5. G. An, B. D. Watson, C. C. Chiang, *Plant Physiol.* **81**, 301 (1986); A. M. Lloyd *et al.*, *Science* **234**, 464 (1986); K. A. Feldmann and M. D. Marks, *Mol. Gen. Genet.* **208**, 1 (1987).
6. E. M. Meyerowitz and R. E. Pruitt, *Science* **229**, 1214 (1985).
7. G. R. Fink, *Genetics* **149**, 473 (1998).
8. C. Koncz, N.-H. Chua, J. Schell, Eds., *Methods in Arabidopsis Research* (World Scientific, River Edge, NJ, 1992); J. M. Martinez-Zapater and J. Salinas, Eds., *Arabidopsis Protocols*, vol. 82 of *Methods in Molecular Biology* (Humana, Totowa, NJ, 1998).
9. For example, see S. A. Kempin *et al.*, *Nature* **389**, 802 (1997).
10. N. Bechtold, J. Ellis, G. Pelletier, *C. R. Acad. Sci. Paris* **316**, 1194 (1993); S. C. Chang *et al.*, *Plant J.* **5**, 551 (1994).
11. V. Sundaresan, *Trends Plant Sci.* **1**, 184 (1996).
12. D. Meinke and M. Koornneef, *Plant J.* **12**, 247 (1997).
13. P. Fransz *et al.*, *ibid.* **13**, 867 (1998).
14. C. Lister and C. Dean, *ibid.* **4**, 745 (1993); C. Alonso-Blanco *et al.*, *ibid.* **14**, 259 (1998).
15. E. J. Finnegan, R. K. Genger, W. J. Peacock, E. S. Dennis, *Annu. Rev. Plant Physiol. Plant Mol. Biol.* **49**, 223 (1998).
16. D. Preuss, S. Y. Rhee, R. W. Davis, *Science* **264**, 1458 (1994); G. P. Copenhaver, W. E. Browne, D. Preuss, *Proc. Natl. Acad. Sci. U.S.A.* **95**, 247 (1998).
17. For recent examples of identifying knockouts in desired genes, see E. C. McKinney *et al.*, *Plant J.* **8**, 613 (1995); P. J. Krysan, J. C. Young, F. Tax, M. R. Sussman, *Proc. Natl. Acad. Sci. U.S.A.* **93**, 8145 (1996).
18. For information, contact http://www.bio.net/hypermail/ARABIDOPSIS/.
19. D. Meinke *et al.*, Eds., "Multinational coordinated *Arabidopsis thaliana* genome research project, progress report, year six" (National Science Foundation Publication 97-131, Arlington, VA, 1997).
20. M. Bevan *et al.*, *Plant Cell* **9**, 476 (1997).
21. H. Hofte *et al.*, *Plant J.* **4**, 1051 (1993); T. Newman *et al.*, *Plant Physiol.* **106**, 1241 (1994).
22. S. Choi, R. A. Creelman, J. E. Mullet, R. A. Wing, *Plant Mol. Biol. Rep.* **13**, 124 (1995); F. Creusot *et al.*, *Plant J.* **8**, 763 (1995).
23. R. Schmidt *et al.*, *Science* **270**, 480 (1995); E. A. Zachgo *et al.*, *Genome Res.* **6**, 19 (1996); R. Schmidt, K. Love, J. West, Z. Lenehan, C. Dean, *Plant J.* **11**, 563 (1997).
24. M. Bevan *et al.*, *Nature* **391**, 485 (1998).
25. M. Sato *et al.*, *DNA Res.* **4**, 215 (1997).
26. C. Chang, S. F. Kwok, A. B. Bleecker, E. M. Meyerowitz, *Science* **262**, 539 (1993); G. E. Schaller and A. B. Bleecker, *ibid.* **270**, 1809 (1995).
27. J. Li, P. Nagpal, V. Vitart, T. C. McMorris, J. Chory, *ibid.* **272**, 398 (1996); M. Szekeres *et al.*, *Cell* **85**, 171 (1996).
28. E. Huala *et al.*, *Science* **278**, 2120 (1997); M. Ahmad, J. A. Jarillo, O. Smirnova, A. R. Cashmore, *Nature* **392**, 720 (1998).
29. H. Guo, H. Yang, T. C. Mockler, C. Lin, *Science* **279**, 1360 (1998); Z. Y. Wang and E. M. Tobin, *Cell* **93**, 1207 (1998).
30. P. H. Quail *et al.*, *Science* **268**, 675 (1995).
31. M. Koornneef, C. Alonso-Blanco, A. J. M. Peeters, W. Soppe, *Annu. Rev. Plant Physiol. Plant Mol. Biol.* **49**, 345 (1998).
32. E. S. Coen and E. M. Meyerowitz, *Nature* **353**, 31 (1991).
33. N. Wei *et al.*, *Curr. Biol.* **8**, 919 (1998).
34. Y. Miyamoto and A. Sancar, *Proc. Natl. Acad. Sci. U.S.A.* **95**, 6097 (1998).
35. A. F. Bent *et al.*, *Science* **265**, 1856 (1994); M. Mindrinos, F. Katagiri, G.-L. Yu, F. M. Ausubel, *Cell* **78**, 1089 (1994); K. S. Century *et al.*, *Science* **278**, 1963 (1997).
36. K. R. Jaglo-Ottosen, S. J. Gilmour, D. G. Zarka, O. Schabenberger, M. F. Thomashow, *Science* **280**, 104 (1998); Z. G. Xin and J. Browse, *Proc. Natl. Acad. Sci. U.S.A.* **95**, 7799 (1998).
37. C. Nawrath, Y. Poirier, C. Somerville, *Proc. Natl. Acad. Sci. U.S.A.* **91**, 12760 (1994).
38. D. Weigel and O. Nilsson, *Nature* **377**, 495 (1995).
39. Between 1971 and 1994 there were 16 patents in the U.S. Patent Database that included the word *Arabidopsis*. From 1995 to present this number increased to 156.
40. M. Schena *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **93**, 10614 (1996); A. Marshall and J. Hodgson, *Nature Biotech.* **16**, 27 (1998).

