accurate because of the limitations of gene-finding programs. But unless the human genome contains a lot of genes that are opaque to our computers, it is clear that we do not gain our undoubted complexity over worms and plants by using many more genes. Understanding what does give us our complexity — our enormous behavioural repertoire, ability to produce conscious action, remarkable physical coordination (shared with other vertebrates), precisely tuned alterations in response to external variations of the environment, learning, memory… need I go on? — remains a challenge for the future.

Complexity

Where do our genes come from? Mostly from the distant evolutionary past. In fact, only 94 of 1,278 protein families in our genome appear to be specific to vertebrates. The most elementary of cellular functions — basic metabolism, transcription of DNA into RNA, translation of RNA into protein, DNA replication and the like — evolved just once and have stayed pretty well fixed since the evolution of single-celled yeast and bacteria. The biggest difference between humans and worms or flies is the complexity of our proteins: more domains (modules) per protein and novel combinations of domains. The history is one of new architectures being built from old pieces. A few of our genes seem to have come directly from bacteria, rather than by evolution from bacteria — apparently bacterial genomes can be direct donors of genes to vertebrates. So DNA chimaeras consisting of the genes from several organisms can arise naturally as well as artificially (opponents of 'genetically modified foods' take note).

The most exciting new vista to come from the human genome is not tackling the question "What makes us human?", but addressing a different one: "What differentiates one organism from another?". The first question, imprecise as it is, cannot be answered by staring at a genome. The second, however, can be answered this way because our differences from plants, worms and flies are mainly a consequence of our genetic endowments. The Celera team[4] presents the more detailed analysis of the numbers of different protein motifs and protein types, in extensive tables. From them, it is easy to see what types of proteins and motifs have been amplified for specific types of organisms. In vertebrates, not surprisingly, we see elaboration and the *de novo* appearance of two types of genes: those for specific vertebrate abilities (such as neuronal complexity, blood-clotting and the acquired immune response), and those that provide increased general capabilities (such as genes for intra- and intercellular signalling, development, programmed cell death, and control of gene transcription). Someday soon we will have the mouse genome, and then those of fish and dogs, and probably the kangaroo genome from the

Australians. Each of these will fill in a piece of the evolutionary puzzle and will provide exciting comparisons.

We wait with bated breath to see the chimpanzee genome. But knowing now how few genes humans have, I wonder if we will learn much about the origins of speech, the elaboration of the frontal lobes and the opposable thumb, the advent of upright posture, or the sources of abstract reasoning ability, from a simple genomic comparison of human and chimp. It seems likely that these features and abilities have mainly come from subtle changes — for example, in gene regulation, in the efficiency with which introns are spliced out of RNA, and in protein–protein interactions — that are not now easily visible to our computers and will require much more experimental study to tease out. Another half-century of work by armies of biologists may be needed before this key step of evolution is fully elucidated.

What is next? Lots of hard work, but with new tools and new aims. First, we have to stay the course and get the most precise representation of the genome that we can: this is a matter of filling the cracks, cleaning up the errors, and getting rid of the uncertainties that plague each of the analytical methods. Second, we need to see more genomes, with each one giving us a deeper insight into our own. Third, we need to learn how to take advantage of this book of life. Tools for scanning the activity levels of genes in different cells, tissues and settings are becoming available and are already revolutionizing how we do biological investigation. But we will have to move back from the general to the particular, because each gene is a story in itself and its full significance can be learned only from concentrating on its particular properties.

Fourth, we need to turn our new genomic information into an engine of pharmaceutical discovery. Individual humans differ from one another by about one base pair per thousand. These 'single nucleotide polymorphisms' (SNPs) are markers that can allow epidemiologists to uncover the genetic basis of many diseases. They can also provide information about our personal responses to medicines — in this way, the pharmaceutical industry will get new targets and new tools to sharpen drug specificity. Moreover, the analysis of SNPs will provide us with the power to uncover the genetic basis of our individual capabilities such as mathematical ability, memory, physical coordination, and even, perhaps, creativity.

