POLICY: GENOMICS

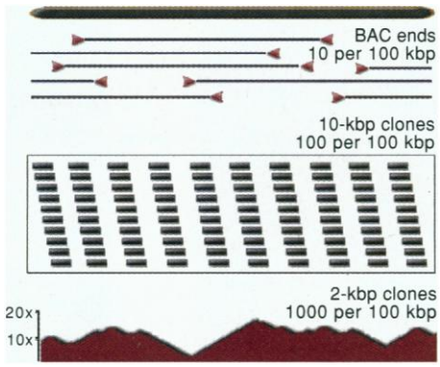
# Shotgun Sequencing of the Human Genome

J. Craig Venter, Mark D. Adams, Granger G. Sutton, Anthony R. Kerlavage, Hamilton O. Smith, Michael Hunkapiller

The Human Genome Project (HGP) was officially launched in the United States on 1 October 1990 as a 15-year program to map and sequence the complete set of human chromosomes and those of several model organisms. The HGP is laying the groundwork for a revolution in medicine and biology. Its importance is underscored by the level of funding from the National Institutes of Health, the Department of Energy (DOE), the Wellcome Trust, and other governments and foundations around the world.

From the inception of the HGP, major technical innovations that would affect its timetable and cost were considered essential to success. The development of bacterial artificial chromosomes (BACs) (*1*) provided a key advance. BACs are propagated in *Escherichia coli* and carry large [~150–kilobase pairs (kbp)] inserts stably. In contrast, ordered cosmid clones that served as the basis of yeast (*2*) and *Caenorhabditis elegans* (*3*) genome sequencing projects are less stable and much shorter (~35 kbp). Fluorescent labeling of DNA fragments generated by the Sanger dideoxy chain termination method has been the mainstay of almost all large-scale sequencing projects since the introduction of the first semi-automated sequencer by Applied Biosystems in 1987 and the development of Taq cycle sequencing in 1990. New models of the sequencer that can process more samples, Taq polymerase engineered especially for sequencing, and higher sensitivity dyes have improved throughput, accuracy, and operating costs. Publication of the first genome from a self-replicating organism, *Haemophilus influenzae*, was based on a whole-genome shotgun (random sequencing) method (*4*). A set of algorithms called the TIGR Assembler (*5*) together with scaffolding sequences from both ends of 18-kbp inserts in bacteriophage lambda clones were critical for determination of correct order and assembly. Eight additional genomes have since been completed by these methods (*4*, *6*, *7*), and several others are nearing completion, including genomes with high GC (~65%) and high AT (~82%) composition, which present special problems for sequencing and assembly.

Current approaches to human genomic sequencing rely on building sequence-ready maps over regions ranging in size from hundreds of kilobase pairs to whole chromosomes and then sequencing individual BACs spanning these regions through a combination of shotgun and directed approaches. This method can produce highly accurate sequence with few gaps, although most sequencing centers have encountered regions that appear to be unsequenceable by current technology. The up-front steps of building and validating the sequence-ready map and subclone library construction and the downstream steps of directed gap filling are generally considered to be rate limiting. About 120 Mbp of human genomic sequence were completed through 1997, and another 200 Mbp are planned for 1998.




**Covering the genome.** A 100-kbp portion of the genome showing expected clone coverage.

The recent announcement by Perkin-Elmer of a new, fully automated sequencer (ABI PRISM 3700) permits a reevaluation of strategies for completing the human genome sequence. This instrument is a capillary-based sequencer that can process ~1000 samples per day with minimal hands-on operator time (~15 min compared with ~8 hours for the same number of samples on ABI PRISM 377s). This reduction in operating labor, coupled with automation of sample purification and sequencing chemistry enabled by the sequencer's improved detection sensitivity, suggests that the tens of millions of sequencing reactions necessary to complete the human genome can be performed more quickly and at lower cost than previously anticipated. The Institute for Genomic Research (TIGR) and Perkin-Elmer have started a program to complete this task within 3 years using this new technology and a whole-genome shotgun strategy that obviates the need for a sequence-ready map before sequencing. We intend to form a new company to carry out this venture and develop a commercial business based on these efforts. The cost of the project is estimated to be between $200 million and $250 million, including the complete computational and laboratory infrastructure to develop the finished sequence and informatics tools to support access to it.

The whole-genome shotgun strategy involves randomly breaking DNA into segments of various sizes and cloning these fragments into vectors. The presence of repeat elements, regions that are unclonable in a particular vector, and the benefit of having more DNA available in clones than is actually sequenced (see figure and table) require that multiple vector libraries be used. A library of pUC18-based plasmids containing ~2-kbp inserts will provide most of the sequencing templates. These clones will be sequenced from both ends to produce pairs of linked sequences representing ~500 bp at the ends of each insert. End sequences from a library of low–copy number plasmid clones containing ~10-kbp inserts will provide medium-range linking, including spanning the common Line-1 and THE repeat elements. Use of multiple cloning systems should help to reduce the effect of sequences that are unclonable or otherwise not present in one of the libraries. The goal is to generate 70 million high-quality DNA sequences totaling ~35 billion bp (10× coverage) of raw human sequence.

