Human Molecular Genetics, 2004, Vol. 13, No. 6 ***577–588***
DOI: 10.1093/hmg/ddh060
Advance Access published on January 20, 2004,

The impact of SNP density on fine-scale patterns of linkage disequilibrium

Xiayi Ke[1], Sarah Hunt[2], William Tapper[3], Robert Lawrence[1], George Stavrides[2], Jilur Ghori[2], Pamela Whittaker[2], Andrew Collins[3], Andrew P. Morris[1], David Bentley[2], Lon R. Cardon[1] and Panos Deloukas[2,*]

[1]Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK, [2]Wellcome Trust Sanger Institute, Hinxton CB10 1SA, UK and [3]University of Southampton, Southampton, UK

Received September 26, 2003; Revised December 23, 2003; Accepted January 13, 2004

Linkage disequilibrium (LD) is a measure of the degree of association between alleles in a population. The detection of disease-causing variants by association with neighbouring single nucleotide polymorphisms (SNPs) depends on the existence of strong LD between them. Previous studies have indicated that the extent of LD is highly variable in different chromosome regions and different populations, demonstrating the importance of genome-wide accurate measurement of LD at high resolution throughout the human genome. A uniform feature of these studies has been the inability to detect LD in regions of low marker density. To investigate the dependence of LD patterns on marker selection we performed a high-resolution study in African-American, Asian and UK Caucasian populations. We selected over 5000 SNPs with an average spacing of ~1 SNP per 2 kb after validating *ca* 12 000 SNPs derived from a dense SNP collection (1 SNP per 0.3 kb on average). Applications of different statistical methods of LD assessment highlight similar areas of high and low LD. However, at high resolution, features such as overall sequence coverage in LD blocks and block boundaries vary substantially with respect to marker density. Model-based linkage disequilibrium unit (LDU) maps appear robust to marker density and consistently influenced by marker allele frequency. The results suggest that very dense marker sets will be required to yield stable views of fine-scale LD in the human genome.

INTRODUCTION

Large-scale linkage disequilibrium (LD) studies conducted at increasing levels of resolution (1–5) have led to the recent launch of a systematic, genome-wide analysis of LD in multiple populations. The project, named HapMap for (human) haplotype map (6–9), aims to develop tools for enhancing the design, interpretation and reproducibility of association studies of common diseases. As the findings of a dense whole-genome association study have yet to be reported, there is some uncertainty at the eventual utility of these maps (6,8,10–13). Nevertheless characterization of LD patterns across the genome is likely to lead to a much better understanding of biological and demographic processes such as recombination, mutation, natural selection and population history (2,14–16).

Simulation studies and empirical data analyses have consistently highlighted extreme variability in LD, extending from a few to several hundred kilobases on average (17–22). Given the high degree of variability in average LD, fine-scale characterization of local LD patterns in the genome remains a serious challenge. This challenge is central to the success of LD maps in facilitating association mapping.

A key difficulty in fine-scale LD characterization concerns the relationship between empirical patterns of LD and operational definitions of them: the highly variable nature of LD renders it difficult to construct objective statistical approaches/algorithms that accurately and consistently reflect the unobserved history of the population sampled (23). Several different statistical methods for considering LD have been developed. The concept of haplotype blocks (3,24) has proven very attractive; in fact the conception of the HapMap project was in part triggered by the suggestion that LD blocks occur ubiquitously throughout the genome (1,3,15) and that a few distinct markers in each block are sufficient to represent the majority of common haplotypes within a population. Different definitions of 'blocks' have since been reported (1,3,24–29). Other, non-block-based descriptions of LD have been developed to predict patterns of pairwise disequilibrium (D')

*To whom correspondence should be addressed. Tel: +44 1223834244; Fax: +44 1223494919; Email: panos@sanger.ac.uk

(19), determine linkage disequilibrium units (LDUs) to construct metric LDU maps (30), estimate recombination rates (31–33), and present average LD against physical and genetic maps (2,22). These approaches all make use of different aspects of genotype data, and thus yield non-identical, though sometimes overlapping, views of LD patterns (4,34).

Empirically, most of the studies of long-range LD have been based on relatively sparse single nucleotide polymorphism (SNP) maps, typically less than one marker per 7 kb (2–4). Notable exceptions to this include a study of chromosome 21 (1) with a density of one marker per 1.3 kb in 10 individuals, and high-density sperm typing studies (35,36) conducted over short genomic regions. Wall and Pritchard (37) also examined a number of short dense regions. At present, the number and distribution of SNPs needed to offer substantive gains for association studies is unknown and debated (13). The initial phase of the HapMap is scheduled to genotype SNP markers at a density of about one marker per 5 kb. No study has evaluated a marker panel of this or higher density across a long chromosome region in population-based samples.

In this paper we assess the variability in LD patterns with respect to marker density and allele frequency along a 10 Mb contiguous segment of chromosome 20q12–13.2 (38). Toward this end, we have genotyped >5000 SNPs at a density of one marker per 2 kb in four different samples (12 CEPH families having 48 founders, 96 UK Caucasians, 97 African Americans and 42 East Asians). Local LD patterns are first assessed using the full marker set and the robustness of the patterns is then examined at lower marker densities (i.e. random marker subsets at one marker per 3, 4, 5, 7.5 and 10 kb). This strategy is depicted in Figure 1.

To describe LD patterns and assess the consistency of local inferences, we use simple sliding window assessments of pairwise LD, the LDMAP program (30) and three haplotype block algorithms: the four-gamete test (25), a D' threshold approach (4) and the popular confidence-limit approach of Gabriel *et al.* (3).

RESULTS

The distributions of marker spacing and minor allele frequencies (m.a.f.) in each of the populations are shown in Figure 2 and descriptive summaries of the samples and marker panels are given in Table 1. The high density of the panel is most clear in Figure 2B; in all samples, more than 80% of the markers are within 3 kb of each other. The largest gap in the CEPH dataset is 36 kb, with only 2% of the gaps exceeding 10 kb (111/5323). The spacing distributions in the other samples are similar, although with slightly wider spaced on average (Table 1).

Descriptive patterns of LD decay are shown in Figure 3. As expected (17,22,39), the rate of decay appears higher in the African Americans than in the other populations. For closely spaced markers (<1 kb: 2659 pairs in the CEPH dataset), the average D' (r^2) is 0.93 (0.47), decreasing with increased spacing, illustrating the well-known pattern of strong but imperfect correspondence between physical spacing and LD for very close markers. Note that more than 5% of such marker pairs are effectively independent ($r^2 < 0.025$), highlighting the extreme variability that is characteristic of pairwise LD and possibly reflecting gene-conversion events (20). To allow a direct comparison of decay rates of the different populations, we constructed datasets having the same number of individuals in each population, using the full East Asian panel and 42 individuals (randomly selected) from each of the African American and UK Caucasian panels. The well-known bias in $|D'|$ according to sample size (12) is clear in these data (Fig. 3A), as D' tends toward a minimum value of ~0.4 for markers that are effectively independent, while r^2 tends towards 0.0 for independent SNPs (Fig. 3B).

For a broad baseline view of the effects of marker density, sliding window plots of r^2 values were constructed for the most extreme densities considered in the three unrelated populations (10 versus 2 kb). In each population, the LD patterns for a 10 kb density are highly similar to those for a 2 kb density (Fig. 4). These similarities are apparent across a number of different window lengths, suggesting that they are not simply a feature of the scale chosen (Supplementary Material Figs S1 and S2). The sliding window plots also reveal good concordance between populations in the regions of high/low LD. High LD regions are most clearly apparent in the UK Caucasian sample, with a similar though slightly diminished profile in the Asian sample. The highest LD regions are also apparent in the African American sample, but at a much reduced level. In general, simple sliding window assessments at coarse marker densities seem to provide a consistent impression of the overall LD trends, which appear useful for broad characterization and comparison of genomic regions, as done initially with microsatellite markers (40). These stable views are useful only for general impressions of LD profiles, as they come at the expense of fine-scale resolution, e.g. the increasing recombination rate across the region is not readily apparent in Figure 4. It is therefore unclear how to directly use these summaries to guide marker selection for disease association studies.

LDU maps offer an alternative summary of LD that can accommodate fine-scale patterns. Lonjou *et al.* (41) compared LDU maps in a number of populations and developed a single 'cosmopolitan' map, which, when appropriately scaled, recovered up to 95% of the information from LD maps constructed from individual samples. Their results suggest that if the map contours are predominantly determined by recombination hot-spots which are co-localized in all populations, population specific LD maps should show similar contours but on different scales.

