

The mosaic that is our genome

Svante Pääbo

Max Planck Institute for Evolutionary Anthropology, Deutscher Platz 6, D-04103 Leipzig, Germany
(e-mail: paabo@eva.mpg.de)

The discovery of the basis of genetic variation has opened inroads to understanding our history as a species. It has revealed the remarkable genetic similarity we share with other individuals as well as with our closest primate relatives. To understand what make us unique, both as individuals and as a species, we need to consider the genome as a mosaic of discrete segments, each with its own unique history and relatedness to different contemporary and ancestral individuals.

The discovery of the structure of DNA¹, and the realization that the chemical basis of mutations is changes in the nucleotide sequence of the DNA, meant that the history of a piece of DNA could be traced by studying variation in its nucleotide sequence found in different individuals and in different species. But it was not until rapid and inexpensive methods became available for probing DNA sequence variation in many individuals that the efficient study of molecular evolution in general — and of human evolution in particular — became feasible. Thus, the development in the 1980s of techniques for efficiently scoring polymorphisms with restriction enzymes and amplifying DNA^{2,3} enabled the study of molecular evolution to become a truly booming enterprise.

What follows is a personal and, by necessity, selective attempt to consider what the accelerating pace of exploration of human genetic variation over the past two decades has taught us about ourselves as a species, as well as some suggestions for what may be fruitful areas for future studies.

Primate relations

The first insight of fundamental importance for our understanding of our origins came from comparisons of DNA sequences between humans and the great apes. These analyses showed that the African apes, especially the chimpanzees and the bonobos, but also the gorillas, are more closely related to humans than are the orangutans in Asia⁴. Thus, from a genetic standpoint, humans are essentially African apes (Fig. 1). Although there had been hints of this from molecular comparisons of proteins^{5,6}, it was a marked shift from the earlier common belief that humans represented their own branch separate from the great apes.

Our sense of uniqueness as a species was further rocked by the revelation that human DNA sequences differ by, on average, only 1.2 per cent from those of the chimpanzees⁷, as a consequence of humans and apes sharing a recent common ancestry. It should be noted

that the dating of molecular divergences has uncertainties of unknown magnitude attached, not least because of calibration based on palaeontological data. Nevertheless, it seems clear that the human evolutionary lineage diverged from that of chimpanzees about 4–6 million years ago, from that of gorillas about 6–8 million years ago, and from that of the orangutans about 12–16 million years ago⁷. Before the advent of molecular data, the human–chimpanzee divergence was widely believed to be about 30 million years old.

In fact, we have recently come to realize that the relationship between humans and the African apes is so close as to be entangled. Although the majority of regions in our genome are most closely related to chimpanzees and bonobos, a non-trivial fraction is more closely related to gorillas⁷. In yet other regions, the apes are more closely related to each other than to us (Fig. 2). This is because the speciation events that separated these lineages occurred so closely in time that genetic variation in the first ancestral species, from which the gorilla lineage diverged, survived

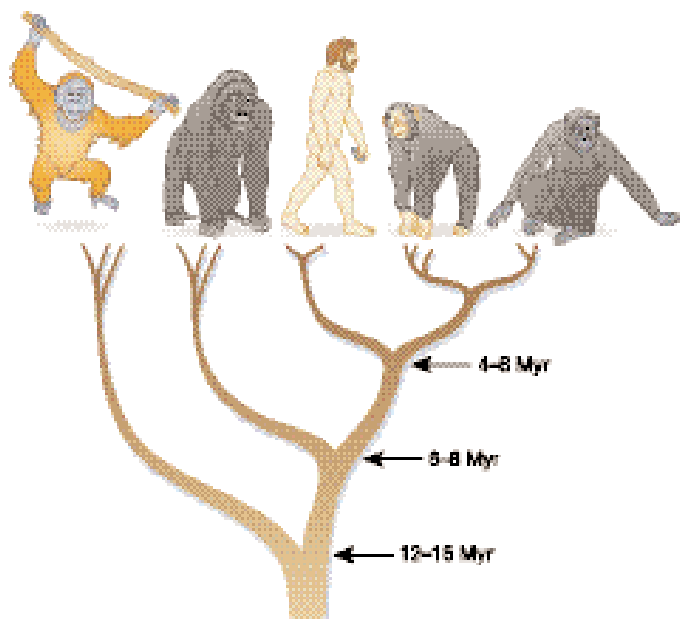
until the second speciation event between the human and chimpanzee lineages⁸. Thus, there is not one history with which we can describe the relationship of our genome to the genomes of the African apes, but instead different histories for different segments of our genome. In this respect, our genome is a mosaic, where each segment has its own relationship to that of the African apes.

