Arabidopsis thaliana: A Model Plant for Genome Analysis

David W. Meinke, J. Michael Cherry,* Caroline Dean, Steven D. Rounsley, Maarten Koornneef

REVIEW

Arabidopsis thaliana is a small plant in the mustard family that has become the model system of choice for research in plant biology. Significant advances in understanding plant growth and development have been made by focusing on the molecular genetics of this simple angiosperm. The 120-megabase genome of *Arabidopsis* is organized into five chromosomes and contains an estimated 20,000 genes. More than 30 megabases of annotated genomic sequence has already been deposited in GenBank by a consortium of laboratories in Europe, Japan, and the United States. The entire genome is scheduled to be sequenced by the end of the year 2000. Reaching this milestone should enhance the value of *Arabidopsis* as a model for plant biology and the analysis of complex organisms in general.

Arabidopsis thaliana has recently become the organism of choice for a wide range of studies in plant sciences (*1*). The current visibility of *Arabidopsis* research reflects the growing realization among biologists that this simple angiosperm can serve as a convenient model not only for plant biology but also for addressing fundamental questions of biological structure and function common to all eukaryotes. While genome projects have documented the extent to which all eukaryotic organisms share a common genetic ancestry, research with *Arabidopsis* has clarified the important role that analysis of plant genomes can play in understanding basic principles of biology relevant to a variety of species, including humans. The emergence of a large, multinational research community devoted to the complete analysis of a single plant represents a dramatic paradigm shift for plant biology. Traditionally, advances in our understanding of plant structure and function were built on research with a wide range of species, particularly those relevant to agriculture. Although an impressive amount of information was collected with this approach, advances in many disciplines were limited by scattered community resources, duplication of effort, and limited funding. Several plants were recognized as model genetic systems, including maize, tomato, pea, rice, barley, petunia, and snapdragon, but research biologists failed to reach a consensus on which species was most suitable for studying processes common to all plants. As a result, our understanding of fundamental aspects of plant growth and development such as flowering, root growth, hormone action, and responses to environmental signals remained limited.

Twenty years ago, plant biologists began to search for another model organism suitable for detailed analysis using the combined tools of genetics and molecular biology. Plants with effective protocols for regeneration in culture (such as petunia and tomato) were logical candidates, particularly for studies involving *Agrobacterium*-mediated cell transformation, but attention gradually shifted toward *Arabidopsis*, a small weed in the mustard family that was first chosen as a model genetic organism by Laibach in Europe and later studied in detail by Rédei in the United States (*2*). The shift toward *Arabidopsis* gained momentum in the early 1980s with the release of a detailed genetic map (*3*) and publications outlining the value of *Arabidopsis* for research in plant physiology, biochemistry, and development (*4*). This was followed by two significant advances, the establishment of transformation protocols (*5*) and the demonstration that *Arabidopsis* had a small genome amenable to detailed molecular analysis (*6*).

The modern era of *Arabidopsis* research began in 1987 with the opening of the Third International *Arabidopsis* Conference at Michigan State University and the subsequent formation of an electronic *Arabidopsis* newsgroup. Many individuals experienced in the analysis of other model organisms soon began to study *Arabidopsis* as a promising model for basic research. One important outgrowth of this increased enthusiasm for *Arabidopsis* research was the drafting in 1990 of a vision statement outlining long-term research goals for the *Arabidopsis* community. These included saturating the genome with mutations, identifying every essential gene, and sequencing the entire genome by the end of the decade. The importance of applying advances with *Arabidopsis* to other plants and to solving practical problems in agriculture, industry, and human health was also stressed. A further commitment to *Arabidopsis* research was made in 1996 with the establishment of the *Arabidopsis* Genome Initiative dedicated to coordinating large-scale sequencing efforts. This initiative has become a model for multinational cooperation and has already resulted in more than 30 Mb of genomic DNA sequence being deposited in public databases. The remainder of the 120-Mb genome is scheduled to be sequenced by the end of 2000. *Arabidopsis* has therefore progressed in 20 years from an obscure weed to a respected member of the "Security Council of Model Genetic Organisms" (*7*). Here we review some recent advances in *Arabidopsis* research and summarize features that have made this simple angiosperm a model for research in plant biology.

