Multinational Coordinated *Arabidopsis thaliana* Genome Research Project Progress Report: Year Seven and Eight December, 1998

The Multinational Science Steering Committee:

Committee Chair: Gerd Jürgens, University of Tübingen, Germany
Michael Bevan, John Innes Centre, Norwich, United Kingdom
Michel Caboche, Lab. Biol. Cellulaire, INRA, Versailles, France
Daphne Preuss, University of Chicago, Chicago, IL, USA
Joseph Ecker, University of Pennsylvania, Philadelphia, PA USA
Fernando Migliaccio, CNR, Monterotondo, Italy
Kiyotaka Okada, Kyoto University, Kyoto, Japan
David Smyth, Monash University, Clayton, Australia
Marc Van Montagu, University of Ghent, Belgium

Table of Contents

Preface
Overview of Genome Analysis
Stock Center Resources and Data Bases
National and Transnational Projects
Appendix 1: NSF Arabidopsis Genome Meeting Report
Appendix 2: Summary of December 1998 AGI Meeting at CSHL
Appendix 3: Database Workshop (Madison, WI 1998) Report
Appendix 4: "Arabidopsis thaliana Information Resource Project" Announcment

Preface

The "Multinational Coordinated Arabidopsis thaliana Genome Research Project" was established in 1990 to promote international cooperation in basic and applied research with Arabidopsis, a model plant species amenable to experimental manipulation in the laboratory. The primary objective of this project has been to understand the molecular basis of plant growth and development and to address fundamental questions in plant genetics, physiology, biochemistry, cell biology, and pathology. Initial plans were outlined in a publication (NSF #90-80) drafted nine years ago by an ad hoc committee of nine scientists from the United States, Europe, Japan, and Australia. In recent years, this project has become a model for widespread participation and effective coordination of multinational research efforts in modern biology.

Arabidopsis thaliana, a small plant in the mustard family, was chosen for this large-scale research effort because it offers many advantages for detailed genetic and molecular studies. Among these features are its small size, short life cycle, small genome, ability to be transformed, availability of numerous mutations, and prolific seed production. By concentrating research efforts on a single model organism, detailed information on specific genes and cellular processes can be readily obtained and rapidly applied to a wide range of plants relevant to agriculture, health, energy, manufacturing, and the environment.

Each year since 1990, the scientific steering committee for the Arabidopsis Genome Project has prepared a progress report summarizing recent advances in Arabidopsis research. This is the seventh annual progress report published by the steering committee in conjunction with the U.S. National Science Foundation. Three years ago the report was a color brochure designed to explain the value and significance of Arabidopsis research to a wide audience. Two years ago the report presented a detailed overview of recent advances in research with Arabidopsis, along with technical information for use by members of the Arabidopsis community. The sixth report presented an updated vision statement for the future to stimulate further advances in the use of Arabidopsis as a model system for the analysis of complex organisms.

This report covers progress for the seventh and eighth years of the project. It is focused on the large-scale analysis of the Arabidopsis genome. Specifically, this report is designed to make the available information accessible to the scientific community in a hands-on format. At the current rate of progress, the genome sequencing project can be expected to be completed within two years. The 1998 genome issue of Science (Meinke et al. 1998) featured Arabidopsis prominently.

Multinational cooperation and communication continue to be an important feature of the Arabidopsis genome project. A brief overview of Arabidopsis research efforts in a number of participating countries is therefore included in this report. Additional information can be obtained through recent publications, electronic news groups and databases, and biological resource centers devoted to Arabidopsis research. As with any document that attempts to summarize the contributions of many individuals, this report may fail to include or misrepresent some significant achievements. The steering committee hopes that members of the Arabidopsis community will overlook such shortcomings and will communicate any concerns to committee members so that future reports will be as accurate as possible. We thank all members of the Arabidopsis community for their many contributions to the success of the initial phase of the Multinational Coordinated Arabidopsis thaliana Genome Research Project.

Overview of Genome Analysis

A Historical Perspective

1983	Publication of first genetic map
1988-89	Publication of RFLP maps
1990	Multinational Coordinated Arabidopsis thaliana Genome Research Project initiated