Modern humans

The mosaic nature of our genome is even more striking when we consider differences in DNA sequence between currently living humans. Our genome sequences are about 99.9 per cent identical to each other. The variation found along a chromosome is structured in 'blocks' where the nucleotide substitutions are associated in so-called haplotypes (Figs 2b and 3). These 'haplotype blocks' are likely to result from the fact that recombination, that is, the re-shuffling of chromosome segments that occurs during formation of sex cells (meiosis), tends to occur in certain areas of the chromosomes more often than in others^{9–11}. In addition, the chance occurrence of recombination events at certain spots and not at others in the genealogy of human chromosomes will influence the structure of these blocks. Thus, any single human chromosome is a mosaic of different haplotype blocks, where each block has its own pattern of variation. Although the delineation of such blocks depends on the methods used to define them, they are typically 5,000–200,000 base pairs in length, and as few as four to five common haplotypes account for most of the variation in each block (Fig. 3).

Of 928 such haplotype blocks recently studied in humans from Africa, Asia and Europe¹², 51 per cent were found on all three

Figure 1 Tree showing the divergence of human and ape species. Approximate dates of divergences are given for, from left to right, orangutan, gorilla, human, bonobo and chimpanzee.



continents, 72 per cent in two continents and only 28 per cent on one continent. Of those haplotypes that were on one continent only, 90 per cent were found in Africa, and African DNA sequences differ on average more among themselves than they differ from Asian or European DNA sequences¹³. Therefore, within the human gene pool, most variation is found in Africa and what is seen outside Africa is a subset of the variation found within Africa.

Two parts of the human genome can be regarded as haplotype blocks where the history is particularly straightforward to reconstruct, as no recombination occurs at all. The first of these is the genome of the mitochondrion (the cellular organelle that produces energy and has its own genetic material), which is passed on to the next generation from the mother's side; the second is the Y chromosome, which is passed on from the father's side. Variation in DNA sequences from both the mitochondrial genome¹⁴⁻¹⁶ and the Y chromosome¹⁷, as well as many sections of the nuclear genome^{13,18-20}, have their geographical origin in Africa. Because other evidence suggest that humans expanded some 50,000 to 200,000 years ago²¹ from a population of about 10,000 individuals, this suggests that we expanded from a rather small African population. Thus, from a genomic perspective, we are all Africans, either living in Africa or in quite recent exile outside Africa.

Ancient humans

What happened to the other hominids that existed in the Old World from about 2 million years ago until about 30,000 years ago? For instance, the Neanderthals are abundant in the fossil record and persisted in western Europe until less than 30,000 years ago.

Analysis of Neanderthal mitochondrial DNA has shown that, at least with respect to the mitochondrial genome, there is no evidence that Neanderthals contributed to the gene pool of current humans²²⁻²⁵. It is possible, however, that some as yet undetected interbreeding took place between modern humans and archaic hominids, such as *Homo erectus* in Asia or Neanderthals in Europe^{22,26,27}.

But any interbreeding would not have significantly changed our genome, as we know that the variation found in many haplotype blocks in the nuclear genome of contemporary humans is older than the divergence between Neanderthals and humans. Thus, the divergence of modern humans and Neanderthals was so recent that Neanderthal nuclear DNA sequences were probably more closely related to some current human DNA sequences than to other Neanderthals. In other words, the overlapping genetic variation that is likely to have existed between different ancient hominid forms makes it difficult to resolve the extent to which any interbreeding occurred.

