High-resolution haplotype structure in the human genome

Mark J. Daly[1], John D. Rioux[1], Stephen F. Schaffner[1], Thomas J. Hudson[1,2] & Eric S. Lander[1,3]
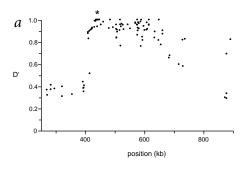
Linkage disequilibrium (LD) analysis is traditionally based on individual genetic markers and often yields an erratic, non-monotonic picture, because the power to detect allelic associations depends on specific properties of each marker, such as frequency and population history. Ideally, LD analysis should be based directly on the underlying haplotype structure of the human genome, but this structure has remained poorly understood. Here we report a high-resolution analysis of the haplotype structure across 500 kilobases on chromosome 5q31 using 103 single-nucleotide polymorphisms (SNPs) in a European-derived population. The results show a picture of discrete haplotype blocks (of tens to hundreds of kilobases), each with limited diversity punctuated by apparent sites of recombination. In addition, we develop an analytical model for LD mapping based on such haplotype blocks. If our observed structure is general (and published data suggest that it may be), it offers a coherent framework for creating a haplotype map of the human genome.

In a companion project, we are studying a 500-kb region on human chromosome 5q31 that is implicated as containing a genetic risk factor for Crohn disease[1]. After high-density SNP discovery, 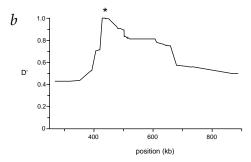we selected 103 common (>5% minor allele frequency) SNPs genotyped in 129 trios from a European-derived population. Our results thus describe 258 chromosomes transmitted to individuals with Crohn disease and 258 untransmitted chromosomes.
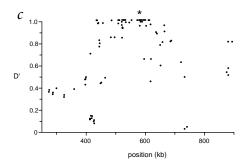
The genotype data provide the highest-resolution picture to date of the patterns of genetic variation across a large genomic region, with a marker density of 1 SNP roughly every 5 kb. For studying both disease association (marker versus disease) and LD (marker versus marker), the traditional approach has been to perform single-marker analysis. Examples of such analysis are shown in Fig. 1. Although there are clearly many strong correlations, the picture is noisy and unsatisfying, and important localization information is obscured by properties of the markers not relevant to the issues under study.
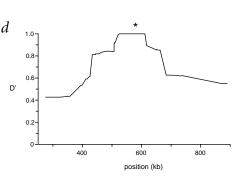
To obtain a clearer picture, we focused on identifying the underlying haplotypes. It became evident that the region could be largely decomposed into discrete haplotype blocks, each with a striking lack of diversity (Fig. 2). Our initial focus was on untransmitted control chromosomes; however, the same haplotype structure was seen in the chromosomes transmitted to individuals with Crohn disease, with the only difference being that one of the haplotypes was enriched in frequency, reflecting its

Fig. 1 Comparison of single-marker LD with haplotype-based LD. *a*, LD between an arbitrary marker (at the 26th position, indicated with an asterisk) and every other marker in the data set using *D′*. *b*, Multiallelic *D′* is used to plot LD between the maximum-likelihood haplotype group assignment at the location of the 26th marker and that assignment at the location of every other marker in the data set. *c*,*d*, Repeat of the comparison in *a* and *b* but with respect to a second marker (at the 61st position) in the map. Both pairs of graphs show the common feature that, when haplotypes rather than individual SNP alleles are considered to be the basic units of variation, the noise (presumably caused by marker history and properties of the specific statistic chosen) essentially disappears, resulting in a clear, monotonic and step-like breakdown of LD by recombination.

association to Crohn disease[1]. Because this structure is the same in both groups, we present combined data from all chromosomes (transmitted and untransmitted).

The haplotype blocks span up to 100 kb and contain multiple (five or more) common SNPs. The blocks have only a few (2–4) haplotypes, which show no evidence of being derived from one another by recombination, and which account for nearly all chromosomes (>90%) in all cases in the sample. For example, an 84-kb block shows only two distinct haplotypes that together account for 95% of the observed chromosomes (Table 1). The lack of diversity is readily seen from the fact that the probability an individual is homozygous for all SNPs genotyped in a block ranges from 30–70%.

