A new strategy for genome sequencing

J. Craig Venter, Hamilton O. Smith and Leroy Hood

Existing approaches to sequencing the human genome are based on the assumption that each region to be sequenced must first be mapped. But there is a simpler strategy in which any number of laboratories can cooperate.

ONE of the main goals of the Human Genome Project is to sequence, in an international cooperative venture, the 3 billion or so base pairs of DNA that make up the 24 different human chromosomes. The order of the bases along each chromosome (the sequence maps) will allow the identification of the 80,000 or so human genes and provide a framework for studying how certain DNA variations among humans predispose towards various diseases. The project was initiated in 1990 by the US government (the National Institutes of Health and the Department of Energy) and has been joined by the United Kingdom, France, Germany and Japan. Its estimated cost is $3 billion over 15 years.

The first five years have focused on genetic and physical mapping[1]. The genetic map is a representation of the order of the genes along the chromosomes; the physical map is a collection of identifiable overlapping fragments of DNA, together with a specification of how they are arranged along the chromosomes.

We are now entering the more complex sequencing phase[2,3]. Existing approaches to sequencing the human genome are based on the assumption that each region to be sequenced must first be mapped. Here we argue that this step is dispensable. The alternative approach we propose makes it possible for any number of laboratories to cooperate in the effort, aiding international collaboration between large genome centres and small groups, and should be simpler and cheaper than map-based strategies. What is more, it will help the sequencing of biologically interesting chromosomal regions such as gene families (encoding, for example, neural and olfactory receptors), as well as smaller genomes of simpler organisms.

The most common approach to sequencing the human genome uses a three-stage divide-and-conquer strategy (see the figure). It involves the construction of three different clone libraries from human chromosomal DNA. This is done by randomly cutting the DNA into fragments, separating these into differing size classes and then inserting the fragments into distinct vectors capable of propagating them in appropriate hosts such as bacteria or yeast (see the table). (A clone here comprises a vector with a single inserted fragment of human DNA, whereas a clone library comprises the entire collection of human DNA fragments each inserted into a vector molecule.)

IMAGE UNAVAILABLE FOR COPYRIGHT REASONS

Hank Morgan/SPL

The Human Genome Project is now entering the more complex sequencing stage.

Low-resolution physical maps of each chromosome are first prepared by identifying shared landmarks (such as unique sites that can be amplified by PCR (sequence-tagged sites or STSs) or restriction-enzyme digestion sites) on overlapping yeast artificial chromosome (YAC) clones. High-resolution or sequence-ready maps are then made by randomly cutting and subcloning YAC inserts into cosmid vectors, and a map is constructed by identifying their landmark overlaps. Next, a minimally overlapping path of cosmid clones is chosen and the DNA from each clone is randomly fragmented into small pieces and subcloned into M13 phage or plasmid vectors. For each cosmid, about 800 M13 phage clones are sequenced (roughly 400 base pairs per clone) and assembled computationally into the sequence of the 40-kilobase (kb) cosmid insert. This random, or 'shotgun', approach ensures a high degree of accuracy because every nucleotide is sequenced about eight times (400 bases per clone × 800 clones = 320,000 bases of sequence). Most genome-wide or chromosome-specific physical maps generated so far are of low resolution and based on YACs[4,5].