The contours of LDU maps for the different populations, marker densities and allele frequencies are shown in Figure 5, following the graphical conventions used previously with LDU maps (42). Alternative representations of these comparisons are given in Figure 6, in which the widths of LDU intervals are plotted against their physical position along the chromosome. In this view, areas of high peaks correspond to steep steps (high recombination) in Figure 5, while flat areas correspond to plateaus.

Figures 5A and 6A show the variation in the strength of LD across the region for the three populations. The data reveal some similarities between populations in regions of extensive LD (e.g. around 1–1.25 and ~9 Mb) as well as regions where

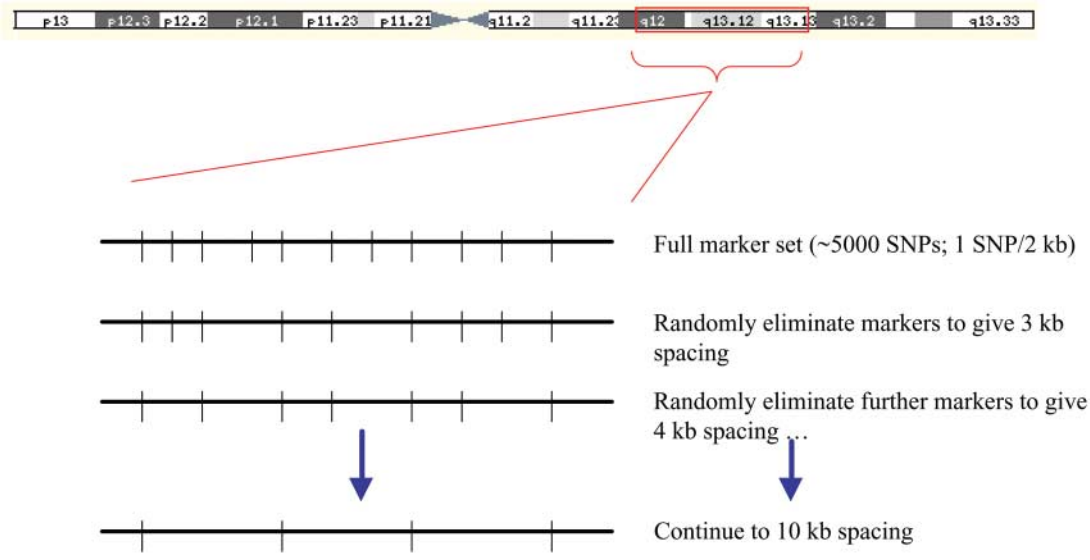

Figure 1. Strategy employed to evaluate the influence of marker density on LD inference. The 10 Mb region of chromosome 20 is highlighted.

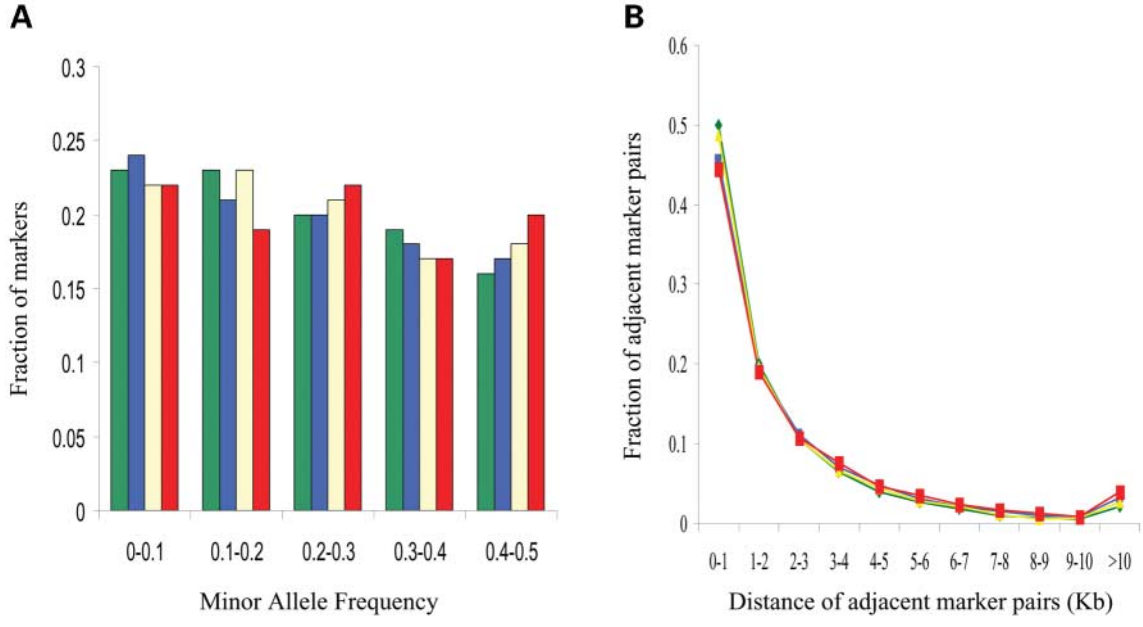

Figure 2. Characteristics of full marker sets on chromosome 20. The color schemes for the different samples are CEPH (green), UK Caucasian (blue), Asian (red) and African-American (yellow) samples. (**A**) Distribution of minor allele frequencies in each sample. (**B**) Distribution of spacing between markers for each sample.

LD is breaking down (~7.5 and ~4.3 Mb). The data also reveal differences (e.g. ~2–4 Mb), many of which reflect the magnitude of steps, with African Americans generally having the steepest steps (greatest apparent recombination), followed by Asians and UK Caucasians, respectively.

LDU maps appear largely insensitive to marker density at the observed resolution as their profiles determined at various densities are highly correlated (Figs 5B, 6B and C, Spearman's rank correlation coefficient $r = 0.78$–0.64). In both the UK Caucasian and the African-American samples, total LDU lengths are generally consistent between 10 and 2 kb (Table 2) with typically less than a 5% change (always $<10\%$) in map length. The maps also reveal a slightly concave pattern that is consistent with the known increasing recombination rate from the left to the right of Figure 5. The stability of LDU maps with respect to marker density reflects an additive effect in distances.

As expected, LDU maps are affected by changes in allele frequencies, with subsets of more common SNPs producing longer maps. This is because under neutral evolution, common alleles tend to be older and thus have had more time to recombine, thereby showing less LD (43). Since LDUs reflect the product of recombination and time, maps made from subsets of common/older SNPs will be longer. Allele frequency influences total length by altering the height of steps, reflecting

Table 1. Characteristics of markers, samples and LD in four samples

Data characteristic	CEPH	UK Caucasian	African American	Asian
Number of samples	93	96	97	42
Number of unrelated chromosomes	94	192	194	84
Number of markers	5324	4427	4938	4160
Average marker density (SNPs/kb)	1.88	2.26	2.03	2.40
Number of blocks (Gabriel)	530	510	620	434
Number of blocks (threshold)	725	603	800	595
Number of blocks (Wang)	1110	1061	1345	967
Sequence coverage (%) (Gabriel)	53	55	39	44
Sequence coverage (%) (threshold)	65	63	52	62
Sequence coverage (%) (Wang)	61	52	44	54
Avg block length (kb) (Gabriel)	10.1	11.1	6.4	10.5
Avg block length (kb) (threshold)	9.0	10.5	6.6	10.5
Avg block length (kb) (Wang)	5.6	5.0	3.3	5.7
LDU map length	250	240	324	301

recombination intensity, while their location and number remain relatively constant. As a result the interval length (LDUs) of these LD maps are positively correlated (Spearman's rank correlation coefficient for African Americans $r = 0.64$; UK Caucasians $r = 0.63$; Fig. 5C). LDU maps made from a subset of SNPs with an average m.a.f. of 20% and density of 4 kb are 11% longer and contain 3% more steps compared with those produced from SNPs with an average m.a.f. of 5% at the same density. Overall, the steps of the m.a.f. 20% density 4 kb LDU map are 8% longer than the m.a.f. 5% density 4 kb LDU map.

For each of the three block definitions, the number of blocks, average block length and proportion of sequence contained within blocks ('sequence coverage') are presented in Table 1 for the full marker sets in each sample. At 2 kb marker density, about 50% of the region is contained within blocks; for example, in the UK Caucasian sample, 54 and 55% of the total sequence was covered by the Gabriel approach and the threshold method, respectively. The average block length varies substantially between methods and populations, ranging from 3.3 kb in the African American samples with the four-gamete test to 11.1 kb in the UK Caucasians with the Gabriel test. Interestingly, while Gabriel *et al.* (3) reported average block lengths of 22 kb in Caucasians and 11 kb in African Americans using an average marker density of 7.8 kb, the average block length using the same statistical method is ~50% lower in this study (10.1–11.1 kb in CEPHs and UK Caucasians versus 6.4 kb in African Americans). This difference could be due in part to the slightly higher recombination rate in this region of chromosome 20 (1.54 cM/Mb) relative to the genome average.

