# DNA Sequence Variation of *Homo sapiens*

D.R. BENTLEY
*The Wellcome Trust Sanger Institute, Hinxton, Cambridge CB10 1SA, United Kingdom*

The finished genome sequence of *Homo sapiens* (Rogers, this volume) provides a starting point for the study of sequence variation in the human population. Every variant that is discovered can be mapped back to the human genome and correlated with genes, regulatory elements, and other functionally important sequences. As we gain a better understanding of the biological information encoded by the human genome sequence, we should aim to define the sequence variants that have biochemical and phenotypic consequences.

The genome sequence enables us to develop targeted strategies to search for disease-related sequence variants. This requires a better understanding of the patterns of variation and of the genetic variants that can be used as reference markers throughout the genome. Genome sequence information will also underpin future surveys of somatic variation and cancer. Studying the DNA sequence variation of *Homo sapiens* will enable us to understand our origins and evolution and will help characterize the genetic basis of our individuality—for example, in our susceptibility or resistance to disease, and our variable response to drugs, toxins, and other environmental factors.

## THE ORIGIN OF SEQUENCE VARIATION

Modern humans are believed to have migrated east out of Africa 50,000–60,000 years ago (Jobling et al. 2004) and subsequently spread across the world, replacing earlier *Homo* species. This pattern was originally deduced largely from archaeological and anthropological evidence (Stringer 2002) but received substantial reinforcement from DNA sequence information. For example, genetic variability is generally higher in Africa than on other continents, and phylogenetic reconstructions of non-recombining regions usually place the root in Africa (Cavalli-Sforza and Feldman 2003; Pääbo 2003). A subset of the genetic variants in Africa at the time were therefore present in the migrant founders of all later subsequent groups, while many variants remained only within population subgroups in Africa. More recent migrations between different parts of the globe have since contributed to admixture between multiple subgroups, and in the last few hundred years, this process has increased substantially.

Sequence variation arises as a result of new mutation and recombination (see Fig. 1). Based on observed variation and simulations, the average human mutation rate has been estimated to be 1–2 bases in 100 million per generation (Drake et al. 1998; Giannelli et al. 1999; Reich et al. 2002) which corresponds to around 30–60 new mutations per gamete. For variants that are neutral (not under selection), the allele frequency in the population will be affected by random drift. Many new mutations will disappear within a few generations; a few may become com-

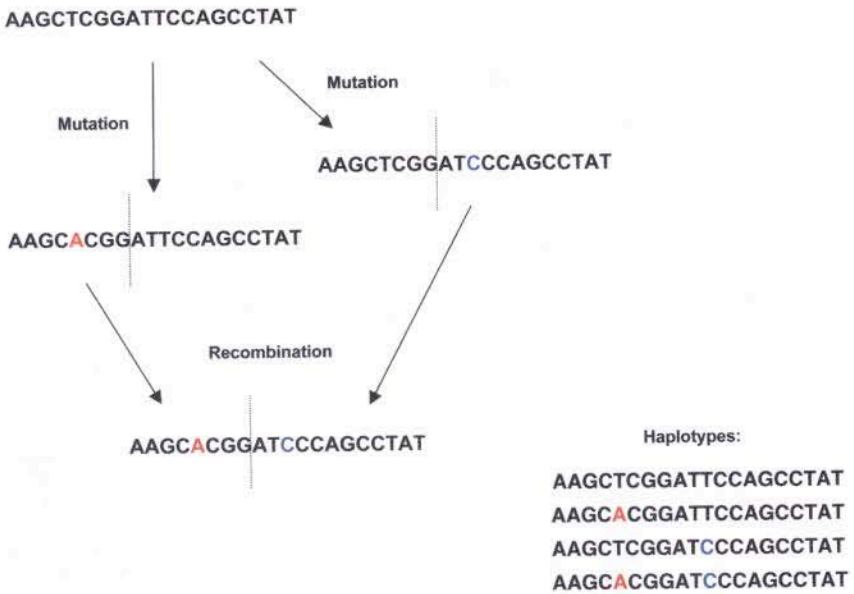


**Figure 1.** Origin of sequence variation. Sequence variation arises by mutation (*colored bases*) and by recombination (*dotted lines*). These processes give rise to individual haplotypes (listed on the right) that coexist in the population.

mon over many generations. Most high-frequency sequence variants therefore arose in Africa, were present in the migrant founders, and are common to multiple population groups around the world (Barbujani et al. 1997). New variants that are under positive selection—i.e., functional variants that confer a survival advantage—will increase in frequency more quickly than expected by random processes. A change in the environment (such as the appearance of a new virus, or release of a new toxin, or change in food source) may impose a new selective pressure on existing functional variants, leading to alterations in allele frequency in a particular population. Balancing or frequency-dependent selection may act to maintain both alleles of a polymorphism stably in the population.

Recombination during meiosis results in cross-overs between homologous chromosomes. This leads to reassortment of previously existing variants into new combinations in the haploid gametes. Assuming a recombination rate of 1 site per 100 million bases per generation (Yu et al. 2001) (i.e., around 30 recombination events per gamete), large segments of each parental homolog are passed on from one generation to the next. Over multiple generations, further recombination events result in progressive fragmentation of the original ancestral segments into more and more pieces. Individual present-day chromosomes are thus mosaics of segments of relatively few ancestral chromosomes, each segment having its own lineage history (Pääbo 2003). The sequence along a chromosome defines the complete set of variants in an individual haploid sequence and represents a particular "haplotype" (from "haploid genotype") (see Fig. 1).

Somatic mutation occurs in individual cells at any stage following the formation of a diploid zygote, during development and throughout adulthood. The mutation process may be enhanced by exposure to particular nongenetic factors such as radiation or toxins. Particular combinations of somatic mutations, sometimes in conjunction with germ-line variants, alter the normal program of cell proliferation, differentiation, and apoptosis, and lead to cancer.

## THE NATURE OF SEQUENCE VARIATION

Individual copies of the haploid human genome differ at approximately one site per kilobase on average (Li and Sadler 1991; Sachidanandam et al. 2001). Single-nucleotide polymorphisms (SNPs) account for around 90% of sequence variants in the human population, the remaining 10% being insertions or deletions ("indels"). It has been estimated that the world's human population contains over 10 million SNPs with a minor allele frequency (m.a.f.) of at least 1% (Kruglyak and Nickerson 2001). Several in-depth studies of sequencing in specific genes have demonstrated that within exons there is a slightly lower density of polymorphic sites (1 site per 2,000 bases), and also typically a lower allele frequency (Cargill et al. 1999; Halushka et al. 1999). Assuming that 1.5% of the genome sequence encodes protein (Lander et al. 2001), the global population contains around 90,000 variants (with m.a.f. >1%) in protein-coding sequence. From previous surveys, the protein-coding variants can be subdivided into ~50% synonymous (causing no amino acid change), 33% nonsynonymous and conservative (i.e., leading to conservative amino acid replacement), and 17% nonsynonymous, nonconservative changes (Cargill et al. 1999; Halushka et al. 1999).

