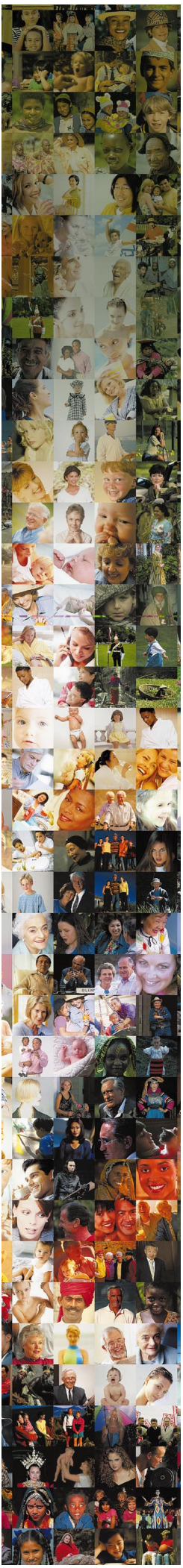

Our genome unveiled

David Baltimore

The draft sequences of the human genome are remarkable achievements. They provide an outline of the information needed to create a human being and show, for the first time, the overall organization of a vertebrate's DNA.

I've seen a lot of exciting biology emerge over the past 40 years. But chills still ran down my spine when I first read the paper that describes the outline of our genome and now appears on page 860 of this issue[1]. Not that many questions are definitively answered — for conceptual impact, it does not hold a candle to Watson and Crick's 1953 paper[2] describing the structure of DNA. Nonetheless, it is a seminal paper, launching the era of post-genomic science.

This milestone of biology's megaproject is the long-promised draft DNA sequence from the International Human Genome Sequencing Consortium (the public project). The sequence itself is available to all those connected to the Internet[3]. In the paper in this issue, we are presented with a description of the strategy used to decipher the structures of the huge DNA molecules that constitute the genome, and with analyses of the content encoded in the genome. It is the achievement of a coordinated effort involving 20 laboratories and hundreds of people around the world. It reflects the scientific community at its best: working collaboratively, pooling its resources and skills, keeping its focus on the goal, and making its results available to all as they were acquired.

Simultaneously, another draft sequence is being published[4]. It is less freely available because it was generated by a company, Celera Genomics, that hopes to sell the information. This week's *Science* contains an account of the history of that project and the analyses of its data, while another of the papers in this issue contains a comparison of the quality of the two sequences[5]. To those who saw this as a competitive sport, the papers make it appear to be roughly a tie. However, it is important to remember that Celera had the advantage of all of the public project's data. Nevertheless, Celera's achievement of producing a draft sequence in only a year of data-gathering is a testament to what can be realized today with the new capillary sequencers, sufficient computing power and the faith of investors.

Answers

What have we learned from all of these AGCTs? The best way to answer the question is to read the analytical sections of the papers. I will only make some general comments. It is important to remember that no statements can be made with high precision because the draft sequences have holes and imperfections, and the tools for analysis remain limited (as described in a further paper[6] in this issue, page 828). However, the answers provided by the draft will be of interest to many investigators, and the value of having the draft published in its imperfect form is unquestionable.

The sequences are about 90% complete for the euchromatic (weakly staining, gene-rich) regions of the human chromosomes. The estimated total size of the genome is 3.2 Gb (that is gigabases, the latest escalation of units needed to contain the fruits of modern technology). Of that, about 2.95 Gb is euchromatic. Only 1.1% to 1.4% is sequence that actually encodes protein; that is just 5% of the 28% of the sequence that is transcribed into RNA. Over half of the DNA consists of repeated sequences of various types: 45% in four classes of parasitic DNA elements, 3% in repeats of just a few bases, and about 5% in recent duplications of large segments of DNA. The amounts in the first and third classes will certainly grow as our ability to characterize them increases in effectiveness and we examine the darkly staining, heterochromatin regions of chromosomes. As the co-discoverer of reverse transcriptase (the enzyme that reverses the common mode of information transfer from DNA to RNA), I find it striking that most of the parasitic DNA came about by reverse transcription from RNA. In places, the genome looks like a sea of reverse-transcribed DNA with a small admixture of genes.

