PAPER

Irruption of genomics in the search for disease related genes

G Thomas, H Cann

Gut 2003;52(Suppl II):ii1–ii5

Genomics was initiated when robotics made possible the characterisation of large numbers of DNA fragments and when ever improving computers with dedicated software were applied to the localisation in the genome of these sequences and to the analysis of their content. By enabling the generation and management of large amounts of DNA based sequences these tools have changed our perception of the genomes of living organisms. These data, as applied to humans, are contributing to the understanding of gene function, disease processes, and evolution of our species. Presently they are changing the research strategies for identifying genetic variations influencing disease susceptibility and response to treatment. These advances will have a profound impact in biomedicine.

Precise knowledge of the genomes of living species is recent. Fifty years ago, genetic information was known to be encoded in DNA. DNA was known to be associated with the chromosomes. From the work of Morgan on *drosophila* in the 1920s, the genes were believed to be linearly organised along these structures. However, the size of the human genome was not evaluated. The exact number of chromosomes present in a human cell was unknown, and the mechanism by which the order of the genes was maintained along the chromosomes remained elusive.

In the late 1950s, a series of technical achievements revealed that the human diploid genome was composed of 46 chromosomes. Soon after, the first human chromosome related abnormalities were recognised, with anomalies at the constitutional level (trisomy 21) and at the somatic level (the Philadelphia chromosome in chronic myeloid leukaemia). In the early 1970s, kinetic analysis of the renaturation of DNA molecules revealed that the human genome, as most eukaryotic genomes, had two components. One component included sequences that were single copy in the haploid genome and the other included repeat elements. The size of the human haploid genome was evaluated to contain about 3 billion base pairs of DNA. At about the same time, experiments based on the viscoelastic properties of elongated molecules showed that human DNA molecules were extremely long. Fifteen years later, the use of pulse field gel electrophoresis revealed that each chromosome of *S cerevisiae* contains a single DNA molecule suggesting that

each human chromosome would also be composed of a single molecule, and providing a simple explanation for maintaining gene order. It was thus inferred that the nuclear haploid genome of humans is composed of only 23 molecules. If these molecules were placed end to end, they would extend over a distance of one metre. Thus a total of two metres of DNA would be present in each diploid nucleated human cell. All regions of the genome were not expected to convey the same density of information. The discovery of repetitive DNA, and subsequently of introns, suggested that only a small portion of the genome codes for proteins. The number of genes was not known with certainty, and it was believed to be around 100 000 or more, an evaluation that, retrospectively, seems grossly overestimated.

The size of the human genome and its number of genes were considered to be enormous and much too large to be studied in an exhaustive manner. Until 1984, no one seriously thought that the sequencing of the entire human genome would be amenable in the near future. No one believed in the early 1980s, that a first draft of the entire human sequence would be available less than two decades later.

The first step to be taken toward the sequencing of the human genome was the building of a genetic map. This was soon followed by the sequencing of transcribed sequences, so called expressed sequence tags. Physical mapping was started in 1992, and the real beginning of systematic sequencing of the human genome started in 1998, preceded by a four year pilot project. In the following paragraphs, we present a brief summary of the human genome project and indicate how the use of the new knowledge arising from the project may be used to identify genetic variations that influence human disease.

THE FIRST STEP: THE GENETIC MAP

A genetic map aims at placing genes for traits or markers along the chromosomes. As initially proposed by Morgan,¹ it is constructed by studying the transmission of two different traits or markers from parents to offspring. Very schematically, the frequency with which the two characters are simultaneously transmitted is a measure of the proximity of their genes on the chromosome. If the two characters are independently transmitted, the corresponding genetic polymorphisms

See end of article for authors' affiliations

Correspondence to:
Dr G Thomas, Centre
d'Etude du Polymorphisme
Humain, 27 rue Juliette
Dodu, 75010, Paris,
France; thomas@cephb.fr

Abbreviations: YAC, yeast artificial chromosome; BAC, bacterial artificial chromosome; SNP, single nucleotide polymorphism; LD, linkage disequilibrium

either are located far away on the same chromosome or are located on two different chromosomes. If the two characters tend to be jointly transmitted their genes are genetically linked.