New Goals for the U.S. Human Genome Project: 1998–2003

Francis S. Collins,* Ari Patrinos, Elke Jordan, Aravinda Chakravarti, Raymond Gesteland, LeRoy Walters, and the members of the DOE and NIH planning groups

REVIEW

The Human Genome Project has successfully completed all the major goals in its current 5-year plan, covering the period 1993–98. A new plan, for 1998–2003, is presented, in which human DNA sequencing will be the major emphasis. An ambitious schedule has been set to complete the full sequence by the end of 2003, 2 years ahead of previous projections. In the course of completing the sequence, a "working draft" of the human sequence will be produced by the end of 2001. The plan also includes goals for sequencing technology development; for studying human genome sequence variation; for developing technology for functional genomics; for completing the sequence of *Caenorhabditis elegans* and *Drosophila melanogaster* and starting the mouse genome; for studying the ethical, legal, and social implications of genome research; for bioinformatics and computational studies; and for training of genome scientists.

The Human Genome Project (HGP) is fulfilling its promise as the single most important project in biology and the biomedical sciences—one that will permanently change biology and medicine. With the recent completion of the genome sequences of several microorganisms, including *Escherichia coli* and *Saccharomyces cerevisiae,* and the imminent completion of the sequence of the metazoan *Caenorhabditis elegans*, the door has opened wide on the era of whole genome science. The ability to analyze entire genomes is accelerating gene discovery and revolutionizing the breadth and depth of biological questions that can be addressed in model organisms. These exciting successes confirm the view that acquisition of a comprehensive, high-quality human genome sequence will have unprecedented impact and long-lasting value for basic biology, biomedical research, biotechnology, and health care. The transition to sequence-based biology will spur continued progress in understanding gene-environment interactions and in development of highly accurate DNA-based medical diagnostics and therapeutics.

Human DNA sequencing, the flagship endeavor of the HGP, is entering its decisive phase. It will be the project's central focus during the next 5 years. While partial subsets of the DNA sequence, such as expressed sequence tags (ESTs), have proven enormously valuable, experience with simpler organisms confirms that there can be no substitute for the complete genome sequence. In order to move vigorously toward this goal, the crucial task ahead is building sustainable capacity for producing publicly available DNA sequence. The full and incisive use of the human sequence, including comparisons to other vertebrate genomes, will require further increases in sustainable capacity at high accuracy and lower costs. Thus, a high-priority commitment to develop and deploy new and improved sequencing technologies must also be made.

Availability of the human genome sequence presents unique scientific opportunities, chief among them the study of natural genetic variation in humans. Genetic or DNA sequence variation is the fundamental raw material for evolution. Importantly, it is also the

F. S. Collins and E. Jordan are with the National Human Genome Research Institute, National Institutes of Health, Bethesda, MD 20892, USA. A. Patrinos is with the Office of Biological and Environmental Research, Department of Energy, Washington, DC 20585, USA. A. Chakravarti is with the Department of Genetics and Center for Human Genetics, Case Western Reserve University and University Hospitals of Cleveland, Cleveland, OH 44106, USA. R. Gesteland is at the Howard Hughes Medical Institute, University of Utah, Salt Lake City, UT 84112, USA. L. Walters is with the Kennedy Institute of Ethics, Georgetown University, Washington, DC 20057, USA.

*To whom correspondence should be addressed: E-mail: fc23a@nih.gov

basis for variations in risk among individuals for numerous medically important, genetically complex human diseases. An understanding of the relationship between genetic variation and disease risk promises to change significantly the future prevention and treatment of illness. The new focus on genetic variation, as well as other applications of the human genome sequence, raises additional ethical, legal, and social issues that need to be anticipated, considered, and resolved.

The HGP has made genome research a central underpinning of biomedical research. It is essential that it continue to play a lead role in catalyzing large-scale studies of the structure and function of genes, particularly in functional analysis of the genome as a whole. However, full implementation of such methods is a much broader challenge and will ultimately be the responsibility of the entire biomedical research and funding communities.

Success of the HGP critically depends on bioinformatics and computational biology as well as training of scientists to be skilled in the genome sciences. The project must continue a strong commitment to support of these areas.

As intended, the HGP has become a truly international effort to understand the structure and function of the human genome. Many countries are participating according to their specific interests and capabilities. Coordination is informal and generally effected at the scientist-to-scientist level. The U.S. component of the project is sponsored by the National Human Genome Research Institute at the National Institutes of Health (NIH) and the Office of Biological and Environmental Research at the Department of Energy (DOE). The HGP has benefited greatly from the contributions of its international partners. The private sector has also provided critical assistance. These collaborations will continue, and many will expand. Both NIH and DOE welcome participation of all interested parties in the accomplishment of the HGP's ultimate purpose, which is to develop and make publicly available to the international community the genomic resources that will expedite research to improve the lives of all people.

The Planning Process

The last 5-year plan for the HGP, published jointly by NIH and DOE in 1993 (*1*), covered fiscal years 1994 through 1998. The current plan is again a joint effort and will guide the project for fiscal years 1999 through 2003.

The goals described below have resulted from a comprehensive planning and assessment process that has taken place over the past year in both agencies. Each agency identified a group of advisors to oversee its process, and eight workshops were held to address specific areas of the plan. A large number of scientists and scholars as well as public representatives participated in these events, including many who had no historical ties to the HGP. Comments were also sought from an extensive list of biotechnology and pharmaceutical companies. A draft of the goals was presented for evaluation at a public meeting in May 1998. Suggestions and comments from that meeting were incorporated into the plan. Finally, the new goals were reviewed and approved by the National Advisory Council for Human Genome Research at NIH and the Biological and Environmental Research Advisory Committee at DOE. Summaries of the workshops that contributed to this plan are available at www.nhgri.nih.gov and www.ornl.gov/hg5yp

Specific Goals for 1998–2003

The following sections outline eight major goals for the HGP over the next 5 years. Table 1 provides an overview of the quantifiable features of these new goals and compares them to the goals from 1993. Information on accomplishment of the 1993 goals is also included. Figure 1 describes the funding the U.S. HGP received to date.

