POLICY FORUM: GENOMICS

The Human Genome Project: Reaching the Finish Line

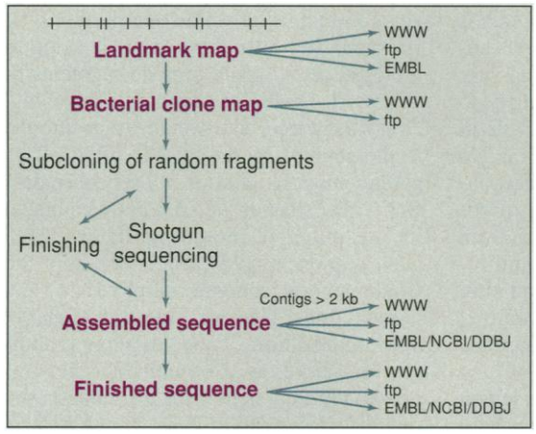
R. Waterston and J. E. Sulston

The human DNA sequence will be the central organizing principle for human genetics in the next century. It will be an essential reference for all biologists, and it is therefore vital that a high standard of accuracy, continuity, and accessibility be achieved as the sequence is determined. Because of the intrinsic excitement of unraveling the human genetic code and its potential to benefit humanity, there is a great temptation to acquire a view of the human genome as fast as possible. New initiatives that accelerate and enhance the program are to be welcomed and integrated into the emerging product, but none must divert us from the central aim of producing the ultimate, complete reference sequence.

Notable milestones in the International Human Genome Project include a comprehensive human genetic map (*1*), on which landmark-based physical maps of the genome have been built (*2*), and a human transcript map, also placed on the genetic framework, which provides any user with the positions of more than 30,000 human genes (*3*). These resources have had an immediate effect in accelerating the identification of human disease genes. Since 1990, roughly 200 disease-associated genes have been identified by strategies that have benefited directly from map or sequence information provided by the Human Genome Project (*4*, *5*). Such studies are leading to advances in the molecular understanding of disease, the development of powerful new diagnostic tests, and increased hopes for preventive strategies, new treatments, and even cures.

The systematic sequencing of the genomes of such model organisms as *Haemophilus influenzae* [1.8 megabases (Mb)] (*6*) and *Saccharomyces cerevisiae* (12.5 Mb) (*7*) have already given us more than a glimpse of the future. The project to sequence the 100-Mb genome of the nematode *Caenorhabditis elegans*, expected to be completed at the end of this year, will reveal all the genes and other information encoded in the DNA of a multicellular organism. The sequences of these genomes, combined with ever-growing amounts of sequence data from a wide variety of other organisms, have opened up new lines of communication among biologists for an explosion of discoveries. Suddenly, a gene discovered in humans can be understood through its relatives in other organisms (*8*).

When the *C. elegans* project began in 1990, there was little consensus about how to sequence genomic DNA on this scale. In fact, most scientists felt that revolutionary technology would be required before sequencing of the human genome could be undertaken. Instead, the continual evolution of sequencing methods and strategies used for model organisms has brought the human genome within reach. As a result, there has been a remarkable convergence by the scientific community on one general process (*9*) (see the figure), outlined as follows: A clone representing the region under study is selected from the map for subcloning and sequencing. Individual sequences of 400 to 1000 bases are generated by gel-based separation of dideoxy-terminated products produced by in vitro DNA synthesis, and the sequencing ladder is detected with fluorescent labels. Sequences are generated first from random fragments of the DNA to be sequenced (the "shotgun" phase). After automatic assembly of these sequence reads, based on sequence overlaps, additional data are collected in a directed fashion. In this "finishing" phase, ambiguities and gaps in the assembled sequence are resolved by editing; by determination of further sequence data on existing templates, using alternative sequencing chemistries, custom primer walks, long reads or reverse reads; and by generating new templates by constructing small-insert libraries of specific restriction fragments or polymerase chain reaction products that span gaps or difficult regions.



Consensus strategy. Flow chart for complete sequencing of the human genome.

The Wellcome Trust, the National Human Genome Research Institute, the U.S. Department of Energy, and other government agencies launched a pilot phase of systematic human sequencing in late 1995 and early 1996 to test the feasibility of applying the methods and strategy of the model organism projects to human genomic DNA. In only 2 years, the world total of finished human genomic sequence has gone from 15 to 180 Mb. If assembled but as-yet-unfinished sequence is added on, then the total available in the public domain exceeds 330 Mb, or 11% of the genome (*10*). With the increases in throughput previously planned, more than 800 Mb would be finished by 2001, and a total of 1500 Mb of finished and assembled sequence (half the genome) would be expected to have appeared in the public domain. This is the result of a combined effort involving large and small groups, coordinating their efforts through a publicly available map (*11*).

