79 (1997); N. W. Toribara and M. H. Sleisenger, *N. Engl. J. Med.* **332**, 861 (1995)].
9. The first report of linkage of familial breast cancer to chromosome 17q21 [J. M. Hall *et al.*, *Science* **250**, 1684 (1990)] involved 23 extended families with an average of 6.3 cases per family. The study describing the risk of breast and ovarian cancer in BRCA1 mutation carriers was based on families in which a combination of at least four cases of ovarian cancer at any age or breast cancer before age 60 were present [D. F. Easton, D. Ford, D. T. Bishop, the Breast Cancer Linkage Consortium, *Am. J. Hum. Genet.* **56**, 265 (1995)].
10. A. Schatzkin, A. Goldstein, L. S. Freedman, *J. Natl. Cancer Inst.* **87**, 1126 (1995). Struewing *et al.* recently found that the risk of developing breast cancer among women drawn from the general population of Ashkenazi Jews before age 70 years was 56% [J. P. Struewing *et al.*, *N. Engl. J. Med.* **336**, 1401 (1997)], which is lower than the 85% risk of breast cancer in families with multiple affected members used in the research studies [D. Easton, *Nature Genet.* **16**, 210 (1997)]. The association of late-onset AD in people with the apoE4 polymorphism is greater in families with multiple affected members than in sporadic cases [S. Seshadri, D. A. Drachman, C. F. Lippa, *Arch. Neurol.* **52**,.1074 (1995); National Institute on Aging, *Lancet* **347**, 1091 (1996)].
11. M. H. Gaston *et al.*, *N. Engl. J. Med.* **314**, 1593 (1986); E. Vichinsky, D. Hurst, A. Earles, K. Kleman, B. Lubin, *Pediatrics* **81**, 749 (1988); N. A. Holtzman, R. A. Kronmal, W. van Doorninck, C. Azen, R. Koch, *N. Engl. J. Med.* **314**, 593 (1986).
12. B. Pasini, I. Ceccherini, G. Romeo, *Trends Genet.* **12**, 138 (1996)
13. E. Treacy, B. Childs, C. R. Scriver, *Am. J. Hum. Genet.* **56**, 359 (1995).
14. W. Burke *et al.*, *J. Am. Med. Assoc.* **277**, 915 (1997); W. Burke *et al.*, *ibid.*, p. 997.
15. N. A. Holtzman, *Adv. Oncol.* **13** (no. 1), 9 (1997).
16. E. S. Tambor *et al.*, *Am. J. Hum. Genet.* **55**, 626 (1994); R. T. Croyle, D. S. Dutson, V. T. Tran, Y. Sun, *Women's Health Res. Gender Behav. Pol.* **1**, 329 (1995).
17. N. S. Wexler, *FASEB J.* **6**, 2820, (1992).
18. K. L. Hudson, K. H. Rothenberg, L. B. Andrews, M. J. E. Kahn, F. S. Collins, *Science* **270**, 391 (1995); K. H. Rothenberg *et al.*, *ibid.* **275**, 1755 (1997); N. A. Holtzman and L. B. Andrews, *Epidemiol. Rev.* **19**, 163 (1997).
19. Until they were confident that locus heterogeneity did not exist, the Huntington researchers refused to make their probes available outside of investigative protocols [J. F. Gusella, *Nature* **320**, 329 (1986); *ibid.* **323**, 118 (1986)]. In addition, concern about the psychological impact of identifying people at risk of a fatal, untreatable disease led to pilot programs with extensive guidelines for counseling and support for those considering predictive testing, many of which have been retained as testing has become available in health care [C. De Somviele *et al.*, *J. Med. Genet.* **27**, 34 (1990)].
20. National Institutes of Health, *N. Engl. J. Med.* **323**, 70 (1990); American Society of Human Genetics, *Am. J. Hum. Genet.* **46**, 393 (1990).
21. O. Smith, *Nature Med.* **2**, 613 (1996).
22. American Society of Clinical Oncology, *J. Clin. Oncol.* **14**, 1730 (1996); American Society of Human Genetics Ad Hoc Committee, *Am. J. Hum. Genet.* **55**, 1 (1994).
23. American College of Medical Genetics and American Society of Human Genetics Working Group on ApoE and Alzheimer Disease, *J. Am. Med. Assoc.* **274**, 1627 (1995); National Institute on Aging, *Lancet* **347**, 1091 (1996).
24. Additional characteristics include the likelihood that a test will have low sensitivity (because of genetic heterogeneity) and low PVP (because of incomplete penetrance), and the absence of an intervention proven to be effective in those with positive test results. Tests for disorders of high prevalence, for screening populations, or for use selectively in ethnic groups with higher incidence or prevalence of the disorder are other characteristics that might increase the priority for review.
25. See, for example, R. R. Howell *et al.*, *NIH Consensus Development Conference Statement: Genetic Testing for Cystic Fibrosis*, in press.
26. Preparation of this paper was supported in part by the National Human Genome Research Institute. The views expressed in this paper are entirely those of the authors and do not necessarily represent the organizations with which they are affiliated.