The discrete blocks are separated by intervals in which several independent historical recombination events seem to have occurred, giving rise to greater haplotype diversity for regions spanning the blocks. The most common recombination events are indicated in Fig. 2 by lines connecting the haplotypes. The recombination events appear to be clustered; multiple obligate exchanges must have occurred between most blocks, with little or no exchange within blocks. For example, in the 84-kb block (Table 1), not a single apparent recombinant between the two major haplotypes was observed (despite the fact that such a recombinant would be obvious because the haplotypes differ at all SNPs examined).

The clustering is suggestive of local hotspots of recombination[2–4], and the same observation of inhomogeneity of recombination is made for the class II region of the MHC elsewhere in this issue[5]. Although there is detectable recombination between blocks, it is modest enough for there to be clear long-range correlation (that is, LD) among blocks. The haplotypes at the various blocks can be readily assigned to one of four ancestral long-range haplotypes. Indeed, 38% of the chromosomes studied carried one of these four haplotypes across the entire length of the region.

Using a hidden Markov model (HMM), we developed an approach to define the block structure formally. The HMM simultaneously assigns every position along each observed chromosome to one of the four ancestral haplotypes and estimates the maximum-likelihood values of the 'historical recombination frequency' (Θ) between each pair of markers. The quantity Θ provides a convenient summary of the degree of haplotype exchange across inter-marker intervals and relates directly to the conventional measures of LD, such as D'. (An alternative measure is the joint probability of homozygosity[6].) In the case at hand, the discrete block structure is evident from the fact that Θ is estimated at less than 1% for 73 of the inter-marker intervals, 1–4% for 14 of the intervals, and more than 4% for only 9 of the intervals.

We considered whether the selection of the SNPs could have significantly influenced the results. The SNPs studied were ascertained by complete resequencing of seven individuals with Crohn disease and one control[1]. To test whether this survey failed to detect much of the common variation, we compared our SNPs in a 100-kb subregion to those identified by
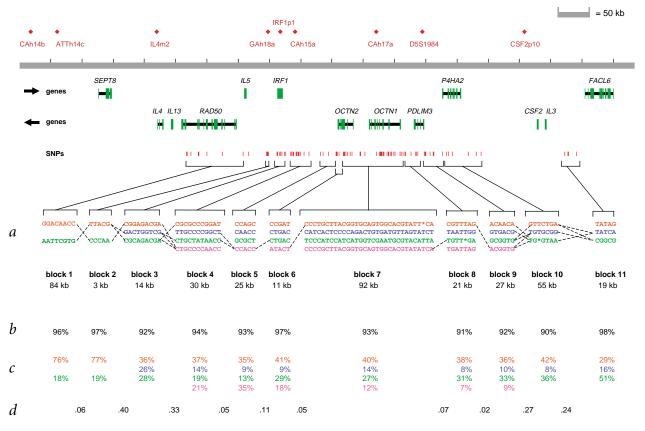

Fig. 2 Block-like haplotype diversity at 5q31. *a*, Common haplotype patterns in each block of low diversity. Dashed lines indicate locations where more than 2% of all chromosomes are observed to transition from one common haplotype to a different one. *b*, Percentage of observed chromosomes that match one of the common patterns exactly. *c*, Percentage of each of the common patterns among untransmitted chromosomes. *d*, Rate of haplotype exchange between the blocks as estimated by the HMM. We excluded several markers at each end of the map as they provided evidence that the blocks did not continue but were not adequate to build a first or last block. In addition, four markers fell between blocks, which suggests that the recombinational clustering may not take place at a specific base-pair position, but rather in small regions.

the International SNP Map Working Group (ISMWG)[7]. We detected 47 of the 54 SNPs (86%) reported by the ISMWG, a rate that exceeds the proportion of ISMWG SNPs (ascertained in a multi-ethnic panel) typically found to be polymorphic in a Caucasian population (roughly 80%; S. Bolk, personal communication). In addition, we discovered 150 SNPs in this region not reported by the ISMWG.

This analysis used SNPs with minor allele frequency greater than 5%. We genotyped six rarer SNPs and found that the rare allele fell exclusively or nearly exclusively on one of the major haplotype patterns and simply created a subtype of that pattern. This underscores that, when we refer to limited haplotype diversity, we are not implying complete sequence identity among chromosomes with the same haplotype, but rather that chromosomes fall into a small number of deep clades. Chromosomes within a clade may differ at one or a few rare SNPs, whereas chromosomes in different clades differ at many SNPs. Finally, we note that we initially eliminated SNPs at CpG sites because the higher mutation rate at such sites[8,9] might introduce recurrent mutation and thereby confound the analysis. Of the 16 high frequency CpG SNPs genotyped, 13 had alleles that aligned perfectly with the haplotype patterns in Fig. 1 and only one added significantly to the overall heterozygosity of the block in which it fell.