An argument for whole-genome shotgun sequencing of the human genome was made (*8*) and rebutted (*9*) in 1997. A year later, we see developments in technology and a new resource for this project consisting of a large database of end sequences of BAC clones. This will provide a framework for linking contigs over larger regions. Currently, the DOE is funding a program at TIGR and the University of Washington to sequence both ends (~500 bp from each end) of 300,000 human BAC clones. This BAC–end sequencing strategy was originally proposed to accelerate genome sequencing by providing markers every 5 kbp throughout the genome (*10*).

The new human genome sequencing facility will be located on the TIGR campus

J. C. Venter, M. D. Adams, G. G. Sutton, A. R. Kerlavage, and H. O. Smith are at The Institute for Genomic Research (TIGR), Rockville, MD 20850, USA. M. Hunkapiller is at Perkin-Elmer Applied Biosystems, Foster City, CA 94404–1128, USA.

in Rockville, Maryland, and will consist of 230 ABI PRISM 3700 DNA sequencers with a combined daily capacity of ~100 Mbp of raw sequence. The facility will also have the infrastructure to produce ~100,000 template preps and ~200,000 sequencing reactions daily. This includes both custom and off-the-shelf robotic devices for picking colonies, pipetting, and thermal cycling. Quality control and assessment procedures will be implemented at each stage of the process.

Accompanying the challenge of obtaining the primary sequence data in a rapid and cost-effective way is the major challenge of assembling raw data into contiguous blocks (contigs) and assigning those to the correct location in the genome. Complete contiguity of the clone map should theoretically be achieved by about 9× coverage, so the 46× coverage (see table) allows for substantial deviation from the statistical model. The pairs of end sequences from each template are constrained by the assembly algorithms to be directed toward one another in the final assembly and located at a given distance apart depending on the insert size of the originating library. Although the BAC end sequences will be the primary scaffold onto which the end sequences from the smaller clones will be assembled, other available resources will be used to verify the alignments and place contigs on individual chromosomes. The most important of these resources is the large number of sequence tagged site (STS) markers that constitute the physical maps that have been produced by many laboratories during the first phase of the HGP. There currently are about 45,000 STS sequences, including about 30,000 that are well ordered along the chromosomes and provide a defined marker approximately every 100 kbp (*11*). Expressed sequence tags (ESTs) that tag 50 to 80% of human genes (*12*) and full-length cDNA sequences spanning up to 5 Mbp of genomic sequence will be used to verify the final assemblies. There are likely to be contigs that are misassembled or incorrectly linked together because of the presence of long, duplicated segments of the genome. We expect to recognize and correct ambiguous or conflicting assembly structures using a combination of manual inspection and directed experimental effort.

The aim of this project is to produce highly accurate, ordered sequence that spans more than 99.9% of the human genome (*13*). The 10× sequence coverage means that the accuracy of the sequence will be comparable to the standard now prevalent in the genome sequencing community of fewer than one error in 10,000 bp. It is likely that several thousand gaps will remain, although we cannot predict with confidence how many unclonable or unsequenceable regions may be encountered. We look forward to working with other genome centers to ensure that the sequence meets the requirements of the scientific community for accuracy and completeness; this will include making clones and electropherograms available.

An essential feature of the business plan is that it relies on complete public availability of the sequence data. The four primary business areas are high-throughput contract sequencing, gene discovery, database services, and high-throughput polymorphism screening. A major consequence of the analysis of data generated by this project will be the creation of a comprehensive human genomic database. It will contain an extensive set of DNA and protein features derived from the primary sequence. DNA features will include identified genes and their regulators, repeats, links with genetic and physical mapping data, synteny with other species, and polymorphisms. Because of the importance of this information to the entire biomedical research community, key elements of this database, including primary sequence data, will be made available without use restrictions. In this regard, we will work closely with national DNA repositories such as National Center for Biotechnology Information. We plan to release contig data into the public domain at least every 3 months and the complete human genome sequence at the end of the project. We also envision providing at a minimum connect fee online access to these data and many of the informatics tools to interpret them. We will also market the database system to commercial companies engaged in pharmaceutical and biotechnology research.