Repeats

By contrast, the puffer fish — another vertebrate — has a genome that contains very few repeats. But it encodes a perfectly functional creature, so it seems likely that most of the repeats are simply parasitic, selfish DNA elements that use the genome as a convenient host. People call this 'junk DNA', but from the DNA's point of view it deserves more respect. In most places in the human genome the selfish elements are tolerated, and in some places —near the ends of chromosomes, for instance, or near the chromosome constrictions called centromeres — it builds up to form huge segments. However, the repeated DNA may have both negative and positive effects. For instance, the paucity of repeats in certain highly regulated regions of the genome suggests that insertions there can disrupt gene regulation and are deleterious. Conversely, the enrichment of the so-called Alu class of repeated sequences in the gene-rich, high-GC regions of the genome implies that they have a positive function. The repeats can also be fodder for evolving new functions and act as loci for gene rearrangements.

In humans, virtually all of the parasitic DNA repeats seem old and enfeebled, with little evidence of continuing reinsertions. However, there has been very little evolutionary scouring of these repeats from the human genome, making it a rich record of evolutionary history. The mouse genome, by contrast, has many actively reinserting parasitic sequences and is scoured more intensely, making it a much younger and more dynamic genome. This difference might reflect the shorter generation time of mice or something about their physiology, but I find it an intriguingly enigmatic observation.

Much of what we learn about the global organization of the genome is an elaboration of previous notions. For instance, we knew that the genome had regions with a relatively high content of GC bases and regions high in AT, but now we have a very complete appreciation of this architecture. What maintains the patchiness of the GC/AT ratio in the genome remains an unanswered question. As was expected,

most genes are located outside the heterochromatic regions; interestingly, however, in regions of the genome rich in GC bases, the gene density is greater and the average intron size is lower. These introns — made up of largely meaningless sequence that breaks up the protein-coding sequences (exons) of genes — are much longer in human DNA than in the genomes previously sequenced. Their dilution of the coding sequence is one element that makes finding genes by computer so difficult in human DNA.

A major interest of the genome sequence to many biologists will be the opportunity it provides to discover new genes in their favourite systems — for instance, cell biologists will search for new genes for signalling proteins, and neurobiologists will look for new ion channels. This data-mining exercise was carried out by various groups which report their initial findings in papers that appear on pages 824–859 of this issue. They found some new and interesting genes, but surprisingly few, and occasionally could not find the full extent of genes that they knew were there. The paucity of discoveries reflects their concentration on systems that were previously heavily studied.

Gene-regulatory sequences are now there for all to see, but initial attempts to find them were also disappointing. This is where the genomic sequences of other species — in which the regulatory sequences, but not the functionally insignificant DNA, are likely to be much the same — will open up a cornucopia. Basically, the human sequence at its present level of analysis allows us to answer many global questions fairly well, but the detailed questions remain open for the future.

What interested me most about the genome? The number of genes is high on the list. The public project estimates that there are 31,000 protein-encoding genes in the human genome, of which they can now provide a list of 22,000. Celera finds about 26,000. There are also about 740 identified genes that make the non-protein-coding RNAs involved in various cell housekeeping duties, with many more to be found. The number of coding genes in the human sequence compares with 6,000 for a yeast cell, 13,000 for a fly, 18,000 for a worm and 26,000 for a plant. None of the numbers for the multicellular organisms is highly

Genome speak

Allele Humans carry two sets of chromosomes, one from each parent. Equivalent genes in the two sets might be different, for example because of *single nucleotide polymorphisms*. An allele is one of the two (or more) forms of a particular gene.

Bacterial artificial chromosome (BAC) A chromosome-like structure, constructed by genetic engineering, that carries genomic DNA to be *cloned*.

Centromere Chromosomes contain a compact region known as a centromere, where sister chromatids (the two exact copies of each chromosome that are formed after replication) are joined.

Cloning The process of generating sufficient copies of a particular piece of DNA to allow it to be sequenced or studied in some other way.