Our analysis of this region of chromosome 5q31 in a European-derived population indicates the following: the region may be largely divided into discrete blocks of 10–100 kb; each block has only a few common haplotypes; and the haplotype correlation between blocks gives rise to long-range LD. Determining whether these are general features of human genetic variation will require studies of other regions with similarly dense genetic maps (increasingly feasible given the availability of human genome sequence[10] and large SNP collections[7]); however, available evidence seems to be consistent with this picture. In numerous data sets, comprehensive SNP genotyping in small regions (2–5 kb upstream from candidate genes) indicates limited haplotype diversity (3 or 4 haplotypes accounting for 80–95% of all observed chromosomes[11–14]), similar to the data presented here. Together with observations of an unexpectedly long extent of LD[15–17], these reports suggest that our description of haplotype diversity in 5q31 may be, in qualitative terms, fairly general.

The structure of LD described here has important implications for the analysis of LD, for association studies to find medically relevant variation, for population genetics, and for the next steps of the Human Genome Project.

Focusing on haplotype blocks greatly clarifies LD analyses. Once the haplotype blocks are identified, they can be treated as alleles and tested for LD (for example, our simple analysis uses Hedrick's multi-allelic extension of D'[18,19], thereby reflecting the underlying population variation more accurately than any individual SNP. The power of the haplotype-based approach can be seen by comparing the noisy single-marker analyses of LD (Fig. 1*a*,*c*) with corresponding analyses performed on the underlying haplotype blocks (Fig. 1*b*,*d*). The latter analyses show that LD decays monotonically (as expected if recombination has the main role in the breakdown), with the decrease occurring in abrupt drops reflecting the sites of significant historical recombination.

In analogous fashion, the haplotype structure provides a crisp approach for testing the association of genomic segments with disease. By contrast, disease association studies traditionally involve testing individual SNPs in and around a gene. This approach is statistically weak and has no clear endpoint: true associations may be missed because of the incomplete information provided by individual SNPs; negative results do not rule out association involving other nearby SNPs; and positive results

Table 1 • Haplotypes of SNPs in block 1 (8 SNPs/84 kb)*

Haplotype	Observations
G G A C A A C C	283 (83.2%) haplotype A
A A T T C G G G	40 (11.8%) haplotype B
G A T T A G C C	2 (0.6%)
G G T C A G C C	2 (0.6%)

*Another 13 chromosomes (3.8%) were observed that matched haplotype A or B at all alleles except one, and might represent gene conversion or an undetected genotyping error.

do not indicate the discovery of the causal SNP but simply a marker in LD with a true causal SNP located some distance (perhaps several genes) away. Once the haplotype blocks are defined, however, it is straightforward to examine a subset of SNPs that uniquely distinguish the common haplotypes in each block (shown elsewhere in this issue)[20]. This allows the common variation in a gene to be tested exhaustively for association with disease (given a specified level of genotype relative risk and disease allele frequency). Although this analysis, such as presented in the companion paper[1], will not always directly result in the identification of the causal gene and mutation, it focuses subsequent functional studies on the critical region of maximum haplotype distortion within which there exists insufficient historical recombination for variation studies to reduce it further. (In addition, although association studies with haplotypes are much clearer than those with individual SNPs, we note that strict monotonic decay of association is not expected, even with perfect haplotype data, for reasons described elsewhere[21].)

The structure and composition of the haplotype blocks have considerable implications for human population genetics. The data here are broadly consistent with coalescent simulations[17], which suggest that models, including both inhomogeneous recombination (reflecting the apparent clustering of major recombinational events) and recent bottlenecks (accounting for the limited number of distinct haplotypes over long distances seen in this European derived sample, may be necessary to explain modern human diversity. Detailed haplotype analysis of many genomic regions in several populations, together with comprehensive simulation studies, will be needed to determine the relative importance of these and other factors.