Biology today enters a new era, mainly with a new methodology for answering old questions. Those questions are some of the deepest and simplest: "Daddy, where did I come from?"; "Mommy, why am I different from Sally?". As these and other questions get robust answers, biology will become an engine of transformation of our society. Instead of guessing about how we differ one from another, we will understand and be able to tailor our life experiences to our inheritance. We will also be able, to some extent, to control that inheritance. We are creating a world in which it will be imperative for each individual person to have sufficient scientific literacy to understand the new riches of knowledge, so that we can apply them wisely. ■

David Baltimore is at the California Institute of Technology, 1200 East California Boulevard, Mail Code 204-31, Pasadena, California 91125, USA.
e-mail: baltimo@caltech.edu

1. International Human Genome Sequencing Consortium *Nature* **409,** 860–921 (2001).
2. Watson, J. D. & Crick, F. H. C. *Nature* **171,** 737–738 (1953).
3. http://genome.cse.ucsc.edu/
4. Venter, J. C. *et al. Science* **291,** 1304–1351 (2001).
5. Aach, J. *et al. Nature* **409,** 856–859 (2001).
6. Birney, E., Bateman, A., Clamp, M. E. & Hubbard, T. J. *Nature* **409,** 827–828 (2001).

The maps

Clone by clone by clone

Maynard V. Olson

The public project's sequencing strategy involved producing a map of the human genome, and then pinning sequence to it. This helps to avoid errors in the sequence, especially in repetitive regions.

This issue of *Nature* celebrates a halfway point in the implementation of the 'map first, sequence later' strategy adopted by the Human Genome Project in the mid-1980s[1]. The results suggest that the strategy was basically sound. It led, as hoped, to a project that could be distributed internationally across many genome-sequencing centres, and that would allow sequenced fragments of the human genome to be anchored to mapped genomic landmarks long before the complete sequence coalesced

into one long string of Gs, As, Ts and Cs.

The centrepiece of the suite of mapping papers in this issue is on page 934, where the International Human Genome Mapping Consortium describes a 'clone-based' physical map of the human genome[2]. A map like this not only charts the genome, giving a structure on which to hang sequence data, but also provides a starting point for sequencing. Figure 1 shows the basics of the approach. I drew this figure in 1981, using India ink and a Leroy lettering set. Both
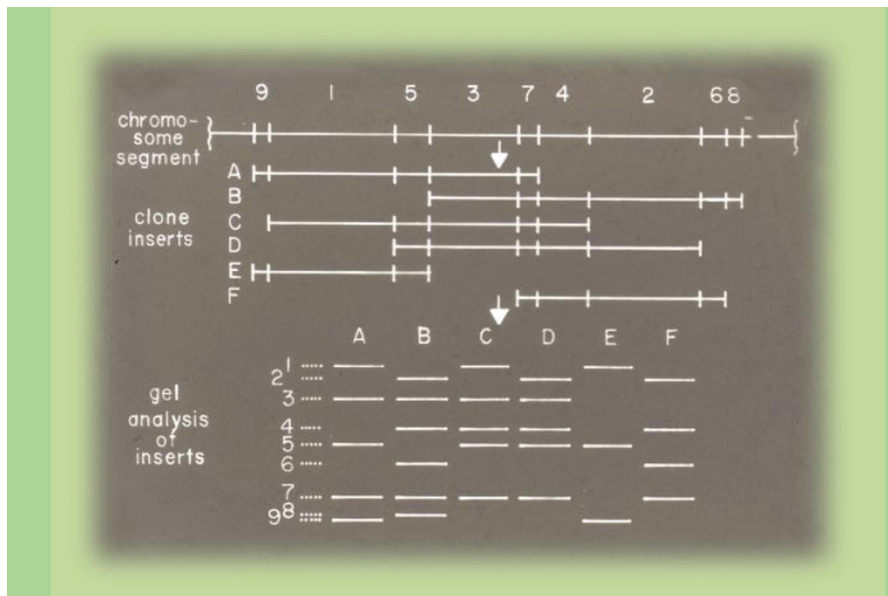
Figure 1 **Clone-based physical mapping. The top line shows the location of 'restriction' sites (vertical bars) in a particular region of the genome. Restriction sites are places at which a site-specific restriction endonuclease cleaves DNA. The fragments produced by cleavage at every possible point in this region are numbered 1 to 9. Below the line are several clones with random end points, labelled A to F. Clones are produced by first partially digesting many copies of the genome with different restriction endonucleases; the resulting large segments are then inserted into bacteria and replicated (cloned). Each clone is digested with a restriction endonuclease, and the resulting fragments are separated, by size, on an electrophoretic gel ('gel analysis of inserts'). This process yields a distinctive pattern ('fingerprint') for each clone. The map-assembly problem requires working backwards (upwards in this figure) from the fingerprints to a clone-overlap map and restriction-site map of the chromosome segment. To finish the analysis of this region of the genome, the natural choice of clones to sequence would be A and B.**