Complementary DNA (cDNA) A DNA sequence made from a *messenger RNA* molecule, using an enzyme called reverse transcriptase. cDNAs can be used experimentally to determine the sequence of messenger RNAs after their introns (non-protein-coding sections) have been *spliced* out.

Conservation Genes that are present in two distinct organisms are said to be conserved. Conservation can be detected by measuring the similarity of the two sequences at the base (RNA or DNA) or amino-acid (protein) level. The more similarities there are, the more highly conserved the two sequences.

Euchromatin The gene-rich regions of a genome (see also *heterochromatin*).

Eukaryote An organism whose cells have a complex internal structure, including a nucleus. Animals, plants and fungi are all eukaryotes.

Expressed sequence tag (EST) A short piece of DNA sequence corresponding to a fragment of a *complementary DNA* (made from a cell's *messenger RNA*). ESTs have been used to hunt for genes, so hundreds of thousands are present in sequence databases.

Genome The complete DNA sequence of an organism.

Genotype The set of genes that an individual carries; usually refers to the particular pair of alleles (alternative forms of a gene) that a person has at a given region of the genome.

Haplotype A particular combination of alleles (alternative forms of genes) or sequence variations that are closely linked — that is, are likely to be inherited together — on the same chromosome.

Heterochromatin Compact, gene-poor regions of a genome, which are enriched in simple sequence repeats. As it can be impossible to *clone*, heterochromatin is often ignored when calculating the percentage of a genome that has been sequenced. Heterochromatin was originally identified as regions of the genome that stained differently to euchromatin (gene-rich regions).

Introns and exons Genes are *transcribed* as continuous sequences, but only some segments of the resulting *messenger RNA* molecules contain information that codes for the gene's protein product. These segments are called exons. The regions between exons are known as introns, and are *spliced* from the RNA before the product is made.

Long and short arms The regions either side of the centromere, a compact part of a chromosome, are known as arms. As the centromere is not in the centre of the chromosome, one arm is longer than the other.

Messenger RNA (mRNA) Proteins are not synthesized directly from genomic DNA. Instead, an RNA template (a precursor mRNA) is constructed from the sequence of the gene. This RNA is then processed in various ways, including *splicing*. Spliced RNAs destined to become templates for protein synthesis are known as mRNAs.

Mutation An alteration in a genome compared to some reference state. Mutations do not always have harmful effects.

Phenotype The observable properties and physical characteristics of an organism.

Polymorphism A region of the genome that varies between individual members of a population. To be called a polymorphism, a variant should be present in a significant number of people in the population.

Prokaryote A single-celled organism with a simple internal structure and no nucleus. Bacteria and archaebacteria are prokaryotes.

Proteome The complete set of proteins encoded by the *genome*.

Pseudogene A region of DNA that shows extensive similarity to a known gene, but which cannot itself function, either because it has lost the signal required for *transcription* (the promoter sequence) or because it carries mutations that prevent it from being *translated* into protein.

Recombination The process by which DNA is exchanged between pairs of equivalent chromosomes during egg and sperm formation. Recombination has the effect of making the chromosomes of the offspring distinct from those of the parents.

Restriction endonuclease An enzyme that cleaves DNA at every location at which a particular short sequence occurs. Different types of restriction endonuclease cleave at different target sequences.

Single nucleotide polymorphism (SNP) A *polymorphism* caused by the change of a single nucleotide. Most genetic variation between individual humans is believed to be due to SNPs.

Splicing The process that removes introns (non-protein-coding portions) from *transcribed* RNAs. Exons (protein-coding portions) can also be removed. Depending on which exons are removed, different proteins can be made from the same initial RNA or gene. Different proteins created in this way are 'splice variants' or 'alternatively spliced'.

Transcription The process of copying a gene into RNA. This is the first step in turning a gene into a protein, although not all transcripts lead to proteins.

Transcriptome The complete set of RNAs *transcribed* from a genome.

Translation The process of using a *messenger RNA* sequence to build a protein. The messenger RNA serves as a template on which transfer RNA molecules, carrying amino acids, are lined up. The amino acids are then linked together to form a protein chain.

Peer Bork and Richard Copley

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