Genetic variants that give rise to nonconservative amino acid changes are particularly good functional variant candidates, but a number of observations suggest that this generalization is likely to be inadequate. Not all nonconservative amino acid changes necessarily alter function. For example, in an exhaustive survey (Green et al. 1999), only half of all possible amino acid changes (233/454) in factor IX cause hemophilia B (Haemophilia B Mutation Database v12: http://www.kcl.ac.uk/ip/petergreen/haemBdatabase.html).

Functionally important amino acids in proteins tend to be conserved between species; on average, about one-third of amino acids are highly or absolutely conserved in vertebrates. Conversely, other types of changes (either at the amino acid or nucleotide level) can alter function. Conservative changes may affect protein function, and any changes may affect transcript function; for example, by introducing or abolishing splice sites. There is no good estimate of the number of functionally important changes outside protein-coding sequence. Changes in promoters or enhancers may affect transcription (Li et al. 1999; Saur et al. 2004). Comparison between multiple genomes suggests that around 5% of the human genome is under selection to conserve sequence (Waterston et al. 2002). These analyses suggest that, in addition to the protein-coding sequence, another 3.7% of the genome is equally strongly conserved and may be important for gene regulation or chromatin behavior (Waterston et al. 2002; Margulies et al. 2003). On this basis, an initial estimate of variants in functional sequence (with m.a.f. < 1%) would be nearer 300,000, and an unknown fraction of these will have phenotypic consequences.

The first genome-wide survey of human variation resulted in detection of 1.42 million candidate SNPs that were mapped to a unique position in the working draft sequence (Sachidanandam et al. 2001). This set was estimated to contain possibly 12% of all SNPs with a m.a.f. of 1% or more (Kruglyak and Nickerson 2001). More recently, the total number of SNPs available with a unique map position has increased to 7 million (dbSNP release 120, http://www.ncbi.nlm.nih.gov/SNP) following generation of additional shotgun data, and may now contain 40% of SNPs with m.a.f. <1% (estimate based on Kruglyak and Nickerson [2001]). A fully characterized catalog of human sequence variation may still be a long way off. In the meantime, however, efforts are focused on how to develop the resources that are available from the finished genome sequence, and how to use them to study the genetics of human disease.

## SEQUENCE VARIATION AND DISEASE

The genes and underlying sequence variants that cause over 1400 diseases have already been identified (data

from Online Mendelian Inheritance in Man; http://www.ncbi.nlm.nih.gov). The majority of these have been rare single-gene disorders. In most cases, linkage studies in large affected families have been used to target a genomic region in order to screen for mutations that disrupt a candidate gene ("positional cloning"; Collins 1992). In a few studies, detection of a cytogenetic abnormality (e.g., a translocation) provided the necessary positional information, and a few genes were discovered using prior knowledge of the protein (e.g., amino acid sequence). The "parametric" linkage approach used in the single-gene disorders has been successful in only a few polygenic conditions, where the gene has a sufficiently strong effect to show a clear familial inheritance pattern. Examples include susceptibility loci for breast cancer (BRCA1 [Miki et al. 1994] and BRCA2 [Wooster et al. 1995]) and the initial study of maturity onset of diabetes of the young (MODY) (for review, see Bell and Polonsky 2001). However, in most studies, this approach has been largely unsuccessful as little or no convincing linkage has been found. Alternative linkage strategies that do not rely on specific transmission patterns initially held great promise for complex trait applications (as seen, for example, in the discovery of NOD2 with Crohn's disease [Hugot et al. 2001; Ogura et al. 2001] and CTLA-4 with Graves' disease [Ueda et al. 2003]), but have not yet yielded many consistent results. Population-based association studies are believed to offer greater statistical power in detecting genetic effects underlying complex traits (Risch 2000).

An association study is used to test for a positive correlation between a sequence variant in the genome and a disease or measurable phenotype (Risch and Merikangas 1996). For example, a genotyping assay is used to determine the frequency of each allele at a particular locus in separate, matched groups of patients and controls (a case-control study). If one of the alleles has a high frequency in cases compared to controls, this is evidence for association of that allele with the disease. A direct association study uses a set of candidate gene variants that are presumed to include the causal variant(s) (Risch and Merikangas 1996). It tests the hypothesis that a particular variant is directly involved in the phenotype. An indirect study uses anonymous variants (such as SNPs) as markers (Collins et al. 1997). It tests the hypothesis that a marker is closely linked to an unknown causative variant. A direct association study was used to demonstrate the association of a 32-base deletion (Δ32) in the cytokine receptor 5 gene (*CKR5*) and resistance to HIV. The CKR5 protein is a G-protein-coupled receptor on the surface of $CD4^+$ T-lymphocytes that appears to be an efficient co-receptor for HIV-1 viral strains. The deletion variant is nonfunctional with respect to both its natural function and its capacity to mediate HIV-1 infection. The homozygous form (Δ32/Δ32) was strongly associated with a protective effect against HIV-1 infection, and there was also evidence that the heterozygous form (+/Δ32) delayed progression to acquired immune deficiency syndrome in some cases (Dean et al. 1996). An indirect association study was used to follow up an earlier linkage to part of Chromosome 5q31 to inflammatory bowel disease (IBD) (Rioux et al. 2001). Using a dense panel of SNPs, the disease was associated with a common haplotype, and the critical region was reduced from ~1 Mb to 250 kb.

The power of a direct association study is influenced by the allele frequency and risk ratio of the causative variant in relationship to the sample size. Detecting causative variants in association studies requires a progressively larger sample size as the allele frequency or risk ratio of the variant decreases. For indirect association studies, an additional factor is the degree of correlation between the anonymous markers used in the study and the causative variant. If one of the markers is completely correlated with the causative variant, the power of the indirect study will match that of the direct approach. In most cases, however, we expect the available anonymous markers to be incompletely correlated, and the indirect approach will have less power. If the allele frequency of the anonymous markers is substantially different from the unknown functional variants, this will also reduce power (Risch 2000; Zondervan and Cardon 2004).