The most direct way to construct a genetic map is to type a collection of highly polymorphic genetic markers on a common set of large families. Such markers, for the most part derived from DNA sequences,^{2,3} possess at least two variations at the same locus (that is, alleles), preferably with equal frequencies. Sequencing of randomly selected fragments of human DNAs had revealed the frequent occurrence of a specific group of highly polymorphic markers called microsatellites.⁴ In the human population, many microsatellites have more than five alleles so that the probability of a person being heterozygous (that is, having two different alleles) at such a locus is high. Heterozygous people enable the direct monitoring of the transmission of each allele to their offspring.

During meiosis, homologous chromosomes (each coming from a different parent) do not physically blend with each other. Instead, they recombine with each other, exchanging large segments of DNA along their length. The number of recombinations (often referred to as "breakpoints") between a pair of chromosome homologs that generate the post-meiotic chromosome transferred to the offspring is rather small. An average of 33 meiotic recombination breakpoints (for a total of 23 chromosome pairs) occur during each meiosis. Thus, a post-meiotic chromosome is usually composed of only two to four fragments coming from the pre-meiotic homolog pair, each fragment containing on average 50 million base pairs. At a low resolution the position of these breakpoints are distributed roughly randomly along all 23 chromosome pairs (with the notorious exception of chromosome Y). Alleles at two polymorphic loci are always jointly transmitted from a parent to an offspring when no meiotic recombination has occurred between the loci. The frequency of joint transmission of two adjacent alleles on the same chromosome enables a direct evaluation of their distance. When the joint transmission of three loci is studied, their order along the chromosome may be determined.

For the human genome project, the genetic map was constructed from genotypes generated with markers from DNA of at least 40 large families, generally of northern European origin, each consisting of an average of eight offspring, both parents and often all four grandparents. The family DNAs were distributed by the Centre d'Etude du Polymorphisme Humain (CEPH) to collaborating investigators throughout the world who typed them with their markers and returned the genotypes to CEPH for inclusion in a central database⁵ (<http://www.cephb.fr/cephdb>). To date 11 successive CEPH family based genetic maps of the human genome have been published, reflecting the increase in numbers of genotypes with time. The most recent CEPH family based map, constructed from some 8000 microsatellites, defines genetic positions on all the chromosomes with a mean distance between positions of less than one million base pairs.⁶ Knowledge of these positions was extremely useful in the subsequent construction of the physical map and remains of great importance for the isolation of functional polymorphisms. Very recently, a whole genome genetic map based on genotypes from many families from Iceland has been published.⁷

An *in vitro* technique for ordering markers on radiation induced human chromosome fragments, similar in principle to genetic mapping, takes advantage of their segregation in cultured human-rodent hybrid cells. The principle here is that two closely placed markers are more likely to be found on the same chromosomal fragment than those far apart. This technique, called radiation hybrid mapping, has provided an independent map with locus order almost identical to that of the genetic map.⁸

THE SECOND STEP: THE PHYSICAL MAP AND THE SEQUENCING OF THE HUMAN GENOME

To sequence large genomes, it is necessary to break the long DNA molecules and clone the resulting smaller fragments. To achieve the sequencing of the human genome, this process was performed in two successive steps. The first step led to isolating and mapping fragments approximately one million or one hundred thousand base pairs long by cloning them into specific vectors, either in yeast (yeast artificial chromosomes or YACs⁹) or in bacteria (bacterial artificial chromosomes or BACs¹⁰). About 3000 YACs and 30 000 BACs are sufficient to contain an entire haploid genome, but about 10 times as many were used to identify regions of overlapping pairs of clones, providing information on their relative positions on the physical map.^{10,11} The absolute positions of the clones were determined by testing them for markers that had been placed on the genetic and on the radiation hybrid map. This technique is called physical mapping. This process was complemented by the systematic sequencing of small human DNA fragments (usually BAC ends¹²) thus generating a large number of unique, precisely localised unique sequences,¹³ that could be used as anchor points in the construction of the physical map.

