

The book of genes

Peter Little

The sequence of human chromosome 22, now published, is the first phase in a biological revolution. Not least, the result will be a transformed appreciation of human individuality.

As 1999 draws to a close and we approach the third millennium, a new book is being written. It will change the way we see ourselves as profoundly as did the momentous books of the first two millennia — the great books of religion and the *Origin of Species*. The first chapter of this book, the book of genes, is summarized on page 489 of this issue¹. It consists of the DNA sequence of most of human chromosome 22 and contains the inaugural results of the Human Genome Project. The whole of the book will probably never be printed in its entirety — it would take about half-a-million pages of this journal to do so — but it will nevertheless inexorably alter our perceptions of human health, attitudes to each other and understanding of our uniqueness. It is fitting that in this age our most telling monument, a culmination of some of the great biological discoveries of our time (Fig. 1), may be an electronic database.

The DNA in all normal human cells is in 23 pairs of pieces, neatly packaged into 46 chromosomes, and the paper summarizes the understanding we have of the 33.4 million or so base pairs that make up the DNA sequence of just one of these pairs (number 22, chosen because it is one of the smallest human chromosomes; only chromosome 21 is smaller). The summary is brief. There are certainly 545 genes encoded in the DNA and there may be as many as 1,000 (see refs 2,3). The sequence is not quite complete because for technical reasons there are 11 gaps; none is larger than 150,000 base pairs, however, and some are just a few thousand. The approach taken, and an alternative one, are described in Box 1 (overleaf).

The sequence tells us about biology in several ways. We do not know whose DNA it is, so we do not learn about a specific person. Instead we draw lessons that apply to any human. All of the constituent parts of the sequence have been generally available for many months^{2,3} — indeed, as fast as they came off the analysers, they were put into free databases — and so the information has already been used by at least seven groups to study newly discovered genes.

There are thought to be at least 27 human disorders associated with changes to genes on chromosome 22, and the causative genes in eight of them have still to be discovered. The conditions range from cancers to disorders of fetal development and of the nervous

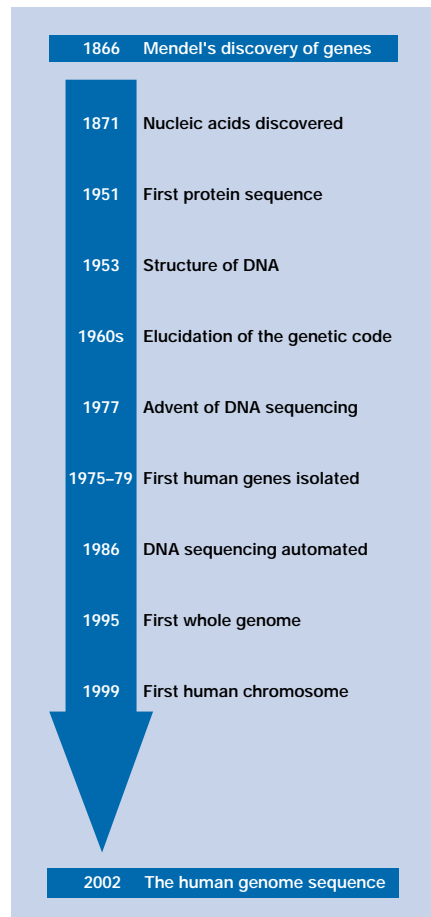


Figure 1 **Sequence of events — from Mendel's discovery of the existence of genes, to 2002 and the projected date for completion of sequencing the human genome. The first protein to be sequenced was insulin; the first genome, that of the bacterium *Haemophilus influenzae*.**

system. The biological jewel in this particular crown is probably the nature of a gene involved in schizophrenia, thought to be located within the sequence but, as yet, unidentified. The sequence also provides tantalizing clues as to how chromosome 22 must have evolved because at least eight regions of the DNA are present in duplicate — a characteristic of evolution of the genetic apparatus. Frustratingly, however, we cannot read its history without comparing its organization with that of the DNA of close animal relatives of humans, and we do not yet have this information.