While the overall levels of sequence coverage appear largely similar across methods, only 43, 41 and 49% of the sequence was common to the Gabriel/threshold, Gabriel/four-gamete and threshold/four-gamete methods, respectively, in the CEPH samples. The number of blocks detected varied even more substantially amongst block definitions; for example, up to 2-fold in the CEPH samples (530 in the Gabriel approach versus 1110 in the Wang method). At this highest marker density (2 kb), there does not appear to be a strong convergence of block-detection methods in delineating specific LD patterns.

We examined the UK Caucasian and African American samples (which have similar sample sizes) with the 4337 markers common to both in order to explore LD block structures

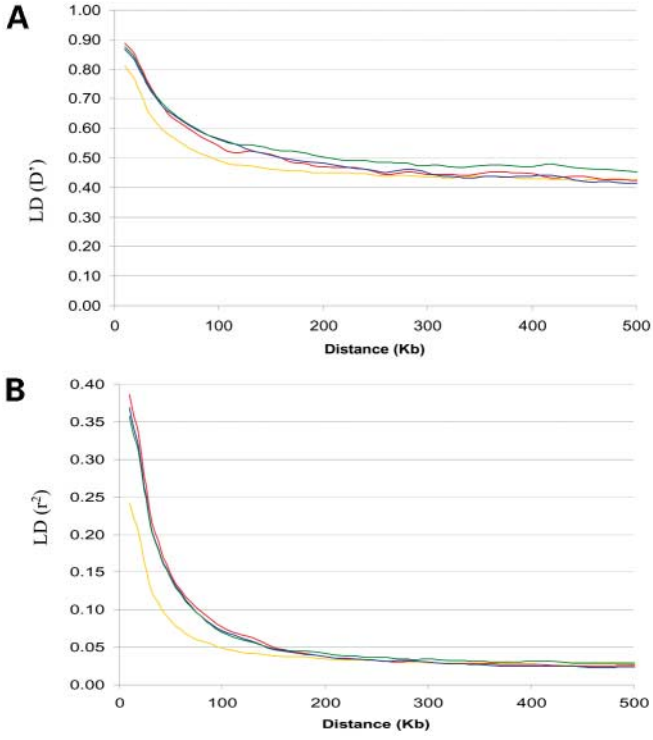

Figure 3. Decay of linkage disequilibrium as a function of physical distance. For direct comparison of decay rates, the same number of individuals were drawn from each sample ($n = 42$). The colour schemes for the different sample are CEPH (green), UK Caucasian (blue), Asian (red) and African American (yellow). The CEPH sample has 48 founders, of which 42 were used. In these families, the pedigree information was used to estimate haplotypes across the full 10 Mb region, and the haplotype frequencies for each pairwise LD coefficient was calculated using the pedigree-derived haplotypes. The decay rates represent the average D' (**A**) or r^2 (**B**) for all marker pairs separated by distance S (for $S = 10, 20, 30, \ldots, 500$ kb). Successive 20 kb bins were incremented by 10 kb overlap to produce the smooth decay plots shown.

at different marker densities and m.a.f. thresholds. At all marker densities, there was more sequence covered by LD blocks and longer average block length in the UK Caucasian than in the African American population (Figure 7A and B), consistent with previous reports (3). Also, in both samples an inverse relationship was observed between sequence coverage and average block length: as marker density increased, the sequence coverage of LD blocks also increased, but the average block-length decreased (Figure 7A and B). Thus, long blocks apparent at low marker densities break up into shorter blocks as more markers are genotyped. Figure 7C shows two regions of 400 kb in which new blocks are established between pre-existing blocks, longer blocks are merged from adjacent shorter blocks and, most commonly, long blocks are broken down into smaller blocks as marker density is increased from one marker per 10 kb to one per 2 kb. This trend, observed throughout the region, does not seem to begin to stabilize at any marker density from 10 to 2 kb. The proportion of blocks which break or remain conserved at each density is shown in Figure 8A. More than one-half of the blocks formed at a density of one marker per 10 kb change boundaries after adding new markers to reach 2 kb. In intermediate densities (7.5, 5, 4 and 3 kb) there is a gradual decrease in the proportion of blocks broken relative to the 2 kb map. Still, nearly 30% of the

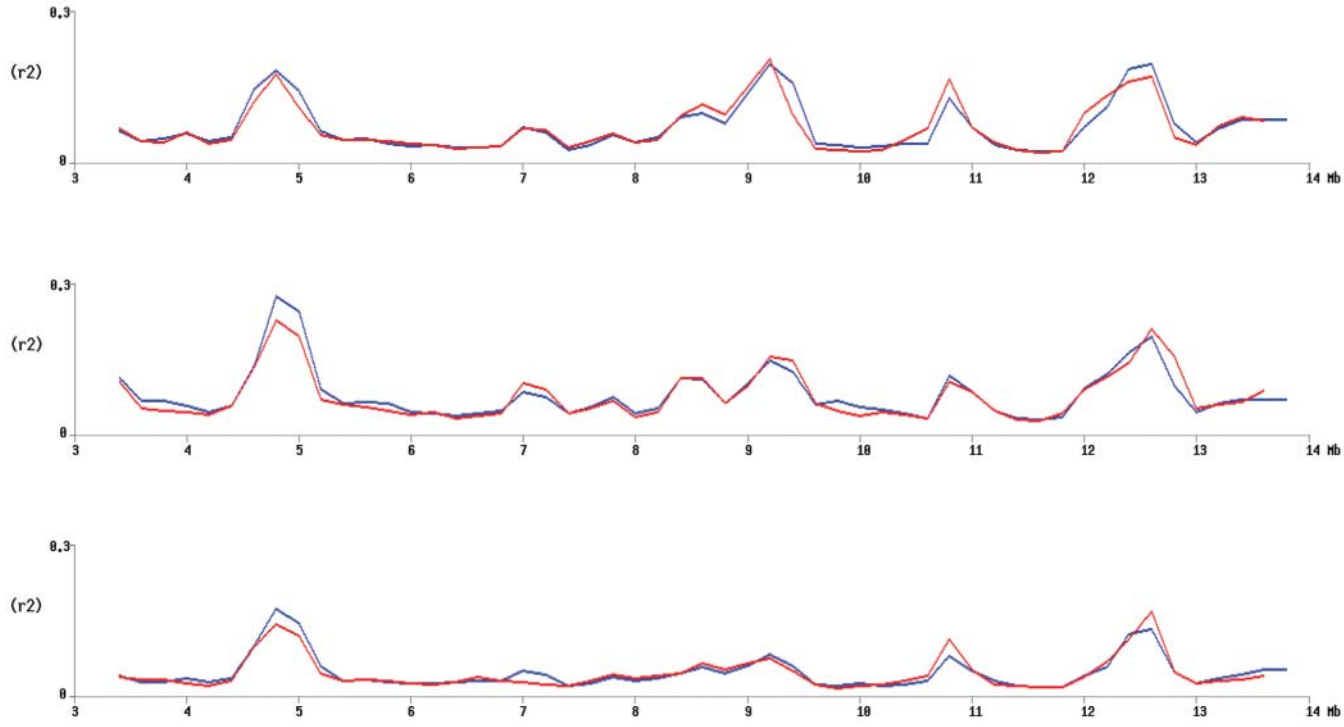

Figure 4. Sliding window plots of average r^2 in each population. Average r^2 was calculated from 25 to 250 kb interspaced SNPs in 500 kb sliding windows (200 kb increment between windows). The top, middle and bottom panel show the trends for the randomly selected 42 UK Caucasians, Asians and African Americans, respectively. In each population sample, two different marker densities were used, i.e. one marker per 2 kb (blue) and one marker per 10 kb (red).

blocks detected at 3 kb have different boundaries when evaluated at 2 kb. The high degree of variability in block length as a function of marker density is consistent with theoretical predictions under various scenarios, including uniform recombination (4,37,44). At all densities, restricting assessments to markers exceeding different m.a.f. thresholds (i.e. 0.05, 0.10, 0.15) did not substantively alter the pattern of change on the block structures (data not shown).

The trend of decreasing block sizes does not seem to be a feature of the specific block-detection algorithm. The D' threshold and four-gamete method showed different absolute levels of block lengths and sequence coverage, but similar patterns of change with increasing marker density. It also does not seem to be a feature of blocks defined by only a few markers, as blocks depicted by six or more markers also revealed a strong dependence on marker density with a clear propensity to breaking upon further genotyping (Supplementary Material Fig. S3).

