Towards a haplotype map of the human genome

Joining forces

A new map of human variation will greatly aid research on the genetic origins of disease.

On 29 October 2002, a group of scientists gathered in Washington DC to launch the International HapMap Project – a major new initiative to create a map of human genetic variation.

Genetically speaking, humans are incredibly similar to one another. Any two unrelated genome sequences differ at only one position in a thousand, on average. The 0.1 per cent difference, which amounts to about three million base pairs of DNA in total, is what makes each of us genetically unique.

Everyone contains two complete copies of the genome sequence, one inherited from each of our parents. Our individual genetic make-up is therefore a unique combination of variations passed down through successive generations in the family.

Much genetic diversity (around 90 per cent) consists of single nucleotide polymorphisms (SNPs), specific positions (or loci) in the genome sequence that are occupied by one nucleotide in some copies and by a different nucleotide in others. For example, someone might inherit a C from one parent and a T from the other. In this case they would carry both variants (also known as alleles; Figure 1).

SNPs are scattered liberally through the genome. While most of them are found outside genes and probably do not have any effect, those located in and around genes may contribute to the genetic basis of our biological individuality, for example affecting how tall we grow, how well we burn fats or carbohydrate, or how we break down toxins in the environment.

In particular, scientists hope to identify SNP alleles that are associated with increased or decreased susceptibility to common diseases. Further investigation can help to isolate the gene, enabling investigation of its normal function, providing insight into disease mechanisms, the opportunity for predictive testing, and offering the ultimate possibility of new therapeutic strategies.

Scientists are also interested in finding SNP alleles that influence the way people respond to drugs. These may offer the promise of personalized medicine, where treatments are tailored to individuals based on their genetic constitution.

Association studies

The approach used to identify SNP alleles associated with disease susceptibility is conceptually quite straightforward. Once a SNP has been found at a particular locus, individuals can be tested to find out which allele they possess – a process known as genotyping. Two large sample populations are screened, a healthy group and one with the disease. Researchers then compare the frequency of the different alleles in each sample population.

If one allele is significantly more common in a disease population than in a healthy group, this suggests that the allele confers susceptibility to the disease (or that the alternative allele confers a degree of resistance). Where an association is found, the SNP is likely to be within or at least near to a gene that influences disease susceptibility.

Although straightforward in principle, such association studies are very demanding in practice. This is because associations can only be found by sifting individually through each of the millions of SNPs in the genome.

Moreover, in order to generate reliable data, large populations need to be screened, and the number of typing assays soon becomes unmanageable. For example, the typing of three million SNPs in 1000 people would involve three billion separate typing assays. Ideally, scientists would like a shortcut that would enable data to be collected from the whole genome with much less effort and expense – and this is where haplotypes may come in.

For many years, researchers have been aware of a phenomenon called linkage disequilibrium – the tendency for alleles at separate sites in the genome (in this case SNP alleles) to be found together more frequently than would be expected by chance. Linkage disequilibrium occurs because neighbouring SNPs are so close together on the chromosome that they are rarely separated by recombination – the shuffling of chromosomes that takes place during sex cell formation (meiosis). Such alleles therefore often behave as a single package when they are inherited, and the combination of alleles travelling in this fashion is known as a haplotype. Where the same pattern of alleles in a region is found in many individuals, it has been described as a haplotype block that is common in the population.

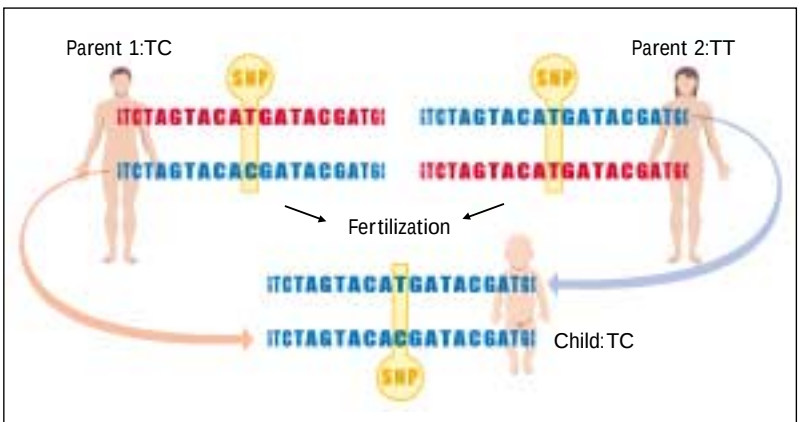


Figure 1: Variation at a single nucleotide (a SNP). The father carries T and C SNP alleles, the mother T and T. Their child has inherited T and C alleles.