Goal 1—The Human DNA Sequence

Providing a complete, high-quality sequence of human genomic DNA to the research community as a publicly available resource continues to be the HGP's highest priority goal. The enormous value of the human genome sequence to scientists and the considerable savings in research costs its widespread availability will allow are compelling arguments for advancing the timetable for completion. Recent technological developments and experience with large-scale sequencing provide increasing confidence that it will be possible to complete an accurate, high-quality sequence of the human genome by the end of 2003, 2 years sooner than previously predicted. NIH and DOE expect to contribute 60 to 70% of this sequence, with the remainder coming from the effort at the Sanger Centre, funded by the Wellcome Trust, and other international partners.

This is a highly ambitious, even audacious goal, given that only

Table 1. A comparison is made of some of the quantitative features of the new goals to the goals developed in 1993, and a snapshot of the worldwide status of completion of the 1993 goals is provided. Other features of the new goals are described in the text; cM, centimorgan.

Area	Goals 1993–98	Status as of Oct. 1998	Goals 1998–2003
Genetic map	Average 2- to 5-cM resolution	1 cM map published Sept. 1994	Completed
Physical map	Map 30,000 STSs	52,000 STSs mapped	Completed
DNA sequence	Complete 80 Mb for all organisms by 1998	180 Mb human plus 111 Mb nonhuman	Finish 1/3 of human sequence by end of 2001 Working draft of remainder by end of 2001 Complete human sequence by end of 2003
Sequencing technology	Evolutionary improvements and innovative technologies	90 Mb/year capacity at ~$0.50 per base Capillary array electrophoresis validated Microfabrication feasible	Integrate and automate to achieve 500 Mb/year at ≤$0.25 per base Support innovation
Human sequence variation	Not a goal	–	100,000 mapped SNPs Develop technology
Gene identification	Develop technology	30,000 ESTs mapped	Full-length cDNAs
Functional analysis	Not a goal	–	Develop genomic-scale technologies
Model organisms	*E. coli*: complete sequence	Published Sept. 1997	–
	Yeast: complete sequence	Released Apr. 1996	–
	C. elegans: most of sequence	80% complete	Complete Dec. 1998
	Drosophila: begin sequencing	9% done	Sequence by 2002
	Mouse: map 10,000 STSs	12,000 STSs mapped	Develop extensive genomic resources Lay basis for finishing sequence by 2005 Produce working draft before 2005

about 6% of the human genome sequence has been completed thus far (Fig. 2). Sequence completion by the end of 2003 is a major challenge, but within reach and well worth the risks and effort. Realizing the goal will require an intense and dedicated effort and a continuation and expansion of the collaborative spirit of the international sequencing community. Only sequence of high accuracy and long-range contiguity will allow a full interpretation of all the information encoded in the human genome. However, in the course of finishing the first human genome sequence by the end of 2003, a "working draft" covering the vast majority of the genome can be produced even sooner, within the next 3 years. Though that sequence will be of lower accuracy and contiguity, it will nevertheless be very useful, especially for finding genes, exons, and other features through sequence searches. These uses will assist many current and future scientific projects and bring them to fruition much sooner, resulting in significant time and cost savings. However, because this sequence will have gaps, it will not be as useful as finished sequence for studying DNA features that span large regions or require high sequence accuracy over long stretches.

Availability of the human sequence will not end the need for large-scale sequencing. Full interpretation of that sequence will require much more sequence information from many other organisms, as well as information about sequence variation in humans (see also Goals 3, 4, and 5). Thus, the development of sustainable, long-term sequencing capacity is a critical objective of the HGP. Achieving the goals below will require a capacity of at least 500 megabases (Mb) of finished sequence per year by the end of 2003.

a) Finish the complete human genome sequence by the end of 2003. The year 2003 is the 50th anniversary of the discovery of the double helix structure of DNA by James Watson and Francis Crick (*2*). There could hardly be a more fitting tribute to this momentous event in biology than the completion of the first human genome sequence in this anniversary year. The technology to do so is at hand, although further improvements in efficiency and cost effectiveness will be needed, and more research is needed on approaches to sequencing structurally difficult regions (*3*). Current sequencing capacity will have to be expanded two- to threefold, but this should be within the capability of the sequencing community.

Reaching this goal will significantly stress the capabilities of the publicly funded project and will require continued enthusiastic support from the Administration and the U.S. Congress. But the value of the complete, highly accurate, fully assembled sequence of the human genome is so great that it merits this kind of investment.

b) Finish one-third of the human DNA sequence by the end of 2001. With the anticipated scale-up of sequencing capacity, it should be possible to expand finished sequence production (Fig. 2) to achieve completion of 1 Gb of human sequence by the worldwide HGP by the

end of 2001. As more than half of the genes are predicted to lie in the gene-rich third of the genome, the finishing effort during the next 3 years should focus on such regions if this can be done without incurring significant additional costs. A convenient, but not the only, strategy would be to finish bacterial artificial chromosome (BAC) clones detected by complementary DNA (cDNA) or EST sequences.

In addition, a rapid peer-review process should be established immediately for prioritizing specific regions to be finished, based on the needs of the international scientific community. This process must be impartial and must minimize disruptions to the large-scale sequencing laboratories.

To best meet the needs of the scientific community, the finished human DNA sequence must be a faithful representation of the genome, with high base-pair accuracy and long-range contiguity. Specific quality standards that balance cost and utility have already been established. One of the most important uses for the human sequence will be comparison with other human and nonhuman sequences. The sequence differences identified in such comparisons should, in nearly all cases, reflect real biological differences rather than errors or incomplete sequence. Consequently, the current standard for accuracy—an error rate of no more than 1 base in 10,000—remains appropriate. Although production of contiguous sequence without gaps is the goal, any irreducible gaps must be annotated as to size and position. In order to assure that long-range contiguity of the sequence will be achievable, several contigs of 20 Mb or more should be generated by the end of 2001. These quality standards should be reexamined periodically; as experience in using sequence data is gained, the appropriate standards for sequence quality may change.

c) Achieve coverage of at least 90% of the genome in a working draft based on mapped clones by the end of 2001. The current public sequencing strategy is based on mapped clones and occurs in two phases. The first, or "shotgun" phase, involves random determination of most of the sequence from a mapped clone of interest. Methods for doing this are now highly automated and efficient. Mapped shotgun data are assembled into a product ("working draft" sequence) that covers most of the region of interest but may still contain gaps and ambiguities. In the second, finishing phase, the gaps are filled and discrepancies resolved. At present, the finishing phase is more labor intensive than the shotgun phase. Already, partially finished, working-draft sequence is accumulating in public databases at about twice the rate of finished sequence.