graphical and mapping technologies have come a long way since then, but the principles behind clone-based physical mapping have not changed.

The clone-based approach works as follows. Many copies of the genome are cut up into segments of about 150,000 base pairs by partial digestion with site-specific restriction endonucleases — enzymes that cleave DNA in specific places. ('Partial' digestion means that the reaction is not carried out for long enough to allow every possible cleavage to be made.) The large DNA segments are plugged into bacterial 'artificial chromosomes' (BACs) and inserted into bacteria, where they are copied exactly each time the bacteria divide. The process produces 'clones' of identical DNA molecules that can be purified for further analysis. Next, each clone is completely digested with a restriction endonuclease, chosen to produce a characteristic pattern of small fragments, or a 'fingerprint', for each clone. Comparison of the patterns reveals overlap between the clones, allowing them to be lined up in order, while the sites in the genome at which the restriction endonuclease cleaves are charted. The result is a physical map.

Individual BAC clones are then sheared into smaller fragments and cloned; the resulting 'small-insert' subclones are sequenced. The sequence of an individual BAC clone is assembled from the sequences of an 'oversampled' set of subclones (in other words, enough subclones are sequenced to ensure that each part of the original clone is analysed several times). Finally, the whole genome sequence is assembled by melding together the sequences of a set of BACs that spans the genome.

This approach is similar to that used in the 1980s and early 1990s to map and sequence the genomes of the nematode *Caenorhabditis elegans* and the yeast *Saccharomyces cerevisiae*[3,4]. What is new in the human project is its staggering scale, and the speed with which it has been completed. By way of comparison, although the nematode and yeast genomes are, respectively, only 3% and 0.5% the size of the human genome, these early mapping projects spanned the better part of a decade, as opposed to two years for the much larger human project.

One weakness of clone-based physical mapping is that the maps often have poor continuity. For example, there is not always a BAC clone to cover every part of the genome; and overlaps between clones can be obscured by data errors or the presence of large-scale repeats in the genome. The current map[2] has more than 1,000 discontinuities. These will cause some difficulties as the Human Genome Project moves to its next phase, which will involve ensuring accuracy and

filling in any gaps in the sequence. Nonetheless, the current map typically maintains continuity for several million base pairs at a stretch. These continuous segments are big enough to allow the clone-based map to be overlaid on various lower-resolution maps. In this way, the mapped segments can be ordered and orientated, much as a discontinuous patchwork of high-resolution maps of the Earth's surface can be orientated by overlaying them on a satellite photograph of the whole Earth.

Two particularly interesting low-resolution maps are the genetic and cytogenetic maps, on pages 951 and 953 of this issue[5,6]. The genetic map[5] is based on the probability of the occurrence of recombination — the swapping of corresponding, nearly identical segments of DNA between maternally and paternally derived chromosomes as the genome is passed from one generation to the next. The cytogenetic map[6] is based on subtle variations in the staining properties of different regions of the genome, as viewed by light microscopy. Yet more papers describe different approaches to clone-based mapping[7–9]. These methods were applied to particular chromosomes simply because different sequencing centres chose to rely on the whole-genome map to different degrees.