In addition to convergence on sequencing methods, the community has agreed on issues that are key to the success of the project and to coordination among groups, as follows:

(i) The release of all data is immediate, without constraints. The finished sequence of each clone is submitted to the public databases without delay and without patenting. All unfinished assembled sequence is released every night (*12*).

(ii) All sequence is finished to high accuracy. The agreed error rate for finished sequence is less than 1 in 10,000 bases, and all sequences are fully contiguous. Without this standard of accuracy, the power and precision of computational searches for genes and sequence variations are compromised, and investigators have to waste time and money to correct and finish the sequence before it is reliable. To achieve this standard, rigorous cross-checking between laboratories has been carried out, and new strategies for gap closure have been developed (*13*), so that even the most difficult (for example, GC-rich) regions in

R. Waterston is at the Genome Sequencing Center, Washington University School of Medicine, St. Louis, MO 63108, USA. E-mail: rw@genetics.wustl.edu. J. E. Sulston is at the Sanger Centre, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK. E-mail: jes@sanger.ac.uk.

the human sequence can be completed. The only exceptions are large arrays of tandem repeats; as these are variable in the population and are unstable in available cloning vectors, attempts to generate perfect sequence from them are unwarranted.

(iii) All clones used for sequencing are physically mapped. This approach ensures good coordination and cooperation between laboratories. Sequencing on a clone-by-clone basis also has the advantage that sequencing problems can be isolated and solved more easily, which becomes critical as the scale of the project increases. In particular, the problem of assembling a consensus sequence from individual reads is greatly simplified.

The absence of a preexisting high-resolution clone map of the human genome has led to proposals that could avoid mapping. A whole-genome shotgun approach was proposed as an alternative by Weber and Myers in 1996 (*14*). This strategy, as recently revived by Venter *et al.* (*15*), entails subcloning random fragments of total human genomic DNA directly and using very high-throughput sequencing technology to generate random sequences to provide at least 10-fold coverage of the genome. No doubt a vast amount of useful sequence information will be derived. But a major shortcoming of this strategy is that even with accurate linking of forward and reverse reads from templates of varying sizes, assembly would likely be woefully inadequate (*16*). This is chiefly because the human genome is highly repetitive (40% or more), and many repeats exceed the length of a single sequence read (*17*). Sequences containing very similar repeats will be automatically assembled into the same contigs, and careful reanalysis will be necessary to detect and correct such misassemblies. Even assemblies of shotgun sequence data of 100- to 200-kb bacterial artificial chromosome (BAC) clones or P1-derived artificial chromosome (PAC) clones sometimes require editing and additional work before they can be completed.

Another strategy to avoid mapping, proposed by Venter in 1996, was to sequence each end of every BAC in a genomic library (*18*). Matching a finished sequence to one or more BAC ends permitted selection of the next clone (*18*). However, this approach is also prone to errors caused by repeats. Since then, the acceleration in physical mapping has removed doubts about the feasibility of taking a map-based approach (*19*).

The success of the international consortium gives us confidence that the human genome sequence can be completed with high quality by the target date of 2005. However, there is a growing urgency worldwide to see the whole human sequence by the millenium. More and more biologists in universities and industries are depending on genomic sequence information. Breakthroughs in understanding human disease have been made with finished and unfinished sequence.

The growing experience of the sequencing centers, plus advances in finishing strategies, suggest that with added funds, these groups could increase the output of finished sequence above the projected rates. It is also clear that the shotgun phase does not need to remain strictly linked with finishing and can be accelerated more readily. The required increase in capacity is being facilitated by advances in technology, with capillary sequencers from at least two commercial sources being developed.

The common view of the sequencing centers is that acceleration of the mapping and shotgun phases will provide a first draft of the entire human genome within 3 years from now, of which one-third will be finished. This would be of great value to biologists as it would provide segments of almost all genes for homology searching. Every sequence read would be positioned in the genome by the map location of the parent clone. In addition, all sequence data from other sources would readily be drawn in and mapped, including all available gene sequences, plus sequences from the whole-genome shotgun initiative (*16*) and possibly other sources. Indeed, the quality and organization of whole-genome shotgun initiatives will be considerably improved with the benefit of the clone-based sequence information. This should, in turn, further accelerate completion of the reference human sequence.

Although the acceleration of parts of the international program will require an initial outlay in expenditure, these are not new costs, but rather the earlier spending of funds already committed to human genome sequencing. More shotgun done now means less shotgun required later. Internationally, by June 2000, one-third of the genome will be covered by already-planned efforts. Shotgun coverage of the remainder would be achieved to a depth determined largely by the amount of available funding. To allow such a plan to proceed, the Wellcome Trust has recently agreed to bring forward part of its commitment to the Sanger Centre.