1991	Arabidopsis Stock Centers at Ohio State (USA) and Nottingham (UK), as well as the Arabidopsis Data Base (TAIR), were established
1991	First YAC libraries and anchoring of YAC clones to RFLP map
1992	Publication of first chromosome walk (local contig)
1993	Recombinant inbred (RI) map
1994-8	Collections of cDNA (EST) clones sequenced linking up genetic and cytogenetic with physical maps
1995-6	CIC-YACs, TAMU-BACs, IGF-BACs, Mitsui-P1, Kazusa-P1 libraries
1995-8	Physical map of all 5 chromosomes delineated
Jan 98	Publication of 1.9 Mb of contiguous DNA sequence from chromosome 4
June 98	29 Mb of genomic DNA sequenced
Oct. 98	*Arabidopsis* featured in genome issue of "Science"
Dec 98	>46 Mb of genomic DNA sequenced and annotated 90 Mb of genomic DNA in edited BAC contigs >41,000 (of 44,000) BAC ends sequenced >11,000 non-redundant (of >37,000) EST clones
2000	Completion of genome sequencing (expected date)

Integration of Genetic and Physical Maps

Two genetic maps were independently developed: a classic map of mutations (Koornneef et al., 1983) and a recombinant inbred (RI) map of molecular markers (Lister and Dean, 1993). As an increasing number of genes originally identified by mutation has been cloned and converted to molecular markers mapped onto the RI map, the two maps are beginning to merge into a unified genetic map. Map distances differ between the two maps, presumably because of the different genetic backgrounds. In addition, map distances are calculated with the Mapmaker program, resulting in local inaccuracies, such as relative order of closely linked markers. These problems will eventually be resolved by physical mapping.

The RI map is now commonly used as the standard reference, enabling new genes identified by mutation to be easily mapped by PCR markers (SSLP, CAPS). The current RI map (November 1998) contains ca. 800 markers which fall into 3 different categories: "framework" (fixed reference location), "unique" (defined location on the map) and "multiple" (several possible locations). RI markers were also used to map a collection of YAC, BAC and P1 clones from which physical maps of the 5 chromosomes were initiated, thus linking genetic and physical maps from the very beginning.

Several physical maps have been established for all 5 chromosomes. Initially, contigs of large YAC clones were assembled and anchored to RI markers (e.g. Schmidt et al., 1997; Bouchez et al., 1998). Corresponding BAC and P1 clones were identified by hybridisation with YAC clones**.** For chromosome 5, a nearly complete physical map was established by P1 and TAC clone contigs (Kazusa homepage; Kotani et al., 1997). BAC contigs have also been established at the global scale by fingerprinting and by hybridisation with BAC endprobes. For example, 9 Mb constituting the bottom arm of chromosome 3 have been covered by a single BAC contig (see http://www.genoscope.cns.fr/externe/English/Projets/projetsindex.html). In addition to whole-chromosome physical mapping with YAC, BAC and P1 clones, chromosome walks in several chromosome regions have yielded local contigs up to 2 Mb long (e.g. Hardtke & Berleth, 1996; Wang et al., 1997; Thorlby et al., 1997), and several hundred EST clones have been PCR-mapped onto YAC clones (Agyare et al., 1997).

Fingerprinting data of BAC clones were used to assemble contigs with FPC software, followed by manual editing to join the initial contigs. At present, ca. 70 BAC contigs encompass ca. 90 Mb of estimated 121 Mb total sequence (M. Marra & M. Sekhon, Washington University, St. Louis; M.A. Marra et al.,1997). High throughput BAC-endprobe hybridization was used as a complementary approach to assemble contigs (Mozo et al., 1998). Information gathered from 2995 hybridization data (including 272 mapped markers) was manually edited after application of the *probeorder* computer program and integrated with the fingerprint data to generate a complete BAC-based physical map consisting of 27 contigs distributed over the 10 chromosome arms that covers approximately 124 Mb (see: http://www.mpimp-golm.mpg.de/101/bac.html). As the genome sequencing project is progressing, many RI markers are mapped physically, resulting in an excellent alignment of genetic and physical maps (see TAIR; see also integrated contig tables by Daphne Preuss and colleagues at the CSHL website). This integration will undoubtedly facilitate gene isolation by map-based cloning.

In addition to the unique-sequence regions of the chromosome arms, both rDNA repeats (NORs on chromosomes 2 and 4) and centromeric regions have been mapped genetically and physically. The centromeric regions were mapped by tetrad analysis (Copenhaver et al., 1998) and localized by in situ hybridization (Brandes et al., 1997). Thus, an outline of the physical organisation of the nuclear genome has emerged.

Agyare FD, Lashkari DA, Lagos A, Namath AF, Lagos G, Davis RW, Lemieux B (1997) Mapping expressed sequence tag sites on yeast artificial chromosome clones of Arabidopsis thaliana DNA. Genome Res. 7: 1-9.

Brandes A, Thompson H, Dean C, Heslop-Harrison JS (1997) Multiple repetitive DNA sequences in the paracentromeric regions of Arabidopsis thaliana L. Chromosome Res. 5: 238-246.

