Using haplotype blocks to map human complex trait loci

Lon R. Cardon[1] and Gonçalo R. Abecasis[2]

[1]Wellcome Trust Centre for Human Genetics, University of Oxford, Roosevelt Drive, Oxford OX3 7BN, UK
[2]Department of Biostatistics, University of Michigan, Center for Statistical Genetics, 1420 Washington Heights, University of Michigan Ann Arbor, MI 48109-2029, Michigan, USA

Understanding of linkage disequilibrium (LD) in human populations could facilitate the discovery of genes that influence complex human diseases. The 'HapMap' project is now underway to characterize patterns of LD in the human genome. A pilot study showed 'haplotype blocks' in 51 regions scattered throughout the genome. These intriguing results raise important questions about the nature of recombination, and highlight practical issues of marker collection, the influence of statistical modelling on apparent block structure, and the levels of genotyping necessary for studies of common diseases. Knowledge of local disequilibrium patterns may help identify common polymorphisms involved in complex disease, but completely new analytical methods and experimental designs will be required to identify important rare variants.

Recent improvements in single nucleotide polymorphism (SNP) genotyping technology have enabled investigators to extend their study of LINKAGE DISEQUILIBRIUM (LD, see Glossary) beyond small genomic regions, such as HLA and β-globin loci, examined previously, and proceed to broader views of the genome. Initial studies focused on average LD levels [1–4] and emphasized the variability in processes generating LD [5]. Broadly speaking, observed genetic–physical distance relationships are compatible with a view of LD as the stochastic outcome of ancestral recombination, mutation, natural selection and population history.

Haplotype blocks

More recently, the original focus on average LD levels has been replaced by interest in specific patterns of LD throughout the genome [6,7]. The genome has been portrayed as a series of high LD regions ('blocks') separated by short discrete segments of very low LD (typically ascribed to 'recombination hotspots') [6,8,9]. As described, such regions exhibit limited haplotype diversity, so that a small number of distinct haplotypes account for most of the chromosomes in the population, and they are now termed HAPLOTYPE BLOCKS (Fig. 1).

Within high LD regions, allelic dependence yields redundancy among markers and improves the chances of detecting association when only a fraction of the markers is typed. By contrast, in intervening low LD regions, low correlation between markers means these regions can only be characterized adequately by typing many (or all) markers.

It is clear that understanding the role of genetic variants within low LD regions will be challenging. It is also clear that understanding LD patterns in common haplotypes will be insufficient for studies of rare variants. Nevertheless, if blocks do prove to be ubiquitous features of the genome [10,11], the simplicity of block structure within high LD regions could enhance association studies of common alleles, because recombination sites could delimit boundaries for candidate genes and indicate how far to extend the search for functional variants. In addition, when all common haplotypes in a region have been evaluated, additional genotyping would provide little additional information.

These prospects would facilitate comprehensive association studies for human COMPLEX DISEASES (Box 1) and

Glossary

Common disease–common variant hypothesis: The unproven hypothesis that genetic susceptibility to common complex disorders is influenced by relatively common polymorphisms. It is likely that this will be the case in some, but not all disorders.
Complex disease: Disease where susceptibility is controlled multiple genetic and environmental risk factors and where each of these risk factors has only a modest effect on susceptibility. Typical examples include asthma, diabetes and obesity.
Genetic heterogeneity: Variation in the effect of genetic risk factors between samples and populations. For example, in many complex diseases, the effect of genetic risk factors can depend on the presence of specific environmental factors.
Linkage disequilibrium (LD): The non-random association of alleles at tightly linked markers. Tight linkage can induce strong correlation between the genetic histories of neighbouring polymorphisms and, when LD is very high, alleles of linked markers can sometimes be used as surrogates for the state of nearby loci.
Phenotypic complexity: A characteristic of many common diseases, which refers to a continuum of multiple symptoms and outcomes measured on different scales. These often preclude a clear-cut definition of the presence or absence of disease.
Punctate recombination: An hypothesized distribution of recombination events characterized by short, precisely localized sites of recurring recombination and long intervening stretches where recombination is rare. Additional biological genetic evidence is still needed to evaluate the prevalence of this phenomenon in the human genome.
Haplotype block: A discrete chromosome region of high linkage disequilibrium and low haplotype diversity. It is expected that all pairs of polymorphisms within a block will be in strong linkage disequilibrium, whereas other pairs will show much weaker association. Blocks are hypothesized to be regions of low recombination flanked by recombination hotspots.

