HAPLOTYPE BLOCKS AND LINKAGE DISEQUILIBRIUM IN THE HUMAN GENOME

Jeffrey D. Wall[*‡] *and Jonathan K. Pritchard*[*]

There is great interest in the patterns and extent of linkage disequilibrium (LD) in humans and other species. Characterizing LD is of central importance for gene-mapping studies and can provide insights into the biology of recombination and human demographic history. Here, we review recent developments in this field, including the recently proposed 'haplotype-block' model of LD. We describe some of the recent data in detail and compare the observed patterns to those seen in simulations.

BOTTLENECK
A temporary reduction in population size that causes the loss of genetic variation.

ADMIXTURE
The mixture of two or more genetically distinct populations.

*Department of Human Genetics, The University of Chicago, 920 East 58th Street, CLSC 507, Chicago, Illinois 60637, USA.
‡Current address: Program in Molecular and Computational Biology, The University of Southern California, 1042 West 36th Place, DRB 289, Los Angeles, California 90089-1113, USA.
e-mails: jwall@genetics.bsd.uchicago.edu; pritch@uchicago.edu
doi:10.1038/nrg1123

Linkage disequilibrium (LD) refers to the fact that particular alleles at nearby sites can co-occur on the same haplotype more often than is expected by chance[1–5] (BOX 1). LD is of fundamental importance in gene mapping because it is used in positional cloning to track down variation that has produced a linkage signal[6,7], and in association studies in which disease variants can be detected through the presence of association at nearby sites[8–10]. Patterns of LD can also be used to infer the distribution of crossing-over events at short scales that are difficult to study experimentally[11,12], and to study gene conversion, about which there are only sparse experimental data in any animal species[13–16]. Finally, patterns of LD are important for untangling the evolutionary history of humans, which includes the identification of demographic effects such as population growth, BOTTLENECKS and ADMIXTURE[15,17–23], and the detection of natural selection[24–26].

In this review, we focus on recent data on the spatial structure of LD and the implications for association mapping. In association mapping, the goal is to identify genetic variants that increase susceptibility to a disease (or other phenotype of interest), and are therefore at a higher frequency among affected individuals than among controls[10]. Under certain assumptions, theoretical arguments indicate that genome-wide association mapping can be a powerful approach for identifying the variants that contribute to complex traits[27,28]. At present, a study that genotyped all the common variants in the genome would be dauntingly expensive. However, as the genotypes at nearby markers are usually correlated (that is, they are in LD), it should be possible to scan the genome using a much smaller marker set, with only a modest loss of power[8]. To design studies that are appropriate for this task, it is necessary to have a detailed understanding of the structure and extent of LD across the genome, both to choose suitable marker sets and to design powerful methods of statistical analysis. Our review describes the data and models of LD in the human genome, and compares these with simulation results. It has been argued that variation in recombination rates is an essential determinant of LD in humans[11,12,29–32], and we also discuss this issue. We do not review the implications for human demography or discuss the experimental data on the extent of PAIRWISE LD, as several review articles have previously discussed these issues[1,4,5,22].

Linkage disequilibrium

Patterns of LD are well known for being noisy and unpredictable. For example, pairs of sites that are tens of kilobases apart might be in 'complete' LD, whereas nearby pairs of sites from the same region might be in weak LD. Similarly, there can be tremendous differences in the extent of LD from one genomic region to another[1,18,33–37]. Much of this apparent randomness is predicted by

population genetic models that describe LD[1,38–40], and some might be the result of fine-scale heterogeneity in recombination rates[11,12,29,30]. Population history also has a large impact on patterns of LD, with factors such as population structure or small population size leading to increased LD[1]. For example, it is consistently observed that LD in non-African populations extends over longer distances than in Africans, which might reflect a population bottleneck at the time when modern humans first left Africa[15,17,18,22,41]. Similarly, there have been reports that certain isolated or admixed populations show LD over large distances[19,21,42,43] (but see REF. 44 for conflicting data).

Despite the apparent complexity of observed patterns, recent studies have proposed that the underlying structure of LD in the human genome can be described using a relatively simple framework in which the data are parsed into a series of discrete haplotype blocks[31,32,45,46] (BOX 2). Neighbouring blocks are separated by regions of numerous recombination events[30–32]. The haplotype-block model has important implications for association mapping because it indicates a simple rationale for how to choose single nucleotide polymorphisms (SNPs) for large-scale association studies. The main haplotypes in each block could be labelled with a small number of 'haplotype-tagging' SNPs, which would provide an efficient mechanism for screening each haplotype-block region for association[32,45,47] (BOX 3). In response to these results, the United States National Human Genome Research Institute has recently initiated a major effort, called the International HapMap Project, which aims to create a genome-wide map of LD and haplotype blocks. The intention is that this project will facilitate large-scale association-mapping studies and positional-cloning studies by cataloguing LD across the genome in many populations.

Haplotype blocks

In 2001, Daly and colleagues[31] reported that the haplotype structure in a 500-kb region on chromosome 5q31 could be broken into a series of discrete haplotype blocks that range in size from 3–92 kb. Each haplotype block corresponded to a region in which there were just a few common haplotypes (2–4 per block), and for which the LD data provided little evidence for recombination. The inferred blocks were separated by regions with several inferred recombination events. Almost simultaneously, Jeffreys *et al.*[30] reported data from single-sperm typing that showed that much of the recombination in the class II major histocompatability complex (MHC) region was restricted to narrow recombination hotspots (see later). Although these observations were restricted to two genomic regions, taken together they suggest the intriguing hypothesis that the genome might be divided into regions of high LD that are separated by recombination hotspots.

Since the publication of those papers, several studies of much larger scope have also reported that the genome can be divided into blocks of high |D′| (BOX 1) or low haplotype diversity[32,45,46,48] (BOX 2). Three of these studies reported LD data for markers that spanned chromosomes 19 (REF. 48), 21 (REF. 45) and 22 (REF. 46), respectively, whereas Gabriel *et al.*[32] surveyed 51 different genomic regions for a total of ~13 Mb. These studies either used PRE-ASCERTAINED SNPS from public databases[32,46,48], which limited the resolution to a mean marker spacing of ~5–20 kb or, in the case of Patil *et al.*[45], used chip-based resequencing on a relatively small sample of 20 chromosomes.

All of these surveys found a small number of extremely long haplotype blocks. Among the published studies, the longest reported block is a region of low haplotype diversity on chromosome 22 that, in individuals

Box 1 | Measuring linkage disequilibrium

Many different measures have been proposed for assessing the strength of linkage disequilibrium (LD). Most capture the strength of association between pairs of biallelic sites. Two important pairwise measures of LD are r^2 (sometimes denoted Δ^2) and |D′|[1,2,82]. Both measures range from 0 (no disequilibrium) to 1 ('complete' disequilibrium), but their interpretation is slightly different. |D′| is defined in such a way that it is equal to 1 if just two or three of the possible haplotypes are present, and it is <1 if all four possible haplotypes are present. So, a value of |D′| that is <1 indicates that historical recombination has occurred between two sites[83] (recurrent mutation can also cause |D′| to be <1, but for single nucleotide polymorphisms (SNPs) this is usually regarded as being less likely than recombination). Intermediate values of |D′| are more difficult to interpret (for example, how different is 0.3 from 0.7?), and even in simulations, values of |D′| can be highly variable for pairs of sites that are separated by a given distance[1,38,39]. The measure r^2 represents the statistical correlation between two sites, and takes the value of 1 if only two haplotypes are present. It is arguably the most relevant measure for association mapping, because there is a simple inverse relationship between r^2 and the sample size required to detect association between susceptibility loci and SNPs. For example, suppose that SNP1 is involved in disease susceptibility, but we genotype cases and controls at a nearby site SNP2. Then, to achieve the same power to detect association at SNP2 as we would have at SNP1, we need to increase our sample size by a factor of $1/r^2$ (REFS 1,70).