Based on recent experience, the rate of production of working draft sequence can be further increased. By continuing to scale up the production of finished sequence at a realistic rate and further scaling up the production of working draft sequence, the combined total of working draft plus finished sequence will cover at least 90% of the genome at an accuracy of at least 99% by the end of 2001. Some areas
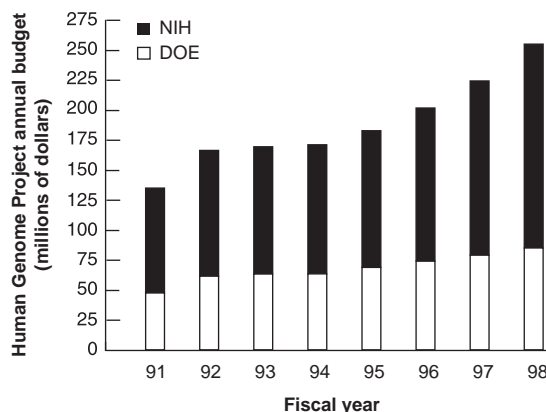
Fig. 1. The annual budget for the U.S. Human Genome Project for NIH and DOE is shown. In fiscal year 1998, NIH devoted 1.25% of its total budget of $13.6 billion to the HGP.
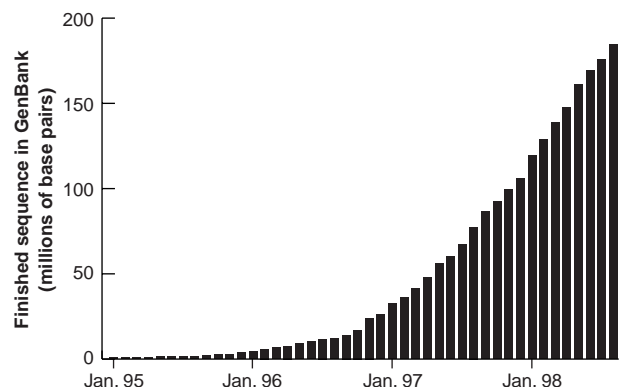
Fig. 2. Worldwide human genome sequencing progress is shown (measured as base pairs of finished sequence deposited with GenBank).

of the genome are likely to be difficult to clone or not amenable to automated assembly because of highly repetitive sequence; thus, coverage is expected to fall short of 100% at this stage. If increased resources are available or technology improves, or both, greater than 90% coverage may be possible.

The individual sequence reads used to generate the working draft will be held to the same high-quality standards as those used for the finished genome sequence. Assembly of the working draft should not create loss of efficiency or increases in overall cost.

Recently, two private ventures announced initiatives to sequence a major fraction of the human genome, using strategies that differ fundamentally from the publicly funded approach. One of these ventures is based upon a whole genome shotgun strategy, which may present significant assembly problems (4). The stated intention of this venture to release data on a quarterly basis creates the possibility of synergy with the public effort. If this privately funded data set and the public one can be merged, the combined depth of coverage of the working draft sequence will be greater, and the mapping information provided by the public data set will provide critically needed anchoring to the private data. The NIH and DOE welcome such initiatives and look forward to cooperating with all parties that can contribute to more rapid public availability of the human genome sequence.

d) Make the sequence totally and freely accessible. The HGP was initiated because its proponents believed the human sequence is such a precious scientific resource that it must be made totally and publicly available to all who want to use it. Only the wide availability of this unique resource will maximally stimulate the research that will eventually improve human health. Public funding of the HGP is predicated on the belief that public availability of the human sequence at the earliest possible time will lead to the greatest public good. Therefore, NIH and DOE continue to strongly endorse the policy for human sequence data release adopted by the international sequencing community in February 1996 (5), and confirmed and expanded to include genomic sequence of all organisms in 1998 (6). This policy states that sequence assemblies 1 to 2 kb in size should be released into public databases within 24 hours of generation and that finished sequence should be released on a similarly rapid time scale.

Goal 2—Sequencing Technology

DNA sequencing technology has improved dramatically since the genome project began. The amount of sequence produced each year is increasing steadily; individual centers are now producing tens of millions of base pairs of sequence annually (Fig. 2). In the future, de novo sequencing of additional genomes, comparative sequencing of closely related genomes, and sequencing to assess variation within genomes will become increasingly indispensable tools for biological and medical research. Much more efficient sequencing technology will be needed than is currently available. The incremental improvements made to date have not yet resulted in any fundamental paradigm shifts. Nevertheless, the current state-of-the-art technology can still be significantly improved, and resources should be invested to accomplish this. Beyond that, research must be supported on new technologies that will make even higher throughput DNA sequencing efficient, accurate, and cost-effective, thus providing the foundation for other advanced genomic analysis tools. Progress must be achieved in three areas:

a) Continue to increase the throughput and reduce the cost of current sequencing technology. Increased automation, miniaturization, and integration of the approaches currently in use, together with incremental, evolutionary improvements in all steps of the sequencing process, are needed to yield further increases in throughput (to at least 500 Mb of finished sequence per year by 2003) and reductions in cost. At least a twofold cost reduction from current levels (which average $0.50 per base for finished sequence in large-scale centers) should be achieved in the next 5 years. Production of the working draft of the human sequence will cost considerably less per base pair.