Causative variants that are known to contribute to common disease include examples of a range of allele frequencies (0.85–0.01 in the examples listed in Table 1). There may be diseases caused by multiple rare variants at the same locus (e.g., the NOD2 locus; Hugot et al. 2001; Ogura et al. 2001). In these cases, each variant is of independent origin and is therefore carried on a different haplotype. The ability to detect these variants by association will depend on the power of the study to detect each individual variant. The case-control association studies reported to date have demonstrated the ability to detect disease-associated variants with modest relative risks. For example, the association of LTA-3 with myocardial infarction was essentially hypothesis-free: 65,671 SNPs distributed in 13,738 genes throughout the genome were tested in 1,133 cases and 1,006 controls (in a two-tier study), and homozygosity with respect to each of two SNPs (both in the LTA-3 gene) showed significant association with the disease (Ozaki et al. 2002). Evidence was obtained suggesting a possible functional significance for each variant that might correlate with phenotype. If so, then this would be a direct association study. Studies such as this provide an indication of the future potential and limitations of the approach. For some of the other disease studies listed in Table 1, prior linkage to a chromosomal region was also detected, and this positional information was used to select genes or variants for use in the association study.

Before starting a direct association study, it is necessary to find the functional variants for the study by sequencing candidate genes in multiple individuals. For maximum sensitivity, the best approach would be to sequence a group of affected individuals (specific to each study) that should be enriched for disease-associated variants. Alternatively, a large number of control samples would need to be sequenced to capture the low-frequency variants that might be expected (for estimates, see Kruglyak and Nickerson 2001). Either option is limited by cost and by the fact that we do not yet know all the functional regions of each gene that would need to be se-

**Table 1.** Genetic Disease Variants

| Reference | Phenotype | Gene | Allele | Frequency (between 0 and 1) | Allelic relative risk |
|---|---|---|---|---|---|
| Bentley et al. (1986) | Hemophilia B | FIX | Arg -4 Gln | $1.7 \times 10^{-9}$ | ∞ |
| Bell et al. (1984) | IDDM | INS | VNTR | 0.7 | 2 |
| Altshuler et al. (2000) | NIDDM | PPARG | Pro 12 Ala | 0.8 | 1.25 |
| Ozaki et al. (2002) | MI | LTA-3 | Thr 26 Asn | 0.5 | 1.8 |
| Ueda et al. (2003) | Graves | CTLA4 | Thr 17 Ala | 0.35 | 1.7 |
| Palmer et al. (1991) | Creutzfeldt-Jacob | PRNP | Met 129 Val | 0.35 | 3 |
| Saunders et al. (1993) | Alzheimer's | APOE | Cys 112 Arg | 0.16 | 4 |
| Dean et al. (1996) | HIV resistance | CCR5 | del 32 | 0.10 | 7 |
| Bertina et al. (1994) | thrombosis | FV | Arg 506 Gln | 0.05 | 7 |
| Ogura et al. (2001) | Crohn's | NOD2 | 1007 fs | 0.04 | 2 |
| Hugot et al. (2001) | Crohn's | NOD2 | Arg 702 Trp | 0.04 | 3 |
| Hugot et al. (2001) | Crohn's | NOD2 | 980 fs | 0.02 | 6 |
| Hugot et al. (2001) | Crohn's | NOD2 | Gly 908 Arg | 0.01 | 6 |

quenced. However, it is important to pursue this strategy, as it will be very effective to search for rare variants, either in control groups or patient collections, in order to explore the full extent of functional sequence variation in the genome.

Before embarking on the indirect approach, we need to develop a comprehensive panel of anonymous markers (SNPs). This can be done for each gene or chromosomal region of interest, but it will be much more valuable to do it systematically across the whole genome and develop a freely available resource. Once created, this SNP panel can be applied universally to any search for associations with any disease or phenotype. It is critical that as near as possible, every part of the genome is tightly linked to at least one marker of the panel. Development of this panel requires a dense SNP map, technology to genotype a large number of SNPs in multiple DNA samples, an effective way to measure linkage between SNPs, a strategy that can be adjusted to characterize regions of high and low linkage disequilibrium (LD), and a way to assess the extent that common patterns of variation are captured in each population group. Recent research on a number of isolated genomic regions and whole chromosomes has led to evaluation of these requirements, as discussed below.

## LINKAGE DISEQUILIBRIUM AND HAPLOTYPES

The degree of correlation between two markers can be determined by population genetic analysis. If the alleles of two neighboring SNPs are in equilibrium in the population, the alleles at each SNP are independent from one another. If this is the case, each haplotype (each combination of two alleles) occurs at a frequency that is the product of the frequency of each individual allele. If, on the other hand, particular SNP alleles occur together more often than expected by the equilibrium model, they are said to be in association, or LD. LD between two loci ("pair-wise" LD) is determined empirically by carrying out genotyping experiments on a population sample and calculating the difference between expected and observed frequencies of each combination of alleles. The most commonly used measures of LD are D′ (Lewontin 1964) and $r^2$ (Hill and Robertson 1968; Ohta and Kimura 1969; for further discussion, see Wall and Pritchard 2003).

In general, pair-wise LD decreases with increasing physical distance between markers. This is because over longer distances there is a higher chance that ancestral recombination events have occurred between the markers. However, LD is also highly variable with respect to physical distance ( Clark et al. 1998; Abecasis et al. 2001; Reich et al. 2001; Dawson et al. 2002). Another observation is that variants which arose recently (for example, most rare variants) tend to exhibit LD over longer distances, reflecting the low probability of a nearby recombination in relatively few generations.

A whole-chromosome study enabled systematic correlation of LD with respect to physical distance and a wide range of structural and genetic features (Dawson et al. 2002). Individual genotypes were collected for 1,504 evenly spaced SNPs along Chromosome 22 (average spacing 1 SNP/20 kb) in a panel of DNAs of North European origin. The profile of LD along the chromosome was determined by averaging pair-wise LD measurements within 1.7-Mb sliding windows. This study revealed highly variable patterns of LD in different chromosome regions, and two notable regions where LD extended over hundreds of kilobases (Fig. 2). LD did not correlate with any features of the gene content, repetitive sequence, or other structural features such as base composition. However, there was a strong correlation between high LD and low recombination frequency (Fig. 2), indicating that historical and contemporary recombination are related, and nonrandom along the chromosome. Within the regions of high LD, it was also possible to observe that a limited number of haplotypes accounted for the majority in the population. For example, over a 700-kb region, 5 extended haplotypes together accounted for 76% of the individual chromosomes studied, the remainder being accounted for in 12 rare haplotypes (Dawson et al. 2002).

