs along the chromosome. To fill gaps, we selected a further 502 SNPs, which resulted in a total of 9,550 SNPs. For each SNP, we chose a set of three primers: we designed two PCR primers to amplify a product of 100–200 bp under standard conditions and an optimized single-base primer extension (SBE) primer to be complementary to the sequence immediately 5′ to the SNP site. For the SNPcode platform (Orchid Biosciences), we assigned universal tag sequences to each SBE primer for use in the tag capture step. We then analyzed these hybrid sequences for secondary structure using an internal algorithm developed from empirical data. Any tag–primer combination found to be unsatisfactory by this algorithm was assigned a new tag sequence *in silico*.

DNA samples. We obtained purified genomic DNA samples from the Coriell Institute for Medical Research after institutional approval. We evaluated 10 CEPH reference pedigrees[17] that consisted of 3 generations and included 4 grandparents, 2 parents and 2 offspring each (80 chromosomes total) against the entire set of chromosome 19 SNPs.

Genotyping. Using a proprietary SBE technology on the SNPcode genotyping platform, we analyzed the genotypes. This technology combines multiplexed assays with high-density oligonucleotide arrays and can screen up to 1,824 SNPs simultaneously on individual DNA samples. SNPcode is a high-throughput genotyping platform that detects a SNP by the specific incorporation of a fluorescent dye. It uses a multiplex thermocycled single-base primer extension, followed by solid-phase sorting using a Universal Tag Array or zip-code chip before readout. The SNPcode platform specifically uses the GenFlex Tag Array chip (Affymetrix), which has 2,000 unique features. Typical SNPcode reactions routinely assay 1,824 SNPs per chip and are carried out using 12-plex PCR reactions.

We used automated liquid handling robotics to set up 10-µl PCR reactions that contained 4.0 ng of genomic DNA. The PCR protocol used on this platform is similar to the one previously described[26] with the exception that only 35 cycles were used. Before starting the SBE genotyping reactions, we removed excess nucleotides and PCR primers using shrimp alkaline phosphatase and exonuclease I (Custom ExoSap-IT; USB Corporation). A cocktail containing a fluorescein-labeled and a biotin-labeled nucleotide terminator (PE-NEN), along with the two remaining unlabeled terminators, was combined with a pool of 12 extension primers and a thermostable polymerase such as ThermoSequenase (Amersham Biosciences) with its appropriate buffer. We then incubated the SBE reactions at 96 °C for 3 min, followed by 46 cycles of 94 °C for 20 s and 40 °C for 11 s.

Before the solid-phase sorting of the multiplexed reactions for readout, 152 12-plex PCR reactions were pooled together and precipitated to concentrate the volume of the reaction for hybridization to the Affymetrix GenFlex chip. We resuspended the pellets in hybridization buffer (100 mM MES, pH 6.6; 1 M NaCl; 20 mM EDTA; 0.01% Tween-20) and injected them onto the GenFlex chips. Chips were incubated at 45 °C for 16 h in the Affymetrix GeneChip system hybridization oven[27]. Arrays were washed with Buffer A (6× saline–sodium phosphate–EDTA buffer (SSPE), 0.01% Tween) at 25 °C, followed by Buffer B (3× SSPE, 0.01% Tween-20) at 45 °C.

Table 2 • Predicted block lengths under uniform recombination for markers with different spacing*

Distance between markers (kb)	Predicted block length (kb)		
	Median	Mean	s.d.
0.1	0.40	0.54	0.54
0.5	1.73	2.17	1.44
1	3.07	3.96	2.36
2	6.02	7.28	4.00
5	14.94	16.82	8.64
10	29.95	32.41	15.97
20	58.94	61.85	28.76
50	132.37	148.56	59.50

*In comparison with a previous analysis of a random selection of genomic regions that were genotyped at a mean of 1 marker per 7.8 kb (ref. 2), the uniform recombination model with constant effective population size of 10,000 predicts haplotype blocks of 22.22 kb (median) or 30.55 kb (mean), in high concordance with the empirical data. At very fine resolution (such as roughly 1 marker per kb as on chromosome 21; ref. 9), the expectations from the empirical and simulated data begin to diverge, but are still within 2 s.d.