But was all this cartography even necessary? Another draft of the human genome sequence is described in this week's *Science* by Celera Genomics[10]. This group adopted a different approach, which involved preparing small-insert clones directly from genomic DNA rather than from mapped BACs. The major rationale for the BAC-by-BAC approach[2] was to make easier the finishing phase of the Human Genome Project, which lies ahead. The consortium now plans to upgrade the 30,000 BAC sequences by sequencing more subclones from each BAC (the 'topping-up' phase) and then resolving internal gaps and discrepancies (the 'finishing' phase). Segmenting the finishing phase into BAC-sized portions provides an enormous advantage in dealing with blocks of sequence that are repeated at many different places within the genome. The power of this strategy is nicely illustrated by the mapping of the Y chromosome, whose repetitive structure is unusually complex (page 943 of this issue[11]).

Nature readers should not expect any real answer to the question of which of these two approaches is the better one. But it is likely that the only players still on the field when the toughest finishing issues are confronted will be the public consortium's BAC brigade. In the future, as genome sequencing moves on to other mammals, the context will have changed; the human sequence will provide an invaluable guide to assembling long stretches of sequence that are shared among all mammalian genomes. So the sequencing of the human genome is likely to be the only large

sequencing project carried to completion by the methods described in this issue. Genome sequencing will get easier from here.

Looking ahead, there are two threats to producing a quality finished product. One is simple exhaustion on the part of the consortium's members: each new round of press conferences announcing that the human genome has been sequenced saps the morale of those who must come to work each day actually to do what they read in the newspapers has already been done.

We may also expect to hear the argument that the current sequence is good enough for most purposes, and that remaining problems should be resolved by users as the need for accurate sequence in specific regions arises. What we have now is certainly a lot better than what we had yesterday. But biologists in the future will be comparing vast data sets to the reference sequence of the human genome. They must be able to do so with confidence that the discrepancies they encounter are due to the limitations of their own data or, more interestingly, to biology. They should not need to expend time, energy and imagination compensating for a failure now to pursue the Human Genome Project to a grand conclusion. We must move on and finish the job, even as the bright lights of media attention shift elsewhere. ∎

Maynard V. Olson is in the Departments of Medicine and Genetics, Fluke Hall, Mason Road, University of Washington, Seattle, Washington 98195-2145, USA.
e-mail: mvo@u.washington.edu

1. National Research Council *Mapping and Sequencing the Human Genome* (National Academy Press, Washington DC, 1988).
2. The International Human Genome Mapping Consortium *Nature* **409**, 934–941 (2001).
3. Coulson, A., Sulston, J., Brenner, S. & Karn, J. *Proc. Natl Acad. Sci. USA* **83**, 7821–7825 (1986).
4. Olson, M. V. *et al. Proc. Natl Acad. Sci. USA* **83**, 7826–7830 (1986).
5. Yu, A. *et al. Nature* **409**, 951–953 (2001).
6. The BAC Resource Consortium *Nature* **409**, 953–958 (2001).
7. Montgomery, K. T. *et al. Nature* **409**, 945–946 (2001).
8. Brüls, T. *et al. Nature* **409**, 947–948 (2001).
9. Bentley, D. R. *et al. Nature* **409**, 942–943 (2001).
10. Venter, J. C. *et al. Science* **291**, 1304–1351 (2001).
11. Tilford, C. A. *et al. Nature* **409**, 943–945 (2001).

The draft sequences

Filling in the gaps

Peer Bork and Richard Copley

Two rough drafts of the human genome sequence are now published. Completion of the sequences lies ahead, but the implications for studying human diseases and for biotechnology are already profound.