The results of several parallel studies revealed similar findings: that LD was highly variable, and that it was possible to define short regions where LD was consistently high between all markers and where there were a limited number of common haplotypes. Following the initial work of Daly et al. (2001), Gabriel et al. (2002) developed a method to define regions of high LD as occurring

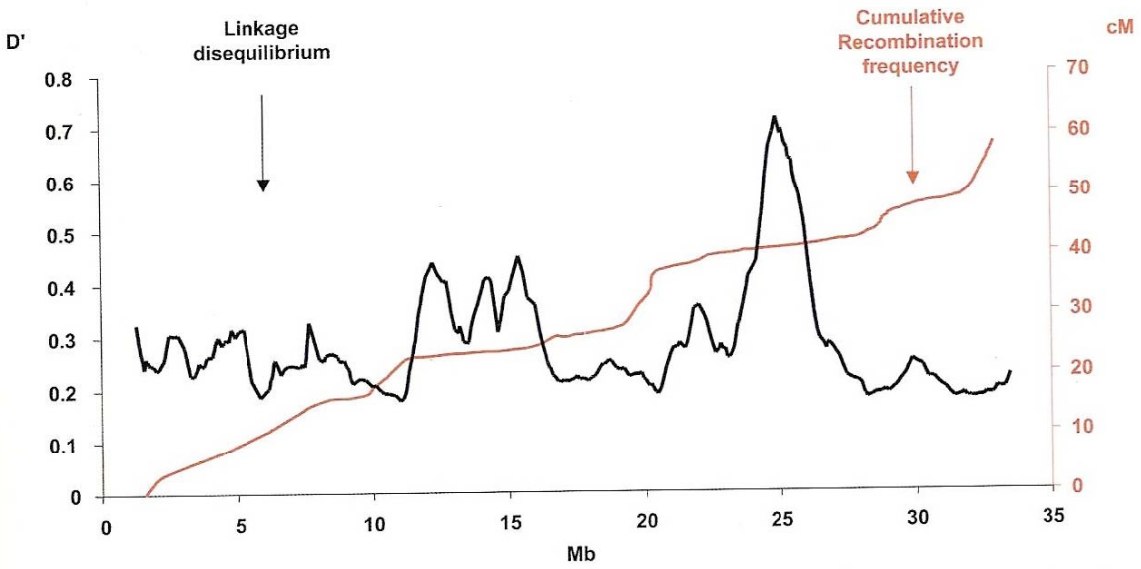


**Figure 2.** Linkage disequilibrium (LD) and meiotic recombination: Chromosome 22. The LD profile is based on average D′ values in sliding windows (see text). LD and cumulative recombination frequency are plotted relative to physical distance along the chromosome, with the telomere of the long arm on the right of the figure.

where pair-wise LD values (based on D′) between three or more adjacent SNPs exceeded a defined threshold (see Fig. 3a–c). Within regions of high LD, it was possible to define a limited number of common haplotypes that represented most of the chromosomes in the study (Fig. 3d). These regions were termed "haplotype blocks" (Gabriel et al. 2002). They were separated by regions where LD had not been detected, either because LD was low or because there were too few markers available. The block method did not fully take into account LD between blocks and did not provide a continuous view of the pattern of LD. Nevertheless, it would be possible to use it as a simple and reliable indicator of regions where data from more SNPs was needed. As expected from previous work

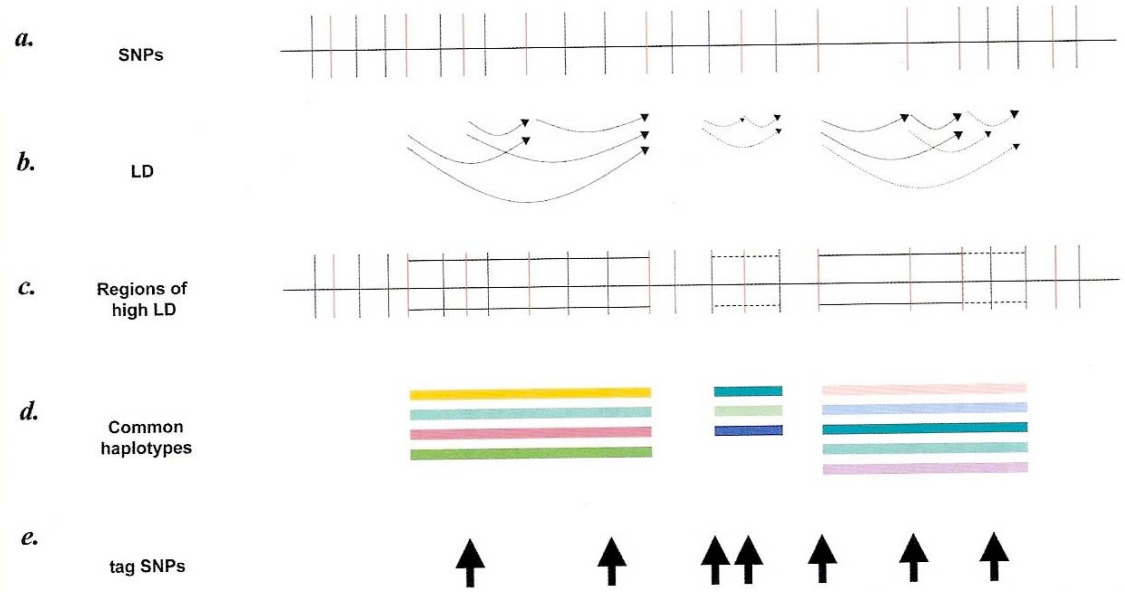


**Figure 3.** SNPs, linkage disequilibrium, and haplotypes. The schematic diagram depicts the course of an experiment carried out in two tiers as follows. *First tier:* From the SNP map of the genome, SNPs are selected (*blue vertical bars* in *a*) and used to genotype DNA samples. The results are used to calculate pair-wise LD (*curved arrows* in *b*), and the results are compiled to define regions of high LD (*diagonally shaded boxes* in *c*, drawn over the SNP map). At this point, overall progress in characterizing LD is assessed, and additional SNPs are selected where more data are needed (*red vertical bars* in *a* and *c*). Common haplotypes (*colored bars* in *d*) are determined from the LD data. A subset of the SNPs ("tag" SNPs; *arrows* in *e*) is selected that captures most or all of the common variation in the region (see text for details).

(Clark et al. 1998; Reich et al. 2001), this study also revealed greater diversity in African-Americans than in North Europeans with respect to ancestral recombination and polymorphism. A lower average block size was estimated in the African-Americans than in North Europeans (11 kb vs. 22 kb), and within blocks, the average haplotype diversity was higher in African-Americans than in North Europeans (5 vs. 4.2 haplotypes per block). The study by Patil et al. (2001) took a different approach. Oligonucleotide sequencing arrays representing the nonrepetitive sequence in Chromosome 21 were used to detect SNPs and common haplotypes by analysis of 20 individuals. 24,047 SNPs detected blocks of common haplotypes of average size 13 kb and 3.2 haplotypes per block covering 80% of the chromosome. (Blocks with 1 or 2 SNPs were excluded from these calculations.)