We stained the chips for 10 min at 25 °C with 6× SSPE, 1× Denhardt's solution (Sigma), 0.01% Tween-20, 5 µg ml^{-1} streptavidin-conjugated R-phycoerythrin and 5 µg µl^{-1} streptavidin for biotin detection, followed by a rinse with Buffer A.

Using a GeneArray scanner (Affymetrix) at 530 nm and 570 nm, we scanned the chips to detect fluorescein and biotin, respectively. We used hybridization controls to normalize the resulting fluorescence intensity scores for signal bleed-through between the two channels. We generated genotyping scores from the ratio of the signal from both channels (fluorescein/(fluorescein + biotin)). The inherently generic design of the tag system allows adaptation of any SNP locus of interest to the assay without any alteration of the chip design or the assay protocol. Therefore, we assayed all 9,550 SNPs with a combination of single chip designs.

Data review and quality control. We genotyped the 10 CEPH pedigrees for all 9,550 SNPs using the SNPcode genotyping platform in a high-throughput setting. Data for each SNP were reviewed independently to verify their quality. We considered each SNP to be validated only if the genotype data met a series of strict criteria, including minimum signal intensity specifications and clear segregation into defined genotype groups or clusters, consistent mendelian inheritance patterns and no significant ($P < 0.01$) departures from Hardy–Weinberg equilibrium. Of the initial 9,550 SNPs, we did not include 46% (4,376 markers) for analysis because they either were classified as non-informative (showed only 1 allele in the CEPH samples tested) or failed to format into useable assays on this platform; 13% (1,266 markers) were monomorphic in all families, and 5% (487 markers) did not consistently map to chromosome 19. Finally, we excluded 1.3% (124 markers) because of deviations from Hardy–Weinberg equilibrium, mendelian inheritance problems or incompatible recombination profiles. A final set of 3,297 SNPs yielded valid genotyping scores, were informative in the CEPH samples, maintained a unique mapping position on chromosome 19 and were included for analysis in this study. The median spacing between markers in this final panel is 1 marker per 5.5 kb, with a mean spacing of 1 marker per 17.65 kb.

DNA sequence assembly. We used build 27 of the human genomic sequence for the chromosome 19 sequence assembly for initial marker selection and mapping. Final analyses were based on build 29 (April, 2002) to take advantage of greater sequence completion. This release has 57 separate contigs, of which 77% of the sequence is completed (56.45 of 73.00 Mb).

LD and haplotype block assessment. LD between pairs of markers was assessed using the $|D'|$ and r^2 measures as described[28]. We estimated the haplotypes for all pedigree founders using Merlin to list all non-recombinant haplotypes, which were then used to estimate haplotype frequencies[29]. Moving averages (sliding windows) of the pairwise LD coefficients were carried out in 1-Mb windows, using the anchor position of each SNP as its reference and considering all markers separated by 50 bp to

500 kb. We assessed the statistical significance of LD runs (Fig. 2) using an adaptation of the Smith–Waterman algorithm as described[7].

For a block definition that could be efficiently simulated, we operationally defined blocks as any series of three or more markers in a contig for which all pairwise values of $|D'|$ exceeded 0.9. Overlapping blocks were excluded by selection of the longer block. This high stringency is similar to that used previously[2]. A block definition that was less stringent (three markers with $|D'| > 0.70$) yielded longer blocks, but was also consistent with the coalescent model outcomes described (data not shown). Also, haplotype networks[7], which are more permissive by relaxing the requirement of a contiguous set of markers, yielded similar block patterns (Fig. 2), suggesting that the stringent, efficient block definition was sufficient to capture the underlying LD trends for evolutionary modeling.

We determined recombination rates directly from the most recent genetic map[21], although microsatellite markers on this map were redefined to their physical locations on the basis of the more recent sequence assembly (build 29). Gene locations and annotations were taken from the Ensembl database.

As an additional tool, we have created a website that allows interactive visualization of the data (see URLs). This includes a basic genome browser for chromosome 19 with zoom capabilities and with features that provide the locations of genes, contigs and LD blocks on the chromosome. It also provides graphs that show $|D'|$ and r^2 values, recombination rates and SNP densities. All the elements in the displayed features are linked to external databases. The chromosome browser provides an interface to obtain contig-by-contig views interactively, with the accompanying external links to resources such as Ensembl and the National Center for Biotechnology Information (NCBI), for information on genes, and dbSNP, for information on SNPs.