With the publication of the human genome sequence — described and analysed on page 860 of this issue[1] and in this week's *Science*[2] — we cross a border on the route to a better understanding of our biological selves. But unlike the previously published sequences of human chromosomes 21 and 22 (refs 3,4), the present sequences of the whole human genome are not considered complete. The bulk of the data make up what is called a 'rough draft'. So what is all the fuss about? What exactly does 'rough draft' mean, and what can we learn from sequences such as this?

In the draft from the publicly funded International Human Genome Sequencing Consortium[1], around 90% of the gene-rich — euchromatic — portion of the genome has been sequenced and 'assembled', the term used to describe the process of using a computer to join up bits of sequence into a larger whole. Each base pair of this 90% was sequenced four times on average, ensuring reasonable precision. Only about a quarter of the whole genome is considered 'finished' — another bit of genomics jargon, which basically means that each base pair has been sequenced eight to ten times on average, with gaps in the sequence existing only because of the limitations of present technology. Nonetheless, the sequence of base pairs in the draft is very accurate, and is unlikely to change much; 91% of the euchromatin sequenced has an error rate of less than one base in 10,000 (ref. 1).

For the other draft, that produced by Celera Genomics[2], a variety of methods suggest that between 88% and 93% of the euchromatin has been sequenced and assembled. But direct comparison of these numbers with the public consortium's draft is almost impossible — different procedures and measures were used to process the data and to estimate accuracy. Both projects also have sequence data that were not used in the assembly process, raising the real level of coverage by a few percentage points.

These numbers might seem rather arbitrary, but even when the first genome of an animal species was published[5], it was clear that simple, practical finish lines do not exist (Box 1, Fig. 1). The present level of coverage of the human genome reflects the point where a shift of focus occurs, from sequencing the genome many times over to producing a high-quality, continuous sequence[6]. There is some way to go yet.

Essentially, 'rough draft' refers to the fact that the sequences are not continuous — there are gaps (Box 1). If there are too many gaps, it can be impossible to order and orientate the many small strings of bases that are the raw products of genome sequencing. This might, for example, hamper projects that seek to identify genes involved in inherited diseases. A first step to finding such genes is to work out which region of which chromosome they are on. The complete genome sequence should be immensely useful for the next step — identifying the relevant gene at that region. But gaps and errors in ordering and placing the strings of sequence will make this difficult.

Another problem of incompleteness is that it is difficult to make definitive

Box 1 What makes a completely sequenced genome?

When is sequencing work on a genome complete? No genome for a eukaryotic organism — roughly, those organisms whose cells contain a nucleus — has been sequenced to 100%. There are regions, often highly repetitive, that are difficult or impossible to clone (one of the initial steps in a sequencing project) or sequence with current technology. Fortunately, such regions are expected to contain relatively few protein-coding genes[4,10].

The extent of these regions varies widely in different species. So, rather than applying a universal gold standard, each sequencing project has made pragmatic decisions as to what constitutes a sufficient level of coverage for a particular genome. For example, as much as one-third of the sequence of the fruitfly *Drosophila melanogaster* was not stable in the cloning systems used, and so was not sequenced. But 97% of the so-called euchromatic portion — where most genes are thought to reside — was sequenced[11] (Fig. 1).

For the human genome, one definition of 'finished' is that fewer than one base in 10,000 is incorrectly assigned[6]; more than 95% of the euchromatic regions are sequenced; and each gap is smaller than 150 kilobases[12]. Such standards represent realistic goals given current technology. By this standard, over a quarter of the public consortium's sequence[1] is considered finished at present, including the previously published long arms of chromosomes 21 and 22 (refs 3,4; Fig. 1). The Celera sequences of chromosomes 21 and 22 are slightly more gappy than those from the public consortium, but the converse seems to be true for the other chromosomes[2]. But again, as different protocols were used, it is not easy to compare the overall status of the two assemblies. In the longer term, as much of the heterochromatin — which is harder to sequence, and contains few genes — as possible must be sequenced, because we might otherwise miss important features.

P.B. & R.C.