Sequencing the Human Genome

Lee Rowen,* Gregory Mahairas, Leroy Hood

At the end of 1997, we are halfway through the time allotted for completing the Human Genome Project. The Human Genome Project aims to sequence the genomes of the human and selected model organisms, identify all of the genes, and develop the technologies required to accomplish these objectives. Significant progress has been made, particularly in identifying and mapping genes, developing a stable DNA-sequencing technology, and in building the computational tools required for the analysis of sequence data. Yet, the large-scale sequencing of the 3 billion base pairs of the human genome has barely begun (Table 1). Approximately 60 million base pairs have been analyzed to date. Of these, the longest contiguous stretch of human DNA sequence in a public database is less than 1.5 million base pairs (*1*). Here we discuss today's challenges for sequencing the human genome.

What Has Been Done So Far?

Gene identification. The expressed genes from hundreds of different human tissues have been partially sequenced after copying the messenger RNAs into complementary DNA libraries. About 800,000 of these so-called expressed sequence tags (ESTs) are available in public databases and at various Web sites (*2*). These represent perhaps 40,000 to 50,000 genes of the estimated total of 70,000 to 100,000 human genes. ESTs from a variety of model organisms are also available.

Mapping. Mapping requires the identification of unique genome markers [for example, ESTs or sequence-tagged sites (STSs)] and their localization to specific chromosomal sites. STSs are unique addresses generated by polymerase chain reaction primers that amplify just a single chromosomal site. Three techniques have been used for marker localization: genetic mapping (generally 1- to 10-Mb resolution), fluorescence in situ hybridization (~1-Mb resolution), and radiation hybrid mapping (down to 50-kb resolution). By means of these techniques, markers have been placed on average every 200 kb across the genome (*3*). Using STS landmarks to identify and order clones, researchers have constructed a framework physical map for most of the human genome from large inserts of human DNA cloned into yeast artificial chromosomes (YACs) in 1993 (*4*). A genetic map with more than 5000 highly polymorphic simple sequence repeats is also available (*5*).

Clone library construction. Human chromosomes cannot be sequenced directly. Rather, human DNA must be isolated, randomly fragmented, and cloned into vectors capable of stable propagation in a suitable host such as the bacterium *Escherichia coli* or yeast. Before sequencing, clones must be selected from libraries with chromosomal markers as probes, verified for their fidelity to the genome, and ordered in a minimal-overlapping tiling path spanning a portion of a chromosome. Several cloning systems with insert sizes varying from hundreds of base pairs to megabases have been successfully developed. The ideal clone library for genomic sequencing has the following features: (i) the clones are highly redundant, covering the entire human genome many times; (ii) the clone coverage is random and not biased toward or against specific regions of the genome; and (iii) the clones are stable, not subject to deletion or rearrangement during the propagation process. A signifi-

Table 1. Current state of genome sequence, as of September 1997.

Organism	Size (Mb)	Sequenced	Percent finished
Microbial genomes (~11)	0.6–4.2	0.6–4.2	100
E. coli	4.6	4.6	100
Yeast	13	13	100
Nematode	100	71	71
Drosophila	130	8	6
Mouse	3000	6	0.2
Human	3000	60	2

The authors are in the Department of Molecular Biotechnology, University of Washington, Post Office Box 357730, Seattle, WA 98195–7730, USA.

*To whom correspondence should be addressed.

cant advance in clone library construction for physical mapping and large-scale sequencing was the development of the bacterial artificial chromosome (BAC) vector in 1992. BAC vectors stably propagate DNA with insert size from 80 to 300 kb (6).