Biology of *Arabidopsis*

Arabidopsis thaliana (Fig. 1) is a member of the mustard family (Cruciferae or Brassicaceae) with a broad natural distribution throughout Europe, Asia, and North America [see (*1*) for detailed reviews]. Many different ecotypes (accessions) have been collected from natural populations and are available for experimental analysis. The Columbia and Landsberg ecotypes are the accepted standards for genetic and molecular studies. The entire life cycle, including seed germination, formation of a rosette plant, bolting of the main stem, flowering, and maturation of the first seeds, is completed in 6 weeks. When it comes to size, almost everything about *Arabidopsis* is small. Flowers are 2 mm long, self-pollinate as the bud opens, and can be crossed by applying pollen to the stigma surface. Seeds are 0.5 mm in length at maturity and are produced in slender fruits known as siliques. Seedlings develop into rosette plants that range from 2 to 10 cm in diameter, depending on growth conditions. Leaves are covered with small unicellular hairs known as trichomes that are convenient models for studying morphogenesis and cellular differentiation.

Plants can be grown in petri plates or maintained in pots located

D. W. Meinke is in the Department of Botany, Oklahoma State University, Stillwater, OK 74078, USA. J. M. Cherry is in the Department of Genetics, Stanford University School of Medicine, Stanford, CA 94305, USA. C. Dean is in the Department of Molecular Genetics, John Innes Centre, Norwich, NR4 7UH, UK. S. D. Rounsley is at The Institute for Genomic Research, Rockville, MD 20850, USA. M. Koornneef is at the Laboratory of Genetics, Wageningen Agricultural University, Wageningen, 6307 HA, Netherlands.

*To whom correspondence should be addressed. E-mail: cherry@genome.stanford.edu

either in a greenhouse or under fluorescent lights in the laboratory. Bolting starts about 3 weeks after planting, and the resulting inflorescence forms a linear progression of flowers and siliques for several weeks before the onset of senescence. Flowers are composed of an outer whorl of four green sepals and inner whorls containing four white petals, six stamens bearing pollen, and a central gynoecium that forms the silique. Mature plants reach 15 to 20 cm in height and often produce several hundred siliques with more than 5000 total seeds. The roots are simple in structure, easy to study in culture, and do not establish symbiotic relationships with nitrogen-fixing bacteria. Natural pathogens include a variety of insects, bacteria, fungi, and viruses.

Genetic Analysis

The *Arabidopsis* research community has developed most of the methods and resource materials expected of a model genetic organism. These include simple procedures for chemical and insertional mutagenesis, efficient methods for performing crosses and introducing DNA through plant transformation, extensive collections of mutants with diverse phenotypes, and a variety of chromosome maps of mutant genes and molecular markers (*8*). The absence of an efficient system for gene replacement through homologous recombination is a limitation shared by other model organisms such as *Drosophila* and *Caenorhabditis elegans*. Promising advances in this important area of *Arabidopsis* research have nevertheless been reported (*9*). Mature seeds are the preferred targets for chemical mutagenesis because millions of progeny seeds homozygous for recessive mutations can be produced by selfing M_1 plants derived from a single experiment. Insertional mutagenesis with transferred DNA (T-DNA) from *Agrobacterium tumefaciens* has become routine through development of whole-plant transformation methods (*10*) that avoid the pitfalls associated with plant regeneration in culture. Thousands of transgenic lines carrying random T-DNA insertions throughout the genome have been deposited in public stock centers. Many additional lines are being produced at private companies interested in functional genomics. Maize transposable elements introduced through *Agrobacterium*-mediated transformation have also been used extensively for gene disruption (*11*).

Several thousand mutants of *Arabidopsis* defective in almost every aspect of plant growth and development have been identified over the past 20 years. The ability to save genetic stocks as seeds has minimized the effort required to maintain these mutants over long periods of time. Mutations that interfere with gametogenesis, seed formation, leaf and root development, flowering, senescence, metabolic and signal transduction pathways, responses to hormones, pathogens, and environmental signals, and many cellular and physiological processes have been identified (*1*). Because mapping and allelism tests have often lagged behind mutant identification, a number of mutants currently being studied in different laboratories are likely to be defective in the same gene. Progress has nevertheless been made toward establishing community standards for gene nomenclature and mutant analysis to minimize duplication of effort (*12*).