Nevertheless, the limited variation among humans outside Africa, as well palaeontological evidence²⁸, suggest that any contribution cannot have been particularly extensive. Thus, it seems most likely that modern humans replaced archaic humans without extensive interbreeding and that the past 30,000 years of human history are unique in that we lack the company of the closely related yet distinct hominids with which we used to share the planet.

Human variation and 'race'

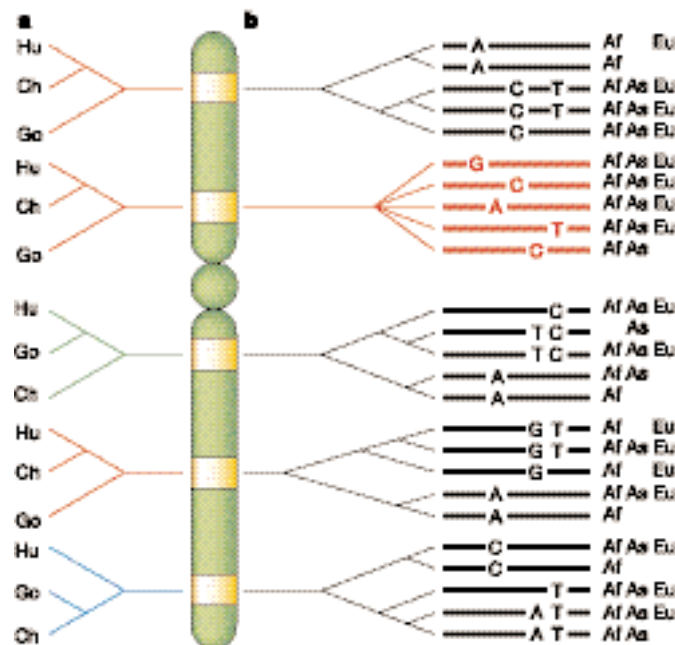
Comparisons of the within-species variation among humans and among the great apes have shown that humans have less genetic

variation than the great apes^{29,30}. Furthermore, early data that only about 10 per cent of the genetic variation in humans exist between so-called 'races'³¹ is borne out by DNA sequences which show that races are not characterized by fixed genetic differences. Rather, for any given haplotype block in the genome, a person from, for example, Europe is often more closely related to a person from Africa or from Asia than to another person from Europe that shares his or her complexion (for example, see ref. 32; Fig. 2).

Claims about fixed genetic differences between races (see ref. 33 for example) have proved to be due to insufficient sampling³⁴. Furthermore, because the main pattern of genetic variation across the globe is one of gene-frequency gradients³⁵, the contention that significant differences between races can be seen in frequencies of various genetic markers³⁶ is very likely due to sampling of populations separated by vast geographical distances. In this context it is worth noting that the colonization history of the United States has resulted in a sampling of the human population made up largely of people from western Europe, western Africa and southeast Asia. Thus, the fact that 'racial groups' in the United States differ in gene frequencies cannot be taken as evidence that such differences represent any true subdivision of the human gene pool on a worldwide scale.

Rather than thinking about 'populations', 'ethnicities' or 'races', a more constructive way to think about human genetic variation is to consider the genome of any particular individual as a mosaic of haplotype blocks. A rough calculation (Fig. 3) reveals that each individual carries in the order of 30 per cent of the entire haplotype variation of the human gene pool. Although not all of our

Figure 2 Within- and between-species variation along a single chromosome. **a**, The interspecies relationships of five chromosome regions to corresponding DNA sequences in a chimpanzee and a gorilla. Most regions show humans to be most closely related to chimpanzees (red) whereas a few regions show other relationships (green and blue). **b**, The among-human relationships of the same regions are illustrated schematically for five individual chromosomes. Most DNA variants are found in people from all three continents, namely Africa (Af), Asia (As) and Europe (Eu). But a few variants are found on only one continent, most of which are in Africa. Note that each human chromosome is a mosaic of different relationships. For example, a chromosome carried by a person of European descent may be most closely related to a chromosome from Asia in one of its regions, to a chromosome from Africa in another region, and to a chromosome from Europe in a third region. For one region (red), the extent of sequence variation within humans is low relative to what is observed between species. The relationship of this sequence among humans is illustrated as star-shaped owing to a high frequency of nucleotide variations that are unique to single chromosomes. Such regions may contain genes that contribute to traits that set humans apart from the apes.