Comparison of the studies listed in Table 2 revealed that for most of them, detection of LD was limited by SNP availability. Caution should be exercised in this comparison, as some data sets were small and the criteria for measuring high LD were slightly different; but the overall trend was clear. As SNP spacing decreased from 20 kb to 8 kb, 5 kb, and 1 kb, the proportion of the studied genomic regions where high LD was detected increased from 20% to 58%, 78%, and 95%, respectively (see Table 2). Future studies of the genome therefore needed more SNPs, particularly in regions where little or no LD had been detected.

Another important observation from these studies is based on the idea that the variation data derived from the initial SNPs could be used to select a subset of the SNPs that distinguished the different haplotypes, and thus captured most or all of the information on common variation in the region. These "haplotype tag SNPs" (htSNPs) (see Fig. 3e) (Johnson et al. 2001; see also Fig. 1 in International HapMap Consortium 2003) would be maximally informative and could be used in future association studies, whereas the other SNPs could be discarded. In three of the studies listed in Table 1, htSNPs were identified comprising between 27% and 20% of the initial SNPs. Determining LD and haplotype patterns empirically could therefore enable fourfold savings in future association studies with little or no loss of power, assuming the patterns of haplotype diversity were the same in the initial population and the subsequent disease sample. The conclusions from these studies provided the motivation for planning a large-scale pilot study on Chromosome 20 (see below), and also the International HapMap Project, to determine common LD and haplotype patterns throughout the human genome, in multiple ethnic groups (International HapMap Consortium 2003).

## CHROMOSOME 20

The experience gained in the previous studies illustrated the need to produce genotype data with very high densities of SNP maps, and to evaluate how SNP density affects measurement of LD and haplotype analysis over large genomic regions. These conclusions formed the basis for a study of Chromosome 20. At the time, the existing map for this Chromosome contained a total of 46,000 candidate SNPs (1 SNP/1.3 kb on average). Further analysis of the SNP distribution along the chromosome revealed that only 30% of the sequence contained 10 or more SNPs per 10-kb window. Given that the density of SNPs used in the study of Jeffreys et al. (2001) (see Table 2) was higher than this, it was necessary to generate many more SNPs in order to obtain this density over the rest of Chromosome 20. Random shotgun sequencing of Chromosome 20 (purified by flow sorting) generated additional SNPs to obtain a minimum density of 10 SNPs per 10 kb (close to the density used in the Jeffreys study) for almost all the chromosome (P. Deloukas et al., unpubl.). The need to supplement the existing SNP map to this level was confirmed in the findings of the subsequent LD analysis and has since been adopted on a genome-wide basis (see above).

SNPs were selected from the new Chromosome 20 map at ~1-kb spacing and typed using the Golden Gate assay (Fan et al., this volume) in samples of North European, African-American, and East Asian origin. Analysis of a 10-Mb region (Ke et al. 2004) revealed variable LD along the chromosome and good correlation between high LD and low recombination frequency, as observed in the Chromosome 22 study. Progressive removal of subsets of the raw data did not alter the LD profile obtained by the sliding windows method (using 500-kb windows), illustrating that this view of LD is robust at SNP densities of 1 SNP/10 kb and above. The profiles were also consistent between all three population groups, although overall LD was lower in the African-Americans than in the other two groups.

In contrast to the LD profiles, it was found that haplotype block patterns were not stable. These differed depending on the SNP densities, and were particularly unstable at low densities both in regions of high and low LD (see Fig. 4) (Ke et al. 2004). This result suggested that future studies should be carried out at a minimum density of 1 SNP/5 kb. More SNPs were also required, particularly in areas of low LD, and more SNPs might also be required to confirm patterns of LD and haplotypes in areas of high LD (see SNPs marked in blue in Fig. 3a and c). At the highest SNP density, the overall coverage of the re-

**Table 2.** LD Studies

| Study | Region studied | kb studied | SNPs (total) | SNP spacing (kb) | LD (% region covered) | tag SNPs (% of total) |
|---|---|---|---|---|---|---|
| Dawson et al. (2002) | Chromosome 22 | 30,000 | 1,504 | 20 | 20 | — |
| Gabriel et al. (2002) | 51 regions | 13,000 | 1,970 | 7.8 | 58 | — |
| Daly et al. (2001) | region 5q31 | 460 | 103 | 5 | 78 | 25 |
| Johnson et al. ( 2001) | selected genes | 135 | 122 | 1.1 | n.d. | 27 |
| Jeffreys et al. (2001) | human MHC | 216 | 179 | 1.2 | 95 | 20 |

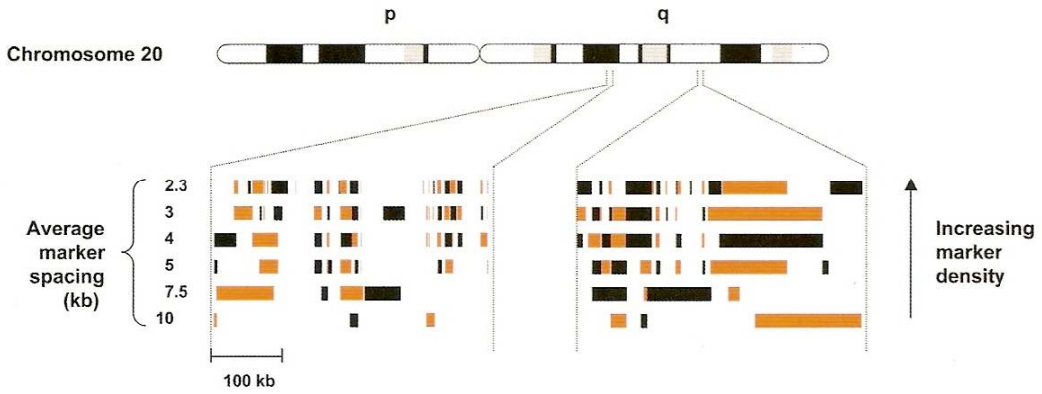


**Figure 4.** LD analysis of Chromosome 20. Haplotype blocks (*red* and *black boxes*) were computed from LD data on Chromosome 20 and are shown for two regions, one each of high and low overall LD. The analysis is repeated using data from different densities of SNPs (average marker spacings in each analysis are listed on the left of the figure, and increasing SNP density is indicated by the vertical arrow).

gion in high LD was estimated to be 65% in the North Europeans and 45% in the African-Americans (Ke et al. 2004). The coverage estimates for North Europeans fits the trend in Table 1. However, the lower coverage figure for the African-Americans suggests that more dense SNP sets may need to be used in these populations—echoing the earlier findings of the Gabriel study (Gabriel et al. 2002). The block patterns also varied significantly when using different analytical methods, indicating the need for other ways to describe LD patterns on a fine scale. For example, LD unit maps (Maniatis et al. 2002) reflect continuous views of LD, although they are to some extent sensitive to varying allele frequency. A promising approach is the development of methods to estimate recombination rates on a fine scale using the same data sets described above (McVean et al. 2004). These approaches could enable a precise assessment of progress in characterization of LD in the genome, identify and prioritize regions that require more data, and help assess the ability of panels of anonymous markers to detect unknown disease-related variants.