The *Arabidopsis* genome is organized into five chromosomes and contains an estimated 20,000 genes. The small size of meiotic chromosomes and the absence of polytene chromosomes have limited cytogenetic studies of chromosome structure, although visualization has improved in recent years with in situ hybridization methods (*13*). Three related maps of each chromosome (classical genetic, recombinant inbred, and physical) are presented on the wall chart included with this genome issue. The classical map shows estimated locations of mutant genes based on recombination frequencies. The original map was produced by analyzing segregating phenotypes in the F_2 generation after self-pollination of F_1 plants. More than 460 mutant genes are included on the current map, which is available through the Internet at http://mutant.lse.okstate.edu. The precise order and distances between many linked genes remain to be determined because map locations are based largely on two-point recombination data. One striking feature of the classical map is the large number of cloned mutant genes included (more than 110 at present). These genes are noted in orange (mapped relative to phenotypic markers) and green (mapped relative to molecular markers) on the attached chart. The recombinant inbred (RI) map illustrates locations of cloned genes and molecular markers based on recombination within a defined mapping population produced through repeated selfing of progeny plants in successive generations (*14*). Markers on this map include restriction fragment length polymorphisms (RFLPs), simple sequence length polymorphisms (SSLPs), cleaved amplified polymorphic sequences (CAPSs), and a variety of cloned genes, expressed sequence tags (ESTs), and the ends of bacterial (BAC) and yeast (YAC) artificial chromosomes. More than 790 markers are included on the current RI map, which can be viewed at http://nasc.nott.ac.uk/new_ri_map.html. The length of each RI chromosome has been adjusted on the chart to match that of the classical chromosome. This facilitates comparison between equivalent regions and

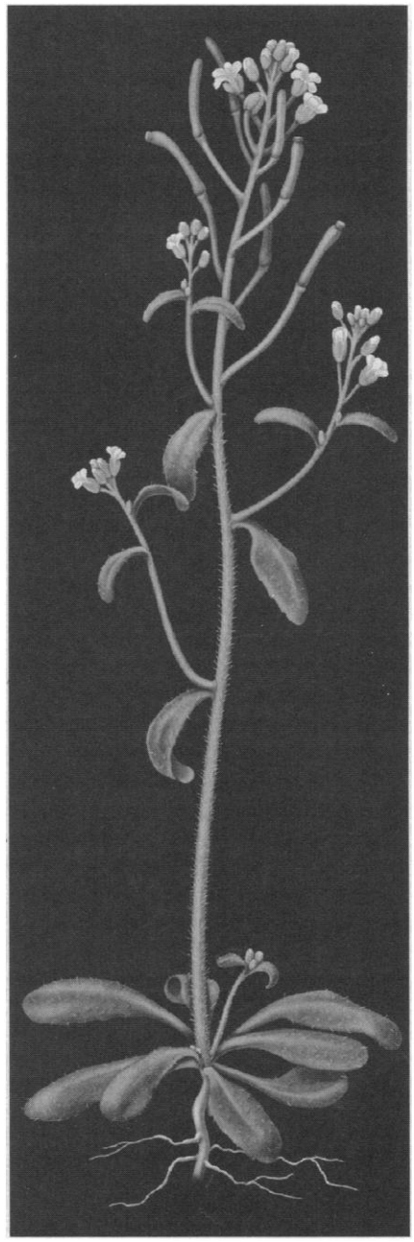

Fig. 1. *Arabidopsis thaliana* at an early stage of flowering. [Drawing by K. Sutliff]

emphasizes the fact that genetic distances between molecular markers on the RI map will eventually become secondary to physical distances measured in base pairs. Mutant genes noted in green and purple on the classical map were first assigned a chromosome position based on recombination frequencies with molecular markers located on the RI map. Updated information on physical maps of the five *Arabidopsis* chromosomes can be found at http://genome-www.stanford.edu/Arabidopsis/.