Corresponding author: Lon R. Cardon (lon.cardon@well.ox.ac.uk).

http://tigs.trends.com 0168-9525/03/$ - see front matter doi:10.1016/S0168-9525(03)00022-2

Box 1. Genetic association studies

Genetic analyses of complex human traits have not met the initial anticipations of widespread success [24,25]. Oft-cited reasons include GENETIC HETEROGENEITY (see Glossary), PHENOTYPIC COMPLEXITY, genotyping error, small sample sizes and over-interpretation of marginal findings. All of these and others are important, but it is also noteworthy that today's typical association-study design involves genotyping a small set of markers in genes or regions of interest and measuring their association with disease status in a small sample of individuals. Even within targeted regions, these studies typically examine only a fraction of human genetic variation and ignore background marker correlations.

With no control over the properties of the variants studied (allele frequencies, type of mutation, extent of genomic region for which variant provides information) it is difficult, if not impossible, to interpret their results [17]. For example, when SNPs and resulting haplotypes are examined in a candidate region they are often used as surrogates for other nearby variants. Allele frequency differences, historical recombination, and recurrent mutation destroy the correlation between neighbouring variants (including haplotypes) and can make genotyped variants unrepresentative of neighbouring polymorphism.

Knowledge of variant frequencies and their relationships can reduce the uncertainty in the design and interpretation of association studies. In particular, it will help investigators decide when they have achieved adequate coverage of a particular genomic region and help point to the probable location of functional variants when association to a SNP marker is identified.

Information on variant frequencies and their relationships is emerging from several large-scale studies and has the potential to revolutionize disease-association analysis [8,12]. Although information on the pattern of linkage disequilibrium will allow the design of association studies that survey a substantial fraction of common variants, it will not account for genetic heterogeneity and other types of phenotypic complexity, inadequate sample sizes or genotyping error. Investigators would be well advised to exercise extreme care in the phenotypic selection and evaluation of samples selected for extensive genotyping. Genome-wide association studies cannot be successful for all diseases [24], so that creative thinking and alternative approaches of gene identification must be encouraged.

have led the National Institutes of Health to fund the Haplotype Mapping ('HapMap') project, a genome-wide catalogue of common haplotype blocks in multiple human populations [12]. Here, we discuss recent data on the haplotype structure of the genome and critically examine the challenges ahead.

Blocks everywhere?

Block-like patterns have now been observed in the immunoglobulin cluster on 5q31 [6], within HLA [13] and throughout chromosomes 21 [9] and 22 [14]. In one study [13], these blocks were flanked by precisely localized recombination hotspots, leading to suggestions that PUNCTATE RECOMBINATION could be a general phenomenon underlying block structure [8,10,11].

Gabriel *et al.* [8] recently evaluated evidence for haplotype blocks in a large dataset. They focused on SNPs with minor allele frequencies >10% in 51 genomic regions (250 kb average, or 1 marker per 7.8 kb). Their sample included individuals of Western European, Japanese, Chinese, Nigerian and African American ancestry. Their results suggest that haplotype blocks can be detected with a few markers, and that within each block, three to five haplotypes can account for 90% of all chromosomes in the population. They also found that blocks are shorter in populations of African ancestry (11 kb on average) than in the other samples examined (average 22 kb). In a provocative extrapolation, Gabriel *et al.* proposed that through careful SNP selection, all common haplotypes in the genome could be