Although the precise boundaries of blocks vary widely, regions of high block coverage correspond to regions of high LD. The average r^2 value between markers that break blocks and markers in the original blocks they break is higher than for random marker pairs (Fig. 8B), suggesting that block-breaking markers often fall just below the arbitrary thresholds imposed by block-detection algorithms.

DISCUSSION

We have genotyped more than 5000 SNPs in four population samples in a 10 Mb region of chromosome 20. This dense 2 kb map offers several main conclusions about linkage disequilibrium patterns, how they are depicted by existing statistical methods, and their possible role in disease association studies.

Firstly, broad views of high and low LD patterns can be observed using a variety of statistical procedures, are clearly distinguishable at coarse marker densities (e.g. 1 SNP/10 kb) and do not change appreciably with increased marker densities. Regions of exceptionally high LD can be delineated by direct observation (e.g. with protracted peaks in sliding window plots, regions of high sequence coverage in blocks or long plateaus in LDU maps). They may also be characterized in a more quantitative sense by comparing long tracts of LD against the empirical distribution of all LD runs (2), evaluating the parameter estimates of LDU maps (30) or assessing measures relating to haplotype diversity in blocks (24). The HapMap project should provide in its first phase (5 kb density) a framework profile of LD to guide further fine-scale assessment.

Secondly, localized patterns of LD are highly dependent upon the density of markers used to discern them. Successive addition of SNPs within close regions often disrupts specific patterns of pairwise LD, which can otherwise (misleadingly) appear very high and consistent at coarser marker densities. This high variability in LD was apparent even at the density of one marker per 2 kb. The variation was also apparent in each of our three population samples. Given the strong dependence of LD measures on allele frequencies and thus mutation age, this effect is somewhat of an expected feature of LD assessments derived from pairwise measures.

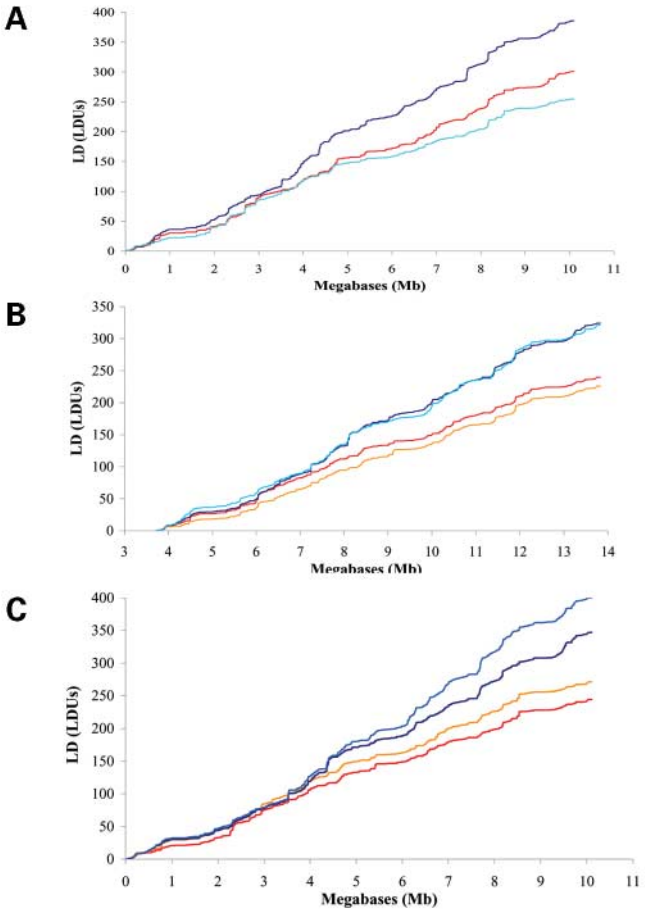

Figure 5. LDU maps across the 10 Mb region of chromosome 20. (**A**) Comparison of LDU maps for African-American (dark blue), Asian (red) and UK Caucasian (green) samples. The full marker panels for each population were used to construct the LDU maps. (**B**) Effects of marker density on LDU maps of chromosome 20. The maps for UK Caucasians are shown in orange (one marker per 10 kb) and red (one marker per 2 kb). The maps for African Americans are shown in light blue (one marker per 10 kb) and dark blue (one marker per 2 kb). (**C**) Effects of minor allele frequency on LDU maps of chromosome 20 region. The orange and red lines show the LDU maps for the 4 kb marker panel in UK Caucasians, subselected for 20 and 5% minimum m.a.f., respectively. The light and dark blue lines show the LDU maps for the 4 kb marker panel in African Americans, subselected for 20 and 5% minimum m.a.f., respectively.

Thirdly, while haplotype blocks reflect high LD in local regions, their precise boundaries are not always robust to the density of markers genotyped. As marker density increases, the sequence coverage of haplotype blocks increases, and the average size of haplotype blocks decreases. Increased marker densities can fill in chromosomal regions which are poorly covered by coarse spacing, but they also often break blocks into smaller pieces, yielding block boundaries which vary according to SNP ascertainment. Variable block boundaries and decreasing block sizes which accompanied increased densities were observed in all populations under each of the three block definitions examined. Given this high degree of dependence on marker density, practical usage of haplotype blocks in association studies using sparse maps does not appear promising.

The use of both a similar marker density and the same block definition algorithm allows a direct comparison between this study (intermediate tier of one SNP per 7.5 kb) and the study of genomic regions by Gabriel *et al.* (3). In our 7.5 kb tier, the average block size was 19 and 12 kb in the UK Caucasian and African-American samples, respectively. These are close to the observed (predicted) average sizes of 18 (22) and 9 (11) kb in the Western European and African samples reported by Gabriel *et al.* (3); thus, despite a slightly higher recombination rate in this region of chromosome 20, the two studies are highly concordant in terms of average block sizes. However, these same blocks decrease by about 40% upon further genotyping to a density of 2 kb, highlighting the variability in block boundaries.

Despite high variability in boundaries, the higher LD in block-rich regions may be amenable to haplotype-tagging strategies for construction of efficient SNP sets (45). In principle, marker sub-selection could be conducted for a region of nearly any complexity, conditional on the sample data (23). The degree to which selected tagging SNPs change as a function of marker density warrants further assessment.

Fourthly, the general pattern of high versus low LD in the LDU map approach appears largely consistent across the range of marker densities examined. The regions of presumed high ancestral recombination, depicted by tall slopes, and presumed low recombination (as in blocks), depicted by plateaus, are generally consistent from densities of 1 marker per 10 kb to one per 2 kb (Table 2), in contrast to the other methods examined. As noted earlier, specific markers often deviate markedly from pairwise LD trends, presumably due to factors other than recurrent recombination (e.g. age of mutations, genetic drift, mutation rate differences, gene-conversion, etc.). LDU maps are more robust to this sort of deviation, as the model-fitting provides a degree of smoothing which seems to remove much of the marker density effects that plagues other fine-scale descriptions of LD.

LDU maps appear largely stable over the densities examined, but they are strongly influenced by allele frequencies. LDU maps derived from markers with common alleles are likely to be longer than those built from rare allele markers. This is not an artefact but a biological feature that LDU maps are capable of depicting since LD is generally higher with rarer (recent) markers than in common (older) markers. This implies that maps derived from studies with non-randomly ascertained SNPs, such as the HapMap project focusing on common allele markers, will require some modification so that they generalize to marker sets having other frequencies. From the LDU map perspective, this allele frequency influence emphasizes probable limitations of the HapMap for diseases with rare alleles.