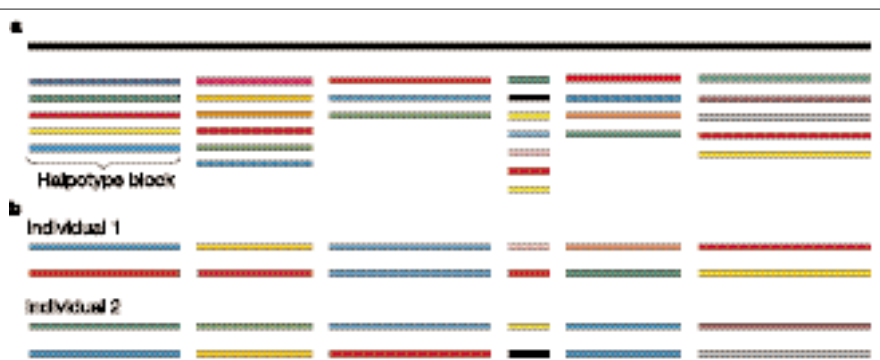


Figure 3 The mosaic structure of human genetic variation. **a**, Each human chromosome is made up of regions, called 'haplotype blocks', which are stretches of DNA sequence where three to seven variants (at frequencies above 5 per cent in the human population) account for most of the variation found among humans. Each such haplotype found in a block is illustrated here as a bar of different colour. The catalogue of haplotypes for every block makes up the 'haplotype map' of the human genome. **b**, The chromosomes of two hypothetical individuals are shown. Each individual carries two copies of each block (as humans carry two sets of chromosomes). As the chance that the two haplotypes carried at a block are identical is about 20 per cent, each of us carries an average of about 1.8 different haplotypes per block. Since there is on average 5.5 haplotypes for every block, each individual carries about 30 per cent of the total haplotype diversity of the entire human species. Haplotype blocks tend to be shorter in Africa than elsewhere; as a result, African variation will probably have to be used to define the species-wide block lengths, which may be an average of around 10,000 base pairs. Note that not all of the human genome may have a clearly definable haplotype-block structure.

genome may show a typical haplotype-block structure and more research is needed to fully understand the haplotype landscape of our genome, this perspective clearly indicates that each of us contain a vast proportion of the genetic variation found in our species. In the future, we therefore need to focus on individuals rather than populations when exploring genetic variation in our species.

Tracking human traits

What are the frontiers ahead of us in human evolutionary studies? One of them, to my mind, is to identify gene variants that have been selected and fixed in all humans during the past few hundred thousand years. These will include genes involved in phenotypic traits that set humans apart from the apes and at least some archaic human forms (for example, genes involved in complex cognitive abilities, language and longevity). However, an important obstacle in this respect is that there is little detailed knowledge of many of the relevant traits in the great apes. For example, only recently has the extent to which apes possess the capability for language³⁷ and culture³⁸ begun to be comprehensively described. As a consequence, we have come to realize that almost all features that set humans apart from apes may turn out to be differences in grade rather than absolute differences.

Many such differences are likely to be quantitative traits rather than single-gene traits. To have a chance to unravel the genetic basis of such traits, we will need to rigorously define the differences between apes and humans — for instance, how we learn, how we communicate and how we age. In the next

few years, geneticists will therefore need to consider insights from primatology and psychology, and more studies will be required that directly compare humans to apes.

There are, however, ways in which we can contribute towards the future unravelling of functionally important genetic differences between humans and apes. For example, we can identify regions of the human genome where the patterns of variation suggest the recent occurrence of a mutation that was positively selected and swept through the entire human population. The sequencing of the chimpanzee genome, as well as the haplotype-map project, will greatly help in this. Further prerequisites include the capability to determine the DNA sequence of many human genomes and the development of tools and methods to analyse the resulting data; in particular, a more realistic model of human demographic history is required.

