tion history demand the analysis of several genes, with the most promising approach involving haplotypes[10], which consist of several closely spaced (linked) polymorphisms. The advantage of haplotypes over simply analysing polymorphisms at random is that there is valuable information in the associations between linked polymorphisms — the whole is greater than the sum of the parts. So the 1.4 million SNPs are a welcome resource that will greatly help in identifying haplotypes for tracing human evolutionary history, especially those that might reveal archaic non-African ancestry.

However, answering all of our questions about human evolutionary history will not be as simple as mining the SNP database and determining haplotypes in a representative sample of worldwide populations. There are four main reasons for that.

First, to be really useful, the SNPs in the database should really be SNPs, and not errors or artefacts, and they should be polymorphic in other samples, not just the sample of individuals used to find the SNPs. An important aspect of the SNP working group's data is that 1,585 SNPs were chosen for further verification, of which about 95% turned out to be true SNPs, which is good news indeed. Moreover, 1,276 SNPs were tested on additional population samples and at least 82% were polymorphic, which is reassuring.

Second, one might ask why only 0.1% of the 1.4 million SNPs were verified and tested. The answer is that our ability to determine allele frequencies efficiently and inexpensively for large numbers of SNPs lags behind our ability to simply identify them. This situation is reminiscent of the beginnings of the Human Genome Project, when developing technology was a primary concern and it was not at all clear how the 3.2 billion nucleotides were going to be determined. But human ingenuity won out then, and given the number of bright and capable minds now wrestling with the SNP-typing problem, one or more solutions should soon be at hand (especially with the motivation of lucrative commercial applications).

Third, a problem known as ascertainment bias can complicate the interpretation of results based on SNPs. For example, SNPs that were found to be polymorphic in European populations will overestimate genetic diversity in European as opposed to non-European populations. Moreover, the probability of finding a SNP, and the frequency of polymorphism at a SNP, depends on how many times a particular DNA segment was sequenced, and from how many individuals. The SNP working group report some intriguing preliminary findings regarding how SNP diversity is apportioned among chromosomes. But further work is required to see if these are truly biological differences, or if they instead reflect ascertainment biases. Ascertainment bias is not an insurmountable problem — statistical geneticists love this sort of challenge and are already coming up with creative solutions[11]. Even so, SNP-finders must keep careful track of how their SNPs were ascertained.

Fourth, the emphasis in the SNP database is on SNPs where both of the alleles occur at high frequency, because these will be most useful for disease-association studies. In general, the higher the frequency of a SNP allele, the older the mutation that produced it, so high-frequency SNPs largely predate human population diversification. But many questions in human evolution involve specific migrations (such as the colonization of Polynesia or the Americas) for which population-specific alleles are most informative — indeed, this is one of the attractions of mitochondrial-DNA and Y-chromosome analyses for such questions, because population-specific alleles can be readily found. It is unlikely that Polynesian-specific SNPs are present in the database, so more work will be required to find such informative, population-specific SNPs.

Still, one can imagine that in the not-too-distant future the details of human population history will have been fleshed out, at least to the extent possible by analysing genetic variation in extant populations. What then? One area that is receiving increasing attention is the detection of the effects of natural selection in human populations[12]. Using SNPs to find chromosomal regions with abnormally low levels of variation is a particularly promising way of detecting the genomic signature of selection for favourable mutations[13].

Another area of increasing interest is identifying the molecular genetic basis of 'normal' phenotypic variation[4] — that is, variation of the old-fashioned, morphological kind, which is a traditional concern of anthropology. Molecular anthropology has for the most part concentrated on the molecules and what their diversity tells us about human evolution. With the advent of the human genome sequence and the SNP database, the ultimate in molecular tools, we are ironically now poised to focus on phenotypes and what their diversity tells us about human evolution — thereby bringing the anthropology back into molecular anthropology.

*Mark Stoneking is at the Max Planck Institute for Evolutionary Anthropology, Inselstrasse 22, D-04103 Leipzig, Germany.*
*e-mail: stoneking@eva.mpg.de*

1. Mourant, A. E. *Blood Relations* p.13 (Oxford Univ. Press, 1983).
2. Hirszfeld, L. & Hirszfeld, H. *Anthropologie* **29,** 505–537 (1919).
3. Crow, J. F. *Genetics* **133,** 4–7 (1993).
4. Weiss, K. M. *Genome Res.* **8,** 691–697 (1998).
5. Collins, F. S., Brooks, L. D. & Chakravarti, A. *Genome Res.* **8,** 1229–1231 (1998).
6. The International SNP Map Working Group *Nature* **409,** 928–933 (2001).
7. Li, W. H. & Sadler, L. A. *Genetics* **129,** 513–523 (1991).
8. Chakravarti, A. *Nature* **409,** 822–823 (2001).
9. Stoneking, M. *Evol. Anthropol.* **2,** 60–73 (1993).
10. Tishkoff, S. A. *et al. Science* **271,** 1380–1387 (1996).
11. Kuhner, M. K., Beerli, P., Yamato, J. & Felsenstein, J. *Genetics* **156,** 439–447 (2000).
12. Przeworski, M., Hudson, R. R. & Di Rienzo, A. *Trends Genet.* **16,** 296–302 (2000).
13. Nurminsky, D., De Aguiar, D., Bustamante, C. D. & Hartl, D. L. *Science* **291,** 128–130 (2001).