These measures are defined for pairs of sites, but for some applications we might instead want to measure how strong LD is across an entire region that contains many polymorphic sites — for example, for testing whether the strength of LD differs significantly among loci or across populations, or whether there is more or less LD in a region than predicted under a particular model. Measuring LD across a region is not straightforward, but one approach is to use the measure ρ, which was developed in population genetics[1,84,85]. Roughly speaking, ρ measures how much recombination would be required under a particular population model to generate the LD that is seen in the data. The development of methods for estimating ρ is now an active research area[12,39,85–90]. This type of method can potentially also provide a statistically rigorous approach to the problem of determining whether LD data provide evidence for the presence of hotspots[12].

PAIRWISE LINKAGE DISEQUILIBRIUM (Pairwise LD). The strength of association between alleles at two different markers.

PRE-ASCERTAINED SINGLE NUCLEOTIDE POLYMORPHISMS (Pre-ascertained SNPs). SNPs that have already been detected in previous studies, usually from an extremely small sample of chromosomes.

Box 2 | Definitions of haplotype blocks

A range of methods have been proposed for defining haplotype blocks. Broadly speaking, they can be classified into two main groups: those that define blocks as regions with limited haplotype diversity[31,45,46,91] and those that make use of pairwise disequilibrium (for example, based on |D′|) to identify transition zones in which there is evidence for extensive historical recombination[32,48,72,78].

The details of the proposed algorithms differ from study to study, which makes the comparison of results from different studies challenging. As examples, we describe one particular definition of each type. The first, from Patil *et al.*[45,91,92], defines a haplotype block as a region in which a fraction 'α' or more of all the observed haplotypes are represented at least *n* times in the sample. So, for example, Patil *et al.*[45] required that in haplotype blocks, at least 80% of the observed haplotypes should be observed two or more times. Clearly, given this rule, there might be many possible ways of dividing the data into blocks. Patil *et al.* used the criterion that (roughly speaking) block boundaries should be defined in a way that minimizes the number of single nucleotide polymorphisms (SNPs) that are required to identify all the haplotypes in a region; Zhang *et al.*[91] have provided an efficient algorithm for doing this.

A different block definition was proposed by Gabriel *et al.*[32]. The authors focused on |D′| and defined haplotype blocks as sets of consecutive sites between which there is little or no evidence of historical recombination. More specifically, for each pair of sites, the data are used to construct a confidence interval on the population value of |D′|. This procedure approximately accounts for the uncertainty owing to finite sample size and UNPHASED DIPLOID DATA, and has the effect of substantially smoothing the estimates of |D′|, which are normally noisy (see REF. 93 for a BAYESIAN APPROACH to the same problem). Values of |D′| are divided into three categories: strong LD (|D′| near 1, which implies little or no evidence of historical recombination); weak LD (|D′| significantly <1, which implies historical recombination); and intermediate/unknown LD. The third category includes pairs of sites with intermediate values of |D′|, as well as pairs for which the confidence intervals are relatively wide. Two or more sites can be grouped together into a block if the outermost pair of sites is in strong LD, and if, for all pairwise comparisons in the block, the number of pairs in strong LD is at least 19-fold greater than the number of pairs in weak LD (for a full version of the original definition, see REF. 32). The authors sought to validate this definition by looking at the properties of sites that were not used to build the blocks, and observed that in blocks, the LD between such sites did not depend on distance. These criteria do not produce a unique assignment of sites to blocks, but in practice the fraction of ambiguous block boundaries is relatively low[71].

Although both approaches have their merits, we prefer the second for several reasons: first, using D′ focuses attention directly on the issue of detecting historical recombination, which seems to be central to the concept of haplotype blocks; second, the pairwise methods are more easily applied to diploid genotype data in which haplotype phase is unknown; and third, it is easy to visualize the pairwise disequilibrium coefficients (examples are shown in FIG. 1).

of European descent (data from the Centre d'Etude du Polymorphisme Humain) stretches across 804 kb[46]. Such long blocks are implausible under population genetic models with uniform recombination rates[1,48,49]; the simplest explanation is that these represent long regions of low recombination — recombination 'coldspots'[46]. Apart from these few long haplotype blocks, most of the reported blocks are much smaller (5–20 kb). As the size of these blocks is similar to the average distance between consecutive markers (except for the Patil *et al.*[45] study) the identification of smaller blocks is beyond the resolution of these studies.

These reports of haplotype blocks raise several questions. Do these results indicate that most recombination in the genome occurs in hotspots that generally correspond to haplotype-block boundaries (or conversely, that haplotype-block boundaries imply hotspots)? Also, to what extent does the haplotype-block model capture the underlying structure of LD, as opposed to being a convenient heuristic description? Finally, given the observed structure of LD, what is the best strategy for choosing SNPs for association mapping (BOX 3)? We consider each of these questions in turn.

UNPHASED DIPLOID DATA
Sequence data in which the phase of double heterozygotes was not determined.

BAYESIAN APPROACH
A statistical approach that, given a set of assumptions about the underlying model, can provide a rigorous assessment of uncertainty.

Experimental evidence for hotspots

As discussed earlier, an important component of the haplotype-block model is the possibility that much of the recombination in the genome might occur in narrow hotspots. Although recombination hotspots are ubiquitous in yeast[50], much less is known about hotspots in humans or other animals. Most of our knowledge about recombination-rate variation in humans is at much longer scales — usually centiMorgan distances or more.

Researchers have traditionally estimated recombination rates by comparing physical maps with genetic maps obtained from pedigree studies[51–53]. There is tremendous variation in recombination rates at centiMorgan scales within chromosomes, between chromosomes and between males and females[51–53]. The resolution of these studies is limited both by the number of meioses and the density of markers that were used to construct the genetic maps. The average distance between consecutive markers in the most accurate genetic map is ~600 kb[53]. So, these studies are not normally informative about variation in recombination rates at shorter scales (<100 kb), although a handful of small regions (<20 kb) have been identified by this approach as having greatly elevated recombination rates[54–56].

Studying variation in recombination rates at fine scales generally requires the examination of many meioses, because the frequency of recombination events in any narrow interval is small. One promising alternative approach has been to estimate recombination rates by genotyping sperm[29,30,57–63]. Although this only estimates male recombination rates (which might differ

Figure 1 | **Pairwise |D′| plots for representative regions from different studies.** Each square in the triangle plots the level of linkage disequilibrium (LD) between a pair of sites in a region; comparisons between neighbouring sites lie along the diagonal. Red colouring indicates strong LD, green indicates weak LD and light brown indicates intermediate or uninformative LD (see BOX 2 and REF. 32 for details). The long diagonal line indicates the physical length of the region, and the short black lines plot the position of each marker in this region. We include the physical length and estimated recombination rate[53] for each region. EGP, Environmental Genome Project; SNP, single nucleotide polymorphism.

Box 3 | Haplotype blocks and association mapping

The haplotype-block model immediately points to a relatively simple approach to designing mapping studies. First, the main haplotypes could be identified in each haplotype block, followed by the determination of the smallest set of single nucleotide polymorphisms (SNPs) that is needed to distinguish among these haplotypes (the haplotype-tagging SNPs)[32,47,92]. It would then be possible to scan across the region of interest by doing a chi-square test of association in each haplotype block, to test for association between phenotype and haplotype status. Gabriel *et al.*[32] estimated that approximately 300,000 and 1,000,000 SNPs would be required to scan the genome in non-African and African populations, respectively, by this approach — an estimate that is surprisingly similar to the theoretical estimate made by Kruglyak in 1999 (REF. 9).

Although this chi-square approach is appealing in its simplicity, it is not clear that this is either the most efficient or powerful statistical approach to the problem. In effect, this approach treats each haplotype block as independent, but in practice there might be substantial (although incomplete) LD from one block to the next[31]. If this is the case, further information about the relationships among chromosomes at one position can potentially be gleaned from the relationships among SNPs in neighbouring blocks[94].