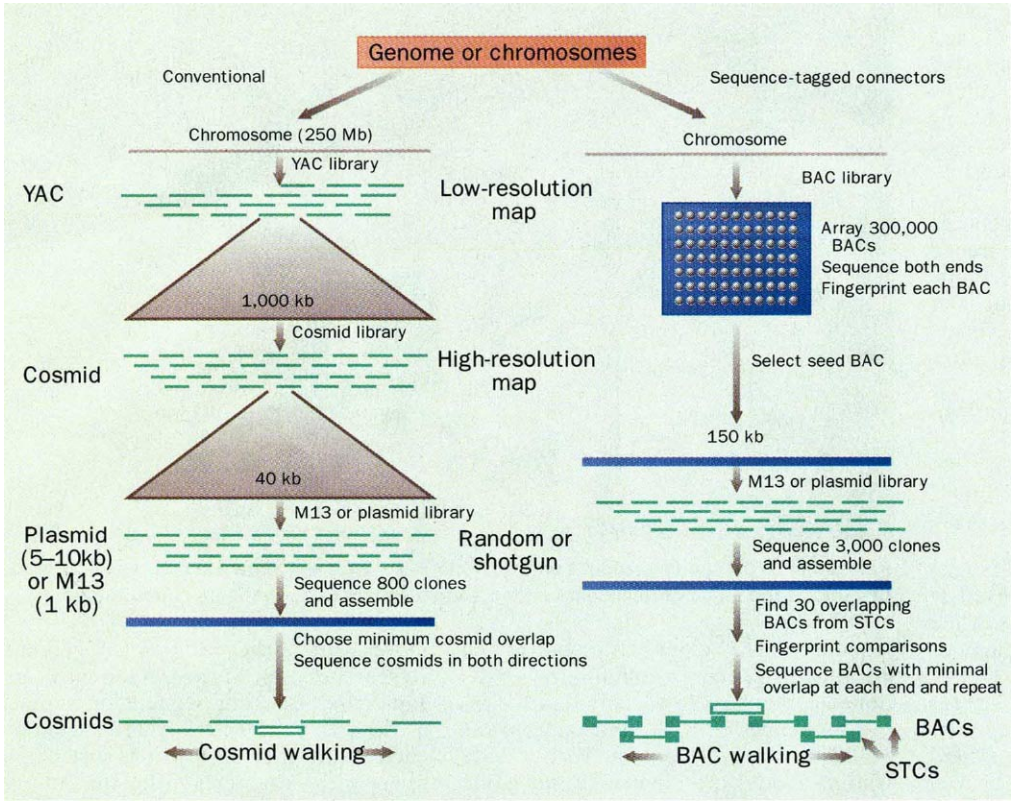
The approach to genome-wide sequencing presents several challenges and has various limitations. First, initial attempts at high-resolution mapping of human chromosomes 16, 19 and 22 have been expensive and are still not finished; the problem is that it is difficult to obtain complete maps without any gaps. Completing sequence-ready maps for these and the other human chromosomes remains a forbidable task.

Second, some 50 per cent of YAC clones show structural instability of inserts, resulting in deletion or rearrangement of portions of the cloned DNA, or are chimaeras in which two or more DNA fragments have become incorporated into one clone. These defective YACs are invariably unsuitable for use as mapping and sequencing reagents, and a great deal of effort is required to identify them. Cosmid inserts sometimes contain the same aberrations and are similarly difficult to detect.

Third, the human genome contains tandem (adjacent) arrays of DNA units with high sequence similarity (for example, five tandem 21-kb arrays) and tandemly arrayed genome-wide repeats (for example, the 7-kb long interspersed nuclear elements) that pose problems for high-resolution mapping when the size of the clone insert is less than that of the tandem array because the landmarks are similar, such as a 40-kb cosmid insert against a 105-kb DNA array.

Fourth, the conventional sequencing procedure is complex and difficult to automate fully for the vast amount of sequencing we hope to achieve in the future.

Finally, because an extensive infrastructure is required for high-resolution physical mapping, collaboration among



The conventional sequencing approach and the newly proposed sequence-tagged connectors (STC) approach. The bacterial artificial chromosome (BAC) clones in the STC approach could be sequenced by any cost-effective strategy.

large and small groups is difficult.

These problems, and two important scientific advances, lead us to propose a new strategy for cooperative sequencing of the human genome. The first advance is that we can now sequence and assemble prokaryote genomes of up to several megabases (Mb) with high accuracy and fidelity[6,7]. The second is the development of bacterial artificial chromosome (BAC) libraries that can accept human inserts of up to 350 kb. BAC clones seem to represent human DNA far more faithfully than their YAC or cosmid counterparts[8,9]. For example, the 1-Mb locus encoding the human αδ T-cell receptor has been mapped using only 17 BAC clones, in contrast to the 75 or so cosmid clones that would have been required for the same coverage. Detailed landmark analyses show that only one of 17 BAC clones had a defect, a small 6-kb deletion (C. Boysen, personal communication).