Finally, our approach provides a precise framework for creating a comprehensive haplotype map of the human genome. By testing a sufficiently large collection of SNPs, it should be possible to define all of the common haplotypes underlying blocks of LD. Once such a map is created, it will be possible to select an optimal reference set of SNPs for any subsequent genotyping study. Such a project is becoming feasible, and this detailed understanding of common human variation represents an important step in the Human Genome Project.

Methods

Individuals and marker selection. The individuals studied, Canadians from metropolitan Toronto of predominantly European descent and the genotyping methodologies are described in the companion paper[1]. To ensure our ability to reconstruct multi-marker haplotypes, SNPs for haplotype analysis were selected from the set of markers for which full genotypes were available for all members of 85 or more trios. To eliminate markers likely to contain significant numbers of undetected genotyping errors, markers not in Hardy–Weinberg equilibrium ($P<0.05$) or those for which more than 10 mendelian inheritance errors were detected were excluded from this analysis. SNPs at CpG sites were not included in the initial analysis to prevent potential confounding of common haplotype patterns from recurrent mutation. In addition, rare SNPs (minor allele frequency <5%) were not included in the initial analysis. The underlying data for this analysis is contained on our website (http://www-genome.wi.mit.edu/humgen/IBD5).

Haplotype counting. Haplotype percentages in Fig. 2 were computed using haplotypes generated by the transmission disequilibrium test (TDT) implementation in Genehunter 2.0 (ref. 22), followed by use of an EM-type algorithm[23,24], to include the minority of chromosomes that had one or more markers with ambiguous phase (that is, where both parents and offspring were heterozygous) or where one marker was missing genotype data. Clark's method[25], or simply counting only fully informative phase-known haplotypes, provided essentially identical answers, because within each block most chromosomes were fully reconstructed without ambiguity from the parental data.

Regions of low-haplotype diversity were initially identified as follows: five-marker haplotypes for all consecutive sets of five markers were generated; the observed haplotypic heterozygosity (HET_{obs}) and expected haplotypic heterozygosity (HET_{exp}) (given allele frequency and assuming equilibrium) were tallied; and each five-marker window was assigned a score, $S_5 = HET_{obs}/HET_{exp}$. A smaller value therefore represents lower diversity of haplotypes compared with expectation. Windows with locally minimal scores were then expanded or contracted by adding or subtracting markers to the ends to find the longest local minimum window. Boundaries between these windows (which we call 'blocks') were examined. The most common connections between haplotypes considered to be the 'ancestral haplotype class' (displayed on the same line in the same color in Fig. 2), and cases in which a high frequency (>2%) haplotype is observed that represents a connection between two different 'ancestral classes' are shown by a line connecting those classes across that interval.

Hidden Markov model. The observations that over long distances most haplotypes can be described either as belonging to one of a small number of common haplotype categories, or as a simple mosaic of those categories, suggested the use of an HMM in which haplotype categories were defined as states. We assigned observed chromosomes to those hidden states (allowing for missing/erroneous genotype data), and simultaneously estimated the transition probability in each map interval by using an EM algorithm and by making the simplifying assumption that there was one transition probability for each map interval (the aforementioned probability of historical recombination Θ) rather than allowing specific transition probabilities from each state to each state. The output of this method was a maximum-likelihood assignment to haplotype category at each position (which can be used to compute, for example, multi-allelic D' and TDT) and maximum-likelihood estimates of Θ indicating how significantly recombination has acted to increase haplotype diversity in each map interval. The use of probabilities of recombination in this context[6] has a simple relationship with the most commonly used measure of gametic disequilibrium (D'). If we consider two SNPs at a time before any recombination (or other type of event) has occurred to create a fourth haplotype (as in the following table):

	SNP 2	
	Allele 1	Allele 2
SNP 1 Allele 1	a	b
Allele 2	$c=0$	d

we can see that D' (which equals $(ad-bc)/[(a+c)(c+d)]$ for this table configuration) is equal to 1 (full disequilibrium). Many generations later, we can collapse all recombination that has occurred between the two markers into a single value: the probability that a modern chromosome has undergone recombination at any time between those two markers. Let $(1-\Theta)$ represent the probability that no recombination has taken place at any time between these two markers. At this time, the table of haplotype frequencies will have changed to

	SNP 2	
	Allele 1	Allele 2
SNP 1 Allele 1	$a-ad\Theta$	$b+ad\Theta$
Allele 2	$ad\Theta$	$d-ad\Theta$

And therefore D' reduces simply to $1-\Theta$. Θ here (Θ_{real}) differs from the observed rates (Θ_{obs}) reported in Fig. 2, as some recombinations occur between chromosomes with identical local haplotypes; however, the observed values are trivially corrected by the local homozygosity to produce the real values.