YACs and BACs were large enough to be individually localised on the chromosomes. They remained too long, however, to be directly sequenced. In addition, some YACs were found to be mosaics made up of fragments from different chromosomes, an artefact resulting from the method used in their construction. BACs could be much more easily manipulated than YACs. Thus, DNA of BACs was broken down again to fragments of about one thousand base pairs in size. These fragments were directly sequenced. Because the fragments are generated by inducing DNA breaks at random positions, many sequenced fragments are overlapping. It is possible from the knowledge of the sequence of these fragments to reconstruct that of the BAC from which they originated. In this process the generation of the sequence of small one Kb fragments is straightforward, and comparatively easy to automate. Their reassembly is more difficult because of the presence of repetitive sequences. Comparing the sequence of two BACs that are predicted from the physical map to be overlapping either would confirm the map or point to an error. In theory, at completion, the sequence of one chromosome should be represented by a continuous series of bases going from one telomere to the other telomere. In practice however, the reconstruction of the centromere sequence is not attempted, because it is almost entirely composed of highly repetitive DNAs containing over millions of base pairs.

The sequencing of the human genome actually started in 1998. It was decided that during this process even if the sequence had not been entirely determined, it should nevertheless be made available to the scientific community. The sequence of a region that has not been entirely ordered is called the draft sequence. Once the order has been determined with high confidence and potential gaps closed, the sequence is said to be finished. The entire draft sequence was announced in early 2001.¹⁴ Now, over 95% of the sequence of the genome has been finished. The entire euchromatic sequences of chromosomes 14, 20, 21, and 22 are finished and published; and those of six others are at least 99% finished (see <http://www.ncbi.nlm.nih.gov/Genome/index.html>). Parallel to the generation of large amounts of sequences, important efforts are devoted to their annotation. In particular the initial effort pioneered by J C Venter on the systematic sequencing of cDNAs has been most useful in the identification of exons.¹⁵ The sequencing of the human genome is being performed by 20 Genome Sequencing Centres located in the United States, England, France, Germany, Japan, and China.

HUMAN GENOME DIVERSITY

The genomes of human individuals are not identical. They differ, thus generating diversity. Because the establishment of the human genome sequence involved overlapping regions from several humans, it has been possible by comparing the sequences of the same region from different individuals to evaluate, in a preliminary fashion, the genetic diversity that exists in mankind.¹⁶ By far the most frequent type of sequence variation is the single nucleotide polymorphism (SNP) because of the substitution, insertion, or deletion at sites of single nucleotides.

Genetic diversity may be quantified by a parameter called nucleotide diversity, which is defined by the average number of nucleotide differences when comparing two sequences selected at random from a population. The nucleotide diversity of the human genome is estimated to be between 10^{-4} and 5×10^{-3} . It is smaller than that of the great apes.¹⁷ This comparatively low diversity is attributable to the size of the ancestral human population that existed until about 100 000 years ago, which has been estimated to be a few tens of thousand of individuals.

Genetic diversity remains substantial within a population (as defined by shared culture, geography, physical appearance, and gene pool).¹⁸ Differences observed among populations, such as those originating from different continents, account for less than 15% of this diversity. Allelic frequencies of most SNPs present moderate geographical variations. The largest genetic diversity is observed in Africa, an observation that supports the concept that this continent is the origin of anatomically modern humans, *homo sapiens sapiens*. The availability of panels of DNA from individuals originating from many different populations throughout the world has enabled a better description of within and between population genetic differences.¹⁹

The nucleotide diversity is not identical throughout the human genome. GC rich regions are more polymorphic than AT rich regions, an observation compatible with the increased mutability of the GC dinucleotide. Natural selection has resulted in a lower nucleotide diversity in sequences that are functionally important (for example, exons versus introns), with the notable exception of the HLA region.