Effective use of information from neighbouring blocks might be of particular value for identifying risk alleles that are at modest frequencies (for example, 1–10%), or loci at which there is modest allelic heterogeneity[95,96]. There is a concern that association mapping with haplotype-tag SNPs will have relatively low power to detect low frequency variants[3]. However, such variants are likely to be young, and hence might lie in conserved haplotypes that extend across several haplotype blocks. One potential signal of such risk alleles might be extended multi-block haplotypes that are shared among affected individuals more than among controls. So far, no methods have been published that can make systematic use of this type of information, and there is a need for new statistical techniques in this area.

Finally, it is clear that to some extent haplotype blocks are a double-edged sword. Large discrete blocks are a bonus in detecting association (the first phase of association mapping), but once a locus of interest has been narrowed down to a single large haplotype block, the patterns of LD might provide no further information about the actual location of disease variant(s)[79]. One possible approach is to first detect association in non-African populations, and then perform fine-mapping in African (or African-American) populations in which LD decays much faster[18,32], assuming that the same disease loci are polymorphic in both groups.

substantially from female recombination rates[51,53]), the advantage of sperm-typing studies is that they facilitate assays of extremely large numbers of meioses and therefore make studies of recombination-rate variation possible at fine scales. Several recent studies have found that recombination tends to cluster in hotspots that are roughly 1–2 kb in length[29,30,61]. However, since sperm typing is laborious and technically challenging, only a handful of regions have been examined so far, and the regions studied have generally been those for which there was previous evidence of recombination hotspots. Determining how representative of the rest of the genome these patterns of rate variation are will be an important avenue of future research.

Little is known about the molecular mechanism of recombination hotspots and how rapidly they appear and disappear over evolutionary time. There is evidence that some hotspots in yeast and humans are allele-specific, with the hotspot allele being more likely to initiate the double-strand break[64,65]. This eventually leads to the loss of the hotspot allele, which points to a mechanism by which hotspots can be lost[66]. Over longer timescales, recombination rates can change substantially: closely related *Drosophila* species can have different genetic maps[67], and the total map length in the baboon is ~20% smaller than the total human map length[68]. However, several studies have found a negative correlation between levels of LD and rates of recombination[11,12,29,30,56,61,62,69], which implies that recombination rates change slowly on time scales of *N* generations. One study found blocks of LD in coldspots that were separated by experimentally-determined recombination hotspots[30]. Although these patterns fit the haplotype-block model well, the particular region studied (MHC class II) is subject to strong diversifying selection and is not necessarily typical of the genome as a whole.

Patterns of LD in human data

In this section, we take a closer look at some of the large-scale genomic data on LD. It is difficult to compare the results of the existing LD studies directly, because of the variation in study designs and the range of methods used to analyse the data (BOX 2). The samples that have been considered range from single[46,48] to multiple populations[32,70], or single mixed samples (see REF. 45 and the Environmental Genome Project (EGP) SNP study), and the average distance between consecutive markers ranges from <1 kb[70] to >22 kb[46]. Moreover, most of the existing studies describe large-scale data sets, and it is difficult to get a good sense of what the data look like. For these reasons, we have used the data from three large studies (REFS 32,70 and the EGP SNP data) to examine some overall properties and create visual summaries of many representative regions. We chose these three studies partly because each provided LD information from many different regions (see later). Our analyses used the Gabriel *et al.*[32] block definition (further details are given in BOX 2).

The data. The Gabriel *et al.*[32] data that we could access consisted of SNP genotypes from 50 genomic regions, which span 12.2 Mb, for European-Americans, African-Americans, East Asians and sub-Saharan Africans. The sample sizes in the four populations ranged from 42–58

independent individuals, and the average marker densities ranged from 1 SNP (with a minor allele frequency of 0.1) per 6.1–6.7 kb. The Seattle SNP study[70] examined variation in and near genes that are candidates for involvement in inflammatory diseases. We analysed 85 loci that were downloaded from the University of Washington and Fred Hutchinson Cancer Research Center UW-FHCRC Variation Discovery Resource web site in October 2002. These loci had an average marker spacing of 665 bp, and covered a total of 1.5 Mb of sequence. The data were obtained by resequencing 24 unrelated African-Americans and 23 unrelated European-Americans[70]. To make the sample size more comparable to the other studies, we pooled all of the samples. The results are similar if the two populations are considered separately[71]. The EGP SNP data came from another large resequencing study based in the Nickerson laboratory at the University of Washington. We accessed 90 loci from the publicly accessible EGP SNPs web site in October 2002. These loci span 1.7 Mb of sequence, with an average marker spacing of 946 bp. The samples consist of 90 unrelated individuals of mixed ethnicity from the DNA Polymorphism Discovery Resource.

Together, the three studies comprise a range of different ethnic groups, sample sizes and marker densities. By comparing them, we can get a sense of the extent to which inferred haplotype-block patterns are affected by study design. For the analyses described next, we only considered sites with a minor allele frequency of 0.1 or greater.

Haplotype blocks. As an initial summary of the data, we tabulated the total proportion of sequence that was contained in haplotype blocks of various sizes. The results show systematic differences in the levels of LD across populations and studies (FIG. 2). In the Gabriel *et al.*[32] study, both the European-American and East Asian samples have more extensive haplotype blocks than the African-American and sub-Saharan African samples[32] (as discussed previously), but it is notable that in all four populations less than half of the total sequence is contained in identified haplotype blocks. By contrast, for the two resequencing studies, in which marker density is much higher, more of the sequence is contained in identified haplotype blocks. As we show later, marker density has a strong impact on the ability to detect small haplotype blocks. Note that the proportion of sequence that is contained in long blocks for the two resequencing studies is underestimated, owing to the limited sizes of the regions that were sequenced. This truncation effect is minimal in the Gabriel *et al.*[32] data, in which most of the regions were much longer.

To get a closer look at the data, we prepared plots that show the values of |D′| for all pairs of sites in a region (representative plots are shown in FIG. 1). Each pair of markers was categorized as being in strong LD (red), weak LD (green) or inconclusive (grey) (see BOX 2 and REF. 32 for details). In this type of figure, haplotype blocks should appear as triangular regions of red (or light brown) squares that sit against the diagonal. Plots for all 225 regions are available online at the Pritchard laboratory web site (by following the 'Data Archive' link).

These plots indicate extensive heterogeneity of LD within regions: areas of strong LD that correspond well to the haplotype-block concept are often bordered by equally large regions with little or no LD. This becomes clearer when patterns of LD are compared across multiple loci. Some regions have extensive blocks of LD, whereas others have only isolated markers in strong LD

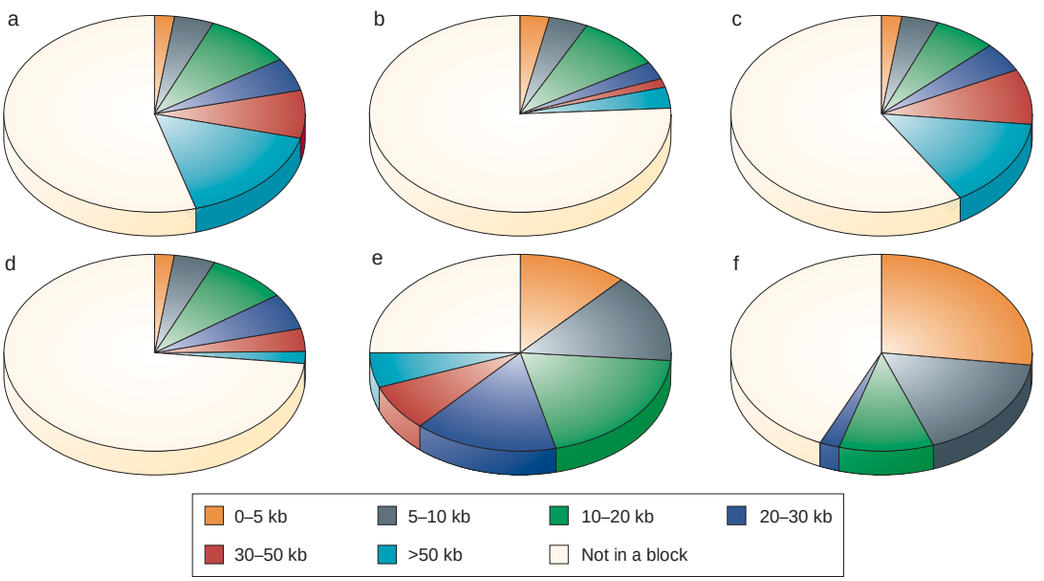

Figure 2 | **The proportion of sequence contained in haplotype blocks of various sizes. a** | European-American sample[32]. **b** | African-American sample[32]. **c** | East Asian sample[32]. **d** | Sub-Saharan African sample[32]. **e** | Environmental Genome Project (EGP) single nucleotide polymorphism (SNP) study. **f** | Seattle SNP study[72].