So what is so revolutionary about all of this? In itself, perhaps not much. But what

book can be appreciated by reading only one chapter? Rather, it is as a foretaste of what is to come that the paper¹ is a milestone in our understanding of our own biology. Within three years, the full sequence of human DNA will be finished and for the first time we will be able to identify the proteins, some 200,000 to 300,000 of them, that are used to make a human being.

The idea that a profound biological problem can have bounds is not part of our usual thinking; all such problems are so complex that research normally touches upon only part of them at any one time. Soon, however, we will have the bounds created by the complete list of proteins that might underpin the process we study. Of course, being bounded is not necessarily the end of the story — the idea that the Atlantic is bounded is no real comfort to anyone who might want to swim across it — and the difficulties of using this protein list effectively are very great indeed. Until now biologists have mainly relied upon intuitive models of how biological systems develop and respond. The new protein world will move the models into areas of analytical complexity where few researchers are trained to go.

A good example of such complexity is in the systems that signal from the outside of a cell to the inside⁴, the end result often being gene activation or deactivation. Individually, one can think of these systems as light circuits with three components — a switch, a wire and a light. The three components have just two interactions, switch to wire and wire to light. If two such signalling circuits can interact (switch with switch, wire with wire, and light with light) there will be 11 interactions. For three circuits, there will be 27 interactions, and so on. The increase in interactions is dramatically nonlinear, and in the face of that intuition is valueless. How will we meet the challenge of the potential interplay of complete sets of proteins drawn from a list that has perhaps a thousand entries?

In more practical terms, what is going to happen next? Where will the DNA sequence take us? First, it will take us to the list of proteins, but not quickly. DNA encodes proteins via RNA, and we have known this code for 30 years. But this does not mean a gene is easy to identify in DNA. Of the 1,000 or so possible genes on chromosome 22, only 545 are easy to spot because they encode proteins that are

similar to ones that have been studied previously in organisms as disparate as humans and bacteria. The remaining, putative genes are predicted by complex computer modelling that is only partly accurate. Showing that these are, or are not, genes will be a continuing challenge.

Second, will we know the functions of these proteins just because we know their sequence? In most cases we will not, but biologists will have huge numbers of proteins and families of proteins to think about and experiment on. At least by a process of exclusion, we will know what we do not know because the list will, in the end, be complete.

This property opens up the third path that the DNA sequence will take us down — the prospect of global analysis of gene activity — that is, the amounts of RNA each gene is, or is not, producing. This is possible using 'arrays' and 'chips' for comprehensive assay of the variation in RNA levels, a wholly novel approach to biology⁵. We do not know what it will tell us about human beings. But the results⁶ of studying just 15,000 genes are already striking, and provide a completely new viewpoint of the life of a cell.

Finally, the DNA sequence will become the framework upon which we will be able to place the changes to the sequence which are present in every one of us. Any two unrelated human beings differ by one base pair in every 1,000 or so. A base pair is chemically defined by the nucleotides it contains and so these differences (polymorphisms) are called single-nucleotide polymorphisms or SNPs (often called 'snips'). About 85% of all DNA differences are SNPs and the population biology of the human species has meant that they are very common — hence the one-in-1,000 difference⁷.

There is a general consensus that SNPs are probably the cause of most common genetic disorders. We all carry many SNPs but if we are unlucky enough to carry the 'wrong' set of changes, we are predisposed to one or other of the common disorders with a genetic component such as diabetes, heart diseases, asthma or cancers⁷. The 'SNP consortium', a group funded by The Wellcome Trust and key pharmaceutical companies, is setting out to find 300,000 of the commonest SNPs in human populations. Once this has been done, we can confidently face the task of identifying which ones contribute to the common disorders — or indeed to any common human trait, be it good or bad.

The reason that each of us is an individual is because we are the products of genes, gene variance and life experiences ('environment'), a combination that defines our unique personalities, abilities and disabilities. If knowledge of gene differences can be combined with an understanding of the richness of environmental influences, we will have the key to unlocking the cause of most of the common disorders that kill or

Box 1 : Two routes to the genome

There is controversy over the best way to sequence the human genome. The public sequencing project, of which the consortium that sequenced chromosome 22 is a member, is jointly supported by the US National Institutes of Health and Department of Energy, and the UK Wellcome Trust. It uses a 'clone-by-clone' approach. In contrast, Celera Genomics, a private company in the United States, has set out to sequence the whole human genome directly by the 'random-shotgun' method.