Evolutionary modeling. We carried out a detailed simulation study to investigate the effects of variability in recombination rate, variability in gap length between adjacent SNPs and variability in SNP ascertainment in terms of allele frequency on the structure of LD blocks. Combinations of the following parameter alternatives were considered: (i) gap length between adjacent SNPs fixed at a median length from chromosome 19 (5.525 kb) versus gap length sampled from the distribution of lengths across chromosome 19 (Fig. 3*a*); (ii) SNPs ascertained if their frequency was >1% in a sample versus SNPs ascertained to reflect distribution of minor allele frequencies across chromosome 19 (Fig. 3*b*); and (iii) recombination rate fixed at the median rate for chromosome 19 (1.63 cM Mb^{-1}) versus a recombination rate sampled from the distribution of rates across chromosome 19 (data not shown).

For each combination of parameter alternatives, we generated 1,000 replicates of SNP haplotypes for a sample of 80 chromosomes (CEPH pedigree founders). Each replicate corresponded to a region of approximately 1 cM, which is roughly equivalent to sequence contigs across chromosome 19.

For each replicate, we generated gap lengths between adjacent SNPs according to our chosen distribution defined in (i) above. The joint ancestry of the SNPs for the sample of chromosomes was represented by an ancestral recombination graph. Each graph consisted of a series of coalescent and recombination events from which segments of the region first appeared in a common ancestral chromosome and were split between two parental chromosomes, respectively. The distribution of time between these events was simulated under the coalescent process with recombination[19,20], with a rate generated according to (iii) defined above.

The observed distribution of minor allele frequencies across chromosome 19 is approximately uniform, reflecting incomplete ascertainment of rare SNPs. To approximately replicate this ascertainment process, we considered a scheme of the form $\Pr(\text{SNP ascertained} \mid p) = p(p + 1.5)$, where p is the minor allele frequency. This scheme corresponds to complete ascertainment of SNPs with minor allele frequency of 0.5.

For each SNP, we selected the position of a single mutation event on a branch of the ancestral recombination graph at random. The frequency of the mutation in the sample of 80 chromosomes was calculated and tested for ascertainment according to our chosen scheme, defined in (ii) above. If the SNP was not ascertained, we rejected the mutation. We then generated alternative positions for the mutation in the ancestral recombination graph until the SNP was ascertained.

For each replicate, we used the block criteria defined above to determine the distribution of non-overlapping blocks for the generated sample of SNP haplotypes.

URLs. We used the following websites for marker and map construction and data analysis: Ensembl database, http://www.ensembl.org; Golden Path sequence assembly, http://www.ncbi.nlm.nih.gov; Orchid Biosciences AutoPrimer, http://www.autoprimer.com; dbSNP database, http://www.ncbi.nlm.nih.gov/SNP; Online Mendelian Inheritance in Man, http://www.ncbi.nlm.nih.gov/omim. The viewer for chromosome 19 LD is available at http://katahdin.cshl.org:9331/chr19/.

Acknowledgments

The authors thank J. Marcella, J. Ball and R. Tomacelli for advice and guidance during the development of this project at Orchid. R.S. thanks the cancer center at Cold Spring Harbor Laboratory for support. R.L., A.P.M., J.B. and L.R.C. were supported by the Wellcome Trust.

Competing interests statement

The authors declare competing finanical interests: see the Nature Genetics website (http://www.nature.com/naturegenetics) for details.

Received 17 September 2002; accepted 16 January 2003.