Large-scale sequencing. Beginning with a source clone, most large-scale sequencing centers use the following steps for sequence determination: randomly fragmenting the source clone into small (~1500 base pairs) pieces, subcloning the small pieces into a sequence-ready vector, sequencing 10 to 30 subclones per kilobase of the source clone (7), and assembling the overlapping sequence reads into a contiguous multiple sequence alignment from which a consensus sequence can be inferred from the highest quality reads.

The development of the automated fluorescent sequencer in the mid-1980s made high throughput genomic sequencing possible (8). One Perkin-Elmer Applied Biosystems 377 sequencer can produce two runs per day of sequence reads averaging 750 bases in length. Each run produces 64 reads (soon to be 96 reads). Improvements in the sequencing chemistries (better polymerases and higher sensitivity dyes) have resulted in higher quality, more accurate sequence data. Finally, more powerful computers combined with more sophisticated assembly programs have facilitated the determination of a consensus sequence from a given set of sequence reads. Using current sequencing methodologies, several laboratories are now producing contiguous stretches of human sequence in the 300-kb to 1.5-Mb range (1).

Sequence analysis. Extensive databases of sequences obtained from expressed genes and genomic clones from hundreds of organisms have been assembled and maintained by the National Center for Biotechnology Information (NCBI), Genome Sequence Data Bank (GSDB), European Molecular Biology Laboratory (EMBL), and the DNA Data Bank of Japan (DDBJ). Powerful search engines, accessible via Web sites, electronic mail, or direct connection to a server and database have enabled biologists to query the sequence data in the context of many different analyses, including gene finding, protein motif identification, regulatory motif analysis, identification of repeated sequences, similarity analyses, nucleotide compositional analyses, and cross-species comparisons. The explosion of data produced by the Human Genome Project has called forth the creation of a new discipline—bioinformatics, whose focus is on the acquisition, storage, analysis, modeling, and distribution of the many types of information embedded in DNA and protein sequence data.

Genomes sequenced. The genomes of *E. coli*, yeast, and 11 microbes have been completely sequenced (9). Those of the worm, fruit fly, mouse, and human have been partially sequenced (10) (Table 1). These sequences have dramatically altered the practice of genetics, molecular biology, developmental biology, immunology, and microbiology.

Challenges for Sequencing the Human Genome

Source material. Over the past decade, numerous clone libraries have been constructed from human sperm and cell lines. With these libraries, physical maps covering significant portions of the genome have been constructed, providing the source material for large-scale sequencing. Virtually all of the existing clone libraries must now be discontinued, however, because the National Institutes of Health (NIH) and the Department of Energy (DOE) have mandated that clone libraries come from donors that have given appropriate consent and are anonymous. This ruling is to prevent possible discrimination against DNA donors or their relatives as information from their genomes becomes available. Sequencing centers must now rebuild physical maps from the new clone libraries that are being constructed with proper internal review board (IRB) approval.

Clone validation. It has been argued that the genome should be sequenced from multiple libraries so that no individual's chromosomes are dominantly represented in the final sequence. Unfortunately, use of multiple libraries for selecting source clones undermines the ability of sequencing centers to validate the fidelity of their clones. This is so because validation is judged by internal consistency among overlapping clones assayed by restriction enzyme fingerprinting. The rate of polymorphism in the human population is about one variation per 500 base pairs with ~15% of the variations being insertions or deletions (11). Sequence polymorphisms lead to differences in fingerprints among overlapping clones that cannot be distinguished easily from differences in the fingerprints that arise from deletions or rearrangements of artifactual clones.

Thus, technical considerations argue for using a minimum number of highly redundant clone libraries, and social considerations argue for diversifying the clone libraries.

Minimum tiling paths of contiguous clones. A rate-limiting step in large-scale sequencing today is the identification of a contiguous array of sequence-ready clones across each human chromosome from which a set of minimally overlapping clones (minimal tiling path) can be sequenced. This problem poses three challenges: (i) What is the most efficient means of obtaining local minimum tiling paths (for example, those around each STS marker)? (ii) How can one identify and cover the gaps between the existing chromosomal markers? and (iii) What can be done when clones are missing from a highly redundant library because of nonrandom coverage or sequence-specific instability? For example: In the nematode, to achieve contiguity in the clone map, both bacterially based and yeast-based cloning systems were required. Only about 80% of the genome is represented in bacterially based cosmid clones, and the average size of the cosmid tiling paths (contigs) is 150 to 200 kb (12). Yeast-based clones (YACs) bridge the cosmid contigs, with only three gaps remaining in the 100-Mb genome. What will be required to achieve this level of contiguity in the human genome is simply unclear. Closing gaps may well be the most difficult challenge for sequencing the human genome as a result of nonrandomness of the clone libraries and STS maps.