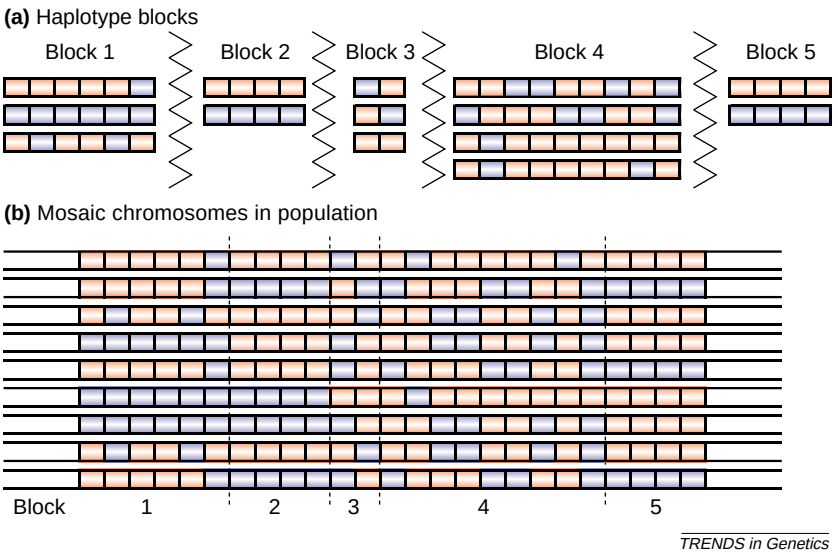

Fig. 1. Haplotype blocks. Blue and red show two alternative alleles. (a) When common variants are considered, the haplotype block paradigm portrays the genome as a series of short segments separated by recombination hotspots (zigzag lines in the figure). Within each block, there is little or no evidence for recombination, only a small number of distinct haplotypes is present in the population, and not all combinations of the various alleles are possible. (b) Most chromosomes in the population are a mosaic arrangement of the variants within each block.

surveyed in an association study using no more than 300 000–1 000 000 SNPs. It is important to recognize that this suggestion depends on several critical variables, such as whether block boundaries and haplotype diversity remain relatively stable as more markers are examined within each region. In the following sections we discuss some of the key issues in comparing these results and future studies.

Methods for block definition

A block-like structure for LD suggests that relatively long chromosome segments have undergone very little historical recombination. Punctate recombination or genetic drift could account for this paucity of recombination. However, ancestral recombination events cannot be directly observed or localized, and this has resulted in a plethora of operational definitions for blocks and their boundaries.

To date, only one group has defined blocks through direct measurement and localization of recombination hotspots [13]. Most studies used *in silico* block definitions that infer block structure and boundaries [6,8,9,14,15]. In general, all of the methods developed aim to identify genomic regions where a few unique haplotypes are distributed widely throughout the population. However, each method carries its own set of assumptions and heuristics. Some methods make use of known [9] or inferred haplotypes [6,15], others use pairwise LD measures [6,8] and others a combination of the two [14].

When pairwise measures are used, a block is defined whenever all pairwise coefficients (adjacent and non-adjacent) within a region exceed some pre-defined threshold. Gabriel *et al.* refined this basic definition by using confidence limits on the pairwise coefficients (which allowed them to account for some of the inaccuracies in estimated disequilibrium coefficients) and imposing constraints on marker number and spacing. This improves on the basic pairwise model by using sampling information to simplify block inference and permits block assessment in samples where haplotypes cannot always be determined accurately (such as unrelated individuals). However, their approach shares the same challenge as any method based on correlated measures: it is unclear how haplotype ancestry is reflected in a matrix of pairwise LD coefficients that are not independent. Moreover, the thresholds of LD and the confidence limits used to define a block are subjective and arbitrary.

Alternatively, when haplotypes are used, a block is usually defined when a small number of haplotypes (e.g. three to five) account for a high proportion of the observations (75–90%). Unfortunately, as with pairwise-derived blocks, precise thresholds for the number of haplotypes and proportion of observations are subjective. Whether pairwise measures or haplotypes are used, population and sample comparisons of block distribution are especially difficult: differences in population ancestry, and thus background haplotype diversity, average LD levels, and allele frequencies (which strongly influence pairwise LD measurements [5]), make it difficult to relate operational definitions of blocks to underlying biological processes.

Blocks identified under either definition typically include a fraction of overlapping blocks. These are usually buried in a summary description of the chromosome as a series of distinct blocks by either picking the longest blocks or picking the set of blocks that maximizes coverage of the chromosome [15]. This process is also subjective, reflecting the need for convenient data description as much as the genetic history of the population.