Single nucleotide polymorphisms

# …to a future of genetic medicine

Aravinda Chakravarti

Single base differences between human genomes underlie differences in susceptibility to, or protection from, a host of diseases. Hence the great potential of such information in medicine.

The beginning of the Human Genome Project, over a decade ago, was accompanied by a cantankerous debate over whose genome was to be sequenced. Would it be a single individual? A celebrity, perhaps (widely rumoured to be Jim Watson, co-discoverer of the structure of DNA)? Or would several genomes, from many individuals, be studied? The discussion struck at the very heart of genetics. As the study of inherited variation between individuals, genetics might not immediately benefit from the sequence of a single genome. But even one genome would be immensely revealing to the science of deciphering the molecular blueprint of a species. Fortunately, geneticists were not forced to make this choice. Papers in this issue describe not only a single, history-making human genome sequence, composed of little bits from many humans[1] (page 860), but also some 1.4 million sites of variation mapped along that reference sequence[2] (page 928).

But why this preoccupation with sequence variation, with the fact that no two humans (except identical twins) are genetically the same? The answer is that such variations, or 'polymorphisms', are markers of genes and genomes with which researchers perform genetic analysis in an outbred species where matings cannot be controlled. The fields of human and medical genetics simply cannot exist without understanding this variation.

It has become clear that the two 'genomes' that each of us carry, inherited

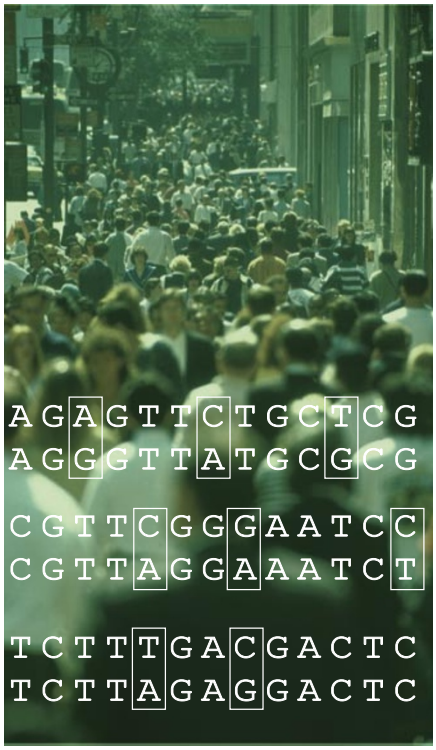


Figure 1 **The most common sources of variation between humans are single nucleotide polymorphisms (SNPs) — single base differences between genome sequences. Fragments of two sequences, with eight SNPs, are shown.**

from our parents, most often differ — from each other, and from the genomes of other humans — in terms of single base changes[1] (Fig. 1). The twentieth century saw the identification of only a few thousand of these so-called single nucleotide polymorphisms (SNPs, or 'snips' to the streetwise). In just the first year of the new century, this number has been increased one-thousand-fold[2]. Beyond the numbers, the excitement today comes from precise knowledge of where these sites of variation are in the genome[2]. The 1.42 million known SNPs are found at a density of one SNP per 1.91 kilobases. This means that more than 90% of any stretches of sequence 20 kilobases long will contain one or more SNPs. The density is even higher in regions containing genes. The International SNP Map Working Group[2] estimates that they have identified 60,000 SNPs within genes ('coding' SNPs), or one coding SNP per 1.08 kilobases of gene sequence. Moreover, 93% of genes contain a SNP, and 98% are within 5 kilobases of a SNP. For the first time, nearly every human gene and genomic region is marked by a sequence variation.

These data provide interesting first glimpses into the pattern of variation across the genome. Variation is commonly assessed by nucleotide diversity — the number of base differences between two genomes, divided by the number of base pairs compared. Nucleotide diversity is a sensitive indicator of biological and historical factors that have affected the human genome[3]. The nucleotide diversity in gene-containing regions has been estimated to be 8 differences per 10 kilobases[4,5]; we now know that the genome-wide average is similar, 7.51 differences per 10 kilobases (ref. 2). The variation between individual non-sex chromosomes is small, and lies in the range 5.19 (for chromosome 21) to 8.79 (for chromosome 15) differences per 10 kilobases (ref. 2).