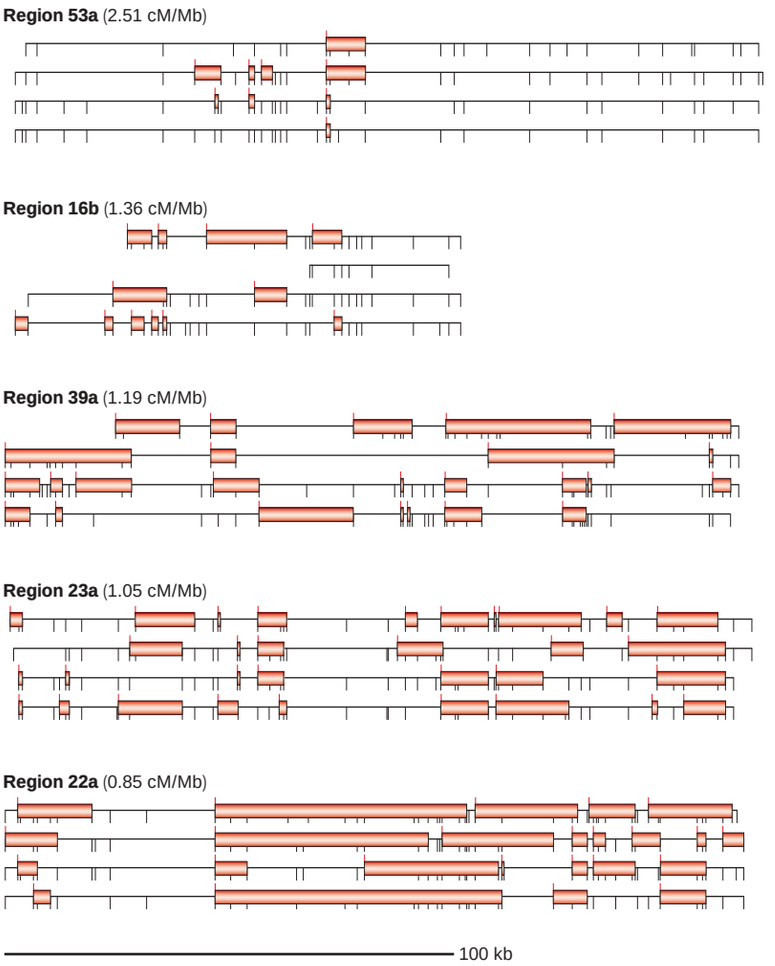

Figure 3 | **Schematic of the haplotype blocks identified in five genomic regions**[32]**.** The four lines for each region represent data from four population samples: European-Americans, East Asians, African-Americans and sub-Saharan Africans, respectively. The horizontal red lines denote haplotype blocks; the left-hand end of each block is marked by an upper red tick mark. The lower tick marks indicate the locations of the markers that were typed in each population.

with each other. It is clear from these figures that the extent of strong LD (red squares) is lower in African and African-American samples than in non-Africans. It is not straightforward to compare the resequencing studies to the Gabriel *et al.* data, because the marker density is different and the resequencing studies treated samples of mixed ethnicity. Our simulations and resampling experiments indicate that in such mixed samples the inferred block characteristics tend to be most similar to the populations with lowest LD (Africans and African-Americans, in this case).

To get a visual sense of the correspondence of haplotype blocks and block boundaries among different populations, we plotted the parts of each region that were contained in haplotype blocks for each of the four populations studied by Gabriel *et al.*[32]. Results for five representative genomic regions are plotted in FIG. 3. As might be expected, there is an inverse correlation between the proportion of sequence that is contained in haplotype blocks and the estimated recombination rate from REF. 53, with relatively fewer and smaller blocks identified in regions of high recombination (for example, see Region 53a in FIG. 3). The block boundaries often line up across populations; this presumably reflects, in part, the shared ancestry of human populations, but to some extent it might also reflect the indirect effect of uneven marker spacing (similar figures for all of the regions are available online at the Pritchard laboratory web site, by following the 'Data Archive' link).

How 'block-like' is LD? Given that any genotype data that show LD can potentially be parsed in haplotype blocks[72], an obvious question is to what extent does the haplotype-block concept provide a natural description of the underlying structure of LD in humans? Elsewhere, we have proposed three criteria to quantify how block-like the structure of LD is[71]. These criteria measure the proportion of sequence that is contained in haplotype blocks (called here the 'coverage'), the extent to which haplotype blocks are internally consistent and the extent of overlap or ambiguity in haplotype-block boundaries. For haplotype blocks to provide a suitable description of LD across a region, it might be expected that the identified blocks would be discrete, consistent and cover most of the region. As noted above, the haplotype-block coverage in existing data is typically not high[48] (FIG. 2), but can potentially be improved by using higher marker densities (and the precise level of the haplotype-block coverage also depends on the block definition). We also found moderate levels of internal inconsistency[71]: given two markers in strong LD with each other, a sizeable fraction of the markers that are in between show historical evidence of recombination[32] with one of the end markers (these are shown as green squares in regions of red in FIG. 1). By contrast, we found that the rate of overlap or ambiguity between blocks was low (however, a study using different methodology concluded that ambiguity in block boundaries was a more serious concern[73]). Taken together, these results indicate that the haplotype-block model might capture some of the prominent features of LD in a simple and intuitive way, but there is also scope for the development of more complex and accurate models of LD that might provide better power for association studies and other applications (see for example REF. 12).

LD and local recombination rates. As discussed above, an important component of the haplotype-block model is the hypothesis that much of the recombination in the genome occurs in narrow hotspots. To examine this issue, we performed COALESCENT SIMULATIONS of patterns of LD under models with and without recombination hotspots. When simulating hotspot models, it is most appropriate to hold the average recombination rate constant (so that the average rate matches pedigree estimates), but to assume that many or most of the recombination events are concentrated into hotspot regions. This means that under hotspot models, the background rate of recombination — for the majority of the sequence that lies outside hotspots — becomes lower than the genome average, and the average extent of LD is longer.

COALESCENT SIMULATION
A method of simulating data under a population genetic model.

Box 4 | The effects of study design on haplotype-block patterns

To explore the effect of study design on observed haplotype-block patterns, we ran simulations that were comparable to the Gabriel *et al.*[32] data from sub-Saharan Africa. We chose a model in which the proportion of sequence contained in haplotype blocks roughly matched the proportion in the actual data (small n, small θ in table below). Using the coalescent with recombination[97], and assuming a population size of $N = 10^4$, we simulated 100 replicates of all 50 regions with $n = 58$ unphased diploids (the same sample size as the sub-Saharan African data from Gabriel *et al.*) in which the mutation parameter θ was set to 7.84×10^{-5} per bp (chosen to produce, on average, one marker with a minor allele frequency of 0.1 per 6.5 kb, as in the actual data). We also ran simulations with an eight-fold greater sample size (large n, small θ), an eight-fold greater marker density (small n, large θ), and both an eight-fold greater sample size and marker density (large n, large θ). The marker density with large θ is less than (but close to) the theoretical maximum marker density that could be obtained by complete resequencing. The underlying simulated genealogies were identical for all four study designs. The average recombination rate for each region was estimated from REF. 53, but the local recombination rate varied across the sequence[12,71] so that ~50% of all recombination events happened in randomly distributed 1-kb hotspots. For the table below, all hotspots were of equal intensity, but rates for FIG. 4 were drawn from an exponential distribution. For all simulations, there was an average of one hotspot per 30 kb. To model ASCERTAINMENT BIAS, we only considered polymorphisms that segregated in the first eight chromosomes. With this model, the marker-allele frequencies in the simulations match the actual marker-allele frequencies reasonably well[71] — this is important because different ascertainment schemes can produce different estimates of linkage disequilibrium (LD)[98].