Although superficially distinct, measures based on pairwise LD and haplotype diversity are highly correlated; regions of low haplotype diversity typically exhibit high pairwise LD and, conversely, regions of high-pairwise LD typically exhibit low haplotype diversity. In this light, Gabriel *et al.*'s finding of limited haplotype diversity is due in part to their choice of a stringent LD threshold.

It is not yet clear how to compare haplotype blocks between studies or whether some of these operational definitions reflect underlying biological processes such as recombination better than others. It is clear that 'observed' haplotype blocks summarize a mixture of the underlying haplotype structure and the whims of the investigators reporting them. Presently, methods for detecting blocks remain subjective and no single approach is appropriate for all datasets.

SNP allele frequency

Although most previous studies have used different block definitions, they do share one common limitation. Specifically, the SNPs currently in public databases and available for genotyping were identified on a small set of chromosomes [16] and are thus not a representative sample of variants in the population. Instead, common alleles are over-represented and rare alleles are under-represented in these samples [8,14,17] (c.f. [9]).

This emphasis on common alleles has been lodged as a criticism of the HapMap project [12,18] because the conserved segments may reflect only the more common variants in the population and miss infrequent variants that could be central to disease. Although this frequency bias is likely to be a minor problem if common diseases are caused by common variants (COMMON DISEASE–COMMON VARIANT HYPOTHESIS) [19], it will seriously limit the usefulness of the HapMap catalogue for diseases caused by rarer alleles [5,20]. Still, if recombination hotspot boundaries are highly specific and their location governed by cellular signals, such as sequence signatures or chromosome structure, as apparent in some organisms [21], some of the intervening block boundaries might be the same whether blocks are defined by common or rare alleles. In this regard, the genomic frequency and location specificity of recombination hotspots in humans could be key determinants of the utility of the HapMap catalogue.

SNP density

Genotyping budgets and technology place practical limitations on large-scale studies, both for the HapMap and for subsequent uses of it in association studies. Thus far, the density of SNPs analysed has ranged from approximately one marker per kilobase of DNA sequence [9,13] to one marker per 15 kb [14]. Although the effect of marker density on inferred block structure is not entirely clear, studies using a denser marker panel have identified more

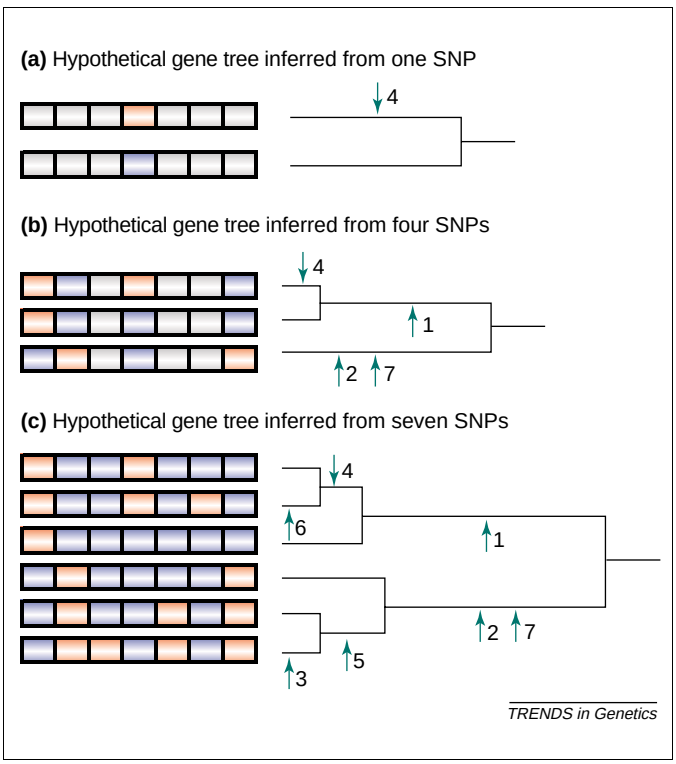

Fig. 2. Gene trees inferred for a hypothetical chromosome segment after genotyping 1, 4 and 7 markers. This genealogy assumes no recombination, and the ancestral state for each polymorphism is represented in blue and the alternative state in red. Green arrows indicate mutation events and are labelled with the corresponding base pair in the haplotype ideograms. Note how genotyping more markers can enhance our understanding of the history of a particular region. The converse is also true. With a small number of markers, it might not be possible to distinguish al haplotypes and coarse genealogy results. These oversimplified genealogies may preclude the identification of recombination events between neighbouring blocks.

short blocks [9], whereas studies using sparser marker panels have identified progressively longer blocks [8,14].