Finally, specific LD patterns in our LDU and block assessments indicate both clear and recurrent transitions from high to low LD and, more often, less intense changes from high to low LD. This situation along with fuzzy block boundaries suggests that assessment of LD by SNP genotyping yields a view of the genome as a few series of discrete jumps from recombination hotspots to extreme haplotype conservation that can be replicated (15,46,47) and many more unstable transitions from regions of high to low LD. It is possible that a higher density of markers than 2 kb would be more stable. It is also possible that localized regions of recurrent recombination exist broadly in the

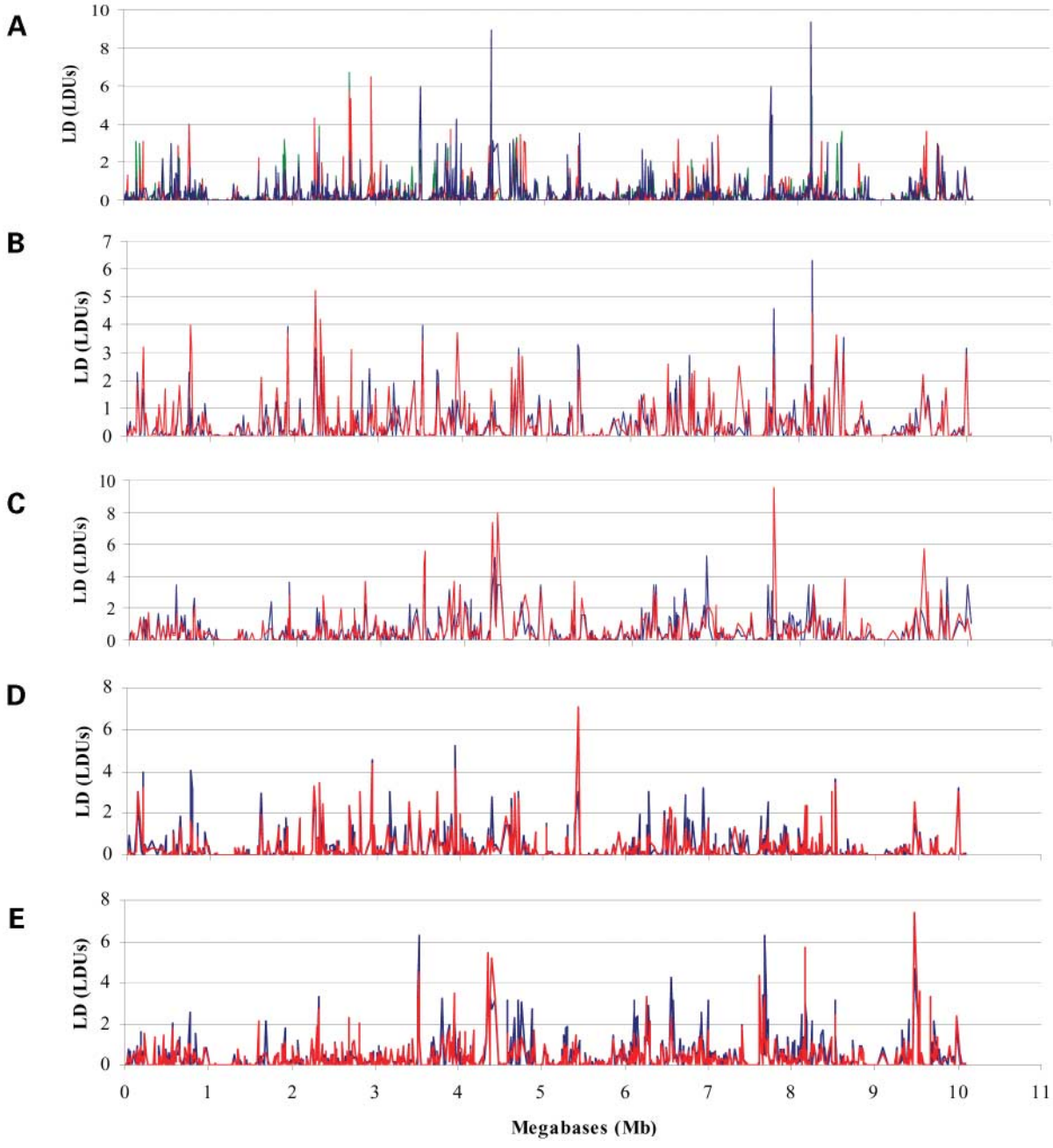

Figure 6. LDU width of SNP intervals across a 10 Mb region of chromosome 20. Intervals are plotted on the physical scale (*x*-axis) by their midpoint. (**A**) Comparison of interval LDU widths for African-American (dark blue), Asian (red) and UK Caucasian (green) samples. (**B**) Effects of marker density on LDU width of SNP intervals for UK Caucasians. The dark blue line represents low density (one marker per 10 kb) whilst the red line indicates high density (one marker per 2 kb). (**C**) Effects of marker density on LDU width of SNP intervals for African Americans. The dark blue line represents low density (one marker per 10 kb) whilst the red line indicates high density (one marker per 2 kb). (**D**) Effects of minor allele frequency on LDU width of SNP intervals for UK Caucasians. The dark blue and red lines are for subsets of SNPs with 20 and 5% minimum m.a.f. respectively and a density of one marker per 4 kb. (**E**) Effects of minor allele frequency on LDU width of SNP intervals for African Americans. The dark blue and red lines are for subsets of SNPs with 20 and 5% minimum m.a.f. respectively and a density of one marker per 4 kb.

genome, but the exact positions of each cross-over event are distributed around each apparent hotspot rather than occurring as discrete points on the chromosome. Approaches which focus on the sites of recombination directly, via experimental techniques such as single sperm typing (35,36,48) or statistical procedures such as estimation of local recombination rates (31–33), are promising for helping to clarify the genealogy of specific chromosome regions (reviewed in 16).

The highly variable continuum of LD along chromosomal regions raises challenges for designing studies and constructing statistical procedures to efficiently use LD information in the context of association studies. For example, how to use LD to

Table 2. LDU lengths at different marker densities

Minimal allele frequency	UK Caucasian Mean interval size (kb)					African American Mean interval size (kb)				
	2	4	5	7.5	10	2	4	5	7.5	10
All	240	240	236	213	226	324	328	330	338	321
5	—[a]	244	251	237	250	—[a]	347	342	335	318
10	—[a]	256	250	249	249	—[a]	377	367	351	331
15	—[a]	269	264	272	264	—[a]	401	403	385	398
20	—[a]	271	267	267	262	—[a]	400	413	405	416

[a]Exclusion of rare SNPs means 2 kb densities are unavailable for subsets with higher m.a.f.

detect causal SNPs remains an open question. Empirical data have the power to shape the development and calibration of new tools; do we need pilot association studies now? The chromosome 20 region described here offer this possibility as it (i) has a marker density higher than the planned end-product of HapMap and (ii) spans a region with linkage to common diseases such as type 2 diabetes and obesity.

MATERIALS AND METHODS

Samples

The following panels were used: HD100AA, panel of 100 unrelated African Americans (Coriell Cell Repositories); 12 three generation CEPH/Utah families (95 individuals); East Asian panel of 32 Japanese (22 from the American Diabetes Association) and 10 Chinese. The CEPH and East Asian panels were originally assembled by the SNP Consortium and are available through Coriell; Human Random Control (HRC) panel1 of 96 UK Caucasians (ECACC).

SNP ascertainment

All publicly available SNPs (dbSNP; release 114) with unique map positions on the chromosome 20 contig NT_011362.7 (3 726 000–13 824 000 bp) were parsed through Illumina's assay design software. Average SNP density along chromosome 20 was at one SNP per 330 bp due to a targeted SNP discovery effort at the Sanger Institute based on shotgun sequencing of flow sorted chromosome 20 DNA (P. Deloukas, unpublished data). Illumina designed assays for ca 65% of the SNPs of which we selected a total of 11 328 SNPs (spacing >100 bp).

The average recombination rate of the region is 1.54 cM/Mb (0.64 cM/Mb in males; 2.45 cM/Mb in females) based on the deCode genetic map (49); note this is lower than the chromosome average of 1.89 cM/Mb (male 1.61cM/Mb; female 2.17 cM/Mb). The recombination rate is lowest at the proximal end of the region (1.22 cM/Mb at 39 Mb) and rises steadily to 1.87 cM/Mb at the distal end.

Genotyping and error checking

SNP genotyping was contracted out to Illumina which used the Golden Gate assay and a multiplex level of 1152 SNPs per well. The methodology including reproducibility and concordance to other platforms will be described elsewhere (50). In brief, all assays (11 328 SNPs) were developed on both strands and used to genotype 384 samples which included 365 DNAs from the panels described above plus controls. Among the DNAs, 32 were blind duplicates. Some 82% of SNPs yielded a working assay on either strand. In addition, we genotyped the CEPH panel with a heavily overlapping set of 2000 SNPs using the Homogeneous Mass Extent assay (Sequenom). We then assessed the 6019 SNPs (both sets) that showed a minor allele frequency ≥4% in either the combined DNA set or in the African American panel.

Of the 5704 SNPs typed on the Illumina platform, 2613 were successful on both strands, 1846 on the optimal and 1245 on the other strand. We used duplicate genotypes (assays on both strands and DNA duplicates) and inheritance (CEPH family panel only) to identify all discrepant genotypes. We first removed five DNAs with poor results (NA10842, NA12046 in CEPH samples, and NA17163, NA17170 and NA17160 in African-American samples). We then looked at the genotype confidence score provided by Illumina and applied a cut-off value above which the retained fraction of genotypes had a discordant rate <0.3%; 92% (1 507 291) of all possible unique genotypes met these criteria. No SNP loci were removed at this point.