Collectively these studies will allow us to identify regions in the human genome that have recently been acted upon by selection and thus are likely to contain genes contributing to human-specific traits (Fig. 2). Other interesting candidate genes for human-specific traits are genes duplicated or deleted in humans³⁹, genes that have changed their expression in humans⁴⁰, and genes responsible for disorders affecting traits unique to humans, such as language⁴¹ and a large brain size⁴².

A problem inherent in studying genes that are involved in traits unique to humans, such as language, is that functional experiments cannot be performed, as no animal model exists, and transgenic humans or chimpanzees cannot be constructed. A further difficulty is that many genes that enable

humans to perform tasks of interest may exert their effects during early development where our ability to study their expression both in apes and humans is extremely limited.

A challenge for the future is therefore to design ways around these difficulties. This will involve *in vitro* as well as *in silico* approaches that study how genes interact with each other to influence developmental and physiological systems. As these goals are achieved, we will be able to determine the order and approximate times of genetic changes during the emergence of modern humans that led to the traits that set us apart among animals. □

doi:10.1038/nature01400

- Watson, J. D. & Crick, F. H. C. A structure for deoxyribose nucleic acid. *Nature* **171**, 737–738 (1953).
- Botstein, D., White, R. L., Skolnick, M. & Davis, R. W. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am. J. Hum. Genet.* **32**, 314–331 (1980).
- Saiki, R. K. *et al.* Enzymatic amplification of β -globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. *Science* **230**, 1350–1354 (1985).
- Miyamoto, M. M., Slightom, J. L. & Goodman, M. Phylogenetic relations of humans and African apes from DNA sequences in the $\psi\eta$ -globin region. *Science* **238**, 369–373 (1987).
- Mayr, E. *Animal Species and Evolution* (Harvard Univ. Press, Cambridge, MA, 1963).
- Wilson, A. C. & Sarich, V. M. A molecular time scale for human evolution. *Proc. Natl Acad. Sci. USA* **63**, 1088–1093 (1969).
- Chen, F. C., Vallender, E. J., Wang, H., Tzeng, C. S. & Li, W. H. Genomic divergence between human and chimpanzee estimated from large-scale alignments of genomic sequences. *J. Hered.* **92**, 481–489 (2001).
- Nei, M. *Molecular Evolutionary Genetics* (Columbia Univ. Press, New York, 1987).
- Daly, M. J., Rioux, J. D., Schaffner, S. F., Hudson, T. J. & Lander, E. S. High-resolution haplotype structure in the human genome. *Nature Genet.* **29**, 229–232 (2001).
- Jeffreys, A. J., Kauppi, L. & Neumann, R. Intensely punctuated meiotic recombination in the class II region of the major histocompatibility complex. *Nature Genet.* **29**, 217–222 (2001).
- Patil, N. *et al.* Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science* **294**, 1719–1723 (2001).
- Gabriel, S. B. *et al.* The structure of haplotype blocks in the human genome. *Science* **296**, 2225–2229 (2002).
- Yu, N. *et al.* Larger genetic differences within Africans than between Africans and Eurasians. *Genetics* **161**, 269–274 (2002).
- Cann, R. L., Stoneking, M. & Wilson, A. C. Mitochondrial DNA and human evolution. *Nature* **325**, 31–36 (1987).
- Vigilant, L., Stoneking, M., Harpending, H., Hawkes, K. & Wilson, A. C. African populations and the evolution of human mitochondrial DNA. *Science* **253**, 1503–1507 (1991).
- Ingman, M., Kaessmann, H., Pääbo, S. & Gyllenstein, U. Mitochondrial genome variation and the origin of modern humans. *Nature* **408**, 708–713 (2000).
- Underhill, P. A. *et al.* Y chromosome sequence variation and the history of human populations. *Nature Genet.* **26**, 358–361 (2000).
- Stoneking, M. *et al.* Alu insertion polymorphisms and human evolution: evidence for a larger population size in Africa. *Genome Res.* **7**, 1061–1071 (1997).
- Tishkoff, S. A. *et al.* Global patterns of linkage disequilibrium at the CD4 locus and modern human origins. *Science* **271**, 1380–1387 (1996).
- Takahata, N., Lee, S. H. & Satta, Y. Testing multiregionality of modern human origins. *Mol. Biol. Evol.* **18**, 172–183 (2001).
- Harpending, H. & Rogers, A. Genetic perspectives on human origins and differentiation. *Annu. Rev. Genomics Hum. Genet.* **1**, 361–385 (2000).
- Krings, M. *et al.* Neandertal DNA sequences and the origin of modern humans. *Cell* **90**, 19–30 (1997).
- Ovchinnikov, I. V. *et al.* Molecular analysis of Neandertal DNA from the northern Caucasus. *Nature* **404**, 490–493 (2000).
- Krings, M., Geisert, H., Schmitz, R. W., Krainitzki, H. & Pääbo, S. DNA sequence of the mitochondrial hypervariable region II