1. Daly, M.J., Rioux, J.D., Schaffner, S.F., Hudson, T.J. & Lander, E.S. High-resolution haplotype structure in the human genome. *Nat. Genet.* **29**, 229–232 (2001).
2. Gabriel, S.B. *et al.* The structure of haplotype blocks in the human genome. *Science* **296**, 2225–2229 (2002).
3. Reich, D.E. *et al.* Human genome sequence variation and the influence of gene history, mutation and recombination. *Nat. Genet.* **32**, 135–142 (2002).
4. Goldstein, D.B. Islands of linkage disequilibrium. *Nat. Genet.* **29**, 109–111 (2001).
5. Jeffreys, A.J., Kauppi, L. & Neumann, R. Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. *Nat. Genet.* **29**, 217–222 (2001).
6. Sachidanandam, R. *et al.* A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409**, 928–933 (2001).
7. Dawson, E. *et al.* A first generation linkage disequilibrium map of human chromosome 22. *Nature* **418**, 544–548 (2002).
8. Roses, A.D. Pharmacogenetics and the practice of medicine. *Nature* **405**, 857–865 (2000).
9. Patil, N. *et al.* Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science* **294**, 1719–1723 (2001).
10. Casci, T. Haplotype mapping: shortcut around the block. *Nat. Rev. Genet.* **3**, 573 (2002).
11. Weiss, K.M. & Clark, A.G. Linkage disequilibrium and the mapping of complex human traits. *Trends Genet.* **18**, 19–24 (2002).
12. Couzin, J. Genomics. New mapping project splits the community. *Science* **296**, 1391–1393 (2002).
13. Ardlie, K.G., Kruglyak, L. & Seielstad, M. Patterns of linkage disequilibrium in the human genome. *Nat. Rev. Genet.* **3**, 299–309 (2002).
14. Jeffreys, A.J., Ritchie, A. & Neumann, R. High resolution analysis of haplotype diversity and meiotic crossover in the human TAP2 recombination hotspot. *Hum. Mol. Genet.* **9**, 725–733 (2000).
15. Johnson, G.C. *et al.* Haplotype tagging for the identification of common disease genes. *Nat. Genet.* **29**, 233–237 (2001).
16. Subrahmanyan, L., Eberle, M.A., Clark, A.G., Kruglyak, L. & Nickerson, D.A. Sequence variation and linkage disequilibrium in the human T-cell receptor β (*TCRB*) locus. *Am. J. Hum. Genet.* **69**, 381–395 (2001).
17. Dausset, J. *et al.* Centre d'etude du polymorphisme humain (CEPH): collaborative genetic mapping of the human genome. *Genomics* **6**, 575–577 (1990).
18. Petes, T.D. Meiotic recombination hot spots and cold spots. *Nat. Rev. Genet.* **2**, 360–369 (2001).
19. Hudson, R.R. Properties of a neutral allele model with intragenic recombination. *Theor. Popul. Biol.* **23**, 183–201 (1983).
20. Griffiths, R.C. & Marjoram, P. Ancestral inference from samples of DNA sequences with recombination. *J. Comput. Biol.* **3**, 479–502 (1996).
21. Kong, A. *et al.* A high-resolution recombination map of the human genome. *Nat. Genet.* **31**, 242–247 (2002).
22. Smith, R.A., Ho, P.J., Clegg, J.B., Kidd, J.R. & Thein, S.L. Recombination breakpoints in the human β-globin gene cluster. *Blood* **92**, 4415–4421 (1998).
23. Hartl, D.L. & Clark, A.G. *Principles of Population Genetics* (Sinauer Associates, Sunderland, Massachusetts, 1997).
24. Thompson, E. *Pedigree Analysis in Human Genetics* (Johns Hopkins University Press, Baltimore, Maryland, 1986).
25. McKeigue, P.M. Mapping genes underlying ethnic differences in disease risk by linkage disequilibrium in recently admixed populations. *Am. J. Hum. Genet.* **60**, 188–196 (1997).
26. Bell, P.A. *et al.* SNPstream UHT: ultra-high throughput SNP genotyping for pharmacogenomics and drug discovery. *Biotechniques (Suppl.)* 70–72, 74, 76–77 (2002).
27. Fan, J.B. *et al.* Parallel genotyping of human SNPs using generic high-density oligonucleotide tag arrays. *Genome Res.* **10**, 853–860 (2000).
28. Weir, B.S. *Genetic Data Analysis II.* 113–138 (Sinauer Associates, Sunderland, Massachusetts, 1996).
29. Abecasis, G.R., Cherny, S.S., Cookson, W.O. & Cardon, L.R. Merlin—rapid analysis of dense genetic maps using sparse gene flow trees. *Nat. Genet.* **30**, 97–101 (2002).
30. Eberle, M.A. & Kruglyak, L. An analysis of strategies for discovery of single-nucleotide polymorphisms. *Genet. Epidemiol.* **19**, S29–S35 (2000).