There are now over two million SNPs that have been assigned to specific regions of the human genome (<http://www.ncbi.nlm.nih.gov/SNP>). Many are located in transcribed regions, but most of them are found in introns or in intergenic regions. It is expected that over 10 million SNPs will be identified each with a world wide frequency of the rarer allele of 5% or more. Such SNPs are referred to as frequent. These represent ancient sequences that have undergone mutations and developed polymorphic frequencies over evolutionary time. Frequent SNPs may be found in more than one major population, reflecting ancient migrations and gene exchange between human groups.

THE NEXT STEP: THE HAPLOTYPE MAP

A haplotype is a combination, on the same DNA molecule, of alleles from two or more polymorphic loci. Two adjacent SNPs are said to be in complete equilibrium if the frequency of each of the four possible haplotypes is equal to the product of the frequencies of the two alleles it contains (that is, when all possible combinations of alleles at each locus have an independent probability of occurring together on the same DNA fragment). When two SNPs are in complete equilibrium, the identification of one allele on a haplotype provides no information on the nature of the other allele present on this haplotype. Complete equilibrium of closely adjacent SNPs, however, is not usually observed. In most cases, these SNPs are in disequilibrium; therefore, the typing of one SNP provides information on the other.

To understand how linkage disequilibrium (LD) comes about, it is helpful to consider the historical ancestor for any given gene. It can be readily seen that the about 3 billion men currently living on the planet are more than the number of their fathers. Likewise, the number of paternal grandfathers of these men is smaller than the number of fathers; hence the number of ancestral Y chromosomes decreases within each previous generation. Returning back far enough through preceding generations, we can expect to find a single male ancestor. This man most probably lived some 100 000 years ago, in some part of Africa. This line of logic holds equally true not just for the Y chromosome, which tends to be transmitted in a single block of DNA from man to man, but also for any chromosome region. In autosomes, because of recombination, smaller fragments of a given region are transmitted, but the same reasoning holds: the number of copies of a given gene has been transmitted from a smaller number of copies throughout succeeding generations. The time that the most recent common ancestors existed for several autosomal genes has been estimated, and found to be about 1 000 000 years BP (before the present era), although there is considerable variation for such estimates. For a given gene, the sequence present in the common ancestor is called the root of the gene for the human species.

From the ancestral “root” of a human gene there have been tens to hundreds of thousands (occasionally more) rounds of DNA replications and accompanying mutations ensue thus generating a phylogenetic tree; when a new mutation occurs, a new branch appears. Thus, the number of branches in the simplest phylogenetic tree is $(n + 1)$, where n is the number of different SNPs that can be found. Occasionally, one mutation is not transmitted, because the individual(s) carrying it have not passed it on to offspring. On these simple trees, only three of the four possible haplotypes defined by two SNPs may be observed. The two loci are thus in strong LD.

There are two possible mechanisms by which the fourth haplotype can come about: (1) homologous recombination or gene conversion (that is, transfer of the mutation without exchange of the rest of the neighboring genetic material), and (2) recurrent mutation. From data on the Y chromosome, where recombination and gene conversion cannot in general occur, recurrent SNP mutations are expected to be infrequent. We can therefore conclude that in most instances a fourth haplotype arises from recombination rather than from recurrent mutation. For SNPs that are distant from each other on the same chromosome, the probability, during meiosis, of the occurrence of a recombination between them is sufficiently high to lead to a rapid decrease in LD.