We also explored the effect of local variation in the recombination rate on block patterns by running simulations similar to those above, but with uniform recombination rates for each region. These simulations still incorporated variation in recombination rates between regions.

Simulation results, averaged across all replicates and all regions, are summarized in the table. We present the range of the middle 90% of simulation replicates for the average marker spacing, sequence coverage, average haplotype-block size and largest haplotype-block size. The values from the actual data are given for comparison. Simulations with uniform recombination rates produce fewer (and shorter) haplotype blocks than are seen in the actual data. Similarly, levels of LD are higher in the actual data than expected under a model with no local variation in recombination rates. As discussed in the text, hotspot models produce more extensive LD (for example, longer haplotype blocks and greater sequence coverage) than comparable uniform recombination models.

Simulation results modelled after African data from the Gabriel *et al.*[32] study

Study design*	Average marker spacing (bp)‡	Sequence coverage§	Average block size (bp)	Largest block size (bp)‖
Actual data	6,480	0.267	9,623	67,765
Simulations				
Uniform recombination	6,024–6,725	0.179–0.217	5,399–6,477	36,674–84,088
Small n, small θ	6,018–6,711	0.236–0.279	6,965–8,170	41,149–84,478
Large n, small θ	6,040–6,781	0.314–0.360	7,638–8,812	45,636–90,214
Small n, large θ	799–872	0.698–0.724	4,895–5,346	41,652–88,395
Large n, large θ	806–872	0.748–0.773	5,521–5,941	43,357–89,506

Results refer to the 5th and 95th percentile of the distribution of haplotype-block summaries across different simulation replicates.*Study design for the simulations, where n is the sample size and θ is proportional to the marker density: 'small' values are as in the actual data, whereas 'large' values are eight-fold larger. The uniform recombination simulations have small n and small θ; all other simulations use the hotspot model described above.‡The average distance between consecutive markers.§The proportion of sequence contained in haplotype blocks.‖The size of the largest block (across regions, for a single replicate), compared with the largest block in the actual data.

Even for data that are simulated in the absence of hotspots, it is possible to identify haplotype blocks[48,71,72], but these will generally be smaller and have lower coverage than simulations in which most recombination is restricted to hotspots. A previous study indicated that patterns of LD from chromosome 19 might fit a model of uniform recombination reasonably well, with just a small excess of large haplotype blocks indicating long recombination coldspots[48]. By contrast, using slightly different methods (BOX 4), our own simulations[71] indicate that models with recombination hotspots provide a substantially better fit to the Gabriel *et al.* data[32]. If we view the proportion of sequence contained in haplotype blocks, and the length distribution of haplotype blocks as global measures of LD, then we find more LD in actual data than expected under a model with no recombination-rate variation. This result holds even in sub-Saharan African populations, which are more likely to fit the simple demographic model used in the simulations[15,74]. By contrast, simulations of a model under which most recombination occurs in hotspots provide a much better fit to the observed LD data[71]. A similar conclusion was reached by Reich *et al.*[37] using correlations in polymorphism rates across the genome.

If it is true that recombination hotspots are a major feature of the genome, then a natural question is whether the boundaries between haplotype blocks usually occur at hotspots. To investigate this further, we simulated data with the same length and average recombination rate as region 23a (REF. 32), in which the average recombination

ASCERTAINMENT BIAS
The bias in patterns of variation that results from using pre-ascertained SNPs.

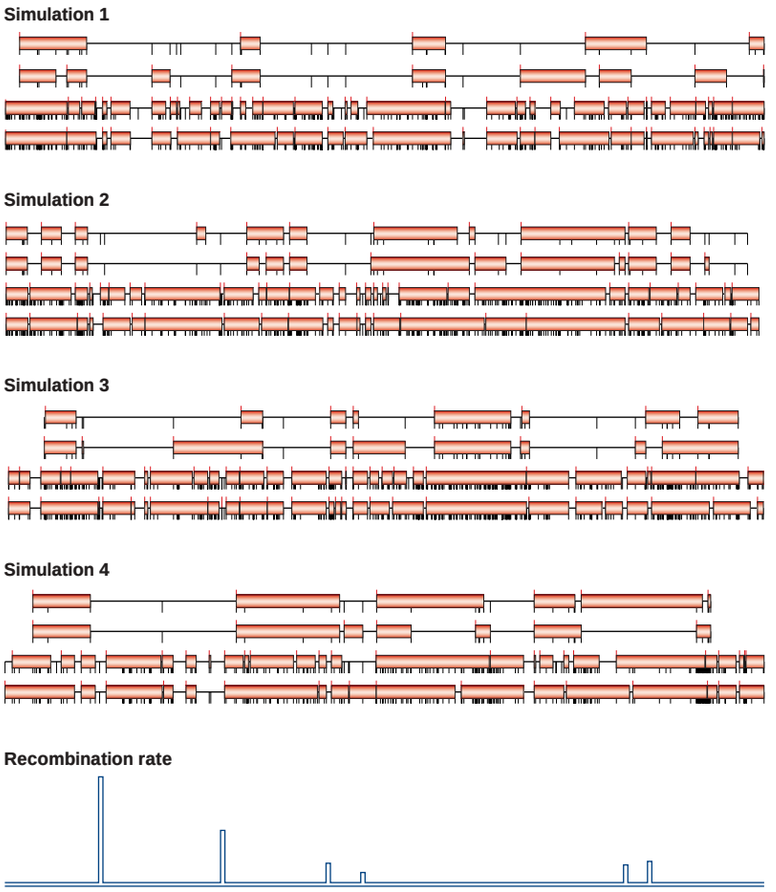

Figure 4 | **Schematic of the haplotype blocks found in simulations.** The parameters in the top line for each simulation were chosen to mimic region 23a from the sub-Saharan African population in Gabriel *et al.*[32]. The parameters in the lower lines have (relative to the first line): eight-fold greater sample size; eight-fold greater marker density; and both eight-folder greater sample size and marker density. The horizontal red bars denote haplotype blocks and the red upper tick marks show the left-most endpoint of each haplotype block. The lower tick marks show the locations of the markers that were typed in each population. In the bottom panel, we show the relative recombination rates (across the sequence) assumed in the simulations. Approximately 50% of all recombination events happen in the six hotspots.

rate (1.05 cM/Mb) is close to the genome average. We assumed a model with local variation in recombination rates that provided a good overall fit[71] to the African data of Gabriel *et al.*[32] — 50% of all recombination events occurred in hotspots. In our simulations there were six hotspots spread over 175 kb. For simplicity, our simulations do not consider GENE CONVERSION, which is believed to be an important feature in disrupting patterns of LD at short scales[15] (see BOX 4 for further simulation details). It should be noted that because the simulations are designed to most closely fit the African data, it is likely that haplotype blocks in non-Africans will be longer and easier to detect than in these simulations.

FIGURE 4 shows the haplotype-block patterns for four different replicates, along with a graph of the relative recombination rates for the region (plots of the extra replicates are available online at the Pritchard laboratory web site). For each replicate, the four lines correspond to different study designs with the same underlying data (described in BOX 4 as simulations 1–4). From these and other simulation results[48,71,72] it is clear that most haplotype-block boundaries do not occur at hotspots, even if the background rate of recombination is low (but not zero). However, the converse is more often true: in these simulation examples, the strong hotspot (on the far left) creates block boundaries in every case, while the weaker hotspots create block boundaries only some of the time. In summary, identifying haplotype-block boundaries is unlikely to be an accurate way of identifying hotspots, although it might be possible to identify them using more thorough analyses of LD patterns[12].