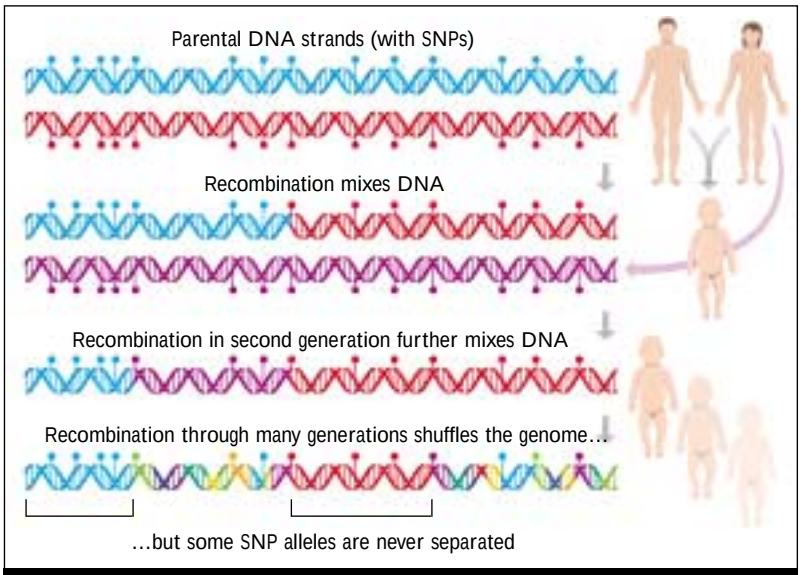


Figure 2: During egg and sperm formation, parental DNA strands exchange DNA by recombination. Over many generations, this will shuffle DNA sequences, but some blocks of DNA appear to maintain their integrity.

Recent studies suggest that the genome may be divided into a remarkably small number of blocks – just 200 000 or so. Recombination seems to be focused between the haplotype blocks, so large groups of alleles end up travelling together (Figure 2). Therefore, rather than the millions of allele combinations potentially available, the human population seems to be made up of a more limited set of haplotype patterns.

This haplotype arrangement appears to be similar in all the different populations around the world, suggesting that many of them represent ancestral haplotypes that existed in the earliest humans. The existence of a restricted set of haplotypes is probably a reflection of our recent evolutionary origins. In simple terms, there haven't been enough generations of human beings to mix up the genome completely. (This characteristic also means that haplotypes are increasingly being used to study changes to human populations – evolution, migrations and so on.)

The relatively low level of shuffling not only reduces the total amount of variation in the human population, but it also considerably simplifies the process of finding disease genes. By carefully selecting individual SNPs that unambiguously define a particular haplotype in a block (haplotype tag SNPs or ht-SNPs), a much smaller number of typing assays will be sufficient to predict the pattern of alleles for all the other SNPs in the same block without having to genotype them individually. Therefore, far fewer SNPs, perhaps only 300 000–500 000, will need to be typed in association studies to provide genome-wide coverage. This is within the capability of today's high-throughput typing technologies.

How the map will be made

The International HapMap Project will be carried out by a consortium of public and private institutions in five countries – Canada, China, Japan, the USA and the UK (led by David Bentley the Wellcome Trust Sanger Institute and Lon Cardon at the Wellcome Trust Centre for Human Genetics in Oxford). It is expected to take three years to complete.

The map will be based on DNA samples obtained from hundreds of people in geographically distinct populations: Nigerian Yorubas, Han Chinese, Japanese, and US residents of European origin. These populations have been selected for their diverse population histories, which may result in differences in haplotype structure and frequencies, and are not meant to be representative of different ethnic or racial groups.

DNA samples will be distributed to the participating organizations who will use high-throughput genotyping technologies to work out the organization of common haplotype patterns in the genome.

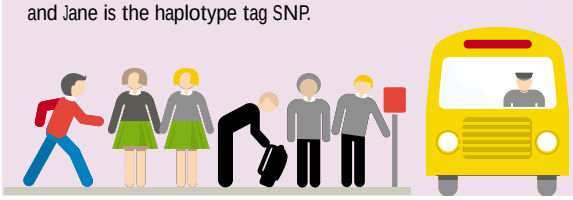
We are only just beginning to get a handle on the nature of variation within the human genome, patterns of variation across the globe, and how variation has changed during human history. The International HapMap project is sure to provide deep insight into the characteristics and distribution of SNP alleles and haplotype blocks – some of the most important aspects of our genetic inheritance. *RT*

SNP tagging

A good way to think about haplotypes and haplotype blocks is to imagine the SNP alleles as children sharing school minibuses.

If there are lots of children and they arrive at the bus stop individually, then the combination of children on any one minibus is going to be random. However, if all the children living in the *same street* walk to the bus stop together, they are likely to catch the *same bus* and they will tend to travel to school every day as the *same group*.

So, if Jane, John, Fred and Annabel live in the same street, they will tend to share the same school bus. Therefore, if Jane is on a particular bus, it follows that John, Fred and Annabel are on it too. In this example, Jane, John, Fred and Annabel are SNPs making up the haplotype, and Jane is the haplotype tag SNP.



HapMap ethics

As with any genetic resource there are two primary concerns:

- consent and privacy
- misuse of the haplotype map, or results obtained from it.

The first issue is being addressed through a sampling strategy that places considerable importance on initial community engagement and information transfer. Potential donors will made aware of how samples will be collected and used, and advisory groups will remain in the communities after the samples have been obtained to provide continuous liaison.

Although each sample will be marked with the population of origin, there will be no information that can be used to identify individual donors. More samples will be collected than are required so that no-one will know whose DNA was actually used to develop the map.

The second issue is important because the HapMap project will help to identify genetic variants associated with disease and drug responses. One danger is that variants more common in certain populations will be mistakenly used to characterize entire populations.

At an individual level, individuals may find themselves 'labelled' according to their haplotype characteristics – with the possibility that too much emphasis will be placed on the presence of a particular haplotype (in most cases, the effects of particular SNP alleles are likely to be small, and their detrimental effects will be neither inevitable nor insuperable).

These problems, common to any human genetic study, call for clarity in the way the project and its results are communicated, and the HapMap Project includes resources specifically set aside for educational purposes. Specifically, a website is being set up to communicate the goals, rationale, and implications of the project, which will be updated regularly.