This relationship between marker density and resulting block length suggests that genotyping more markers can break seemingly large blocks into smaller segments. This possibility is not unexpected, as the reconstructed genealogy of any particular region becomes increasingly detailed as more markers are genotyped (Fig. 2). If this is true in general, blocks detected at low marker density might not reflect all the diversity in the data and might not be reliable for association studies. Gabriel *et al.* suggested that their density of one marker per 7.8 kb was sufficient to reflect the block pattern throughout the genome. Empirical studies using higher marker densities are required to evaluate this conclusion for the HapMap project.

If blocks defined at low marker densities prove reliable, an efficient strategy for genome-wide characterization would involve hierarchical genotyping, in which long haplotype blocks are initially detected using a sparse map and the remaining regions then genotyped at higher density to identify the shorter blocks. At present, the results are unclear as to whether this could produce results comparable to a full-scale high-density map. Again, further information clearly is needed; for example, a deep re-sequencing study to examine most or all SNPs over several genomic regions (~1 Mb each) would permit quantification of the loss in information that occurs as the marker density is progressively lowered [17]. In the absence of such information, the optimal map density and genotyping strategy remains unclear.

Haplotype tagging: the 1 000 000 marker question

The observation that fine-scale genetic variation can be described as a sequence of haplotype blocks has suggested practical alternatives to re-genotyping all markers in each disease study. As an example, consider that ten chromosomes were genotyped resulting in the pattern of variation illustrated in Fig. 3a (a total of 22 di-allelic polymorphisms are listed and, for each one, the common allele is red and the rare allele is blue). These polymorphisms can be grouped into two sets of 11 polymorphisms, as in Fig. 3b, such that there is no evidence for ancestral recombination within each group. Clearly, most of the information provided by the 11 SNPs in each set is redundant.

A list of the common haplotypes in each block could be used to select a minimal set of SNPs that allows one to distinguish all common haplotypes in the list and, therefore, infer the state of all other genetic variants in a block. SNPs within this minimal set have been christened haplotype-tagging SNPs (htSNPs [7]). A possible set of htSNPs for each block is labelled with the letter T in Fig. 3b. If such htSNPs identify precisely the state of all variants within each block, they provide an appealing way to accommodate differences in allele frequency between neighbouring SNPs. Matching of marker and trait allele frequencies is crucial for powerful association studies [22].

Although htSNPs provide a simple and natural way to use a catalogue of common haplotypes in association studies, efficient strategies can be designed to select smaller sets of representative SNPs. For example, neighbouring blocks can be correlated [6] or overlap [14]. SNP picking strategies that allow for interblock disequilibrium or ignore block boundaries altogether could be more effective than those that treat chromosomes as a series of discrete and independent blocks.

In Fig. 3, there are two haplotype blocks and the two most common haplotypes within each block occur in all possible combinations (Fig. 3c). However, the rare haplotype variants from each block always occur in tandem. By haplotype tagging SNPs within each block, a total of four markers is required. If, instead, SNPs are selected directly from the set of 22 markers (ignoring block boundaries), only three htSNPs are required to distinguish all common variants – a 33% saving in genotyping effort. The benefits of htSNP selection are not dependent on the concept of blocks, but rather on more general patterns of LD and limited haplotype diversity.

DNA pooling

Many case-control studies proceed by simply comparing allele frequencies at representative SNPs between a set of affected ('case') and unaffected ('control') individuals. In this setting, DNA pooling has been proposed as an alternative to individual genotyping that would further reduce genotyping costs and allow examination of the large samples required for complex disease mapping [23].