b) Support research on novel technologies that can lead to significant improvements in sequencing technology. New conceptual approaches to DNA sequencing must be supported to attain substantial improvements over the current sequencing paradigm. For example, microelectromechanical systems (MEMS) may allow significant reduction of reagent use, increase in assay speed, and true integration of sequencing functions. Rapid mass spectrometric analysis methods are achieving impressive results in DNA fragment identification and offer the potential for very rapid DNA sequencing. Other more revolutionary approaches, such as single-molecule sequencing methods, must be explored as well. Significant investment in interdisciplinary research in instrumentation, combining chemistry, physics, biology, computer science, and engineering, will be required to meet this goal. Funding of far-sighted projects that may require 5 to 10 years to reach fruition will be essential. Ultimately, technologies that could, for example, sequence one vertebrate genome per year at affordable cost are highly desirable.

c) Develop effective methods for the advanced development and introduction of new sequencing technologies into the sequencing process. As the scale of sequencing increases, the introduction of improvements into the production stream becomes more challenging and costly. New technology must therefore be robust and be carefully evaluated and validated in a high-throughput environment before its implementation in a production setting. A strong commitment from both the technology developers and the technology users is essential in this process. It must be recognized that the advanced development process will often require significantly more funds than proof-of-principle studies. Targeted funding allocations and dedicated review mechanisms are needed for advanced technology development.

Goal 3—Human Genome Sequence Variation

Natural sequence variation is a fundamental property of all genomes. Any two haploid human genomes show multiple sites and types of polymorphism. Some of these have functional implications, whereas many probably do not. The most common polymorphisms in the human genome are single base-pair differences, also called single-nucleotide polymorphisms (SNPs). When two haploid genomes are compared, SNPs occur every kilobase, on average. Other kinds of sequence variation, such as copy number changes, insertions, deletions, duplications, and rearrangements also exist, but at low frequency and their distribution is poorly understood. Basic information about the types, frequencies, and distribution of polymorphisms in the human genome and in human populations is critical for progress in human genetics. Better high-throughput methods for using such information in the study of human disease is also needed.

SNPs are abundant, stable, widely distributed across the genome, and lend themselves to automated analysis on a very large scale, for example, with DNA array technologies. Because of these properties, SNPs will be a boon for mapping complex traits such as cancer, diabetes, and mental illness. Dense maps of SNPs will make possible genome-wide association studies, which are a powerful method for identifying genes that make a small contribution to disease risk. In some instances, such maps will also permit prediction of individual differences in drug response. Publicly available maps of large numbers of SNPs distributed across the whole genome, together with technology for rapid, large-scale identification and scoring of SNPs, must be developed to facilitate this research. The early availability of a working draft of the human genome should greatly facilitate the creation of dense SNP maps (see Goal 1).

a) Develop technologies for rapid, large-scale identification or scoring, or both, of SNPs and other DNA sequence variants. The study of sequence variation requires efficient technologies that can be used on a large scale and that can accomplish one or more of the following tasks: rapid identification of many thousands of new SNPs in large numbers of samples; and rapid and efficient scoring of large numbers of samples for the presence or absence of already known

SNPs. Although the immediate emphasis is on SNPs, ultimately technologies that can be applied to polymorphisms of any type must be developed. Technologies are also needed that can rapidly compare, by large-scale identification of similarities and differences, the DNA of a species that is closely related to one whose DNA has already been sequenced. The technologies that are developed should be cost-effective and broadly accessible.

b) Identify common variants in the coding regions of the majority of identified genes during this 5-year period. Initially, association studies involving complex diseases will likely test a large series of candidate genes; eventually, sequences in all genes may be systematically tested. SNPs in coding sequences (also known as cSNPs) and the associated regulatory regions will be immediately useful as specific markers for disease. An effort should be made to identify such SNPs as soon as possible. Ultimately, a catalog of all common variants in all genes will be desirable. This should be cross-referenced with cDNA sequence data (see Goal 4).

c) Create an SNP map of at least 100,000 markers. A publicly available SNP map of sufficient density and informativeness to allow effective mapping in any population is the ultimate goal. A map of 100,000 SNPs (one SNP per 30,000 nucleotides) is likely to be sufficient for studies in some relatively homogeneous populations, while denser maps may be required for studies in large, heterogeneous populations. Thus, during this 5-year period, a map of at least 100,000 SNPs should be created. If technological advances permit, a map of greater density is desirable. Research should be initiated to estimate the number of SNPs needed in different populations.

d) Develop the intellectual foundations for studies of sequence variation. The methods and concepts developed for the study of single-gene disorders are not sufficient for the study of complex, multigene traits. The study of the relationship between human DNA sequence variation, phenotypic variation, and complex diseases depends critically on better methods. Effective research design and analysis of linkage, linkage disequilibrium, and association data are areas that need new insights. Questions such as which study designs are appropriate to which specific populations, and with which population genetics characteristics, must be answered. Appropriate statistical and computational tools and rigorous criteria for establishing and confirming associations must also be developed.

e) Create public resources of DNA samples and cell lines. To facilitate SNP discovery it is critical that common public resources of DNA samples and cell lines be made available as rapidly as possible. To maximize discovery of common variants in all human populations, a resource is needed that includes individuals whose ancestors derive from diverse geographic areas. It should encompass as much of the diversity found in the U.S. population as possible. Samples in this initial public repository should be totally anonymous to avoid concerns that arise with linked or identifiable samples.

DNA samples linked to phenotypic data and identified as to their geographic and other origins will be needed to allow studies of the frequency and distribution of DNA polymorphisms in specific populations and their relevance to disease. However, such collections raise many ethical, legal, and social concerns that must be addressed. Credible scientific strategies must be developed before creating these resources. (see Goal 6)