Over large distances the meiotic recombination frequency is not equivalent throughout the genome and demonstrates small variations.⁶ What is becoming apparent however is that over small distances, the probability of recombination is quite variable.^{20, 21} There seem to be “hot spots” or regions of frequent recombination or gene conversion that separate discrete blocks (“islands”) within which strong LD between many, frequent SNPs is maintained. The average size of blocks of strong LD is of the order of 40 Kb in the Asian and European population and may be twice smaller in the African population. However, their individual sizes differ widely varying from less than a few Kb to hundreds of Kb. The LD blocks, demonstrate comparatively low diversity, often four to five haplotypes accounting for over 90% of all observed haplotypes. The phylogenetic trees of these common haplotypes is therefore simple.

The aim of the haplotype map is to identify within the genomes of several major populations the position of each recombination hot spot and associated blocks of high LD, to list and organise along a phylogenetic tree the most frequent haplotypes and to identify the small number of diagnostic SNPs²² that would enable their unambiguous identification. Presently, it is estimated that, when the haplotype map will be

completed, the typing of about 500 000 SNPs will be required to ascertain the entire haplotype composition of an individual.

IDENTIFICATION OF DISEASE RELATED ALLELES

The number of polymorphic loci is large. However, only a small fraction is expected to measurably influence biological characters. Identification of functional polymorphisms and more specifically those that influence occurrence, age of onset, the clinical course, or response to treatment of frequent diseases that have an impact on public health is becoming an important endeavour and the most important challenge that faces human genetics.

A SNP allele functionally involved in a trait should be more frequently observed in those individuals with the trait than in those without it. As can be understood from the previous paragraph, the reciprocal proposition is not true. When an allele is more frequently observed in disease cases than in controls that are adequately matched (that is, they have the same ethnic and demographic composition), it is said to be associated with the disease. The observation of an association with a disease does not demonstrate that an allele is functionally implicated in the disease, but it strongly suggests that the functional polymorphism lies in the same block of LD.

Knowledge of the haplotype map will substantially change the strategy used to identify among 10^7 frequent SNPs those that are involved in the definition of a trait.²⁰ More specifically, it will enable a two step strategy aimed at answering two specific questions: (1) "Which one(s), among the 200 000 LD blocks may be associated with the trait?" and (2) "Which SNP(s) among those that lie within the associated block(s) is (are) functionally implicated in the trait?". In theory, the first question may be answered solely by genetics; the response to the second question will often require other approaches such as *in vitro* functional studies and development of experimental animal models.

Although, it has been considered as a possible approach,²³ and will probably be used in the future, genome wide association studies of all existing LD blocks to answer the first question are not yet technically feasible, requiring the genotyping of some 500 000 SNPs on phenotypically characterised large cohorts of individuals. It is still necessary to apply additional methods to select a group of blocks with substantial probability to harbour a functional SNP for the trait under study. There are at presently two possible methods, that may be combined, to reach such a goal:

- Linkage studies leading to positional candidate genes
- Functional genomics leading to functional candidate genes

Linkage studies

If during meiosis, the rate of recombination between several LD blocks in the same family is rather small, when two family members have inherited the same block from an ancestor, there is a high probability that they have also inherited identical adjacent blocks. The physical distance over which this probability remains high depends on the number of meioses that separates the two relatives from their common ancestor. Typically for sibs, this distance is of the order of 50 million base pairs so that by testing about 400 highly polymorphic loci evenly distributed throughout the genome, it is possible to infer with a high probability which regions of the genome transferred from either parent are identical in two sibs.