Simulations can also provide insight into the impact of differences in study design (BOX 4). Several long haplotype blocks in simulations 3 and 4 in BOX 4, which have an eight-fold greater marker density, are missed completely when sparser marker density is used. Overall, an eight-fold greater sample size increases haplotype-block coverage levels only slightly, whereas an eight-fold greater marker density more than doubles coverage levels — consistent with the results from real data shown in FIG. 2. Most of this coverage gain comes from the identification of smaller blocks that were missed with sparser marker density. BOX 4 shows that the average block size decreases by ~30% when the marker density is increased eight-fold[48,71]. The contrast between the different simulations in BOX 4 highlights the strong effects of study design on the apparent fit of the haplotype-block model. Even though each group of four lines in FIG. 4 was produced from the same underlying simulation, the interpretation of haplotype-block structure for this region would be different depending on the sample size and marker density used.

Conclusions

Understanding the structure of LD across the human genome is a vital task on the road to unravelling the genetics of complex traits in humans. Interpreting patterns of LD is important both for large-scale association mapping and for the final stages of positional-cloning studies. Just a few years ago, there were few empirical data on the average extent of LD and our best information came from simulation studies[9]. Since then, a series of large empirical studies have greatly augmented our knowledge of the extent and structure of LD[18,32,45,46,48].

Some of the key observations on the LD patterns are as follows. First, the average extent of LD in non-African populations is much greater than in Africans[15,17,18,32]. LD in non-Africans also extends further than expected from simple models[1,9,15,18], which possibly reflects the impact of a population bottleneck associated with the founding and spread of fully modern humans from Africa[17,18,75–77], whereas LD in Africans seems to fit a simple demographic model more closely[15,71] (it should be noted that most of these results are based on samples from just a handful of populations: Europeans, East Asians, African Americans and two west-African populations). Second, the level of LD varies a great deal among different regions of the genome[18,46]. Part of this variability can be explained by variation in large-scale recombination rates derived from genetic maps (see FIG. 3 for

GENE CONVERSION
Recombination that involves the nonreciprocal transfer of information from one sister chromatid to another.

example), or other genomic features[18,46], but much of the variability is not yet accounted for. Some of the remaining variability presumably stems from fine-scale variation in recombination rates that is not detectable by genetic maps, and some from the inherent stochastic nature[1] of LD. Third, all of the large-scale studies have detected some large blocks of LD (for example, 804 kb[46]). These probably reflect large coldspots of recombination (alternatively, if it is true that most recombination in the genome occurs in hotspots, these might be large regions without hotspots). Fourth, there are a handful of well-characterized recombination hotspots, especially in the class-II MHC region[30], in which most recombination occurs in just a handful of narrow hotspots. It is not yet clear whether this region is typical of the genome as a whole and conclusions drawn from studies of LD are inconsistent[37,48,71].

This brings us to the question of whether the haplotype-block model provides a 'good' description of LD in the human genome? This is not a completely well-posed question: first, because the idea of haplotype blocks has been interpreted in many ways[31,32,45,46,48,72,73,78]; and second, as no model is perfect, it is unclear what represents an acceptable fit between model and data. Undaunted, we offer the following observations.

One way forward is to define formal criteria that can be used to decide whether haplotype blocks accurately describe LD data. One choice of criteria is described above (see also REF. 71). According to these criteria, the available data show non-trivial departures from the haplotype-block model, but they still fit the criteria substantially better than expected under models of uniform recombination. Whether the observed departures are large enough to invalidate the haplotype-block model is a matter of personal choice.

Application of these criteria provides an overall view of the structure of LD across many regions. However, this summary analysis hides the tremendous variability across loci in the extent and nature of LD. To get a qualitative view of patterns of LD across the genome, we created pairwise LD plots of all 225 regions in the Gabriel *et al.*[32], Seattle SNP[70] and EGP data sets (see FIG. 1 for examples). We have found these plots to be extremely valuable for getting a general sense of LD patterns, and we encourage readers to scroll through them (available online at the Pritchard laboratory web site, by following the 'Data Archive' link). What becomes clear from looking at these plots is that there are some regions that seem to fit the haplotype-block concept well, and some regions in which the structure of LD is more complicated and the block description seems less natural. Another feature of the data is that in many regions, the overall extent of LD is limited, so that many or most of the markers are not in identified haplotype blocks.

So, how do the available data indicate that we should think about either positional cloning or large-scale association studies? Certainly, where there are large and well-defined haplotype blocks, their presence provides important information for mapping studies. In large blocks, a small number of well-chosen haplotype-tag SNPs can potentially capture much of the available information about association across many kilobases[47] (of course, the downside is that within blocks, LD provides no information to help localize the actual variants of interest[79]). But what of regions in which there are no large well-defined haplotype blocks? For example, in REF. 32, less than half of the total sequence was assigned to haplotype blocks. Increasing the marker density would allow much more of the sequence to be assigned to blocks, but most of the added blocks would be small (REF. 48; FIG. 4; BOX 4). So, a mapping strategy that aims to completely cover the genome by tagging every haplotype block would be wasteful. Instead, it makes most sense to envision a dual strategy whereby we use haplotype-tagging SNPs to mark large haplotype blocks, but elsewhere we need to be more flexible and make efficient use of multipoint information with partial LD between markers (BOX 3; REFS 80,81). The development of analytical methods to do this should be valuable not only in disease association studies but also in human evolutionary studies.

1. Pritchard, J. K. & Przeworski, M. Linkage disequilibrium in humans: models and data. *Am. J. Hum. Genet.* **69**, 1–14 (2001).
This paper discusses ways of quantifying LD, and explores how LD is affected by different demographic models.
2. Devlin, B. & Risch, N. A comparison of linkage disequilibrium measures for fine-scale mapping. *Genomics* **29**, 311–322 (1995).
3. Cardon, L. R. & Abecasis, G. R. Using haplotype blocks to map human complex trait loci. *Trends Genet.* **19**, 135–140 (2003).
4. Jorde, L. B. Linkage disequilibrium and the search for complex disease genes. *Genome Res.* **10**, 1435–1444 (2000).
5. Ardlie, K. G., Kruglyak, L. & Seielstad, M. Patterns of linkage disequilibrium in the human genome. *Nature Rev. Genet.* **3**, 299–309 (2002).
6. Kerem, B. *et al.* Identification of the cystic fibrosis gene: genetic analysis. *Science* **245**, 1073–1080 (1989).
7. Hastbacka, J. *et al.* Linkage disequilibrium mapping in isolated founder populations: diastrophic dysplasia in Finland. *Nature Genet.* **2**, 204–211 (1992).
8. Collins, F. S., Guyer, M. S. & Chakravarti, A. Variations on a theme: cataloging human DNA sequence variation. *Science* **278**, 1580–1581 (1997).
9. Kruglyak, L. Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nature Genet.* **22**, 139–144 (1999).
10. Risch, N. J. Searching for genetic determinants in the new millennium. *Nature* **405**, 847–856 (2000).
11. Chakravarti, A. *et al.* Nonuniform recombination within the human β-globin gene cluster. *Am. J. Hum. Genet.* **36**, 1239–1258 (1984).
12. Li, N. & Stephens, M. A new multilocus model for linkage disequilibrium, with application to exploring variations in recombination rate. *Genetics* (in the press).
This study provides an innovative approach to modelling LD, and introduces a powerful new method for quantifying local variation in levels of LD.
13. Hilliker, A. J. *et al.* Meiotic gene conversion tract length distribution within the *rosy* locus of *Drosophila melanogaster*. *Genetics* **137**, 1019–1026 (1994).
14. Przeworski, M. & Wall, J. D. Why is there so little intragenic linkage disequilibrium in humans? *Genet. Res.* **77**, 143–151 (2001).
15. Frisse, L. *et al.* Gene conversion and different population histories may explain the contrast between polymorphism and linkage disequilibrium levels. *Am. J. Hum. Genet.* **69**, 831–843 (2001).
This paper quantifies differences in levels of LD across populations, and provides the first estimates of gene-conversion rates in humans.
16. Ardlie, K. *et al.* Lower-than-expected linkage disequilibrium between tightly linked markers in humans suggests a role for gene conversion. *Am. J. Hum. Genet.* **69**, 582–589 (2001).
17. Tishkoff, S. A. *et al.* Global patterns of linkage disequilibrium at the *CD4* locus and modern human origins. *Science* **271**, 1380–1387 (1996).
18. Reich, D. E. *et al.* Linkage disequilibrium in the human genome. *Nature* **411**, 199–204 (2001).
This paper is the first genomic-scale study to document the variability in levels of LD across different populations and genetic regions.
19. McKeigue, P. M., Carpenter, J. R., Parra, E. J. & Shriver, M. D. Estimation of admixture and detection of linkage in admixed populations by a Bayesian approach: application to African-American populations. *Ann. Hum. Genet.* **64**, 171–186 (2000).
20. Falush, D. *et al.* Traces of human migrations in *Helicobacter pylori* populations. *Science* **299**, 1582–1585 (2003).
21. Falush, D, Stephens, M. & Pritchard, J. K. Inference of population structure: extensions to linked loci and correlated allele frequencies. *Genetics* (in the press).
22. Wall, J. D. Insights from linked single nucleotide polymorphisms: what we can learn from linkage disequilibrium. *Curr. Opin. Genet. Dev.* **11**, 647–651 (2001).
23. Wall, J. D. Detecting ancient admixture in humans using sequence polymorphism data. *Genetics* **154**, 1271–1279 (2000).