## IMPLICATIONS FOR FUTURE STUDIES

The first global study of patterns of common variation in the human genome has crystallized in the form of the International HapMap Project (International HapMap Consortium 2003). The density of SNPs that are now available in the public domain is likely to be sufficient for this initial analysis of LD and common haplotypes, although some targeted SNP discovery to fill gaps may be necessary. This project will produce a freely available genome-wide panel of tested SNPs that can be assessed for use in association studies. It is intended to reduce the need for researchers to search for their own SNPs, and to enable better searches for disease variants in genes, chromosomal regions, or ultimately, the whole genome using the indirect association approach. Evaluation of the parameters that govern association studies are still required, and the results of pilot studies might help calibrate the HapMap and other studies to fit more precisely with their anticipated applications. In parallel, it is also necessary to address the challenge of developing sufficiently large sample collections with accurate, accessible phenotype information. Without this, even a perfect SNP panel that covers 100% of the genome will lack the power to establish an association between genotype and phenotype.

We should continue sequence-based discovery of variants on a genome-wide basis; this would give us an unbiased and properly quantified picture of the extent of common sequence variation in the human population, replacing the current simulation-based estimates. Achieving the first step at this level would benefit from upward of 100 individual human genome sequences, a task that requires a greater than tenfold increase in sequence output and is likely to demand the successful implementation of new technologies for cheap, accurate, high-throughput resequencing. Discovering rare variants (m.a.f. < 1%) would require improvements of a further one or two orders of magnitude in sequencing. At present, therefore, pursuing rare variants remains in the domain of individual candidate genes and may be best carried out in conjunction with a genetic disease study. Both candidate gene and whole-genome sequencing will also be instrumental in characterizing somatic variants that cause cancer, as demonstrated, for example, by Davies et al. (2002). Systematic pursuit of these areas of research will help us gain a view of the complete spectrum of variant allele frequencies that contributes to human disease.

As we learn more about the full extent of functional sequences in the genome, our knowledge of the functionally important subset of variants will increase (for further discussion, see Bentley 2004). Developing a panel of functional variants for all genes would increase the power of association studies. However, we must be aware of the unknown; any association study based on the direct approach needs to take account of the possibility that the true causative variant is not included in the study. It is therefore necessary to establish the underlying mechanism and explain the functional significance of the variant. Within the medical field, understanding the mechanism that underlies disease will be the most effective way to make progress in diagnosis, alleviation, and the effective treatment of disease.

## ACKNOWLEDGMENTS

The author especially thanks Ines Barroso, Lon Cardon, Panos Deloukas, Peter Donnelly, Christine Rees, and others for discussion and assistance. The author gratefully acknowledges the support of the Wellcome Trust.

## REFERENCES

Abecasis G.R., Noguchi E., Heinzmann A., Traherne J.A., Bhattacharyya S., Leaves N.I., Anderson G.G., Zhang Y., Lench N.J., Carey A., Cardon L.R., Moffatt M.F., and Cookson W.O. 2001. Extent and distribution of linkage disequilibrium in three genomic regions. *Am. J. Hum. Genet.* **68:** 191.

Altshuler D., Hirschhorn J.N., Klannemark M., Lindgren C.M., Vohl M.C., Nemesh J., Lane C.R., Schaffner S.F., Bolk S., Brewer C., Tuomi T., Gaudet D., Hudson T.J., Daly M., Groop L., and Lander E.S. 2000. The common PPARgamma Pro12Ala polymorphism is associated with decreased risk of type 2 diabetes. *Nat. Genet.* **26:** 76.

Barbujani G., Magagni A., Minch E., and Cavalli-Sforza L.L. 1997. An apportionment of human DNA diversity. *Proc. Natl. Acad. Sci.* **94:** 4516.

Bell G. I. and Polonsky K.S. 2001. Diabetes mellitus and genetically programmed defects in beta-cell function. *Nature* **414:** 788.

Bell G.I., Horita S., and Karam J.H. 1984. A polymorphic locus near the human insulin gene is associated with insulin-dependent diabetes mellitus. *Diabetes* **33:** 176.

Bentley A.K., Rees D.J., Rizza C., and Brownlee G.G. 1986. Defective propeptide processing of blood clotting factor IX caused by mutation of arginine to glutamine at position -4. *Cell* **45:** 343.

Bentley D.R. 2004. Genome vision. *Nature* (in press).

Bertina R.M., Koeleman B.P., Koster T., Rosendaal F.R., Dirven R.J., de Ronde H., van der Velden P.A., and Reitsma P.H. 1994. Mutation in blood coagulation factor V associated with resistance to activated protein C. *Nature* **369:** 64.

Cargill M., Altshuler D., Ireland J., Sklar P., Ardlie K., Patil N., Shaw N., Lane C.R., Lim E.P., Kalyanaraman N., Nemesh J., Ziaugra L., Friedland L., Rolfe A., Warrington J., Lipshutz R., Daley G.Q., and Lander E.S. 1999. Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nat. Genet.* **22:** 231.

Cavalli-Sforza L.L. and Feldman M.W. 2003. The application of molecular genetic approaches to the study of human evolution. *Nat. Genet.* (suppl.) **33:** 266.

Clark A.G., Weiss K.M., Nickerson D.A., Taylor S.L., Buchanan A., Stengard J., Salomaa V., Vartiainen E., Perola M., Boerwinkle E., and Sing C.F. 1998. Haplotype structure and population genetic inferences from nucleotide-sequence variation in human lipoprotein lipase. *Am. J. Hum. Genet.* **63:** 595.

Collins F.S. 1992. Positional cloning: Let's not call it reverse anymore. *Nat. Genet.* **1:** 3.