Finally, from each of the four population samples, i.e. African Americans, East Asians, CEPH families and UK Caucasians, we removed loci with less than 80% of all possible genotypes (157, 165, 190 and 131, respectively), out of Hardy–Weinberg ($\chi^2 \geq 10$) (118, 30, 65 and 52), zero heterozygosity (23, 541, 268 and 273) and double recombinant (698) (CEPH samples only).

Final SNP panels

Per population, the final analyses were based on: CEPH, 5324 markers (average spacing one marker per 1.88 kb); UK Caucasians, 4427 markers (one marker per 2.26 kb); African Americans, 4938 markers (one marker per 2.03 kb) and 4160 markers for the East Asians (one marker per 2.40 kb). Unless noted otherwise, all analyses use these sets of markers or subsets of them (see Data Repository URL).

The UK Caucasian and African American panels were analysed based on a series of decreasing marker densities generated from a total of 4337 markers common to both. The following marker densities were created by randomly removing a subset of markers from the full dataset: full (one marker per 2.3 kb), one marker per 3, 4, 5, 7.5 and 10 kb. These subsets of data were then subjected to block analysis and comparison. This process was repeated 30 times and only very small variances were observed between repeats in terms of sequence coverage of blocks, average block size and percentage of blocks being broken (data not shown).

Block algorithms

Haplotype block detection was based on three block definitions:

(i) The Gabriel *et al.* (3) approach. This method estimates lower (CL) and upper (CU) confidence limits of D' coefficients between each pair of markers and uses a series

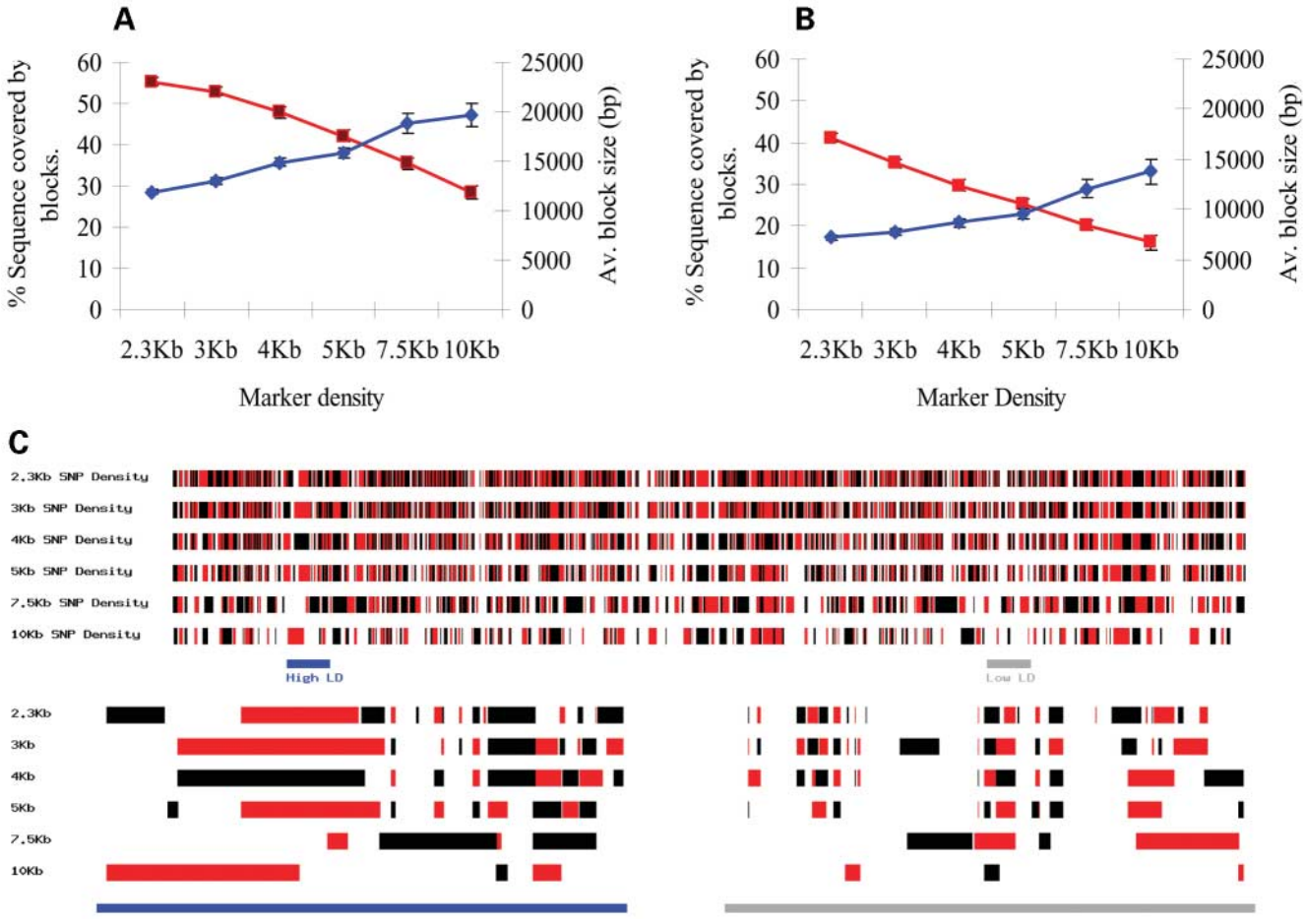

Figure 7. Average block sizes and sequence coverage for the 96 unrelated UK Caucasian and 97 African American individuals. Blocks were identified using the Gabriel *et al.* (3) approach at an overall marker density of 10 kb (one marker per 10 kb), 7.5 kb or 5, 4, 3 and 2.3 kb for UK Caucasian (**A**) and African American (**B**), respectively. The top two panels show the percentage of genomic sequence covered by blocks on the *y*-axis (red lines) and the average block size on the *z*-axis (blue lines). (**C**) A graphic depiction of blocks across the entire 10 Mb region. Within each marker density (row) blocks are alternately coloured black and red to show the points of change. Below the full region image, two illustrative 400 kb segments of high (rs2066906–rs967083) and low (rs2183794–rs932675) LD are expanded in greater detail to indicate specific block compositions.

of heuristics to define a block. To assess the accuracy of our software implementation of this approach (supplementary information), we compared the data set distributed with 'HaploView' (see Electronic Database listing for URL), two other datasets from the Gabriel *et al.* (3) data, and two small regions (<200 markers) from chromosome 20. All confidence limits, block boundaries, numbers of blocks and block lengths were identical with HaploView.

(ii) The D' threshold approach (3). A block is defined whenever all pairwise D' coefficients between three or more successive markers exceed 0.90. A greedy algorithm is applied to avoid block overlaps. The algorithm is designed to recover the longest sequence coverage by blocks. Pairwise D' coefficients were calculated using the programs ldmax and haploXT in the GOLD package (51), together with Merlin (52) for the CEPH families.

(iii) The four-gamete test. This approach detects runs of markers with no evidence for historical recombination (25). For the CEPH families and the three unrelated population samples, we use Merlin (25) and snphap (see Electronic Database section for URL) to estimate haplotypes, which are then used to conduct the extended four gamete test. Recombination was considered to have occurred if there were four haplotypes for any marker pair. A region was only considered as part of a block if none of the marker pairs in the region showed evidence of recombination. A block contained a minimum number of two markers.

For all block analyses, we present the results for the commonly-used Gabriel *et al.* approach in the first instance, and then highlight any noteworthy similarities/differences with the other methods.