In addition to mutagenesis and mapping efforts, genetic analysis of *Arabidopsis* has expanded in recent years to include specialized topics of broad interest such as epigenetics, gene silencing, tetrad analysis, centromere mapping, and reverse genetics. The history of maize genetics is filled with elegant studies of epigenetics and paramutation. Research with *Arabidopsis* has offered molecular details on some of the genes involved within a functional genomics context (*15*). Tetrad analysis became possible in *Arabidopsis* with the isolation of the *quartet* mutant in which four pollen grains derived from a single meiotic event remain attached when released from the anther but nevertheless participate in fertilization (*16*). The precise number of insertional mutants available in *Arabidopsis* is difficult to determine because some collections are available through public stock centers whereas others are being produced in the private sector. However, plans are under way to improve community access to insertional mutants and to make it possible to obtain a knockout of virtually any gene of interest with only minimal effort (*17*). Thus, with continued advances in mutant analysis, genome sequencing, and production of knockouts, *Arabidopsis* may soon become the higher eukaryote of choice for studying many fundamental concepts of modern genetics.

Research Community

The *Arabidopsis* community is a diverse group of scientists representing more than 30 different countries. Almost every major university, research institute, and private company active in plant research has at least one individual working on *Arabidopsis*. This wide involvement, reflected in increased attendance at annual *Arabidopsis* meetings, attracted more than 900 participants to the summer 1998 meeting held in Madison, Wisconsin. Community resources include a centralized database, two stock centers, established EST projects, and several large-scale sequencing laboratories associated with the *Arabidopsis* Genome Initiative (AGI) (Table 1). Rapid communication on scientific matters is facilitated through broad participation in the electronic *Arabidopsis* newsgroup (*18*). For the past 6 years, annual progress toward goals set forth in the Multinational Coordinated *Arabidopsis thaliana* Genome Research Project has been summarized in a document published by the U.S. National Science Foundation (NSF) (*19*).

Community decisions are coordinated by two representative groups: the multinational science steering committee and the North American steering committee. Advanced courses and workshops such as those offered by Cold Spring Harbor Laboratory and the European Molecular Biology Organization have played an important role in training an entire generation of *Arabidopsis* biologists. The contributions of Asian scientists to *Arabidopsis* research have become increasingly apparent at recent meetings, particularly the Fifth International Congress of Plant Molecular Biology held in Singapore. Funding agencies have also played a significant role in supporting and promoting *Arabidopsis* research. Significant investments in basic plant research have been made throughout Europe, Japan, Australia, and the United States, where the NSF continues to play a leadership role in funding sequencing efforts and a wide range of individual investigator awards.

Genome Sequencing Initiative

The AGI was established in 1996 to facilitate coordinated sequencing of the *Arabidopsis* genome (*20*). This initiative followed advances in EST sequencing projects (*21*), construction of standardized YAC and BAC libraries (*22*), establishment of physical maps for limited regions of the genome (*23*), and molecular analysis of many individual genes. AGI participants from Europe, Japan, and the United States agreed on a strategy that combined BAC end sequencing, fingerprinting, hybridization with anchored YACs and molecular markers, and starting points spread across the genome to begin sequencing contiguous clusters of BACs with minimal overlaps. The Japanese group proceeded with sequencing P1 artificial chromosome (PAC) clones because they had already invested in this approach. Each group was assigned a chromosomal region to begin sequencing with the understanding that assignments could be adjusted later to reflect progress and availability of funding. Updated information on sequencing efforts can be obtained from the Internet addresses listed in Table 1.

By 1 July 1998, the total amount of random BAC end sequence generated by the TIGR, SPP, and Genoscope groups was 13.6 Mb from 18,746 clones. By this same date, the entire AGI consortium had deposited in GenBank another 28 Mb of annotated genomic sequence from defined chromosomal regions. This included 4 to 5 Mb each from chromosomes 1 and 2, 9 to 10 Mb each from chromosomes 4 and 5, and less than 0.5 Mb from chromosome 3. Analysis of a contiguous 1.9-Mb region of chromosome 4 was recently published by the ESSA group (*24*) and several sequenced regions on chromosome 5 have been published by the Kazusa group in Japan (*25*). In addition, the CSHL consortium has made available extensive fingerprinting data and initial results of organizing BAC clones into a genome-wide contig map. Approximately 70 Mb of the genome was contained in 66 BAC contigs by 1 July, and plans were under way to complete the analysis of all 22,000 BAC clones by the end of the year. The combined availability of BAC end sequences and a genome-wide contig map should have an immediate impact on *Arabidopsis* research, particularly in the widespread use of chromosome walk-

Table 1. Community resources for *Arabidopsis* genome analysis.