What is more, BAC clones are excellent substrates for shotgun sequence analysis. For example, five out of five BAC clones, ranging in size from 89 to 210 kb, have been sequenced using this approach (C. Boysen, personal communication); other laboratories have had similar success rates. So BAC clones seem ideal for producing an accurate contiguous sequence.

Our new approach to genomic sequencing eliminates the need for any prior physical mapping and uses BAC clones as the basic sequencing reagent (see the figure). A human BAC library with an average insert size of 150 kb and about 15-fold coverage of the human genome contains 300,000 clones. These are arrayed into microtitre wells. Both ends (starting at the vector-insert points) of each BAC clone are then sequenced to generate 500 bases from each end. The 600,000 BAC end sequences are scattered roughly every 5 kb across the genome and make up 10 per cent of the genome sequence. We denote them 'sequence-tagged connectors', or STCs, because they allow any one BAC clone to be connected to about 30 others (for example, a 150-kb insert 'divided' by 5 kb will be represented in 30 BACs). The STCs would be made immediately available electronically on the World Wide Web.

Next, each BAC clone is fingerprinted using one restriction enzyme to provide the insert size and detect artefactual clones by comparing the fingerprints with those of overlapping clones. A seed BAC of interest is sequenced by any method and checked against the database of STCs to identify the 30 or so overlapping BAC clones. The two BAC clones showing internal consistency among the fingerprints and minimal overlap at either end are then sequenced. In this way, the entire human genome could be sequenced with just over 20,000 BAC clones (see the table).

Our approach has several distinct advantages:

- The cost and effort of obtaining complete low- and high-resolution maps are virtually eliminated, so the front-end automation is greatly simplified (including clone arraying, DNA purification, fingerprinting and sequence reactions).
- The BAC clones can be made readily available to sequencing groups worldwide through resource centres and commercial distributors. Large centres could sequence many different BAC clones forming major contiguous regions of DNA, while small groups could contribute one or a few BAC sequences.
- As improved techniques for generating BAC or other yet-to-be-developed libraries appear, reasonable numbers of these new clones could easily be added to the clone collection.
- It is likely that our approach will eliminate the problem of generating complete maps without gaps for high-resolution physical mapping.

CLONE LIBRARIES USED FOR GENOME MAPPING AND SEQUENCING

Vector	Human-DNA insert size range	Number of clones required to cover the human genome
Yeast artificial chromosome (YAC)	100–2,000 kb	3,000 (1,000 kb)
Bacterial artificial chromosome (BAC)	80–350 kb	20,000 (150 kb)
Cosmid	30–45 kb	75,000 (40 kb)
Plasmid	3–10 kb	600,000 (5 kb)
M13 phage	1 kb	3,000,000 (1 kb)

■ The existing chromosomal landmarks, STSs (PCR-specific sites) and expressed sequence tags or ESTs (partial complementary DNA sequences), can be easily placed on the BAC clones, adding additional markers for BAC clones and taking advantage of any associated biological information.
■ The 10 per cent of the genome obtained in the STCs can be searched against the sequence database to identify many interesting landmarks (such as genes, STSs and ESTs) that could locate the BAC clone on the preexisting chromosomal maps.
■ Chromosomal regions of key biological interest can be sequenced first.
■ The human genome can be sequenced earlier and for less money (with savings on, for example, high-resolution physical mapping).
■ The STC approach will provide useful clones for biological studies even at the early STC sequencing stages when only three- to fourfold coverage is achieved.
■ This would be an efficient strategy for sequencing genomes of smaller organisms such as single-cell eukaryotes and the malarial and other parasites, as well as the model organisms that the genome project is already committed to sequencing: the bacterium *Eschericia coli*, the nematode, the flowering plant *Arabidopsis*, the fruitfly *Drosophila* and the mouse. (The yeast genome is finished.)