LDU maps

The LDMAP program (30) was used to construct metric LDU maps from these data sets. This program employs the Malecot model (53,54) to predict association as $\rho = (1 - L)Me^{-\varepsilon d} + L$, where L is the residual association at large distance, M is the

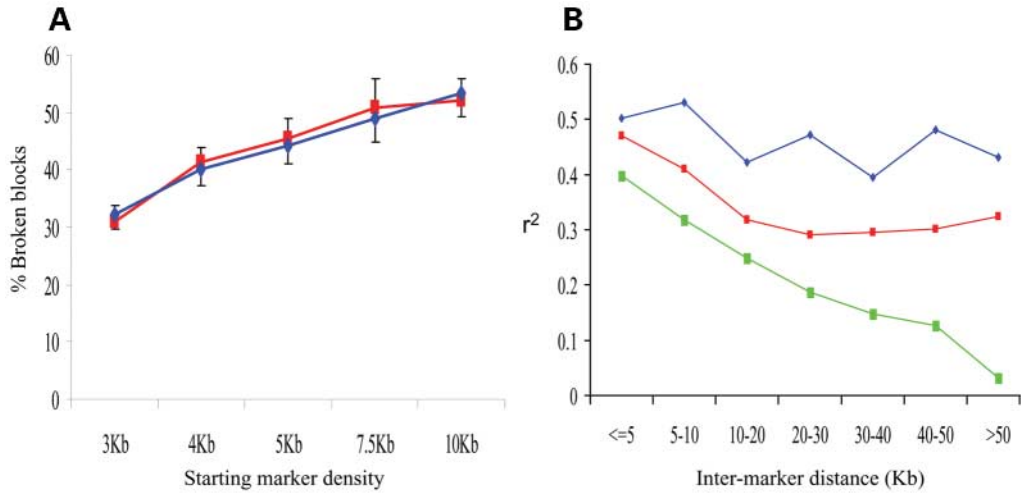

Figure 8. The effects of increasing marker density on block boundaries. (**A**) shows the results percentage of blocks that break as marker density is increased to one marker per 2 kb (e.g. the 10 kb point on the *x*-axis reflects a change from 10 to 2 kb; the 7.5 kb point reflects a change from 7.5 to 2 kb, etc.) These results were obtained using the full set of block rules as described by Gabriel *et al.* (3), and repeated 30 times; African-American samples are shown in blue; UK Caucasian samples are in red. (**B**) Inter-marker disequilibrium for all marker pairs at 2 kb density across the 10 Mb region (green), marker pairs of the original blocks (at the density of one marker per 10 kb; blue), and pairs between markers of the original block (at one marker per 10 kb) and the markers that broke the block (at one marker per 2 kb) (red). The Gabriel *et al.* (2002) block definition was used in these analyses.

association at zero distance and a measure of phylogeny where 1 represents a monophyletic origin and is less than 1 otherwise, and ε is the exponential decline of ρ with physical distance *d*. For marker by marker association, in unrelated individuals, ρ equates to the absolute value of D' (55). The program estimates values of M and L for the entire sample, while ε is estimated in each marker-defined interval using all of the pairwise data that is informative for that interval. Convergence of parameter estimates is achieved by maximizing the composite likelihood $[-\sum K_\rho(\hat{\rho} - \rho)^2/2]$, where $\hat{\rho}$ is an empirical estimate of ρ, the information (K_ρ) satisfies $\chi^2{}_1 = \rho^2 K_\rho$, and the length of the ith interval in LDUs is given by $\varepsilon_i d_i$, where ε_i is the Malecot parameter and d_i is the length of the interval on the physical map in kb. The LDU map identifies regions of high LD as plateaus (similar to 'blocks') and characterizes steps which reflect recombining areas or events and define the relationship between blocks (56). As a result, whole chromosome LDU maps (42) are highly correlated with linkage maps (49), suggesting that much of the variation in the extent LD across the genome is due to recombination.

Sliding windows

Sliding window assessments were conducted by considering all pairwise LD (r^2) values for markers spaced between 25 and 250 kb. Average r^2 values were plotted for all marker pairs within each successive window of 500 kb. The overlap length of each window was 200 kb. The window length/overlap characteristics were chosen to exhibit general trends in the data. Similar trends were apparent with various window lengths and marker composition, showing more specific features of LD patterns (Supplementary Material Figs S1 and S2). As with any sliding window model, window length was negatively correlated with variability in LD trend (i.e. the longer the window, the less distinct the peaks).

ELECTRONIC DATABASE INFORMATION

Data Repository (this study): www.well.ox.ac.uk/~xiayi/data/chr20/10Mb/index.html. Ensembl chromosome 20: www.ensembl.org. NHGRI Haplotype Map site: www.genome.gov/page.cfm?pageID=10001688. ECACC: www.ecacc.org.uk/. Corriell Cell Repositories: http://coriell.umdnj.edu/ccr/ccrsumm.html. dbSNP: www.ncbi.nlm.nih.gov/SNP. Merlin: www.sph.umich.edu/csg/abecasis/Merlin/. HaploView: www-genome.wi.mit.edu/personal/jcbarret/haplo/. Gabriel *et al.* (3) data: www-genome.wi.mit.edu/mpg/hapmap/hapstruc.html. snphap: www-gene.cimr.cam.ac.uk/clayton/software/. LDMAP: http://cedar.genetics.soton.ac.uk/public_html/.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at HMG Online.

ACKNOWLEDGEMENTS

This work was supported by the Wellcome Trust, and in part by grants from The SNP Consortium and NIH (EY-126562; L.R.C.) and Medical Research Council (G9801327). We like to thank the genotyping team at Illumina, in particular Drs A. Oliphant and L. Galver.

REFERENCES

1. Patil, N., Berno, A.J., Hinds, D.A., Barrett, W.A., Doshi, J.M., Hacker, C.R., Kautzer, C.R., Lee, D.H., Marjoribanks, C., McDonough, D.P. *et al.* (2001) Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science*, **294**, 1719–1723.
2. Dawson, E., Abecasis, G.R., Bumpstead, S., Chen, Y., Hunt, S., Beare, D.M., Pabial, J., Dibling, T., Tinsley, E., Kirby, S., Carter, D. *et al.* (2002) A first generation linkage disequilibrium map of human chromosome 22. *Nature*, **418**, 544–548.