Although the advantages of DNA pooling in this simple setting are obvious, applying it to htSNPs is more

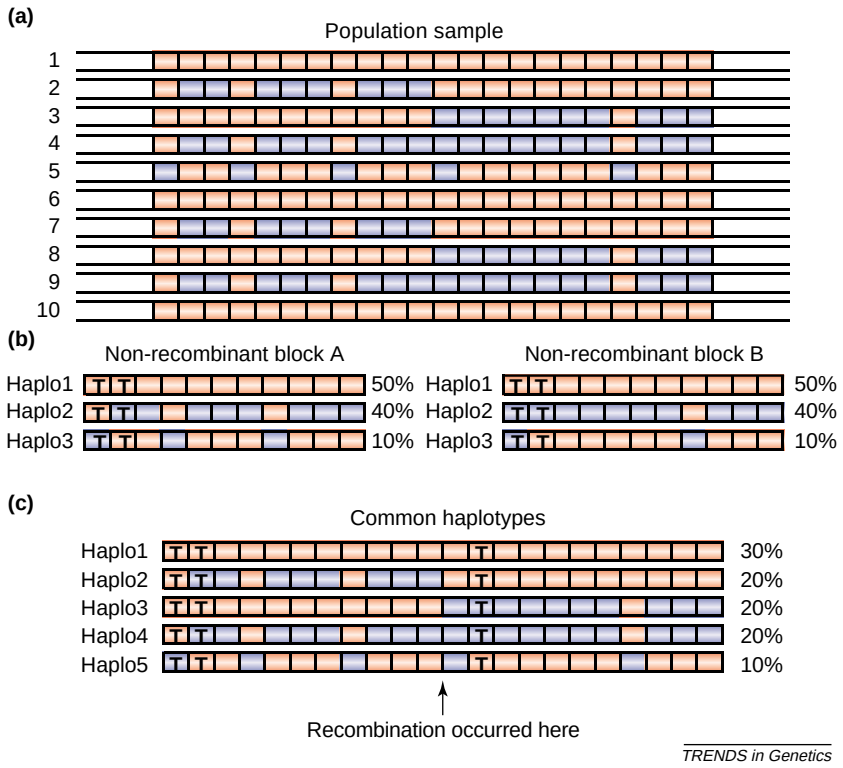

Fig. 3. How to choose haplotype-tagging single nucleotide polymorphisms (htSNPs). (a) A hypothetical sample of chromosomes (one chromosome per row) from a population of interest. The common alleles (one marker per column) are red, blue shows alternative, rare alleles. (b) Variation within this sample can be summarized as two non-recombining haplotype blocks. Within each block there are three distinct haplotypes and there is no evidence of recombination. Genotyping only two htSNPs (labelled T) per block is sufficient to determine the state of all other variants. (c) If variation over the entire region is considered jointly, ignoring the block boundaries, only three markers are required to determine the state at all other variants.

challenging. Together, a set of htSNPs for a particular region should be capable of distinguishing common haplotypes and inferring the state of neighbouring variants. However, inferring the distribution of haplotypes and neighbouring variants using only frequency information for individual htSNPs is less efficient than when individual genotypes are available. When pooling is used to reduce genotyping costs, it could be necessary to genotype additional htSNPs to ensure that information collected from pools can be used to reconstruct the genetic variation in each region and construct powerful association tests.

Future directions

With rapid technical advances in sequencing and genotyping technology, we might expect that in the not too distant future it will be possible to use re-sequencing to evaluate the genetic complement of any individual or study subject. For now, genotyping costs are a limiting factor in the design of association studies and selecting markers that can capture much of the variation in the genome in a cost efficient manner is an important part of study design. Increased understanding of local patterns of linkage disequilibrium in the human genome, provided by experiments such as the one reported by Gabriel *et al.*, is required for the design of practical association studies during the next decade. We cannot expect that these studies will be successful in identifying genes for all complex disorders, but they could help identify some genes for some disorders.

Geneticists are only beginning to study the advances offered by haplotype blocks and large-scale LD assessment for unravelling the genetic etiology of complex traits by association mapping. Currently, as a result of the subjectivity in block definition, the high-level of population variability, and the fact that apparent blocks are in part determined by the properties and spacing of the markers selected for analysis, it seems most appropriate to assign 'observed' blocks the status of 'draft' quality. As the LD patterns become more closely refined and more-objective statistical methods for block definitions are developed, blocks could move closer to 'finished' level, at least for the specific samples assessed and common markers genotyped.