Arraying and DNA preparation facilities could readily make the end-sequenced BAC clones available to the worldwide genome community. BAC clones could be easily mailed and BAC end sequences or STCs and fingerprints would be available on the World Wide Web, as would the identity of any clone selected for sequencing. Several research teams could therefore work on the same chromosomal region without unintended duplication of effort. This would aid international cooperation. With our proposed strategy, participating laboratories could sequence the BAC inserts by any cost-effective method they choose. Similarly, any DNA sequencing chemistries could be used, including those not yet developed.

Peter Menzel/SPL

IMAGE UNAVAILABLE FOR COPYRIGHT REASONS

With the STC strategy, participating laboratories could sequence the BAC inserts using any DNA sequencing chemistries, including those not yet developed.

The complete set of BAC end sequences and fingerprints could be obtained in two years or less using, for example, 30 Applied Biosystems 377 sequencers at a total cost of $5–10 million. This is a small fraction of the cost of sequence-ready physical mapping that has yet to be incurred.

A highly cooperative combination of large genome centres and small groups could finish sequencing the entire human genome in under ten years. The current cost of DNA sequencing is around $0.30 per finished base pair in the most efficient laboratories, and it is expected that it will fall to $0.10–0.25 per base pair in the next one to three years. At these rates, the total sequencing cost for the entire genome would be less than the genome funds spent so far. The Wellcome Trust in the United Kingdom recently announced that it was funding the Sanger Centre to sequence a sixth or more of the human genome. Researchers in France, Germany and Japan are discussing sequencing as much as 10 per cent of the human genome each, while the US effort is just beginning with the establishment of six genome sequencing centres for pilot scale-up studies[11–14]. These large centres around the world, together with many smaller groups, require an improved strategy for genome coordination. The STC strategy proposed here offers a powerful new approach to sequencing human and other genomes, with maximum international cooperation and with all participants working on an equal basis in a self-regulating, open scientific effort. □

Philippe Plailly/SPL

IMAGE UNAVAILABLE FOR COPYRIGHT REASONS

Release of STCs and fingerprints on the World Wide Web will allow several research teams to work on the same chromosome region and so enhance international cooperation.

J. Craig Venter is at the Institute for Genomic Research, 9712 Medical Center Drive, Rockville, Maryland 20850, USA. Hamilton O. Smith is in the Department of Molecular Biology and Genetics, Johns Hopkins Medical School, 725 Wolfe Street, Baltimore, Maryland 21205-2185, USA. Leroy Hood is in the Department of Molecular Biotechnology, University of Washington, Box 357730, Seattle, Washington 98195-7730, USA.

1. Watson, J. D. *Science* **248,** 49 (1990).
2. National Research Council *Mapping and Sequencing the Human Genome* (National Academy Press, Washington DC, 1988).
3. Marshall, E. *Science* **268,** 1270–1271 (1995).
4. Chumakov, I. M. *et al. Nature* **377** (suppl.), 175–298 (1995).
5. Olsen, M. V. *Science* **270,** 394–396 (1995).
6. Fleischmann, R. D. *et al. Science* **269,** 496–512 (1995).
7. Fraser, C. M. *et al. Science* **270,** 397–403 (1995).
8. Kim, U. J. *et al. Nucleic Acids Res.* **20,** 1,083–1,085 (1992).
9. Shizuya, H. *et al. Proc. natn. Acad. Sci U.S.A.* **89,** 8,794–8,797 (1992).
10. Dickson, D. *Nature* **378,** 120 (1995).
11. Marshall, E. & Pannisi, E. *Science* **272,** 188–189 (1996).
12. Macilwain, C. *Nature* **380,** 471 (1996).
13. Craig, C. *BioWorld Today* **7,72,** 1,3 (1996).
14. Marshall, E. *Science* **272,** 477–478 (1996).