Because the whole-genome shotgun approach will contain data from multiple individuals (the exact number has not yet been determined), we will generate a large number of precisely located single-nucleotide polymorphic (SNP) sites spanning the genome. Using technology being developed at Perkin-Elmer, we will generate assay systems to validate these markers and select a highly informative set of at least 100,000 SNPs. We plan to work with commercial partners to screen DNA samples associated with diseases or other conditions in an effort to link them with particular genetic loci. The assay systems will also be marketed by Perkin-Elmer to third parties for in-house research. Although we do not plan to seek patent protection for the randomly selected SNPs, we may seek patents on diagnostic tests based on the association of particular SNPs with important phenotypic traits.

We also do not plan to seek patents on primary human genome sequences. However, we expect that we and others will be able to use these primary data as a starting point for additional biological studies that could identify and define new pharmaceutical and diagnostic targets. Once we have fully characterized important structures (including, for example, defining biological function), we expect to seek patent protection as appropriate. Given both the complexity and scope of the information contained in human genome sequence, as well as its public availability, we would expect to focus our own biological research efforts on 100 to 300 novel gene systems from among the thousands of potential targets. If we are successful in these efforts, the patents would be available for licensing to interested parties.

| Vector type | Insert size (kbp) | Number of | | Coverage (×) | |
|---|---|---|---|---|---|
| | | Clones | Sequences | Sequences | Clones |
| High-copy plasmid | 2 | 30,000,000 | 60,000,000 | 8.5 | 17 |
| Low-copy plasmid | 10 | 5,000,000 | 10,000,000 | 1.4 | 14 |
| BAC | 150 | 300,000 | 600,000 | 0.1 | 15 |
| **Total** | | **35,300,000** | **70,600,000** | **10** | **46** |

**Analysis of coverage.** As each clone is not completely sequenced, there is a greater coverage of clones than sequences in the assembly. We assume a 500-bp average read length and 3.5-Gbp genome size.

Although it is clear that shotgun sequencing at this scale has never been attempted, it is our hypothesis that the desired result is achievable. While building the human genome sequencing infrastructure we plan to attempt to demonstrate the effectiveness of the shotgun strategy on a large and complex genome, in collaboration with Gerald Rubin (Howard Hughes Medical Institute/University of California Berkeley) and the Berkeley *Drosophila* Genome Project (BDGP). *Drosophila melanogaster* represents a good system for testing the whole-genome shotgun strategy because of the extensive physical and genetic maps that exist, the presence of about 12% of the genome as high-quality finished sequence with which to compare shotgun assembly results, and its importance as a model organism. We will work fully with the BDGP to facilitate the final closure process (which includes making clones and electropherograms available), with the expected result being a highly accurate and contiguous set of chromosome sequences. The *Drosophila*

genome sequence will be deposited in GenBank both while in progress and at completion. An international workshop is being organized for September 1998 to develop a plan for completing the *Drosophila* genome that encourages participation of all groups currently working on this project.

It is our hope that this program is complementary to the broader scientific efforts to define and understand the information contained in our genome. It owes much to the efforts of the pioneers both in academia and government who conceived and initiated the HGP with the goal of providing this information as rapidly as possible to the international scientific community. The knowledge gained will be key to deciphering the genetic contribution to important human conditions and justifies expanded government investment in further understanding of the genome. We look forward to a mutually rewarding partnership between public and private institutions, which each have an important role in using the marvels of molecular biology for the benefit of all.

### References and Notes


1. H. Shizuya *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **89**, 8794 (1992).
2. A. Goffeau *et al.*, *Nature* **387** (suppl.), 5 (1997).
3. J. Sulston *et al.*, *ibid.* **356**, 37(1992).
4. R. Fleischmann *et al.*, *Science* **269**, 496 (1995).
5. G. G. Sutton, O. White, M. D. Adams, A. R. Kerlavage, *Genome Sci. Technol.* **1**, 9 (1995).
6. C. M. Fraser *et al.*, *Science* **270**, 397 (1995); C. J. Bult *et al.*, *ibid.* **273**, 1058 (1996); J.-F. Tomb *et al.*, *Nature* **388**, 539 (1997); H.-P. Klenk *et al.*, *ibid.* **390**, 364 (1997); C. M. Fraser *et al.*, *ibid.*, p. 580; C. M. Fraser *et al.*, *Science*, in press.
7. D. R. Smith *et al.*, *J. Bacteriol.* **179**, 7135 (1997); G. Deckert *et al.*, *Nature* **392**, 353 (1998).
8. J. Weber and E. W. Myers, *Genome Res.* **7**, 401 (1997).
9. E. Green, *ibid.*, p. 410.
10. J. C. Venter, H. O. Smith, L. Hood, *Nature* **381**, 364 (1996).
11. T. Hudson *et al.*, *Science* **270**, 1945 (1995); C. Dib *et al.*, *Nature* **380**, 152 (1996); G. D. Schuler *et al.*, *Science* **274**, 540 (1996).
12. M. Adams *et al.*, *Science* **252**, 1651 (1991); M. Adams *et al.*, *Nature* **37** (suppl.), 3 (1995); L. Hillier *et al.*, *Genome Res.* **6**, 807 (1996).
13. Considerable sequence will be generated from regions of heterochromatin including centromeres, telomeres, and ribosomal DNA arrays, which are not targeted by HGP sequencing laboratories. We will make unique assemblies where possible in these regions.