1. Rioux, J.D. *et al.* Hierarchical linkage disequilibrium mapping of a susceptibility gene for Crohn's disease to the cytokine cluster on chromosome 5. *Nature Genet.* **29**, 223–228 (2001).
2. Templeton, A.R. *et al.* Recombinational and mutational hotspots within the human lipoprotein lipase gene. *Am. J. Hum. Genet.* **66**, 69–83 (2000).
3. 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).
4. 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).
5. Jeffreys, A.J., Kauppi, L. & Neumann, R. Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. *Nature Genet.* **29**, 217–222 (2001).
6. Sved, J.A. Linkage disequilibrium and homozygosity of chromosome segments in finite populations. *Theor. Pop. Biol.* **2**, 125–141 (1971).
7. The International SNP Map Working Group. A map of the human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409**, 928–933 (2001).
8. Krawczak, M., Ball, E.V. & Cooper, D.N. Neighboring-nucleotide effects on the rates of germ-line single-base-pair substitution in human genes. *Am. J. Hum. Genet.* **63**, 474–488 (1998).
9. Nachman, M.W. & Crowell, S.L. Estimate of the mutation rate per nucleotide in humans. *Genetics* **156**, 297–304 (2000).
10. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
11. Drysdale, C.M. *et al.* Complex promoter and coding region β2-adrenergic receptor haplotypes alter receptor expression and predict *in vivo* responsiveness. *Proc. Natl Acad. Sci. USA* **97**, 10483–10488 (2000).
12. Park, H.Y. *et al.* Identification of new single–nucleotide polymorphisms in the thrombin receptor gene and their effects on coronary artery diseases in Koreans. *Clin. Exp. Pharmacol. Physiol.* **27**, 690–693 (2000).
13. Jordanides, N., Eskdale, J., Stuart, R. & Gallagher, G. Allele associations reveal four prominent haplotypes at the human interleukin-6 (IL-6) locus. *Genes Immun.* **1**, 451–455 (2000).
14. D'Alfonso, S., Rampi, M., Rolando, V., Giordano, M. & Momigliano–Richiardi, P. New polymorphisms in the IL–10 promoter region. *Genes Immun.* **1**, 231–233 (2000).
15. Bonnen, P.E. *et al.* Haplotypes at ATM identify coding–sequence variation and indicate a region of extensive linkage disequilibrium. *Am. J. Hum. Genet.* **67**, 1437–1451 (2000).
16. Moffatt, M.F., Traherne, J.A., Abecasis, G.R. & Cookson, W.O. Single nucleotide polymorphism and linkage disequilibrium within the TCR α/δ locus. *Hum. Mol. Genet.* **9**, 1011–1019 (2000).
17. Reich, D.R. *et al.* Linkage disequilibrium in the human genome. *Nature* **411**, 199–204 (2001).
18. Lewontin, R.C. The interaction of selection and linkage. General considerations; heterotic models. *Genetics* **49**, 49–67 (1964).
19. Hedrick, P.W. Gametic disequilibrium measures: proceed with caution. *Genetics* **117**, 331–341 (1987).
20. Johnson, G.C.L. *et al.* Haplotype tagging for the identification of common disease genes. *Nature Genet.* **29**, 233–237 (2001).
21. Kruglyak, L. & Lander, E.S. High-resolution genetic mapping of complex traits. *Am. J. Hum. Genet.* **56**, 1212–1223 (1995).
22. Daly, M.J. *et al.* Genehunter 2.0—a complete linkage analysis system. *Am. J. Hum. Genet.* **63**, A286 (1998).
23. Dempster, A.P., Laird, N.M. & Rubin, D.B. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc.* **39**, 1–38 (1977).
24. Excoffier L. & Slatkin M. Maximum–likelihood estimation of molecular haplotype frequencies in a diploid population. *Mol. Biol. Evol.* **12**, 921–927 (1995).
25. Clark A.G. Inference of haplotypes from PCR-amplified samples of diploid populations. *Mol. Biol. Evol.* **7**, 111–122 (1990).