Most sequencing centers currently build minimal tiling paths by relying on STSs or genetic markers that have already been mapped to a 200-kb to 2-Mb region of a chromosome (3, 4). With these markers as probes, clones are selected from a library and ordered by restriction enzyme fingerprints. To obtain a minimal tiling path, gaps between contigs must be filled by additional rounds of library screening, after the identification of new markers from the contig ends. Because the rounds of library screening required to fill gaps are time-consuming, the construction of physical maps that contain minimal tiling paths may not be able to match the required throughput of sequencing, when the scale of sequencing ramps up to 75 Mb/year or more for each sequencing center.

In 1996, it was proposed that an extensive up-front characterization of a highly redundant BAC clone library would provide a simple and easily automatable approach to the construction of minimal tiling paths (13). This characterization involves arraying the clones into 384-well plates, fingerprinting each clone, and sequencing the two vector-insert ends of each clone. With a library that covers the genome 15-fold (300,000 clones), a BAC-end sequence (sequence-tagged connector, or STC) would be found in the genome on average every 5 kb. The STC sequences are ideal potential chromosomal markers for creating a more dense physical map. For example, 30,000 random STCs could be

localized to human chromosomes by radiation hybrid mapping at an average spacing of 100 kb. This would facilitate the identification of gaps in the tiling path so additional markers in the gap regions could be identified.

Any sequenced BAC clone of 150 kb would permit the alignment of 30 STCs across the sequenced region. Indeed, the minimally overlapping clones at the 5′ and 3′ ends of any stretch of preexisting sequence could be selected for sequencing, thereby extending the length of contiguous sequence. A comparison of the fingerprints from clones overlapping those selected for sequencing would provide the validation of genomic integrity. Because the STC markers would be tied directly to characterized clones, local minimal tiling paths of clones could be rapidly identified by the computer.

Sequence scale-up. To complete the genome by 2005, starting in 1998, seven large-scale sequencing centers, for example, would each have to complete on the order of 75 Mb/year. Sequencing centers now have a throughput of 2 to 30 Mb/year. If the genome is to be sequenced on time and within budget, sequencing must become significantly faster and cheaper. Technology improvements in mapping, sequencing, and informatics leading to largely automated sequencing production lines will be critical for this scale-up. Scaling-up sequencing requires laboratory information management systems (LIMS) that track clones and data produced from the clones and also record parameters such as reagents, protocols, and machine performance. Process-oriented quality measures will be essential for maintaining the efficiency of sequencing and the generation of data of adequate quality. Proper training of personnel is crucial to the success of a sequencing operation, as is the integration of effective new technologies in a manner that does not disrupt production.

Standards. The current standards for sequencing set by the NIH and DOE require that three conditions be met: (i) an error rate of not more than 1 in 10,000; (ii) sequence contiguity, that is, sequence without gaps; and (iii) clone validation, that is, a demonstration that clones faithfully represent the genome. The precision of sequencing and, by inference, the probable error, is estimated by comparing the sequences of overlapping clones. It is also gauged by programs that assign a quality measure to each base in a consensus sequence. Contiguity is gauged by the sequence assembly programs and by the successful overlapping of sequences from adjacent clones. Fingerprint comparison of the clones against other overlapping clones, and in some cases directly against genomic DNA, verifies clone integrity and validity (but see the discussion under "Clone validation" above).

Sequencing centers are challenged to meet these standards in the context of a high-throughput, largely automated factory-style operation. To date, sequencing operations have relied on a pool of skilled and experienced "finishers" who examine data, resolve discrepancies between conflicting sequence reads, direct any required additional sequencing, and evaluate the precision of a consensus sequence by comparing its consistency with fingerprints and sequences of overlapping clones. Finishing is made significantly easier when the overall quality of sequence reads is high (long reads, few errors or ambiguities). However, even with good data, there are usually gaps that must be filled, and often there are difficulties in the assembly of sequence reads into the correct alignment owing to the presence of repeats—two or more highly similar copies of a stretch of sequence. The process of finishing is more difficult to automate than bulk sequencing and is, for many sequencing centers, the rate-limiting step of their sequencing operation.