Camilleri C, Lafleuriel J, Macadre C, Varoquaux F, Parmentier Y, Picard G, Caboche M, Bouchez D (1998) A YAC contig map of Arabidopsis thaliana chromosome 3. Plant J. 14:633-642.

Copenhaver GP, Browne WE, Preuss D (1998) Assaying genome-wide recombination and centromere functions with Arabidopsis tetrads. Proc. Natl. Acad. Sci. USA 95: 247-252.

Hardtke CS, Berleth T (1996) Genetic and contig map of a 2200-kb region encompassing 5.5 cM on chromosome 1 of Arabidopsis thaliana. Genome 39: 1086-1092.

Kotani H, Sato S, Fukami M, Hosouchi T, Nakazaki N, Okumura S, Wada T, Liu YG, Shibata D, Tabata S (1997) A fine physical map of Arabidopsis thaliana chromosome 5: construction of a sequence-ready contig map. DNA Res. 4:371-378.

Marra MA, Kucaba TA, Dietrich NL, Green ED, Brownstein B, Wilson RK, McDonald KM, Hillier LW,

McPherson JD, Waterston RH (1997) High throughput fingerprint analysis of large-insert clones. Genome Res. 7: 1072-1084.

Meinke, DW, Cherry JC, Dean C, Rounsley SD, Koornneef M (1998) Arabidopsis thaliana: A model plant for genome analysis. Science 282: 662-682.

McPherson JD, Waterston RH (1997) High throughput fingerprint analysis of large-insert clones. Genome Res. 7:1072-1084.

Mozo T, Fischer S, Maier-Ewert S, Lehrach H, Altmann T (1998) Use of the IGF BAC library for physical mapping of the Arabidopsis thaliana genome. Plant J. 16, 377-384.

Round EK, Flowers SK, Richards EJ (1997) Arabidopsis thaliana centromere regions: genetic map positions and repetitive DNA structure. Genome Res 1997 Nov;7(11):1045-53

Sato S, Kotani H, Hayashi R, Liu YG, Shibata D, Tabata S (1998) A physical map of Arabidopsis thaliana chromosome 3 represented by two contigs of CIC YAC, P1, TAC and BAC clones. DNA Res.5:163-168.

Schmidt R, Love K, West J, Lenehan Z, Dean C (1997) Description of 31 YAC contigs spanning the majority of Arabidopsis thaliana chromosome 5. Plant J. 11: 563-572.

Thorlby GJ, Shlumukov L, Vizir IY, Yang CY, Mulligan BJ, Wilson ZA (1997) Fine-scale molecular genetic (RFLP) and physical mapping of a 8.9 cM region on the top arm of Arabidopsis chromosome 5 encompassing the male sterility gene, ms1. Plant J. 12: 471-479.

Wang ML, Huang L, Bongard-Pierce DK, Belmonte S, Zachgo EA, Morris JW, Dolan M, Goodman HM (1997) Construction of an approximately 2 Mb contig in the region around 80 cM of Arabidopsis thaliana chromosome 2. Plant J. 12: 711-730.

Useful web sites

RI maps by chromosome (text or graphic), including access to marker mapping data.	http://nasc.nott.ac.uk/new_ri_map.html /ww/Nov98RImaps/index.html
CAPS markers:	/aboutcaps.html

SSLP markers:	http://genome.bio.upenn.edu/SSLP_info/SSLP.html
Classic map:	http://mutant.lse.okstate.edu/
Display of genetic and physical maps (includes classic map)	/chromosomes/
Physical map of chromosome 1:	http://genome.bio.upenn.edu/physical-mapping/physmaps.html
Physical map of chromosome 2:	http://weeds.mgh.harvard.edu/goodman.html
Physical map of chromosome 3:	http://www.kazusa.or.jp/arabi/ /chromosomes/ (genetic&physical map)
Physical map of chromosome 3 bottom arm (a combination of YAC map, Wash U BAC contigs and location of BAC end sequences):	http://www.genoscope.cns.fr/externe/English/Projets/projetsindex.html
Physical map of chromosome 4:	http://nucleus.cshl.org/protarab/ (top arm) http://mips.gsf.de/proj/thal/db/gv/gv_g_chr4.html (bottom arm)
Physical map of chromosome 5:	http://www.kazusa.or.jp/arabi/
Physical map of genome overview:	/cgi-bin/maps/Pchrom
BAC contigs by fingerprinting:	http://nucleus.cshl.org/protarab/edited_bac_contigs.htm (overview) http://genome.wustl.edu/gsc/cgi-bin/findgif.pl (physical map contigs display)
BAC contigs by endprobe hybridisation:	http://www.mpimp-golm.mpg.de/101/bac.html
ESTs to YACs:	/EST2YAC.html
ESTs to CIC-YACs:	/EST2CIC.html
chromosome 4&5 YAC contigs:	http://nasc.nott.ac.uk/JIC-contigs/JIC-contigs.html
integrated contig tables (by Daphne Preuss and colleagues):	http://nucleus.cshl.org/arabmaps/get_started.htm
Arabidopsis thaliana links:	http://www.nsf.gov/cgi-bin/getpub?nsf9950