Our understanding of the patterns of recombination and disequilibrium in the genome is still limited. Even if many common variants in the genome appear to be organized in relatively simple patterns, a substantial number of variants (both common and rare) will be organized in complex patterns that require alternative study designs and considerations of population diversity. In these cases, the notion of discrete blocks will probably prove too rigid to account for the complexity of linkage disequilibrium. Still, the empirical data emerging from large-scale projects such as the HapMap should increase our understanding of the variability in LD and how it might influence disease association studies.

Acknowledgements
L.R.C. is supported by a Wellcome Trust Principal Research Fellowship. The authors were also supported by NIH grant EY-12562.

References

1 Reich, D.E. *et al.* (2001) Linkage disequilibrium in the human genome. *Nature* 411, 199–204
2 Abecasis, G.R. *et al.* (2001) Extent and distribution of linkage disequilibrium in three genomic regions. *Am. J. Hum. Genet.* 68, 191–197
3 Taillon-Miller, P. *et al.* (2000) Juxtaposed regions of extensive and minimal linkage disequilibrium in human Xq25 and Xq28. *Nat. Genet.* 25, 324–328
4 Eaves, I.A. *et al.* (2000) The genetically isolated populations of Finland and Sardinia may not be a panacea for linkage disequilibrium mapping of common disease genes. *Nat. Genet.* 25, 320–323
5 Weiss, K.M. and Clark, A.G. (2002) Linkage disequilibrium and the mapping of complex human traits. *Trends Genet.* 18, 19–24
6 Daly, M.J. *et al.* (2001) High-resolution haplotype structure in the human genome. *Nat. Genet.* 29, 229–232
7 Johnson, G.C. *et al.* (2001) Haplotype tagging for the identification of common disease genes. *Nat. Genet.* 29, 233–237
8 Gabriel, S.B. *et al.* (2002) The structure of haplotype blocks in the human genome. *Science* 296, 2225–2229
9 Patil, N. *et al.* (2001) Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science* 294, 1719–1723
10 Goldstein, D.B. (2001) Islands of linkage disequilibrium. *Nat. Genet.* 29, 109–111
11 Reich, D.E. *et al.* (2002) Human genome sequence variation and the influence of gene history, mutation and recombination. *Nat. Genet.* 32, 135–142
12 Couzin, J. (2002) Genomics. New mapping project splits the community. *Science* 296, 1391–1393
13 Jeffreys, A.J. *et al.* (2001) Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. *Nat. Genet.* 29, 217–222
14 Dawson, E. *et al.* (2002) A first generation linkage disequilibrium map of human chromosome 22. *Nature* 418, 544–548
15 Zhang, K. *et al.* (2002) A dynamic programming algorithm for haplotype block partitioning. *Proc. Natl. Acad. Sci. U. S. A.* 99, 7335–7339
16 Sachidanandam, R. *et al.* (2001) A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* 409, 928–933
17 Ardlie, K.G. *et al.* (2002) Patterns of linkage disequilibrium in the human genome. *Nat. Rev. Genet.* 3, 299–309
18 Lai, E. *et al.* (2002) Medical applications of haplotype-based SNP maps: learning to walk before we run. *Nat. Genet.* 32, 353–354
19 Reich, D.E. and Lander, E.S. (2001) On the allelic spectrum of human disease. *Trends Genet.* 17, 502–510
20 Pritchard, J.K. and Przeworski, M. (2001) Linkage disequilibrium in humans: Models and data. *Am. J. Hum. Genet.* 69, 1–14
21 Petes, T.D. (2001) Meiotic recombination hot spots and cold spots. *Nat. Rev. Genet.* 2, 360–369
22 Abecasis, G.R. *et al.* (2001) The power to detect linkage disequilibrium with quantitative traits in selected samples. *Am. J. Hum. Genet.* 68, 1463–1474
23 Sham, P. *et al.* (2002) DNA pooling: A tool for large-scale association studies. *Nat. Rev. Genet.* 3, 862–871
24 Weiss, K.M. and Terwilliger, J.D. (2000) How many diseases does it take to map a gene with SNPs? *Nat. Genet.* 26, 151–157
25 Cardon, L.R. and Bell, J.I. (2001) Association study designs for complex diseases. *Nat. Rev. Genet.* 2, 91–99