Goal 4—Technology for Functional Genomics

The HGP is revolutionizing the way biology and medicine will be explored in the next century and beyond. The availability of entire genome sequences is enabling a new approach to biology often called functional genomics—the interpretation of the function of DNA sequence on a genomic scale. Already, the availability of the sequence of entire organisms has demonstrated that many genes and other functional elements of the genome are discovered only when the full DNA sequence is known. Such discoveries will accelerate as sequence data accumulate. However, knowing the structure of a gene or other element is only part of the answer. The next step is to elucidate function, which results from the interaction of genomes with their environment. Current methods for studying DNA function on a genomic scale include comparison and analysis of sequence patterns directly to infer function, large-scale analysis of the messenger RNA and protein products of genes, and various approaches to gene disruption. In the future, a host of novel strategies will be needed for elucidating genomic function. This will be a challenge for all of biology. The HGP should contribute to this area by emphasizing the development of technology that can be used on a large scale, is efficient, and is capable of generating complete data for the genome as a whole. To the extent that available resources allow, expansion of current approaches as well as innovative technology ideas should be supported in the areas described below. Large-scale characterization of the gene transcripts and their protein products underpins functional analysis. Therefore, identifying and sequencing a set of full-length cDNAs that represent all human genes must be a high priority.

a) Develop cDNA resources. Complete sets of full-length cDNA clones and sequences for both humans and model organisms would be enormously useful for biologists and are urgently needed. Such resources would help in both gene discovery and functional analysis. Unfortunately, neither cloning full-length cDNAs nor identifying rare transcripts is yet a routine task. High priority should therefore be placed on developing technology for obtaining full-length cDNAs and for finding rare transcripts. Complete and validated inventories of full-length cDNA clones and corresponding sequences should be generated and made available to the community once such technology is at hand.

b) Support research on methods for studying functions of non–protein-coding sequences. In addition to the DNA sequences specifying protein structure, there are numerous sequences responsible for other functions, such as control of gene expression, RNA splicing, formation of chromatin domains, maintenance of chromosome structure, recombination, and replication. Other sequences specify the numerous functional untranslated RNAs. Improved technologies are needed for global approaches to the study of non–protein-coding sequences, including production of relevant libraries, comparative sequencing, and computational analysis.

c) Develop technology for comprehensive analysis of gene expression. Information about the spatial and temporal patterns of gene expression in both humans and model organisms offers one key to understanding gene expression. Efficient and cost-effective technology needs to be developed to measure various parameters of gene expression reliably and reproducibly. Complementary DNA sequences and validated sets of clones with unique identifiers will be needed for array technologies, large-scale in situ hybridization, and other strategies for measuring gene expression. Improved methods for quantifying, representing, analyzing, and archiving expression data should also be developed.

d) Improve methods for genome-wide mutagenesis. Creating mutations that cause loss or alteration of function is another prime approach to studying gene function. Technologies, both gene- and phenotype-based, which can be used on a large scale in vivo or in vitro, are needed for generating or finding such mutations in all genes. Such technologies should be piloted in appropriate model systems, including both cell culture and whole organisms.

e) Develop technology for global protein analysis. A full understanding of genome function requires an understanding of protein function on a genome-wide basis. Development of experimental and computational methods to study global spatial and temporal patterns of protein expression, protein-ligand interactions, and protein modification needs to be supported.

Goal 5—Comparative Genomics

Because all organisms are related through a common evolutionary tree, the study of one organism can provide valuable information

about others. Much of the power of molecular genetics arises from the ability to isolate and understand genes from one species based on knowledge about related genes in another species. Comparisons between genomes that are distantly related provide insight into the universality of biologic mechanisms and identify experimental models for studying complex processes. Comparisons between genomes that are closely related provide unique insights into the details of gene structure and function. In order to understand the human genome fully, genomic analysis on a variety of model organisms closely and distantly related to each other must be supported.

Genome sequencing of *E. coli* and *S. cerevisiae*, two of the five model organisms targeted in the first 5-year plan, has been completed. Availability of these sequences has led to the discovery of many new genes and other functional elements of the genome. It has allowed biologists to move from identifying genes to systematic studies to understand their function. Completion of the DNA sequence of the remaining model organisms, *C. elegans*, *D. melanogaster*, and mouse, continues to be a high priority and should proceed as rapidly as available resources allow. Additional model organisms will need to be analyzed to allow the full benefits of comparative genomics to be realized. This ongoing need is a major rationale for building sustainable sequencing capacity (see Goals 1 and 2).

a) Complete the sequence of the C. elegans *genome in 1998*. The DNA sequence of the *C. elegans* genome is well on the way to completion, with a target date of December 1998. Some difficult-to-close regions may remain at the end of this year and should become the subject of research projects aimed at closing them. The lessons learned from this project will be crucial in devising strategies for larger genomes.

b) Complete the sequence of the Drosophila *genome by 2002*. The wealth of information accumulated about *Drosophila* over many decades makes it a critically important genetic model. Its DNA sequence is eagerly awaited by all biologists. A significant increase in investment in *Drosophila* sequencing capacity will be needed to achieve this goal, and the benefits of early completion to comparative biology will be tremendous. Anticipated contributions from the private sector may enable the completion of this goal even earlier than 2002.

c) The mouse genome. The mouse is currently the best mammalian model for studies of a broad array of biomedical research questions. The complete mouse genome sequence will be a crucial tool for interpreting the human genome sequence, because it will highlight functional features that are conserved, including noncoding regulatory sequences as well as coding sequences. Comparisons between mouse and human genomes will also identify functionally important differences that distinguish mouse from human. Therefore, this is the time to invest in a variety of mouse genomic resources, culminating eventually in full-genome sequencing, to allow development of whole-genome approaches in a mammalian system.

1) Develop physical and genetic mapping resources. The integrated mouse yeast artificial chromosome (YAC)/STS map that has been developed provides a useful framework for the more detailed mapping resources now needed for positional cloning and sequencing projects. These resources should include mapped STSs, polymorphic markers, cDNA sequences, and BACs. The usefulness of SNPs as polymorphic markers in the mouse should also be explored in the near term.

2) Develop additional cDNA resources. More cDNA libraries and cDNA sequences are needed. These should derive from a variety of tissues and developmental stages and have good representation of rare transcripts. The mouse offers an opportunity to capture cDNA sequences from developmental stages, anatomical sites, and physiological states that are under-represented in human cDNA collections, and these should receive particular attention. Full-length cDNAs should be developed and sequenced once the technology for doing this efficiently becomes available (see also Goal 4).