24. Saunders M. A., Hammer, M. F. & Nachman, M. W. Nucleotide variability at *G6PD* and the signature of malarial selection in humans. *Genetics* **162**, 1849–1861 (2002).
25. Tishkoff, S. A. *et al.* Haplotype diversity and linkage disequilibrium at human *G6PD*: recent origin of alleles that confer malarial resistance. *Science* **293**, 455–462 (2001).
This paper, along with references 24 and 26, shows how recent natural selection can affect patterns of LD.
26. Sabeti, P. C. *et al.* Detecting recent positive selection in the human genome from haplotype structure. *Nature* **419**, 832–837 (2002).
27. Risch, N. & Merikangas, K. The future of genetic studies of complex human diseases. *Science* **273**, 1516–1517 (1996).
28. Camp, N. J. Genomewide transmission/disequilibrium testing — consideration of the genotypic relative risks at disease loci. *Am. J. Hum. Genet.* **61**, 1424–1430 (1997).
29. 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).
30. Jeffreys, A. J., Kauppi, L. & Neumann, R. Intensely punctuate meiotic recombination in the class II region of the major histocompatibility complex. *Nature Genet.* **29**, 217–222 (2001).
This high-resolution experimental analysis shows that most recombination events in the class II MHC region occur in just a handful of narrow hotspots.
31. Daly, M., Rioux, J. D., Schaffner, D. F., Hudson, T. J. & Lander, E.S. High-resolution haplotype structure in the human genome. *Nature Genet.* **29**, 229–232 (2001).
The notable patterns of LD in this study spurred interest in the haplotype-block concept.
32. Gabriel, S. B. *et al.* The structure of haplotype blocks in the human genome. *Science* **296**, 2225–2229 (2002).
This study explores haplotype-block patterns across many populations and genomic regions.
33. Taillon-Miller, P. *et al.* Juxtaposed regions of extensive and minimal linkage disequilibrium in human Xq25 and Xq28. *Nature Genet.* **25**, 324–328 (2000).
34. Dunning, A. M. *et al.* The extent of linkage disequilibrium in four populations with distinct demographic histories. *Am. J. Hum. Genet.* **67**, 1544–1554 (2000).
35. Abecasis, G. R. *et al.* Extent and distribution of linkage disequilibrium in three genomic regions. *Am. J. Hum. Genet.* **68**, 191–197 (2001).
36. Bonnen, P. E., Wang, P. J., Kimmel, M., Chakraborty, R. & Nelson, D. L. Haplotype and linkage disequilibrium architecture for human cancer-associated genes. *Genome Res.* **12**, 1846–1853 (2002).
37. Reich, D. E. *et al.* Human genome sequence variation and the influence of gene history, mutation and recombination. *Nature Genet.* **32**, 135–142 (2002).
38. Hudson, R. R. The sampling distribution of linkage disequilibrium under an infinite allele model without selection. *Genetics* **109**, 611–631 (1985).
39. Hudson, R. R. Two-locus sampling distributions and their application. *Genetics* **159**, 1805–1817 (2001).
40. Nordborg, M. & Tavare, S. Linkage disequilibrium: what history has to tell us. *Trends Genet.* **18**, 83–90 (2002).
41. Weiss, K. M. & Clark, A. G. Linkage disequilibrium and the mapping of complex human traits. *Trends Genet.* **18**, 19–24 (2002).
42. Laan, M. & Pääbo, S. Demographic history and linkage disequilibrium in human populations. *Nature Genet.* **17**, 435–438 (1997).
43. Kaessmann, H. *et al.* Extensive linkage disequilibrium in small human populations in Eurasia. *Am. J. Hum. Genet.* **70**, 673–685 (2002).
44. Eaves, I. A. *et al.* The genetically isolated populations of Finland and Sardinia may not be a panacea for linkage disequilibrium mapping of common disease genes. *Nature Genet.* **25**, 320–323 (2000).
45. Patil, N. *et al.* Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science* **294**, 1719–1723 (2001).
46. Dawson, E. *et al.* A first generation linkage disequilibrium map of human chromosome 22. *Nature* **418**, 544–548 (2002).
47. Johnson, G. C. *et al.* Haplotype tagging for the identification of common disease genes. *Nature Genet.* **29**, 233–237 (2001).
This paper explores how haplotype tag SNPs might aid future association studies.
48. Phillips, M. S. *et al.* Chromosome-wide distribution of haplotype blocks and the role of recombination hotspots. *Nature Genet.* **33**, 382–387 (2003).
49. Innan, H., Padhukasahasram, B. & Nordborg, M. The pattern of polymorphism on human chromosome 21. *Genome Res.* **13**, 1158–1168 (2003).
50. Petes, T. D. Meiotic recombination hot spots and cold spots. *Nature Rev. Genet.* **2**, 360–369 (2001).
51. Broman, K. W., Murray, J. C., Sheffield, V. C., White, R. L. & Weber, J. L. Comprehensive human genetic maps: individual and sex-specific variation in recombination. *Am. J. Hum. Genet.* **63**, 861–869 (1998).
52. Yu, A. *et al.* Comparison of human genetic and sequence-based physical maps. *Nature* **409**, 951–953 (2001).
53. Kong, A. *et al.* A high-resolution recombination map of the human genome. *Nature Genet.* **31**, 241–247 (2002).
54. 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).
55. Yip, S. P., Lovegrove, J. U., Rana, N. A., Hopkinson, D. A. & Whitehouse, D. B. Mapping recombination hotspots in human phosphoglucomutase (*PGM1*). *Hum. Mol. Genet.* **8**, 1699–1706 (1999).
56. Badge, R. M., Yardley, J., Jeffreys, A. J. & Armour, J. A. Crossover breakpoint mapping identifies a subtelomeric hotspot for male meiotic recombination. *Hum. Mol. Genet.* **9**, 1239–1244 (2000).
57. Li, H. H. *et al.* Amplification and analysis of DNA sequences in single human sperm and diploid cells. *Nature* **335**, 414–417 (1988).
58. Hubert, R., MacDonald, M., Gusella, J. & Arnheim, N. High resolution localization of recombination hot spots using sperm typing. *Nature Genet.* **7**, 420–424 (1994).
59. Jeffreys, A. J., Murray, J. & Neumann, R. High-resolution mapping of crossovers in human sperm defines a minisatellite-associated recombination hotspot. *Mol. Cell* **2**, 267–273 (1998).
60. Lien, S., Szyda, J. Schechinger, B., Rappold, G. & Arnheim, N. Evidence for heterogeneity in recombination in the human pseudoautosomal region: high resolution analysis by sperm typing and radiation-hybrid mapping. *Am. J. Hum. Genet.* **66**, 557–566 (2000).
61. May, C. A., Shone, A. C., Kalaydjieva, L., Sajantila, A. & Jeffreys, A. J. Crossover clustering and rapid decay of linkage disequilibrium in the Xp/Yp pseudoautosomal gene *SHOX*. *Nature Genet.* **31**, 272–275 (2002).
62. Schneider, J. A., Peto, T. E., Boone, R. A., Boyce, A. J. & Clegg, J. B. Direct measurement of the male recombination fraction in the human β-globin hot spot. *Hum. Mol. Genet.* **11**, 207–215 (2002).