**BOOKS: PALEOBIOLOGY**

# Palaeobiography


Paul Copper


**Life.** A Natural History of the First Four Billion Years of Life on Earth. RICHARD FORTEY. Knopf, New York, 1998. xiv, 347 pp., + plates. $30 or C$42. ISBN 0-375-40119-9.

A portentous book title as bold as this—*Life*—is bound to raise a few eyebrows. It is also almost certain to catch the eye of the book browser. In a drama bolder and more sweeping than *Gone with the Wind*, Richard Fortey sketches the full story of life on Earth, the stage and the actors, over more than four billion years. Originally published in Britain as *Life: An Unauthorized Biography* (Harper Collins, 1997), this bright brown volume, plastered with the imprint of *Archaeopteryx* (the oldest known bird), is as encompassing as its title suggests. Fortey, senior palaeontologist at the Natural History Museum, London, takes us on a roller coaster from the spawning of the simplest unicellular organisms during violent infancy of the Earth; through monumental crustal upheavals, voyages of continents, and mass extinctions; to an ending at the dawn of human-recorded history.

The key to this book, a layperson's guide to the secrets of fossils and environments most ancient, is the way the author has magically transposed and integrated his academic biography and intellectual growth into the natural history of life. I know of no other "autobiography"—if the book can be called one—quite like this, where the author's life is stitched into such an immense stretch of time. Neatly and adroitly, Fortey weaves his personal observations, his encounters with scientists (famous and less well known), and his introductions to controversies (century-old and contemporary) into a chronological tapestry of life on Earth.

The text literally begins with *Salterella*, the vessel that in 1967 carried Fortey, then a young Cambridge undergraduate, to his first field season in Spitsbergen. *Salterella* is also one of the oldest shelly fossils, a curious Early Cambrian genus named after the pioneering trilobite specialist John W. Salter. First described in 1861 from the shores of Labrador (where I have collected thousands of the little conical shells around some of the earliest metazoan reefs), its affinities can only be guessed: is it a worm, a coral, a mollusk?

CHIP CLARK / NMNH / SMITHSONIAN INSTITUTION

**Ordovician "sea beetle."** Guaranteed an excellent fossil record by their calcite carapaces, trilobites are the characteristic creatures of the Early Paleozoic. (*Ceraurus pleurexanthemus*, from Ontario.)

Coincidence, circumstance, and chance, and their effects on the global gene pool through time, are pervasive themes articulated throughout the book. At the personal level, Fortey explores how one chooses a career path, who happens to win the prizes and scholarships, and who loses out to disappear from sight. In the fossil record we learn about the luck of the gene draw, evolution through the trials of mass extinctions, the consequences of changing climates, continental drift, and cosmic impacts.

The book has many strengths. Fortey lyrically raises fossils from the dead, re-creating vibrant, vivid organisms that absorb light, breathe, eat, function, and interact with their ecosystems. Read his descriptions of the Middle Cambrian Burgess Shale from Canada ("on the dark shales there was a fishmonger's slabful of arthropods"), a Carboniferous rainforest ("the air is so humid that the moisture congeals upon your shoulders"), and the Eocene Messel Grube from Germany ("imagine a delicate bat, *Palaeochiropteryx*, as fragile as a paper kite, with every bone laid out upon a dark slab, as if it had been waiting its turn as an extra in a Dracula movie"). The author presents bites of life's story sequentially, from oldest to newest, as if to suggest (probably rightly so) that the past is the key to understanding the present and the future. He moves continents about like cardboard cut-outs to explain migration paths of continental tetrapods and plants. He lucidly spells out the "rules of the evolutionary game" (which organisms needed to follow to succeed, compete, and survive over millenia), and how these are displayed in the fossil record. Fortey provides a bird's eye view of the science of paleontology, and an insider's perspective of the "psycho-cultural" shenanigans that often come with the paleopriesthood: the cladist cult, the mass extinction dichotomy of catastrophists and uniformitarians, the taxonomic schism of splitters and lumpers, the heretic leaders, and the hermits who wait in isolation to reach


The author is at the Department of Earth Sciences, Laurentian University, Sudbury, Ontario, Canada P3E 2C6. E-mail: pcopper@nickel.laurentian.ca