Clone-by-clone sequencing is sequencing applied to 150,000-base-pair fragments of human DNA (the clones), one at a time. The entire sequence of a chromosome is then reassembled. The random-shotgun approach simply goes directly to the totality of our DNA by sequencing at random. Why the two strategies? The human genome is full of duplications of DNA sequences and contains about a million 'interspersed repeats'. These are short, near-identical copies of the same sequence, and they make assembly of DNA sequences difficult because the wrong repeats can get matched together when sequences are being assembled by computer. Advocates of the clone-by-clone technique argue that

working on one DNA clone at a time simplifies the repeats problem by concentrating on just those that are found in a single clone. Backers of the random-shotgun approach believe that the problem can be overcome with massive computing power.

At the time of writing, the fruitfly genome — about one-tenth the size of the human — has nearly been completed by Celera using random sequencing. This is sometimes held to be a model for the human project, but it is not: not only is the fruitfly genome smaller, it is much less rich in interspersed repeats and duplications. There will probably be very many gaps in the final human sequence assembled by the random approach.

The arguments for and against^{8,9} the two techniques are not easy to reconcile. One advantage of the clone-by-clone approach is that useful information is generated well before completion of the entire sequence, as has been shown with chromosome 22 (ref. 1). But the work that goes into the clone maps is considerable and taxing (and there are still ten small gaps in the chromosome 22 map). In March of this year the public consortium announced that it would

sequence 90% of human DNA by spring 2000. The goal will be achieved by continuing a clone-by-clone approach, as before, but not completing each 150,000-base-pair fragment. The result will be a 'draft' sequence that will not be continuous, in that each 150,000 base pairs will be contained in 10,000–15,000 base-pair pieces. This is less useful than the fully 'finished' sequence achieved for chromosome 22 but will still be enormously powerful for writing the book of genes. The project will then move into a finishing phase, and has a projected completion date of 2002.

Who is right? We do not know but can hope that both sets of data can be combined to give us the very clearest picture of the genome. One crucial difference remains: only the public consortium is making the DNA sequence freely available, as it comes off the machines, to anyone who wants it. The consortium's 'Bermuda agreement' is a binding set of rules for immediate and free release of sequence, and is to be applauded — it would be a terrible blow for science and humanity if the human genome became a commercial property.

P. L.

otherwise cause suffering.

This is the field of genetic epidemiology, the analysis in large populations of the interactions of gene, genetic variation and environment. Its study will require new techniques for DNA testing, new statistical tools for analysing hundreds of thousands of people and the utmost care to protect individual privacy. The goal is worth attaining because genetic epidemiology will contribute to discovering new drug treatments, will enable those treatments to be tailored to individual genetic and life profiles, and will predict some of our futures and contribute to the advice needed to avoid the unpleasant ones. It will change the emphasis of health care from treatment to prevention. But ethical dilemmas will also arise as we discover some of the sources of the differences in individual health, behaviour and personality.

Biology is entering a new world; not only

do we face a revolutionary leap in what we know, we also face radical changes in the tools we must use to understand that information. I am not sure that we are prepared for the full impact of either but we have already made our first tentative steps into the new world of the genome. The challenge is now to translate the new biology into tangible benefits for humanity. ■

Peter Little is in the Department of Biochemistry, Imperial College, Prince Consort Road, London SW7 2AZ, UK.
e-mail: p.little@ic.ac.uk

1. Dunham, I. *et al.* *Nature* **402**, 489–495 (1999).
2. <http://www.sanger.ac.uk/HGP/Chr22>
3. <http://www.genome.ou.edu/human.html>
4. Weng, G. *et al.* *Science* **284**, 92–96 (1999).
5. "The Chipping Forecast" *Nature Genet.* **21** (Suppl.) (1999).
6. Iyer, V. R. *et al.* *Science* **283**, 83–87 (1999).
7. Chakravarti, A. *Nature Genet.* **21** (Suppl.) 56–60 (1999).
8. Green, P. *Genome Res.* **7**, 410–417 (1997).
9. Weber, J. L. & Myers, E. W. *Genome Res.* **7**, 401–409 (1997).