Resource available	Contact person	Internet address for information
Arabidopsis database (AtDB)	M. Cherry	http://genome-www.stanford.edu/Arabidopsis/
ABRC* Stock Center (USA)	R. Scholl	http://aims.cps.msu.edu/aims
NASC† Stock Centre (UK)	M. Anderson	http://nasc.nott.ac.uk
AGI Sequencing Laboratories:		
TIGR‡ (USA)	S. Rounsley	http://www.tigr.org/tdb/at/at.html
SPP§ Consortium (USA)	R. Davis	http://sequence-www.stanford.edu/ara/SPP.html
CSHL‖ Consortium (USA)	R. McCombie	http://nucleus.cshl.org/protarab/
ESSA¶ Consortium (Europe)	M. Bevan	http://muntjac.mips.biochem.mpg.de/arabi/index.html
Genoscope (France)	F. Quetier	http://www.genoscope.cns.fr/externe/arabidopsis/Arabidopsis.html
Kazusa Institute (Japan)	S. Tabata	http://www.kazusa.or.jp/arabi/

**Arabidopsis* Biological Resource Center, Ohio State University. The ABRC database (AIMS) is maintained at Michigan State University. †Nottingham *Arabidopsis* Stock Centre, University of Nottingham. ‡The Institute for Genomic Research. §Stanford University (N. Federspiel), University of Pennsylvania (J. Ecker), USDA Plant Gene Expression Center (S. Theologis). ‖Cold Spring Harbor Laboratory (R. McCombie), Washington University (R. Wilson), Perkin-Elmer–Applied Biosystems (E. Chen). ¶European Scientists Sequencing *Arabidopsis*. Informatics for ESSA is provided by the Munich Information Center for Protein Sequences (MIPS).

ing to clone genes identified by mutation.

Representatives from AGI sequencing laboratories met again later in July to discuss strategies for completing the genome in 2.5 years, several years ahead of the schedule established in 1996. This accelerated timetable was made possible in part by additional funding from the European Commission and from the NSF Plant Genome Research Program. Participants agreed that completion of a given chromosome would be defined as the full sequence of each arm as a single contig from subtelomeric repeat to centromeric tandem repeats, with acceptable gaps defined as internal tandem repeat regions (including ribosomal DNA) of known length. It became apparent during the meeting that experience gained from completing the 100-Mb genome of *C. elegans* will be helpful in finishing the *Arabidopsis* project and that lessons learned with *Arabidopsis* could be applied to the analysis of more complex genomes in the future. Improved technology such as the automated template procedures developed by the SPP consortium for use with *Arabidopsis* may also find broad application in future genome projects.

The AGI and EST projects described above have provided a wealth of information on gene identity and genome organization in plants. The *Arabidopsis* genome is highly enriched for coding sequences, with one gene every 5 kb on average (*24*). About half of these genes appear to be closely related in sequence to genes found in other organisms ranging from bacteria to humans. In striking contrast to maize, where repetitive DNA rich in transposons constitutes a large percentage of the genome, *Arabidopsis* has a relatively small amount of interspersed repetitive DNA. Sequencing the *Arabidopsis* genome has therefore proven to be a cost-effective method of identifying every gene in a representative flowering plant.

Examples of Research Advances

Research with *Arabidopsis* has provided valuable insights into all aspects of modern biology. In some cases, long-standing questions in plant physiology and biochemistry were first resolved through genetic and molecular analysis of *Arabidopsis* mutants. For example, elucidation of ethylene signal transduction pathways in *Arabidopsis* provided the first unequivocal identification of a hormone receptor in plants (*26*). The developmental significance of another class of plant hormones, the brassinosteroids, was revealed by analyzing *Arabidopsis* mutants defective in brassinosteroid synthesis (*27*). This work had the additional benefit of providing insights into the biochemistry of related steroids important for human health. In the area of light perception, mutant analysis with *Arabidopsis* has led to the identification of plant receptors and signal transduction components for phototropism (*28*) and circadian rhythms (*29*) in addition to advancing our understanding of phytochrome action (*30*). Several genes that regulate the transition to flowering have been identified (*31*) and elegant models constructed for the genetic control of pattern formation during floral development (*32*). Advances in biochemistry and cell biology have covered topics ranging from ion transport and fatty acid biosynthesis to cell wall formation and chloroplast maintenance (*1*).