63. Arnheim, N., Calabrese, P. & Nordborg, M. 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 (2003).
64. Nicolas, A., Treco, D., Schultes, N. P. & Szostak, J. W. An initiation site for meiotic gene conversion in the yeast *Saccharomyces cerevisiae*. *Nature* **338**, 35–39 (1989).
65. Jeffreys, A. J. & Neumann, R. Reciprocal crossover asymmetry and meiotic drive in a human recombination hot spot. *Nature Genet.* **31**, 267–271.
66. Boulton, A., Myers, R. S. & Redfield, R. J. The hotspot conversion paradox and the evolution of meiotic recombination. *Proc. Natl Acad. Sci. USA* **94**, 8058–8063 (1997).
67. True, J. R., Mercer, J. M. & Laurie, C. C. Differences in crossover frequency and distribution among three sibling species of *Drosophila*. *Genetics* **142**, 507–523 (1996).
68. Rogers, J. *et al.* A genetic linkage map of the baboon (*Papio hamadryas*) genome based on human microsatellite polymorphisms. *Genomics* **67**, 237–247 (2000).
69. Kauppi, L., Sajantila, A. & Jeffreys, A. J. Recombination hotspots rather than population history dominate linkage disequilibrium in the MHC class II region. *Hum. Mol. Genet.* **12**, 33–40 (2003).
70. Carlson, C. S. *et al.* Additional SNPs and linkage-disequilibrium analysis in whole-genome association studies in humans. *Nature Genet.* **33**, 518–521 (2003).
71. Wall, J. D. & Pritchard, J. K. Assessing the performance of the haplotype block model of linkage disequilibrium. *Am. J. Hum. Genet.* (in the press).
72. Wang, N. Akey, J. M., Zhang, K., Chakraborty, R. & Jin, L. 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 (2002).
73. Schwartz, R., Halldorsson, B. V., Bafna, V., Clark, A. G. & Istrail, S. Robustness of inference of haplotype block structure. *J. Comp. Biol.* **10**, 13–19 (2003).
74. Pluzhnikov, A., Di Rienzo, A. & Hudson, R. R. Inferences about human demography based on multilocus analyses of noncoding sequences. *Genetics* **161**, 1209–1218 (2002).
75. Cann, R. L., Stoneking, M. & Wilson, A. C. Mitochondrial DNA and human evolution. *Nature* **325**, 31–36 (1987).
76. Stringer, C. B. & Andrews, P. Genetic and fossil evidence for the origin of modern humans. *Science* **239**, 1263–1268 (1988).
77. Wall, J. D., Andolfatto, P. & Przeworski, M. Testing models of selection and demography in *Drosophila simulans*. *Genetics* **162**, 203–216 (2002).
78. Stumpf, M. P. & Goldstein, D. B. Demography, recombination hotspot intensity, and the block structure of linkage disequilibrium. *Curr. Biol.* **13**, 1–8 (2003).
79. Rioux, J. D. *et al.* Genetic variation in the 5q31 cytokine gene cluster confers susceptibility to Crohn disease. *Nature Genet.* **29**, 223–228 (2001).
80. McPeek, M. S. & Strahs, A. Assessment of linkage disequilibrium by the decay of haplotype sharing, with application to fine-scale genetic mapping. *Am. J. Hum. Genet.* **65**, 858–875 (1999).
81. Morris, A. P., Whittaker, J. C. & Balding, D. J. Fine-scale mapping of disease loci via shattered coalescent modeling of genealogies. *Am. J. Hum. Genet.* **70**, 686–707 (2002).
82. Lewontin, R. C. The interaction of selection and linkage. I. General considerations: heterotic models. *Genetics* **49**, 49–67 (1964).
83. Hudson, R. R. & Kaplan, N. L. Statistical properties of the number of recombination events in the history of a sample of DNA sequences. *Genetics* **111**, 147–164 (1985).
84. Long, A. D. & Langley, C. H. The power of association studies to detect the contribution of candidate genetic loci to variation in complex traits. *Genome Res.* **9**, 720–731 (1999).
85. Hudson, R. R. Estimating the recombination parameter of a finite population model without selection. *Genet. Res.* **50**, 245–250 (1987).
86. McVean, G., Awadalla, P. & Fearnhead, P. A coalescent-based method for detecting and estimating recombination from gene sequences. *Genetics* **160**, 1231–1241 (2002).
87. Fearnhead, P. & Donnelly, P. Estimating recombination rates from population genetic data. *Genetics* **159**, 1299–1318 (2001).
88. Wall, J. D. A comparison of estimators of the population recombination rate. *Mol. Biol. Evol.* **17**, 156–163 (2000).
89. Kuhner, M. K., Yamato, J. & Felsenstein, J. Maximum likelihood estimation of recombination rates from population data. *Genetics* **156**, 1393–1401 (2000).
90. Griffiths, R. C. & Marjoram, P. Ancestral inference from samples of DNA sequences with recombination. *J. Comp. Biol.* **3**, 479–502 (1996).
91. Zhang, K., Deng, M., Chen, T., Waterman, M. S. & Sun, F. A dynamic programming algorithm for haplotype block partitioning. *Proc. Natl Acad. Sci. USA* **99**, 7335–7339 (2002).
92. Zhang, K., Calabrese, P., Nordborg, M. & Sun, F. Haplotype block structure and its applications to association studies: power and study designs. *Am. J. Hum. Genet.* **71**, 1386–1394 (2002).
93. Ayres, K. L. & Balding, D. J. Measuring gametic disequilibrium from multilocus data. *Genetics* **157**, 413–423 (2001).
94. Vermeire, S. *et al. CARD15* genetic variation in a Quebec populations: prevalence, genotype-phenotype relationship, and haplotype structure. *Am. J. Hum. Genet.* **71**, 74–83 (2002).
95. Pritchard, J. K. & Cox, N. J. The allelic architecture of human disease genes: common disease-common variant... or not? *Hum. Mol. Genet.* **11**, 2417–2423 (2002).
96. Pritchard, J. K. Are rare variants responsible for susceptibility to complex diseases? *Am. J. Hum. Genet.* **69**, 124–137 (2001).
97. Hudson, R. R. Properties of a neutral allele model with intragenic recombination. *Theor. Popul. Biol.* **23**, 183–201 (1983).
98. Akey, J. M., Zhang, K., Xiong, M. & Jin, L. The effect of single nucleotide polymorphism identification strategies on estimates of linkage disequilibrium. *Mol. Biol. Evol.* **20**, 232–242 (2003).

Acknowledgements

We thank D. Nickerson, S. Gabriel, M. Daly, D. Altshuler and S. Schaffner for help in accessing and interpreting their data, and A. DiRienzo and S. Zoellner for discussions. We also thank M. Przeworski and the anonymous reviewers for comments on an earlier version of this manuscript. This work was supported by a National Institutes of Health grant to J.K.P.

Online Links

FURTHER INFORMATION
EGP SNPs web site: http://egp.gs.washington.edu
International HapMap Project:
http://www.genome.gov/page.cfm?pageID=10001688
Jonathan K. Pritchard's laboratory:
http://pritch.bsd.uchicago.edu
UW-FHCRC Variation Discovery Resource web site:
http://pga.gs.washington.edu
Access to this interactive links box is free online.