- from the Neandertal type specimen. *Proc. Natl Acad. Sci. USA* **96**, 5581–5585 (1999).
25. Krings, M. *et al.* A view of Neandertal genetic diversity. *Nature Genet.* **26**, 144–146 (2000).
26. Nordborg, M. On the probability of Neandertal ancestry. *Am. J. Hum. Genet.* **63**, 1237–1240 (1998).
27. Pääbo, S. Human evolution. *Trends Cell Biol.* **9**, M13–M16 (1999).
28. Stringer, C. Modern human origins: progress and prospects. *Phil. Trans. R. Soc. Lond. B* **357**, 563–579 (2002).
29. Deinard, A. & Kidd, K. Evolution of a HOXB6 intergenic region within the great apes and humans. *J. Hum. Evol.* **36**, 687–703 (1999).
30. Kaessmann, H., Wiebe, V., Weiss, G. & Pääbo, S. Great ape DNA sequences reveal a reduced diversity and an expansion in humans. *Nature Genet.* **27**, 155–156 (2001).
31. Lewontin, R. C. The problem of genetic diversity. *Evol. Biol.* **6**, 381–398 (1972).
32. Kaessmann, H., Heissig, F., von Haesler, A. & Pääbo, S. DNA sequence variation in a non-coding region of low recombination on the human X chromosome. *Nature Genet.* **22**, 78–81 (1999).
33. Harris, E. E. & Hey, J. X chromosome evidence for ancient human histories. *Proc. Natl Acad. Sci. USA* **96**, 3320–3324 (1999).
34. Yua, N. & Li, W.-H. No fixed nucleotide difference between Africans and non-Africans at the pyruvate dehydrogenase E1 α -subunit locus. *Genetics* **155**, 1481–1483 (2000).
35. Cavalli-Sforza, L. L., Menozzi, P. & Piazza, A. *The History and Geography of Human Genes* (Princeton Univ. Press, Princeton, NJ, 1993).
36. Risch, N., Burchard, E., Ziv, E. & Tang, H. Categorization of humans in biological research: genes, race and disease. *Genome Biol.* **3**, 2007.1–2007.12 (2002).
37. Tomasello, M. & Call, J. *Primate Cognition* (Oxford Univ. Press, New York, 1997).
38. Whiten, A. *et al.* Cultures in chimpanzees. *Nature* **399**, 682–685 (1999).
39. Eichler, E. E. Recent duplication, domain accretion and the dynamic mutation of the human genome. *Trends Genet.* **17**, 661–669 (2001).
40. Enard, W. *et al.* Intra- and interspecific variation in primate gene expression patterns. *Science* **296**, 340–343 (2002).
41. Enard, W. *et al.* Molecular evolution of *FOXP2*, a gene involved in speech and language. *Nature* **418**, 869–872 (2002).
42. Jackson, A. P. *et al.* Identification of microcephalin, a protein implicated in determining the size of the human brain. *Am. J. Hum. Genet.* **71**, 136–142 (2002).

Acknowledgements

My work is funded by the Max Planck Society, the Bundesministerium für Bildung und Forschung and the Deutsche Forschungsgemeinschaft. I thank B. Cohen, H. Kaessmann, D. Serre, M. Stoneking, C. Stringer, L. Vigilant and especially D. Altshuler for helpful comments on the manuscript.