It is important to state explicitly that production of the first draft must not deflect us from the real goal of finishing the job, nor should it add to its overall cost. Acceleration of the intermediate stage must be carried out alongside the already planned commitment to increase the rate of finished sequence production. Highly accurate, fully contiguous sequence is the correct and most valuable product of the Human Genome Project (*20*). It will be required to find all the genes in human DNA, and it will be critical in mouse-human comparisons to refine gene predictions and find regulatory and other functional elements. Such studies will be vital to a fuller understanding of the human genome, as we have already seen with the nematode and yeast genomes. Furthermore, in areas where nothing is known now, accurate finished sequence is essential to facilitate future discovery.

References and Notes

1. C. Dib *et al.*, *Nature* **300**, 152 (1996); J. C. Murray *et al.*, *Science* **265**, 2049 (1994).
2. For whole-genome physical maps, see I. M. Chumakov *et al.*, *Nature* **377**, 175 (1995); T. J. Hudson *et al.*, *Science* **270**, 1945 (1995). Detailed physical maps of some individual chromosomes (21, 22, 16, 12, 3, and 7) appear in the following: M. Chumakov *et al.*, *Nature* **359**, 380 (1992); F. S. Collins *et al.*, *ibid.* **377**, 367 (1995); N. A. Doggett *et al.*, *ibid.*, p. 335; K. S. Krauter *et al.*, *ibid.*, p. 321; R. M. Gemmill *et al.*, *ibid.*, p. 299; G. G. Bouffard *et al.*, *Genome Res.* **7**, 673 (1997).
3. A transcript map of the human genome is available at http://ncbi.nlm.nih.gov/genemap. P. Deloukas *et al.*, *Science*, in press (Oct. 23, 1998).
4. A list of inherited disease genes identified by positional cloning is available at http://www.nhgri.nih.gov/DIR/GTB/CLONE/.
5. F. S. Collins, *Nature Genet.* **9**, 347 (1995).
6. The genome sequence of *H. influenzae* was determined by whole-genome random sequencing and assembly [R. D. Fleischmann *et al.*, *Science* **269**, 496 (1995)].
7. A. Goffeau *et al.*, *Nature* **387** (suppl.),1 (1997).
8. A database cross-referencing human genes and related genes in model organisms is available at http://www.ncbi.nlm.nih.gov/XREFdb and also in D. Bassett *et al.*, *Trends Genet.* **11**, 372 (1996).
9. For methods, see E. Mardis and R. Wilson, in *Genome Analysis, A Laboratory Manual, Vol. 1: Analyzing DNA*, B. Birren, E. D. Green, S. Klapholz, R. M. Myers, J. Roskams, Eds. (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, 1997), pp. 397–453; R. Wilson *et al.*, *Nature* **368**, 32 (1994).
10. These data are from the National Center for Biotechnology Information and Genome Monitoring Table at the European BioInformatic Institute. Currently, the largest sequence contig is over 2 Mb, and the largest mapped contig is 12.9 Mb (chromosome 22, available at http://webace.sanger.ac.uk/HGP/Chr22).
11. Current and planned efforts of the international genome sequencing consortium are listed in the human genome sequence index (http://ncbi.nlm.nih.gov/HUGO). Graphic displays of the information can be viewed at http://webace.sanger.ac.uk or at http://www.ncbi.nlm.nih.gov/Entrez/Genome/org.html. D. Bentley *et al.*, *Trends Genet.*, in press.
12. Participants at the first international strategy meeting on human genome sequencing unanimously endorsed the "Bermuda statement" that "all human genomic information produced at large-scale sequncing centres should be freely available and in the public domain, in order to encourage research and development and to maximise its benefit to society." For details of access, see (*8*).
13. A. A. MacMurray, J. E. Sulston, M. A. Quail, *Genome Res.* **8**, 562 (1998).
14. J. L. Weber and E. W. Myers, *ibid.* **7**, 401 (1997).
15. J. C. Venter *et al.*, *Science* **280**, 1540 (1998).
16. P. Green, *Genome Res.* **7**, 410 (1997).
17. A. F. A. Smit, *Curr. Opin. Gen. Dev.* **6**, 743 (1996).
18. J. C. Venter, H. O. Smith, L. Hood, *Nature* **381**, 364 (1996).
19 Over 400 Mb of mapped clones for sequencing have been produced in our centers (see http://www.sanger.ac.uk or http://genetics.wustl.edu).
20. M. V. Olson and P. Green, *Genome Res.* **8**, 414 (1998).

21. We are grateful for discussions with F. Collins, M. Morgan, E. Lander, D. Cox, R. Gibbs, and M. Olson, who have expressed support. We also to thank the staff of the Sanger Centre and the Genome Sequencing Center for assistance in preparing the manuscript.