3) Complete the sequence of the mouse genome by 2005. Mouse genomic sequence is an essential resource for interpreting human DNA sequence. For this reason, the centers sequencing human DNA are encouraged to devote up to 10% of their capacity to sequencing mouse DNA. Additional capacity for mouse DNA sequencing should be built up over the next few years with a goal of finishing the mouse sequence by 2005. Initially, a working draft of the mouse genome should be produced even sooner (see Goal 1 for a discussion of a working draft of the human genome sequence).

d) Identify other model organisms that can make major contributions to the understanding of the human genome and support appropriate genomic studies. As DNA sequencing capacity becomes available, new model organisms that can contribute to understanding human biology should be identified for genomic sequencing. Even if such capacity is not available during this 5-year period, development of other useful genomic resources should be considered. The scientific community will need to establish criteria for choosing those models that can make the greatest contribution. Characteristics such as phylogenetic distance from other models, genome size, transfection capability, ability to mutagenize, and availability of experimental material should all be considered. Because different characteristics will be useful for different purposes, organisms that are phylogenetically distant from each other and those that are close should be studied.

Goal 6—Ethical, Legal, and Social Implications (ELSI)

While recognizing that genetics is not the only factor affecting human well-being, the NIH and DOE are acutely aware that advances in the understanding of human genetics and genomics will have important implications for individuals and society. Examination of the ethical, legal, and social implications of genome research is, therefore, an integral and essential component of the HGP. In a unique partnership, biological and social scientists, health care professionals, historians, legal scholars, and others are committed to exploration of these issues as the project proceeds. The ELSI program has generated a substantial body of scholarship in the areas of privacy and fair use of genetic information, safe and effective integration of genetic information into clinical settings, ethical issues surrounding genetics re-
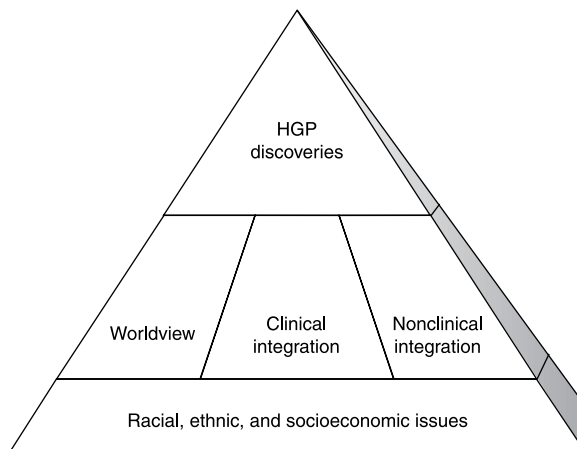
Fig. 3. The pyramid depicts the Ethical, Legal, and Social Implications (ELSI) Research Program goals for 1998–2003. The first goal, at the top of the pyramid, deals with the issues around the completion of the first human DNA sequence and the study of human genetic variation, making concrete the vision that advances in genome science will be an important factor contributing to the ELSI research agenda. The second and third goals focus on the integration of the information generated by these new discoveries into clinical, nonclinical, and research settings. The fourth goal examines the interaction of this information with philosophical, theological, and ethical perspectives. Finally, providing the foundation for all of these explorations is the fifth goal, examining how the understanding and use of genetic information are affected by socioeconomic factors and concepts of race and ethnicity.

search, and professional and public education. The results of this research are already being used to guide the conduct of genetic research and the development of related health professional and public policies. The ELSI program has also stimulated the examination of similar issues in other areas of the biological and medical sciences.

Continued success of the ELSI program will require attention to the new challenges presented by the rapid advances in genetics and its applications. As the genome project draws closer to completing the first human genome sequence and begins to explore human sequence variation on a large scale, it will be critical for biomedical scientists, ELSI researchers, and educators to focus attention on the ethical, legal, and social implications of these developments for individuals, families, and communities. The new goals for ELSI research and education can be visualized as a pyramid of interrelated issues and activities (Fig. 3). Given the complexity of the issues encompassed by the ELSI goals, only a summary of the major areas is presented here. To illustrate more fully the breadth and range of the issues that will be addressed, a Web site has been created that provides examples of the types of research questions and education activities envisioned within each goal (www.nhgri.nih.gov/98plan/elsi/).

The major ELSI goals for the next 5 years are:

a) Examine the issues surrounding the completion of the human DNA sequence and the study of human genetic variation.

b) Examine issues raised by the integration of genetic technologies and information into health care and public health activities.

c) Examine issues raised by the integration of knowledge about genomics and gene-environment interactions into nonclinical settings.

d) Explore ways in which new genetic knowledge may interact with a variety of philosophical, theological, and ethical perspectives.

e) Explore how socioeconomic factors and concepts of race and ethnicity influence the use, understanding, and interpretation of genetic information, the utilization of genetic services, and the development of policy.

Goal 7—Bioinformatics and Computational Biology

Bioinformatics support is essential to the implementation of genome projects and for public access to their output. Bioinformatics needs for the genome project fall into two broad areas: (i) databases and (ii) development of analytical tools. Collection, analysis, annotation, and storage of the ever increasing amounts of mapping, sequencing, and expression data in publicly accessible, user-friendly databases is critical to the project's success. In addition, the community needs computational methods that will allow scientists to extract, view, annotate, and analyze genomic information efficiently. Thus, the genome project must continue to invest substantially in these areas. Conservation of resources through development of portable software should be encouraged.

a) Improve content and utility of databases. Databases are the ultimate repository of HGP data. As new kinds of data are generated and new biological relationships discovered, databases must provide for continuous and rapid expansion and adaptation to the evolving needs of the scientific community. To encourage broad use, databases should be responsive to a diverse range of users with respect to data display, data deposition, data access, and data analysis. Databases should be structured to allow the queries of greatest interest to the community to be answered in a seamless way. Communication among databases must be improved. Achieving this will require standardization of nomenclature. A database of human genomic information, analogous to the model organism databases and including links to many types of phenotypic information, is needed.

b) Develop better tools for data generation, capture, and annotation. Large-scale, high-throughput genomics centers need readily available, transportable informatics tools for commonly performed tasks such as sample tracking, process management, map generation, sequence finishing, and primary annotation of data. Smaller users

urgently need reliable tools to meet their sequencing and sequence analysis needs. Readily accessible information about the availability and utility of various tools should be provided, as well as training in the use of tools.