Nature, nurture and human disease

Aravinda Chakravarti* & Peter Little†

*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)

†School of Biotechnology and Biomolecular Sciences, University of New South Wales, Sydney, New South Wales 2052, Australia (e-mail: p.little@unsw.edu.au)

What has been learnt about individual human biology and common diseases 50 years on from the discovery of the structure of DNA? Unfortunately the double helix has not, so far, revealed as much as one would have hoped. The primary reason is an inability to determine how nurture fits into the DNA paradigm. We argue here that the environment exerts its influence at the DNA level and so will need to be understood before the underlying causal factors of common human diseases can be fully recognized.

“We used to think our fate was in our stars. Now we know, in large measure, our fate is in our genes.” J. D. Watson, quoted in *Time* magazine, 20 March 1989 (ref. 1).

The double helix, in its simplicity and beauty, is the ultimate modern icon of contemporary biology and society. Its discovery provided the bridge between the classical breeding definition and the modern functional definition of genetics, and permanently united genetics with biochemistry, cell biology and physiology. The DNA structure provided an immediate explanation for mutation and variation, change, species diversity, evolution and inheritance. It did not, however, automatically provide a mechanism for understanding how the environment interacts at the genetic level.

One gene, one disease

Recognition that genes have a role in human disease dates back to the rediscovery of the rules that govern the inheritance of genes by Gregor Mendel — the so-called Mendelian

laws of inheritance. So far, human geneticists have been most successful at understanding single-gene disorders, as their biological basis, and thus presumed action, could be predicted from inheritance patterns. Mendelian diseases are typically caused by mutation of a single gene that results in an identifiable disease state, the inheritance of which can readily be traced through generations.

The landmark sequencing of the human genome provided some important lessons about the role of genes in human disease. Notably, mutations in specific genes lead to specific biological changes, and rarely do mutations in multiple genes lead to an identical set of characteristics that obey ‘Mendelian inheritance’. Additionally, sequence diversity of mutations is large and, consequently, individual mutations are almost always rare, showing relatively uniform global distributions.

But a few exceptions do exist. Some recessive mutations (mutations that influence a person only if both copies of the gene are altered) are surprisingly common in specific populations. This defiance of general mutation patterns arises either from chance increases in frequency in isolated populations, such as in the Old Order Amish², or from the protective effect of a deleterious mutation in a single copy, such as the genetic mutation that on the one hand causes sickle-cell anaemia, but on the other hand offers protection against malaria³. These examples show that human history, geography and ecology of a particular people are relevant to understanding their present-day molecular disease burden⁴.

For over 90 years, the association between DNA mutations and a vast variety of single-gene disorders has repeatedly emphasized the notion that human disease results from faults in the DNA double helix (see, for example, the Online Mendelian Inheritance in Man database at www.ncbi.nlm.nih.gov/omim/, which provides a catalogue of human genes and genetic disorders). Is it then too extrapolative to suggest that all diseases and traits, each of which has some familial and imputed inherited component, will be caused by a corrupted piece of double helix?

Is our fate encoded in our DNA?

Is Watson’s genetic aphorism of human disease really true? The excitement of genetics, and the perceived medical importance of the human genome sequence, is pegged to the promise of an understanding of common chronic disease and not rare Mendelian diseases. In theory, one might hope that approaches used successfully to identify single-gene diseases could simply be applied to the common causes of world-wide morbidity and mortality, such as cancer, heart disease, psychiatric illness and the like. This would enable a boon for diagnosis, understanding and the eventual treatment of these common maladies⁵.

The reality is that progress towards identifying common disease mutations has been slow, and only recently have there been some successes⁶. It is now appreciated that although genes are one contributor to the origin of common diseases, the mutations they contain must have properties that are different from the more familiar, deterministic features of single-gene mutations. Indeed, the underlying genes are likely to be numerous, with no single gene having a major role, and mutations within these genes being common and imparting small genetic effects (none of which are either necessary or sufficient⁷).

Moreover, there is a suspicion that these mutations both interact with one another and with the environment and lifestyle, although the molecular specificity of inter-