

to determine the structure of genes — which parts (the exons) make their way into a functional mRNA molecule and which do not (the introns). The high degree of alternative splicing in vertebrates makes this comparative approach particularly important. Gene-finding computational algorithms cannot easily predict the existence of alternative forms of an mRNA without experimental information, but this information is difficult to come by in the case of rare mRNAs. For example, an exon that is used in only a few cells of the human brain might never be experimentally detected in an mRNA. But that exon's sequence would probably be conserved in the mouse genome.

Comparing the genomes of closely related species can also help in identifying gene-control regions. This approach has been used for over two decades¹¹, and has been validated by showing that the conserved sequences indeed correspond to functional control elements in individual genes¹². But this computational problem is more difficult than identifying exons, and it will be challenging to scale up to a genome-wide level. The proteins that control gene expression by recognizing regulatory regions often detect sequence features that elude the best computer algorithms, and may use information from contacts with other proteins that is difficult to model. Proteins are simply cleverer than computers.

That said, our knowledge of the DNA-

binding properties of individual proteins, as well as the structural features of the DNA sites to which they bind, continues to increase. Moreover, we can use experimental evidence; for example, genes that are expressed together might be expected to share control elements. And, as methods for comparing sequences continue to improve, we can expect to learn more about elusive features of the genome, such as genes encoding RNAs that do not encode proteins¹³, start points of DNA replication, and genetic elements that control chromosome structure. ■

Gerald M. Rubin is in the Department of Molecular and Cell Biology, University of California at Berkeley, Berkeley, California 94708-3200, and the Howard Hughes Medical Institute, 4000 Jones Bridge Road, Chevy Chase, Maryland 20815-6789, USA.

e-mail: gerry@fruitfly.berkeley.edu

1. International Human Genome Sequencing Consortium *Nature* **409**, 860–921 (2001).
2. Venter, J. C. *et al.* *Science* **291**, 1304–1351 (2001).
3. Adams, M. D. *et al.* *Science* **287**, 2185–2195 (2000).
4. The *C. elegans* Sequencing Consortium *Science* **282**, 2012–2018 (1998).
5. The Arabidopsis Genome Initiative *Nature* **408**, 796–815 (2000).
6. Rubin, G. M. *et al.* *Science* **287**, 2204–2215 (2000).
7. Spring, J. *FEBS Lett.* **400**, 2–8 (1997).
8. Hentze, M. W. & Kulozik, A. E. *Cell* **96**, 307–310 (1999).
9. Ashburner, M. *et al.* *Genetics* **153**, 179–219 (1999).
10. Fraser, A. G. *et al.* *Nature* **408**, 325–330 (2000).
11. Ravetch, J. V., Kirsch, I. R. & Leder, P. *Proc. Natl Acad. Sci. USA* **77**, 6734–6738 (1980).
12. Fortini, M. E. & Rubin, G. M. *Genes Dev.* **4**, 444–463 (1990).
13. Lee, R. C., Feinbaum, R. L. & Ambros, V. *Cell* **75**, 843–854 (1993).

Single nucleotide polymorphisms

From the evolutionary past...

Mark Stoneking

Single nucleotide polymorphisms are the bread-and-butter of DNA sequence variation. They provide a rich source of information about the evolutionary history of human populations.

Studies of genetic variation in human populations began inauspiciously¹. The first such study — of ABO blood-group frequencies — was carried out by two Polish immunologists, Ludwik and Hanka Hirszfeld, at the end of the First World War. This work was notable for its broad coverage of the world's populations, large sample sizes and scrupulous attention to anthropological details. Yet the Hirszfelds still ran into difficulties in publishing in *The Lancet*, the premier medical journal of the time. The editor could not see the relevance of their work, and so this seminal study of human genetic variation first appeared in an obscure anthropological journal². The relevance became abundantly clear when Felix Bernstein subsequently used the Hirszfelds' data to demonstrate that the ABO blood-group frequencies were better explained by a single gene with three variants (alleles), and not —

as prevailing wisdom then held — two genes each with two alleles³.

Happily, times have changed, diversity is now all the rage^{4,5}, and editors have become more appreciative of the importance of human genetic variation. The latest evidence of that is the paper on page 928 of this issue⁶, which reports the identification and mapping of 1.4 million single nucleotide polymorphisms (SNPs, pronounced 'snips') in the human genome. The paper is the result of the labours of a large collaboration, The International SNP Map Working Group.

So, what are SNPs? Quite simply, they are the bread-and-butter of DNA sequence variation — polymorphism, to those in the business. A DNA sequence is a linear combination of four nucleotides; compare two sequences, position by position, and wherever you come across different nucleotides at the same position, that's a SNP (see Fig. 1 on

page 823). So SNPs reflect past mutations that were mostly (but not exclusively) unique events, and two individuals sharing a variant allele are thereby marked with a common evolutionary heritage. In other words, our genes have ancestors, and analysing shared patterns of SNP variation can identify them.

However, the real importance of SNPs is that there are so many of them. One estimate⁷ is that comparing two human DNA sequences results in a SNP every 1,000–2,000 nucleotides. That may not sound like much until you realize that there are 3.2 billion nucleotides in the human genome, which translates into 1.6 million–3.2 million SNPs. And that's just from comparing two sequences — the total number of SNPs in humans is obviously much more. Most human variation that is influenced by genes can be traced to SNPs, especially in such medically (and commercially) important traits as how likely you are to become afflicted with a particular disease, or how you might respond to a particular pharmaceutical treatment, as discussed by Chakravarti⁸ on the following page. And even when a SNP is not directly responsible, the sheer number of SNPs means they can also be used to locate genes that influence such traits.

The deluge of SNPs reported by the SNP working group⁶ also promises great things for those of us who analyse patterns of molecular genetic variation to reconstruct the evolutionary history of human populations. Our genes contain the signature of an expansion from Africa within the past 150,000 years or so⁹. But there is still debate as to whether the modern humans from Africa completely replaced archaic non-African populations with no interbreeding, or whether we perhaps carry the vestiges of Neanderthal or other archaic non-African genes.

Demonstrating a recent African origin for every single one of our 3.2 billion nucleotides goes beyond the bounds of reason or necessity, but there is still much to be learned. For a start, most of our insights into molecular anthropology arise from DNA in mitochondria and (more recently) polymorphisms of the Y chromosome. This is because these DNA sequences are haploid — that is, represented just once in each cell, in contrast to the other chromosomes, which are represented twice — and they are inherited from just one parent, so they do not undergo the usual sequence shuffling (recombination) during egg and sperm production. This makes them easier to analyse and extremely informative. But both suffer from the drawback that, in the absence of recombination, they behave as single genes, and the history of any single gene can differ from that of a population or species because of natural selection or chance events involving that gene.

Accurate inferences concerning popula-

tion history demand the analysis of several genes, with the most promising approach involving haplotypes¹⁰, 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 solutions¹¹. 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 populations¹². Using SNPs to find chromosomal regions with abnormally low levels of varia-

tion is a particularly promising way of detecting the genomic signature of selection for favourable mutations¹³.

Another area of increasing interest is identifying the molecular genetic basis of 'normal' phenotypic variation⁴ — 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. Hirschfeld, L. & Hirschfeld, 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. & Siedler, 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 humans¹ (page 860), but also some 1.4 million sites of variation mapped along that reference sequence² (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