c) Develop and improve tools and databases for comprehensive functional studies. Massive amounts of data on gene expression and function will be generated in the near future. Databases that can organize and display this data in useful ways need to be developed. New statistical and mathematical methods are needed for analysis and comparison of expression and function data, in a variety of cells and tissues, at various times and under different conditions. Also needed are tools for modeling complex networks and interactions.

d) Develop and improve tools for representing and analyzing sequence similarity and variation. The study of sequence similarity and variation within and among species will become an increasingly important approach to biological problems. There will be many forms of sequence variation, of which SNPs will be only one type. Tools need to be created for capturing, displaying, and analyzing information about sequence variation.

e) Create mechanisms to support effective approaches for producing robust, exportable software that can be widely shared. Many useful software products are being developed in both academia and industry that could be of great benefit to the community. However, these tools generally are not robust enough to make them easily exportable to another laboratory. Mechanisms are needed for supporting the validation and development of such tools into products that can be readily shared and for providing training in the use of these products. Participation by the private sector is strongly encouraged.

Goal 8—Training

The HGP has created the need for new kinds of scientific specialists who can be creative at the interface of biology and other disciplines, such as computer science, engineering, mathematics, physics, chemistry, and the social sciences. As the popularity of genomic research increases, the demand for these specialists greatly exceeds the supply. In the past, the genome project has benefited immensely from the talents of nonbiological scientists, and their participation in the future is likely to be even more crucial. There is an urgent need to train more scientists in interdisciplinary areas that can contribute to genomics. Programs must be developed that will encourage training of both biological and nonbiological scientists for careers in genomics. Especially critical is the shortage of individuals trained in bioinformatics. Also needed are scientists trained in the management skills required to lead large data-production efforts. Another urgent need is for scholars who are trained to undertake studies on the societal impact of genetic discoveries. Such scholars should be knowledgeable in both genome-related sciences and in the social sciences. Ultimately, a stable academic environment for genomic science must be created so that innovative research can be nurtured and training of new individuals can be assured. The latter is the responsibility of the academic sector, but funding agencies can encourage it through their grants programs.

a) Nurture the training of scientists skilled in genomics research. A number of approaches to training for genomics research should be explored. These include providing fellowship and career awards and encouraging the development of institutional training programs and curricula. Training that will facilitate collaboration among scientists from different disciplines, as well as courses that introduce scientists to new technologies or approaches, should also be included.

b) Encourage the establishment of academic career paths for genomic scientists. Ultimately, a strong academic presence for genomic science is needed to generate the training environment that will encourage individuals to enter the field. Currently, the high demand for genome scientists in industry threatens the retention of genome scientists in academia. Attractive incentives must be devel-

oped to maintain the critical mass essential for sponsoring the training of the next generation of genome scientists.

c) Increase the number of scholars who are knowledgeable in both genomic and genetic sciences and in ethics, law, or the social sciences. As the pace of genetic discoveries increases, the need for individuals who have the necessary training to study the social impact of these discoveries also increases. The ELSI program should expand its efforts to provide postdoctoral and senior fellowship opportunities for cross-training. Such opportunities should be provided both to scientists and health professionals who wish to obtain training in the social sciences and humanities and to scholars trained in law, the social sciences, or the humanities who wish to obtain training in genomic or genetic sciences.

References and Notes

1. F. Collins and D. Galas, *Science* **262**, 43 (1993).
2. J. D. Watson and F. H. C. Crick, *Nature* **171**, 737 (1953).
3. The finished genome sequence refers to the portion of human DNA that can be stably cloned and sequenced by current technology. The small proportion of highly repeated sequence represented by the centromeres and other constitutive heterochromatic regions of the genome may not be finished by 2003. In addition, it is possible that a small fraction of other parts of the genome may present unanticipated and serious challenges. Such regions are expected to be rare.
4. J. C. Venter *et al.*, *Science* **280**, 1540 (1998). A whole-genome shotgun strategy has been proposed previously [J. Weber and E. W. Myers, *Genome Res.* **7**, 401, but major concerns have been raised (P. Green, *ibid.*, p. 410), about the difficulties expected in obtaining correct long-range contig assemblies. It will not be possible to evaluate the feasibility, impact, or quality of the product of this approach until more data are available, which is not estimated to occur for about 12 to 18 months. See also R. Waterston and J. E. Sulston, *Science* **287**, 53 (1998).
5. D. R. Bentley, *Science* **274**, 533 (1996).
6. M. Guyer *et al.*, *Genome Res.* **8**, 413 (1998).
7. The members of the planning groups are: NIH scientific planning subcommittee–Aravinda Chakravarti, chair, Eric Fearon, Lee Hartwell, Charles H. Langley, Richard A. Mathies, Maynard Olson, Anthony J. Pawson, Thomas Pollard, Alan Williamson, Barbara Wold; DOE planning subcommittee–Raymond Gesteland, chair, Kenneth Buetow, Elbert Branscomb, Mario Capecchi, George Church, Harold Garner, Richard A. Gibbs, Trevor Hawkins, Keith Hodgson, Michael Knotek, Miriam Meisler, Gerald M. Rubin, Lloyd M. Smith, Randall F. Smith, Monty Westerfield; NIH-DOE ELSI Research Planning and Evaluation Group–LeRoy Walters, chair, Ellen Wright Clayton, Nancy L. Fisher, Caryn E. Lerman, Joseph D. McInerney, William Nebo, Nancy Press, David Valle. The members of these groups have contributed substantially to this plan but have not necessarily approved every word in it. In addition to the authors, this plan represents the combined efforts of a large number of consultants, including all the participants in the numerous workshops that were conducted as part of the planning process. We are deeply indebted to them for the time and expertise they contributed so generously to this effort.

Editor's Summary

Article Tools   Visit the online version of this article to access the personalization and article
tools:
http://science.sciencemag.org/content/282/5389/682

Permissions   Obtain information about reproducing this article:
http://www.sciencemag.org/about/permissions.dtl