For family members that share the same trait, linkage studies aim at identifying chromosomal regions that tend to be more frequently inherited from a common ancestor than predicted by Mendel's law. This strategy has been most effective in identifying genes implicated in monogenic, highly penetrant diseases for which functional studies had failed to suggest candidate genes, such as cystic fibrosis and adenomatous polyposis coli. It is presently being used with some success for

more frequent multifactorial diseases. Its main disadvantage for use with multifactorial diseases is low statistical power because of decreased penetrance and to the presence of phenocopies—that is, patients who show a pathognomonic manifestation in the absence of an underlying genetic determinant. Thus in order to be applied to common complex disorders effectively, this method requires the recruitment of a large number (typically several hundreds to thousands) of families with multiple affected individuals. Another difficulty in extending this method to multifactorial disorders is that, in contrast with monofactorial, fully penetrant, diseases, the localisation of the functional SNP(s) remains imprecise, typically found in a 10 to 20 megabase region. Such a region would contain in the order of a 1000 blocks.

Thus, it is becoming increasingly clear that prerequisites to identifying functional polymorphisms implicated in common diseases, in addition to genomic tools, are large, preferably family based, cohorts, containing thousands of patients rigorously selected according to standardised criteria, together with matched controls. Cohorts based on the families with multiple affected individuals can be used initially for linkage analysis, and subsequently for search of association.²⁴

Functional genomics

On many previous occasions, studies of a given trait have identified a biological process that would participate in its determination. Some knowledge has been gained on the protein or the family of proteins that participate in the determination of the trait. The genes encoding such proteins are called functional candidates and may be tested for the presence of functional DNA variants. From the knowledge of the haplotype map it will be possible to identify the block or blocks in which functional candidate genes lie and test them directly for association.

However, there have appeared recently new functional approaches that may be used in a more systematic fashion. The sequencing of the human genome has revealed that the number of human genes is much smaller than anticipated, of the order of 30 000–40 000. This number is about twice that for the fly *Drosophila* and for the worm *C. elegans*. Development of high density arrays of cDNAs (DNA chips), has now made it possible to evaluate in a single experiment the expression at the transcriptional level of almost the entire set of human genes.^{25–26} Such experiments entail the simultaneous hybridisation of DNA copies of messenger RNA from a given cell population to thousands to tens of thousands of different cDNA molecules arrayed on microscope slides of other suitable supports. This expression profiling can be most profitably applied to the comparison of the transcriptome of relevant cells of individuals possessing or not possessing a trait. It can also be applied to the comparison of cancer cells demonstrating different characteristics.²⁷ The objective of such studies is to identify a set of genes with different expression levels in the two groups of individuals, thus providing potential functional candidate genes and suggesting underlying molecular mechanisms. The advantage of this approach is that, as for the linkage method, it requires no *a priori* functional knowledge or hypothesis.

CONCLUSION

The genomic approach will very soon present mankind with the complete sequence of the human genome. To date this approach has resulted in the construction of the genome genetic map, which was essential for the genome physical map upon which the entire sequence is being determined. These maps, including the sequences already available, have proved to be key tools for the identification of genes implicated in monogenic disorders. These tools are becoming more precise and powerful, and new tools, such as the haplotype map, are being generated. In parallel, technical improvements provide

tremendous high throughput for genotyping and expression profiling. Together these advances offer dazzling promises for the future of genetic medicine, especially for the identification of genetic variations influencing multifactorial diseases, particularly those of public health importance, and response to treatment. Such identification will provide avenues for the search of specific treatment and prevention that could be tailored to the genetic constitution of each individual. The detailed knowledge of the human genome as applied to populations throughout the world will also permit understanding their origins and the history of genes implicated in modern day diseases. Genomics has initiated a new era in biomedicine.