Some research with *Arabidopsis* has provided unexpected insights into cellular mechanisms shared with other organisms. For example, a protein complex initially identified through genetic analysis of the *constitutive photomorphogenic* class of *Arabidopsis* mutants has been found throughout eukaryotes and may provide clues to complex signal transduction networks active in humans (*33*). A retinal photoreceptor that may serve to entrain the circadian clock in mammals was recently identified on the basis of, in part, similarity to the *CRY2* photoreceptor of *Arabidopsis* (*34*). Plant biologists have long realized that cellular mechanisms common to eukaryotes are often characterized first in yeast or animal systems and then later extended to plants. The advent of *Arabidopsis* functional genomics and the availability of large numbers of *Arabidopsis* mutants defective in known gene products provides a unique opportunity for plant biologists to contribute to research efforts in a variety of related disciplines. As a result, it will become increasingly important for those studying other groups of organisms to keep abreast of continuing advances in plant biology.

Many biotechnology companies are counting on *Arabidopsis* research to help solve practical problems related to agriculture, energy, and the environment. Significant advances have already been reported in applied research efforts including molecular cloning of disease resistance genes (*35*), engineering of plants resistant to cold temperatures (*36*), production of specialized hydrocarbons (*37*), and stimulation of premature flowering in trees and other plants with extended life cycles (*38*). If patent applications are any indication of the practical benefits of *Arabidopsis* research, then the economic value of this simple weed has already been demonstrated (*39*). One of the original ideas behind using *Arabidopsis* as a model system was to facilitate the identification of related genes of importance in crop plants. At the moment there is every indication that this strategy is working as planned.

Vision for the Future

A new vision statement for the future of *Arabidopsis* research was recently articulated in the annual report for the Multinational *Arabidopsis* Genome Project (*19*). The short-term goals were to complete the genomic sequence and screens for informative mutations, obtain insertional knockouts of every major class of gene, continue detailed characterization of cellular, physiological, and developmental pathways, continue the widespread use of *Arabidopsis* as a model to study basic principles of genetics, establish improved computing systems to organize information on cellular processes involved in plant growth and development, and make advances obtained through the *Arabidopsis* genome project available to those working on other projects. Technological innovations such as the use of DNA chips and microarrays to study global patterns of gene expression (*40*) should play an important role in *Arabidopsis* research during this period. The more long-term goals are to determine the functions and locations of key gene products identified through large-scale sequencing efforts, uncover mechanisms by which complex networks of gene products become established and localized, combine information on gene products with advances in plant physiology and biochemistry to establish a comprehensive picture of plant structure and function, and use *Arabidopsis* to resolve questions concerning evolutionary relationships among eukaryotic organisms and the evolution of common cellular and developmental pathways (*19*).

Meeting these goals will place increasing demands on the development of databases designed to present massive amounts of information to *Arabidopsis* experts and the diverse audience of biologists. Representatives of the *Arabidopsis* and informatics communities met in the summer of 1998 to discuss options for designing and supporting a new generation of databases for *Arabidopsis* in particular and plant biology in general. Although a number of problems remain to be addressed, there was agreement that developing innovative methods for providing access to information represents one of the principal long-term challenges of the *Arabidopsis* genome project. With continued progress in genomics, biology, and database management, it nevertheless appears likely that *Arabidopsis* will soon become a model not only for understanding plant structure and function, but also for addressing more universal questions concerning the nature and origin of biological complexity.

References and Notes

1. E. M. Meyerowitz and C. R. Somerville, *Arabidopsis* (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, 1994).
2. F. Laibach, *Bot. Arch.* **44**, 439 (1943); G. P. Rédei, *Bibliogr. Genet.* **20**, 1 (1970).
3. M. Koornneef *et al.*, *J. Hered.* **74**, 265 (1983).
4. D. W. Meinke and I. M. Sussex, *Dev. Biol.* **72**, 50 (1979); C. R. Somerville and W. L. Ogren, *Nature* **280**, 833 (1979); M. Koornneef, E. Rolff, C. J. P. Spruit, *Z. Pflanzenphysiol.* **106**, 147 (1980).

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. S. 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