Strikingly, humans vary least in their sex chromosomes. The variation between different X chromosomes is about 4.69 differences per 10 kilobases, and it is very much lower for the Y chromosome (1.51 differences per 10 kilobases). This is because the sex chromosomes have patterns of mutation and recombination (the swapping of similar DNA segments during the generation of eggs and sperm) that differ both from each other and from the non-sex chromosomes. Moreover, fewer ancestors have contributed to the sex chromosomes, which are therefore less variable than the non-sex chromosomes.

Perhaps not surprisingly, some genomic regions have significantly lower or higher diversity than the average. For example, the HLA locus, which encodes proteins that present antigens to the immune system, shows the greatest diversity. Such comparisons within genomes will be essential to our understanding of how variation shapes biochemical and cellular functions, and in illuminating past human evolution, as discussed in ref. 3, and by Stoneking in the preceding article (page 821; ref. 6).

But the main use of the human SNP map will be in dissecting the contributions of individual genes to diseases that have a complex, multigene basis. Knowledge of genetic variation already affects patient care to some degree. For example, gene variants lead to tissue and organ incompatibility, affecting the success of transplants. And the mainstay of medical genetics has been the study of the rare gene variants that lie behind inherited diseases such as cystic fibrosis.

But variations in genome sequences underlie differences in our susceptibility to, or protection from, all kinds of diseases; in the age of onset and severity of illness; and in the way our bodies respond to treatment. For example, we already know that single base differences in the *APOE* gene are associated with Alzheimer's disease, and that a simple deletion within the chemokine-receptor gene *CCR5* leads to resistance to HIV and AIDS. The benefit of the SNP map is that it covers the entire genome. So, by comparing patterns and frequencies of SNPs in patients and controls, researchers can identify which SNPs are associated with which diseases[7–9]. Such research will bring about 'genetic medicine', in which knowledge of our uniqueness will alter all aspects of medicine, perceptibly and forever.

Studies of SNPs and diseases will become more efficient when a few more problems are solved[3]. First, although 82% of SNP variants are found at a frequency of more than 10% in the global human population, the 'microdistribution' of SNPs in individual populations is not known. Second, not all SNPs are created equal, and it will be essential to know as much as possible about their effects from computational analyses before studying their involvement in disease. For example, each SNP can be classified by whether it is coding or not. Coding SNPs can be classified by whether they alter the sequence of the protein encoded by the altered gene. Changes that alter protein sequences can be classified by their effects on protein structure. And non-coding SNPs can be classified according to whether they are found in gene-regulating segments of the genome[10] — many complex diseases may arise from quantitative, rather than qualitative, differences in gene products. Third, the technology for assaying thousands of SNPs, in thousands of patients and controls[7], is not yet fully developed, although there are some creative ideas around.

In the twentieth century, humans were not the geneticists' species of choice. The emphasis then was on understanding gene structure and function. Now, geneticists will concentrate increasingly on understanding physical and behavioural characteristics. Here, our species, with its obsession with self-examination, will make a superior subject. We will also see more studies of how natural variation leads to each one of our qualities. To some, there is a danger of genomania, with all differences (or similarities, for that matter) being laid at the altar of genetics[11]. But I hope this does not happen. Genes and genomes do not act in a vacuum, and the environment is equally important in human biology. By identifying variation across the whole genome, the SNP map[2] may be our best route yet to a better understanding of the roles of nature and (not versus) nurture. ■

*Aravinda Chakravarti is at the McKusick–Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, 600 North Wolfe Street, Jefferson Street Building 2-109, Baltimore, Maryland 21287, USA.*
*e-mail: aravinda@jhmi.edu*

1. International Human Genome Sequencing Consortium *Nature* **409,** 860–921 (2001).
2. The International SNP Map Working Group *Nature* **409,** 928–933 (2001).
3. Chakravarti, A. *Nature Genet.* **21** (suppl.), 56–60 (1999).
4. Halushka, M. K. *et al. Nature Genet.* **22,** 239–247 (1999).
5. Cargill, M. *et al. Nature Genet.* **22,** 231–238 (1999).
6. Stoneking, M. *Nature* **409,** 821–822 (2001).
7. Risch, N. & Merikangas, K. *Science* **273,** 1516–1517 (1996).
8. Lander, E. S. *Science* **274,** 536–539 (1996).
9. Collins, F. S., Guyer, M. S. & Chakravarti, A. *Science* **278,** 1580–1581 (1997).
10. Loots, G. G. *et al. Science* **288,** 136–140 (1999).
11. Lewontin, R. *It Ain't Necessarily So: The Dream of the Human Genome and Other Illusions* (New York Review of Books, 2000).