Authors' affiliations

G Thomas, H Cann, Centre d'Etude du Polymorphisme Humain, Paris, France

REFERENCES

- 1 **Morgan TH**. *The theory of genes*. New Haven, CT: Yale University Press, 1928.
- 2 **Botstein D**, White RL, Skolnick M, *et al*. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am J Hum Genet* 1980;**32**:314–31.
- 3 **Jeffreys AJ**, Royle NJ, Wilson V, *et al*. Spontaneous mutation rates to new length alleles at tandem-repetitive hypervariable loci in human DNA. *Nature* 1988;**332**:278–81.
- 4 **Weber JL**, May PE. Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction. *Am J Hum Genet* 1989;**44**:388–96.
- 5 **Dausset J**, Cann H, Cohen D, *et al*. Centre d'étude du polymorphisme humain (CEPH): collaborative genetic mapping of the human genome. *Genomics* 1990;**6**:575–7.
- 6 **Broman KW**, Murray JC, Sheffield VC, *et al*. Comprehensive human genetic maps: individual and sex-specific variation in recombination. *Am J Hum Genet* 1998;**63**:861–9.
- 7 **Kong A**, Gudbjartsson DF, Sainz J, *et al*. A high-resolution recombination map of the human genome. *Nat Genet* 2002;**31**:241–7.
- 8 **Olivier M**, Aggarwal A, Allen J, *et al*. A high-resolution radiation hybrid map of the human genome draft sequence. *Science* 2001;**291**:1298–302.
- 9 **Albertsen HM**, Abderrahim H, Cann HM, *et al*. Construction and characterization of a yeast artificial chromosome library containing seven haploid human genome equivalents. *Proc Natl Acad Sci USA* 1990;**87**:4256–60.
- 10 **McPherson JD**, Marra M, Hillier L, *et al*. A physical map of the human genome. *Nature* 2001;**409**:934–41.
- 11 **Chumakov IM**, Rigault P, Le Gall I, *et al*. A YAC contig map of the human genome. *Nature* 1995;**377**:175–297.
- 12 **Venter JC**, Smith HO, Hood L. A new strategy for genome sequencing. *Nature* 1996;**381**:364–6.
- 13 **Olson M**, Hood L, Cantor C, *et al*. A common language for physical mapping of the human genome. *Science* 1989;**245**:1434–5.
- 14 **Lander ES**, Linton LM, Birren B, *et al*. Initial sequencing and analysis of the human genome. *Nature* 2001;**409**:860–921.
- 15 **Adams MD**, Dubnick M, Kerlavage AR, *et al*. Sequence identification of 2,375 human brain genes. *Nature* 1992;**355**:632–4.
- 16 **Sachidanandam R**, Weissman D, Schmidt SC, *et al*. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* 2001;**409**:928–33.
- 17 **Kaessmann H**, Wiebe V, Paabo S. Extensive nuclear DNA sequence diversity among chimpanzee. *Science* 1999;**286**:1159–62.
- 18 **Owens K**, King MC. Genomic views of human history. *Science* 1999;**286**:451–3.
- 19 **Rosenberg NA**, Pritchard JK, Weber JL, *et al*. Genetic structure of human populations. *Science* 2002;**298**:2381–5.
- 20 **Goldstein DB**. Islands of linkage disequilibrium. *Nat Genet* 2001;**29**:109–11.
- 21 **Gabriel SB**, Schaffner SF, Nguyen H, *et al*. The structure of haplotype blocks in the human genome. *Science* 2002;**296**:2225–9.
- 22 **Johnson GCL**, Esposito L, Barratt BJ, *et al*. Haplotype tagging for the identification of common disease genes. *Nat Genet* 2001;**29**:233–7.
- 23 **Weiss KM**, Clark AG. Linkage disequilibrium and the mapping of complex human traits. *Trends Genet* 2002;**18**:19–24.
- 24 **Hugot JP**, Chamaillard M, Zouali H, *et al*. Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. *Nature* 2001;**411**:599–603.
- 25 **Brown PO**, Botstein D. Exploring the new world of the genome with DNA microarrays. *Genetics* 1999;**21** (suppl 1):33–7.
- 26 **Lockhart DJ**, Winzler EA. Genomics, gene expression and DNA arrays. *Nature* 2000;**405**:827–36.
- 27 **van't Veer LJ**, Hongyue D, Van De Vijver MJ, *et al*. Gene expression profiling predicts clinical outcome of breast cancer. *Nature* 2002;**415**:530–6.