3. Gabriel, S.B., Schaffner, S.F., Nguyen, H., Moore, J.M., Roy, J., Blumenstiel, B., Higgins, J., DeFelice, M., Lochner, A., Faggart, M. *et al.* (2002) The structure of haplotype blocks in the human genome. *Science*, **296**, 2225–2229.
4. Phillips, M.S., Lawrence, R., Sachidanandam, R., Morris, A.P., Balding, D.J., Donaldson, M.A., Studebaker, J.F., Ankener, W.M., Alfisi, S.V., Kuo, F.S. *et al.* (2003) Chromosome-wide distribution of haplotype blocks and the role of recombination hot spots. *Nat. Genet.*, **33**, 382–387.
5. Walsh, E.C., Mather, K.A., Schaffner, S.F., Farwell, L., Daly, M.J., Patterson, N., Cullen, M., Carrington, M., Bugawan, T.L., Erlich, H. *et al.* (2003) An integrated haplotype map of the human major histocompatibility complex. *Am. J. Hum. Genet.*, **73**, 580–590.
6. Couzin, J. (2002) Genomics. New mapping project splits the community. *Science*, **296**, 1391–1393.
7. Harris, R.F. (2002) Hapmap flap. *Curr. Biol.*, **12**, R827.
8. Clark, A.G., Nielsen, R., Signorovitch, J., Matise, T.C., Glanowski, S., Heil, J., Winn-Deen, E.S., Holden, A.L. and Lai, E. (2003) Linkage disequilibrium and inference of ancestral recombination in 538 single-nucleotide polymorphism clusters across the human genome. *Am. J. Hum. Genet.*, **73**, 285–300.
9. Schmidt, C.W. (2003) HapMap: building a database with blocks. *EHP Toxicogenomics*, **111**, A16.
10. Weiss, K.M. and Terwilliger, J.D. (2000) How many diseases does it take to map a gene with SNPs? *Nat. Genet.*, **26**, 151–157.
11. Lai, E., Bowman, C., Bansal, A., Hughes, A., Mosteller, M. and Roses, A.D. (2002) Medical applications of haplotype-based SNP maps: learning to walk before we run. *Nat. Genet.*, **32**, 353.
12. Weiss, K.M. and Clark, A.G. (2002) Linkage disequilibrium and the mapping of complex human traits. *Trends Genet.*, **18**, 19–24.
13. Carlson, C.S., Eberle, M.A., Rieder, M.J., Smith, J.D., Kruglyak, L. and Nickerson, D.A. (2003) Additional SNPs and linkage-disequilibrium analyses are necessary for whole-genome association studies in humans. *Nat. Genet.*, **33**, 518–521.
14. Ardlie, K.G., Kruglyak, L. and Seielstad, M. (2002) Patterns of linkage disequilibrium in the human genome. *Nat. Rev. Genet.*, **3**, 299–309.
15. Reich, D.E., Schaffner, S.F., Daly, M.J., McVean, G., Mullikin, J.C., Higgins, J.M., Richter, D.J., Lander, E.S. and Altshuler, D. (2002) Human genome sequence variation and the influence of gene history, mutation and recombination. *Nat. Genet.*, **32**, 135–142.
16. Arnheim, N., Calabrese, P. and Nordborg, M. (2003) Hot and cold spots of recombination in the human genome: the reason we should find them and how this can be achieved. *Am. J. Hum. Genet.*, **73**, 5–16.
17. Clark, A.G., Weiss, K.M., Nickerson, D.A., Taylor, S.L., Buchanan, A., Stengard, J., Salomaa, V., Vartiainen, E., Perola, M., Boerwinkle, E. and Sing, C.F. (1998) Haplotype structure and population genetic inferences from nucleotide-sequence variation in human lipoprotein lipase. *Am. J. Hum. Genet.*, **63**, 595–612.
18. Kruglyak, L. (1999) Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nat. Genet.*, **22**, 139–144.
19. Abecasis, G.R., Noguchi, E., Heinzmann, A., Traherne, J.A., Bhattacharyya, S., Leaves, N.I., Anderson, G.G., Zhang, Y., Lench, N.J., Carey, A. *et al.* (2001) Extent and distribution of linkage disequilibrium in three genomic regions. *Am. J. Hum. Genet.*, **68**, 191–197.
20. Ardlie, K., Liu-Cordero, S.N., Eberle, M.A., Daly, M., Barrett, J., Winchester, E., Lander, E.S. and Kruglyak, L. (2001) Lower-than-expected linkage disequilibrium between tightly linked markers in humans suggests a role for gene conversion. *Am. J. Hum. Genet.*, **69**, 582–589.
21. Pritchard, J.K. and Przeworski, M. (2001) Linkage disequilibrium in humans: models and data. *Am. J. Hum. Genet.*, **69**, 1–14.
22. Reich, D.E., Cargill, M., Bolk, S., Ireland, J., Sabeti, P.C., Richter, D.J., Lavery, T., Kouyoumjian, R., Farhadian, S.F., Ward, R. and Lander, E.S. (2001) Linkage disequilibrium in the human genome. *Nature*, **411**, 199–204.
23. Cardon, L.R. and Abecasis, G.R. (2003) Using haplotype blocks to map human complex trait loci. *Trends Genet.*, **19**, 135–140.
24. Daly, M.J., Rioux, J.D., Schaffner, S.F., Hudson, T.J. and Lander, E.S. (2001) High-resolution haplotype structure in the human genome. *Nat. Genet.*, **29**, 229–232.
25. Wang, N., Akey, J.M., Zhang, K., Chakraborty, R. and 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.
26. Zhang, K., Deng, M., Chen, T., Waterman, M.S. and Sun, F. (2002a) A dynamic programming algorithm for haplotype block partitioning. *Proc. Natl Acad. Sci. USA*, **99**, 7335–7339.
27. Anderson, E.C. and Novembre, J. (2003) Finding haplotype block boundaries by using the minimum-description-length principle. *Am. J. Hum. Genet.*, **73**, 336–354.
28. Mannila, H., Koivisto, M., Perola, M., Varilo, T., Hennah, W., Ekelund, J., Lukk, M., Peltonen, L. and Ukkonen, E. (2003) Minimum description length block finder, a method to identify haplotype blocks and to compare the strength of block boundaries. *Am. J. Hum. Genet.*, **73**, 86–94.
29. Zhang, K., Sun, F., Waterman, M.S. and Chen, T. (2003b) Haplotype block partition with limited resources and applications to human chromosome 21 haplotype data. *Am. J. Hum. Genet.*, **73**, 63–73.
30. Maniatis, N., Collins, A., Xu, C.F., McCarthy, L.C., Hewett, D.R., Tapper, W., Ennis, S., Ke, X. and Morton, N.E. (2002) The first linkage disequilibrium (LD) maps: delineation of hot and cold blocks by diplotype analysis. *Proc. Natl Acad. Sci. USA*, **99**, 2228–2233.
31. Fearnhead, P. and Donnelly, P. (2001) Estimating recombination rates from population genetic data. *Genetics*, **159**, 1299–1318.
32. Hudson, R.R. (2001) Two-locus sampling distributions and their application. *Genetics*, **159**, 1805–1817.
33. McVean, G., Awadalla, P. and Fearnhead, P. (2002) A coalescent-based method for detecting and estimating recombination from gene sequences. *Genetics*, **160**, 1231–1241.
34. Schwartz, R., Halldorsson, B.V., Bafna, V., Clark, A.G. and Istrail, S. (2003) Robustness of inference of haplotype block structure. *J. Comput. Biol.*, **10**, 13–19.
35. Jeffreys, A.J., Kauppi, L. and Neumann, R. (2001) Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. *Nat. Genet.*, **29**, 217–222.
36. Kauppi, L., Sajantila, A. and Jeffreys, A.J. (2003) Recombination hotspots rather than population history dominate linkage disequilibrium in the MHC class II region. *Hum. Mol. Genet.*, **12**, 33–40.
37. Wall, J.D. and Pritchard, J.K. (2003) Assessing the performance of the haplotype block model of linkage disequilibrium. *Am. J. Hum. Genet.*, **73**, 502–515.
38. Deloukas, P., Matthews, L.H., Ashurst, J., Burton, J., Gilbert, J.G., Jones, M., Stavrides, G., Almeida, J.P., Babbage, A.K., Bagguley, C.L. *et al.* (2001) The DNA sequence and comparative analysis of human chromosome 20. *Nature*, **414**, 865–871.
39. Jorde, L.B. (2000) Linkage disequilibrium and the search for complex disease genes. *Genome Res.*, **10**, 1435–1444.
40. Huttley, G.A., Smith, M.W., Carrington, M., O'Brien, S.J. (1999) A scan for linkage disequilibrium across the human genome. *Genetics*, **152**, 1711–1722.
41. Lonjou, C., Zhang, W., Collins, A., Tapper, W.J., Elahi, E., Maniatis, N. and Morton, N.E. (2003) Linkage disequilibrium in human populations. *Proc. Natl Acad. Sci. USA*, **100**, 6069–6074.
42. Tapper, W., Maniatis, N., Morton, N.E. and Collins, A. (2003) A metric linkage disequilibrium map of a human chromosome. *Ann. Hum. Genet.*, **67**, 487–494.
43. Kimura, M. (1983) *The Neutral Theory of Molecular Evolution.* Cambridge University Press, Cambridge.
44. Zhang, K., Akey, J.M., Wang, N., Xiong, M., Chakraborty, R. and Jin, L. (2003) Randomly distributed crossovers may generate block-like patterns of linkage disequilibrium: an act of genetic drift. *Hum. Genet.*, **113**, 51–59.
45. Johnson, G.C., Esposito, L., Barratt, B.J., Smith, A.N., Heward, J., Di Genova, G., Ueda, H., Cordell, H.J., Eaves, I.A., Dudbridge, F. *et al.* (2001) Haplotype tagging for the identification of common disease genes. *Nat. Genet.*, **29**, 233–237.
46. Goldstein, D.B. (2001) Islands of linkage disequilibrium. *Nat. Genet.*, **29**, 109–111.
47. Paabo, S. (2003) The mosaic that is our genome. *Nature*, **421**, 409–412.
48. Jeffreys, A.J., Ritchie, A. and Neumann, R. (2000) High resolution analysis of haplotype diversity and meiotic crossover in the human TAP2 recombination hotspot. *Hum. Mol. Genet.*, **9**, 725–733.
49. Kong, A., Gudbjartsson, D.F., Sainz, J., Jonsdottir, G.M., Gudjonsson, S.A., Richardsson, B., Sigurdardottir, S., Barnard, J., Hallbeck, B., Masson, G., Shlien, A., Palsson, S.T., Frigge, M.L., Thorgeirsson, T.E., Gulcher, J.R. and Stefansson, K. (2002) A high-resolution recombination map of the human genome. *Nat. Genet.*, **31**, 242–247.

50. Fan, J.B., Oliphant, A., Shen, R., Kermani, B.G., Garcia, F., Gunderson, K.L., Hansen, M., Steemers, F., Butler, F.L., Deloukas, P., *et al.* (2004) Highly parallel SNP genotyping. *Cold Spring Harb. Symp. Quant. Biol.*, **68**, in press.
51. Abecasis, G.R. and Cookson, W.O. (2000) GOLD—graphical overview of linkage disequilibrium. *Bioinformatics*, **16**, 182–183.
52. Abecasis, G.R., Cherny, S.S., Cookson, W.O. and Cardon, L.R. (2002) Merlin—rapid analysis of dense genetic maps using sparse gene flow trees. *Nat. Genet.*, **30**, 97–101.
53. Malecot, G. (1948) *Les Mathématiques de l'Hérédité*. Maison et Cie, Paris.
54. Collins, A, Morton, N.E. (1998) Mapping a disease locus by allelic association. *Proc. Natl Acad. Sci. USA*, **95**, 1741–1745.
55. Lewontin, R.C. (1964) The interactions of selection and linkage. I. General considerations: heterotic models. *Genetics*, **49**, 49–67.
56. Zhang, W., Collins, A., Maniatis, N., Tapper, W. and Morton, N.E. (2002) Properties of linkage disequilibrium (LD) maps. *Proc. Natl Acad. Sci. USA*, **99**, 17004–17007.