Collins F.S., Guyer M.S., and Charkravarti A. 1997. Variations on a theme: Cataloging human DNA sequence variation. *Science* **278:** 1580.

Daly M.J., Rioux J.D., Schaffner S.F., Hudson T.J., and Lander E.S. 2001. High-resolution haplotype structure in the human genome. *Nat. Genet.* **29:** 229.

Davies H., Bignell G.R., Cox C., Stephens P., Edkins S., Clegg S., Teague J., Woffendin H., Garnett M.J., Bottomley W., Davis N., Dicks E., Ewing R., Floyd Y., Gray K., Hall S., Hawes R., Hughes J., Kosmidou V., Menzies A., Mould C., Parker A., Stevens C., Watt S., and Hooper S., et al. 2002. Mutations of the BRAF gene in human cancer. *Nature* **417:** 949.

Dawson E., Abecasis G.R., Bumpstead S., Chen Y., Hunt S., Beare D.M., Pabial J., Dibling T., Tinsley E., Kirby S., Carter D., Papaspyridonos M., Livingstone S., Ganske R., Lohmussaar E., Zernant J., Tonisson N., Remm M., Magi R., Puurand T., Vilo J., Kurg A., Rice K., Deloukas P., and Mott R., et al. 2002. A first-generation linkage disequilibrium map of human chromosome 22. *Nature* **418:** 544.

Dean M., Carrington M., Winkler C., Huttley G.A., Smith M.W., Allikmets R., Goedert J.J., Buchbinder S.P., Vittinghoff E., Gomperts E., Donfield S., Vlahov D., Kaslow R., Saah A., Rinaldo C., Detels R., and O'Brien S.J. 1996. Genetic restriction of HIV-1 infection and progression to AIDS by a deletion allele of the CKR5 structural gene. Hemophilia Growth and Development Study, Multicenter AIDS Cohort Study, Multicenter Hemophilia Cohort Study, San Francisco City Cohort, ALIVE Study. *Science* **273:** 1856.

Drake J.W., Charlesworth B., Charlesworth D., and Crow J.F. 1998. Rates of spontaneous mutation. *Genetics* **148:** 1667.

Gabriel S.B., Schaffner S.F., Nguyen H., Moore J.M., Roy J., Blumenstiel B., Higgins J., DeFelice M., Lochner A., Faggart M., Liu-Cordero S.N., Rotimi C., Adeyemo A., Cooper R., Ward R., Lander E.S., Daly M.J., and Altshuler D. 2002. The structure of haplotype blocks in the human genome. *Science* **296:** 2225.

Giannelli F., Anagnostopoulos T., and Green P.M. 1999. Mutation rates in humans. II. Sporadic mutation-specific rates and rate of detrimental human mutations inferred from hemophilia B. *Am. J. Hum. Genet.* **65:** 1580.

Green P.M., Saad S., Lewis C.M., and Giannelli F. 1999. Mutation rates in humans. I. Overall and sex-specific rates obtained from a population study of hemophilia B. *Am. J. Hum. Genet.* **65:** 1572.

Halushka M.K., Fan J.B., Bentley K., Hsie L., Shen N., Weder A., Cooper R., Lipshutz R., and Chakravarti A. 1999. Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. *Nat. Genet.* **22:** 239.

Hill W.G. and Robertson A. 1968. Linkage disequilibrium in finite populations. *Theor. Appl. Genet.* **38:** 226.

Hugot J.P., Chamaillard M., Zouali H., Lesage S., Cezard J.P., Belaiche J., Almer S., Tysk C., O'Morain C.A., Gassull M., Binder V., Finkel Y., Cortot A., Modigliani R., Laurent-Puig P., Gower-Rousseau C., Macry J., Colombel J.F., Sahbatou M., and Thomas G. 2001. Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. *Nature* **411:** 599.

International HapMap Consortium. 2003. The International HapMap Project. *Nature* **426:** 789.

Jeffreys A.J., Kauppi L., and Neumann R. 2001. Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. *Nat. Genet.* **29:** 217.

Jobling M.A., Hurles M.E., and Tyler-Smith C. 2004. *Human evolutionary genetics*. Garland Science, New York.

Johnson G.C., Esposito L., Barratt B.J., Smith A.N., Heward J., Di Genova G., Ueda H., Cordell H.J., Eaves I.A., Dudbridge F., Twells R.C., Payne F., Hughes W., Nutland S., Stevens H., Carr P., Tuomilehto-Wolf E., Tuomilehto J., Gough S.C., Clayton D.G., and Todd J.A. 2001. Haplotype tagging for the identification of common disease genes. *Nat. Genet.* **29:** 233.

Ke X., Hunt S., Tapper W., Lawrence R., Stavrides G., Ghori J., Whittaker P., Collins A., Morris A.P., Bentley D., Cardon L.R., and Deloukas P. 2004. The impact of SNP density on fine-scale patterns of linkage disequilibrium. *Hum. Mol. Genet.* **13:** 577.

Kruglyak L. and Nickerson D.A. 2001. Variation is the spice of life. *Nat. Genet.* **27:** 234.

Lander E.S., Linton L.M., Birren B., Nusbaum C., Zody M.C., Baldwin J., Devon K., Dewar K., Doyle M., FitzHugh W., Funke R., Gage D., Harris K., Heaford A., Howland J., Kann L., Lehoczky J., LeVine R., McEwan P., McKernan K., Meldrim J., Mesirov J.P., Miranda C., Morris W., and Naylor J., et al. (International Human Genome Sequencing Consortium). 2001. Initial sequencing and analysis of the human genome. *Nature* **409:** 860.

Lewontin R.C. 1964. The interaction of selection and linkage. I. General considerations: Heterotic models. *Genetics* **49:** 49.

Li W.D., Reed D.R., Lee J.H., Xu W., Kilker R.L., Sodam B.R., and Price R.A. 1999. Sequence variants in the 5′ flanking region of the leptin gene are associated with obesity in women. *Ann. Hum. Genet.* **63:** 227.

Li W.H. and Sadler L.A. 1991. Low nucleotide diversity in man. *Genetics* **129:** 513.

Maniatis N., Collins A., Xu C.F., McCarthy L.C., Hewett D.R., Tapper W., Ennis S., Ke X., and Morton N.E. 2002. The first linkage disequilibrium (LD) maps: Delineation of hot and cold blocks by diplotype analysis. *Proc. Natl. Acad. Sci.* **99:** 2228.

Margulies E.H., Blanchette M., Haussler D., and Green E.D. 2003. Identification and characterization of multi-species conserved sequences. *Genome Res.* **13:** 2507.