Dissemination of data to the community. Genome centers are currently required to release their data to the community in a timely manner. Currently, finished sequence is available from four large databases (GenBank and GSDB in the United States; EMBL and DDDJ in Europe and Japan, respectively). Unfinished sequence is usually available and searchable from the HTG (high-throughput genomes) division of GenBank (and other databases) or the Web pages of the individual genome centers, or both. Genome centers generally provide a minimum of annotation for finished sequence—identification of repeat sequences, similarity analyses, and some indication of the quality of the data. If the data are really to be useful to the biological and medical communities, it is essential that biological information in these sequences be far more completely annotated in the future. Indeed, the complete human genome sequence must be seen as a starting point for new biological investigations, not as an end in itself.

REFERENCES AND NOTES

1. The longest submitted sequence to date, 1.492 Mb, produced by the Olson laboratory at the University of Washington, derives from the metabotropic glutamate receptor 8 gene on chromosome 7 (GenBank accession number AE000668). Other sequences >1 Mb include the immunoglobulin λ locus (1.025 Mb, accession numbers D86989-D87024 and D88268-D88271); the T cell receptor α locus (1.071 Mb, accession numbers AE000658-AE000662); the DiGeorge critical region (1.25 Mb, http://www.genome.ou.edu/maps/dgcr.gb); and a region of chromosome 19q13 (1.02 Mb, http://www-bio.llnl.gov/sequence-bin/seq_19?band=q13.1). Numerous sequences in the several hundred kilobase range are now being produced by the large sequencing centers in the United States and elsewhere (for example, the Sanger Centre, Cambridge, UK).
2. M. D. Adams *et al.*, *Nature* **377** (suppl.), 3 (1995). EST databases can be accessed via the NCBI BLAST server (http://www.ncbi.nlm.nih.gov/BLAST/).
3. T. J. Hudson *et al.*, *Science* **270**, 1945 (1995).
4. D. Cohen, I. Chumakov, J. Weissenbach, *Nature* **366**, 698 (1993).
5. C. Dib *et al.*, *ibid.* **380**, 152 (1996).
6. H. Shizuya *et al.*, *Proc. Natl. Acad. Sci U.S.A.* **89**, 8794 (1992).
7. The number of reads per kilobase is a function of the desired redundancy of coverage (typically 5- to 10-fold) and the average read length (typically 400 to 800 bases). A comprehensive review of sequencing technology is found in Adams *et al.* [M. D. Adams, C. Fields, J. C. Venter, Eds., *Automated DNA Sequencing and Analysis* (Academic Press, London, 1994)].
8. L. M. Smith *et al.*, *Nature* **321**, 674 (1986).
9. *Escherichia coli*: F. R. Blattner *et al.*, *Science* **277**, 1453 (1997); yeast: A. Goffeau *et al.*, *ibid.* **274**, 546 (1996); representative microbial genomes: R. D. Fleischmann *et al.*, *ibid.* **269**, 496 (1995); C. M. Fraser *et al.*, *ibid.* **270**, 397 (1995); C. J. Bult *et al.*, *ibid.* **273**, 1058 (1996); R. Himmelreich *et al.*, *Nucleic Acids Res.* **24**, 4420 (1996); T. Kaneko *et al.*, *DNA Res.* **3**, 109 (1996).
10. The sequencing status of *C. elegans* can be accessed from http://genome.wustl.edu/gsc/gschmpg.html; the *Drosophila melanogaster* sequencing status can be accessed from http://fruitfly.berkeley.edu/. There is no central location for the overall status of human and mouse sequencing progress.
11. Truly reliable estimates of sequence variation rates across the human genome require more data than currently exist. The numbers cited derive from an analysis of 290 kb of the human β T cell receptor locus for which more than one haplotype was sequenced (L. Rowen, unpublished results). These variations are annotated in GenBank accession numbers U66059, U66060, and U66061.
12. R. Wilson and R. Waterston, personal communication.
13. C. J. Venter, H. O. Smith, L. Hood, *Nature* **381**, 364 (1996).