McVean G.A.T., Myers S.R., Hunt S.E.H., Deloukas P., Bentley D.R., and Donnelly P. 2004. The fine-scale structure of recombination rate variation in the human genome. *Science* (in press).

Miki Y., Swensen J., Shattuck-Eidens D., Futreal P.A., Harshman K., Tavtigian S., Liu Q., Cochran C., Bennett L.M., and Ding W., et al. 1994. A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. *Science* **266:** 66.

Ogura Y., Bonen D.K., Inohara N., Nicolae D.L., Chen F.F., Ramos R., Britton H., Moran T., Karaliuskas R., Duerr R.H., Achkar J.P., Brant S.R., Bayless T.M., Kirschner B.S., Hanauer S.B., Nunez G., and Cho J.H. 2001. A frameshift mutation in NOD2 associated with susceptibility to Crohn's disease. *Nature* **411:** 603.

Ohta T. and Kimura M. 1969. Linkage disequilibrium due to random genetic drift. *Genet. Res.* **13:** 47.

Ozaki K., Ohnishi Y., Iida A., Sekine A., Yamada R., Tsunoda T., Sato H., Hori M., Nakamura Y., and Tanaka T. 2002. Functional SNPs in the lymphotoxin-alpha gene that are associated with susceptibility to myocardial infarction. *Nat. Genet.* **32:** 650.

Pääbo S. 2003. The mosaic that is our genome. *Nature* **421:** 409.

Palmer M.S., Dryden A.J., Hughes J.T., and Collinge J. 1991. Homozygous prion protein genotype predisposes to sporadic Creutzfeldt-Jakob disease. *Nature* **352:** 340.

Patil N., Berno A.J., Hinds D.A., Barrett W.A., Doshi J.M., Hacker C.R., Kautzer C.R., Lee D.H., Marjoribanks C., McDonough D.P., Nguyen B.T., Norris M.C., Sheehan J.B., Shen N., Stern D., Stokowski R.P., Thomas D.J., Trulson M.O., Vyas K.R., Frazer K.A., Fodor S.P., and Cox D.R. 2001. Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science* **294:** 1719.

Reich D.E., Schaffner S.F., Daly M.J., McVean G., Mullikin J.C., Higgins J.M., Richter D.J., Lander E.S., and Altshuler D. 2002. Human genome sequence variation and the influence of gene history, mutation and recombination. *Nat. Genet.* **32:** 135.

Reich D.E., Cargill M., Bolk S., Ireland J., Sabeti P.C., Richter D.J., Lavery T., Kouyoumjian R., Farhadian S.F., Ward R., and Lander E.S. 2001. Linkage disequilibrium in the human genome. *Nature* **411:** 199.

Rioux J.D., Daly M.J., Silverberg M.S., Lindblad K., Steinhart H., Cohen Z., Delmonte T., Kocher K., Miller K., Guschwan S., Kulbokas E.J., O'Leary S., Winchester E., Dewar K., Green T., Stone V., Chow C., Cohen A., Langelier D., Lapointe G., Gaudet D., Faith J., Branco N., Bull S.B., and McLeod R.S., et al. 2001. Genetic variation in the 5q31 cytokine gene cluster confers susceptibility to Crohn disease. *Nat. Genet.* **29:** 223.

Risch N.J. 2000. Searching for genetic determinants in the new millennium. *Nature* **405:** 847.

Risch N. and Merikangas K. 1996. The future of genetic studies of complex human diseases. *Science* **273:** 1516.

Sachidanandam R., Weissman D., Schmidt S.C., Kakol J.M., Stein L.D., Marth G., Sherry S., Mullikin J.C., Mortimore B.J., Willey D.L., Hunt S.E., Cole C.G., Coggill P.C., Rice C.M., Ning Z., Rogers J., Bentley D.R., Kwok P.Y., Mardis E.R., Yeh R.T., Schultz B., Cook L., Davenport R., Dante M., and Fulton L., et al. 2001. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409:** 928.

Saunders A.M., Strittmatter W.J., Schmechel D., George-Hyslop P.H., Pericak-Vance M.A., Joo S.H., Rosi B.L., Gusella J.F., Crapper-MacLachlan D.R., and Alberts M.J., et al. 1993. Association of apolipoprotein E allele epsilon 4 with late-onset familial and sporadic Alzheimer's disease. *Neurology* **43:** 1467.

Saur D., Vanderwinden J.M., Seidler B., Schmid R.M., De Laet M.H., and Allescher H.D. 2004. Single-nucleotide promoter polymorphism alters transcription of neuronal nitric oxide synthase exon 1c in infantile hypertrophic pyloric stenosis. *Proc. Natl. Acad. Sci.* **101:** 1662.

Stringer C. 2002. Modern human origins: Progress and prospects. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **357:** 563.

Ueda H., Howson J.M., Esposito L., Heward J., Snook H., Chamberlain G., Rainbow D.B., Hunter K.M., Smith A.N., Di Genova G., Herr M.H., Dahlman I., Payne F., Smyth D., Lowe C., Twells R.C., Howlett S., Healy B., Nutland S., Rance H.E., Everett V., Smink L.J., Lam A.C., Cordell H.J., and Walker N.M., et al. 2003. Association of the T-cell regulatory gene CTLA4 with susceptibility to autoimmune disease. *Nature* **423:** 506.

Wall J.D. and Pritchard J.K. 2003. Haplotype blocks and linkage disequilibrium in the human genome. *Nat. Rev. Genet.* **4:** 587.

Waterston R.H., Lindblad-Toh K., Birney E., Rogers J., Abril J.F., Agarwal P., Agarwala R., Ainscough R., Alexandersson M., An P., Antonarakis S.E., Attwood J., Baertsch R., Bailey J., Barlow K., Beck S., Berry E., Birren B., Bloom T., Bork P., Botcherby M., Bray N., Brent M.R., Brown D.G., and Brown S.D., et al. (Mouse Genome Sequencing Consortium). 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420:** 520.

Wooster R., Bignell G., Lancaster J., Swift S., Seal S., Mangion J., Collins N., Gregory S., Gumbs C., and Micklem G. 1995. Identification of the breast cancer susceptibility gene BRCA2. *Nature* **378:** 789.

Yu A., Zhao C., Fan Y., Jang W., Mungall A.J., Deloukas P., Olsen A., Doggett N.A., Ghebranious N., Broman K.W., and Weber J.L. 2001. Comparison of human genetic and sequence-based physical maps. *Nature* **409:** 951.

Zondervan K.T. and Cardon L.R. 2004. The complex interplay among factors that influence allelic association. *Nat. Rev. Genet.* **5:** 89.