Sequencing of ESTs and Genomic Regions

More than 37,000 partial cDNA (EST) sequences have been deposited in the public databases while the total number of genes is most likely about 20,000. Building EST "contigs", i.e. larger cDNA sequences from overlapping ESTs, reduces the number of ESTs to those representing different genomic sequences (Rounsley et al., 1996; Cooke et al., 1997). The current estimate of non-redundant ESTs is about 11,000 or approximately half the total number of genes.

Large-scale high-throughput genomic sequencing makes use of the physical maps and the available BAC (TAMU, IGF), P1 and TAC (Mitsui, Kazusa) libraries (see AGI). BAC, TAC and P1 clones are mapped onto YAC, and their ends are sequenced to determine minimum tiling paths for sequencing large regions. More than 41,000 BAC ends (of a total of 22,000 BAC clones) have been sequenced, yielding stretches of ca. 400 bp every 4 kb on average (total sequence ca. 14 Mb). The largest contiguous region sequenced to date is nearly 1.9 Mb long (Bevan et al., 1998). This region around *FCA* on chromosome 4 contains 389 genes of which 46% could not be assigned a putative function by sequence comparisons with the databases. On average, one gene (ORF) was found every 4.8 kb, and similar values were observed for other genomic regions (Quigley et al., 1996; Sato et al., 1997; Kotani et al., 1997). For many ORFs no corresponding EST was found in the databases. To identify expressed genes within contig regions, a novel cDNA selection method has been proposed (Seki et al., 1997).

Bevan M et al. (1998) Analysis of 1.9 Mb of contiguous sequence from chromosome 4 of Arabidopsis thaliana. Nature 391: 485-488.

Cooke R, Raynal M, Laudie M, Delseny M (1997) Identification of members of gene families in Arabidopsis thaliana by contig construction from partial cDNA sequences: 106 genes encoding 50 cytoplasmic ribosomal proteins. Plant J. 11: 1127-1140.

Kotani H, Nakamura Y, Sato S, Kaneko T, Asamizu E, Miyajima N, Tabata S (1997) Structural analysis of Arabidopsis thaliana chromosome 5. II. Sequence features of the regions of 1,044,062 bp covered by thirteen physically assigned P1 clones. DNA Res. 4: 291-300.

Quigley F, Dao P, Cottet A, Mache R (1996) Sequence analysis of an 81 kb contig from Arabidopsis thaliana chromosome III. Nucl. Acids Res. 24: 4313-4318.

Rounsley SD, Glodek A, Sutton G, Adams MD, Somerville CR, Venter JC (1996) The construction of Arabidopsis expressed sequence tag assemblies. Plant Phys. 112: 1177-1183.

Sato S, Kotani H, Nakamura Y, Kaneko T, Asamizu E, Fukami M, Miyajima N, Tabata S (1997) Structural analysis of Arabidopsis thaliana chromosome 5. I. Sequence features of the 1.6 Mb regions covered by twenty physically assigned P1 clones. DNA Res. 4:215-230

Seki M, Hayashida N, Kato N, Yohda M, Shinozaki K (1997) Rapid construction of a transcription map for a cosmid contig of Arabidopsis thaliana genome using a novel cDNA selection method. Plant J. 12: 481-487.

Useful web sites

EST database:	http://www.cbc.umn.edu/ResearchProjects/Arabidopsis/index.html http://www.tigr.org/tdb/agi/index.html
BAC end sequences:	http://www.tigr.org/tigr_home/tdb/at/atgenome/bac_end_search/bac_end_search.html http://www.genoscope.cns.fr/externe/English/Projets/projetsindex.html
P1, TAC, CIC-YAC end sequences:	http://www.kazusa.or.jp/arabi/endseq/

Arabidopsis Genome Initiative (AGI): Current State of High-Throughput Genome Sequencing

The AGI was established on August 20-21, 1996 when representatives of six research groups (3 from USA and one each from EU, Japan and France) committed to sequencing the Arabidopsis genome met in Arlington, VA to discuss strategies for facilitating international cooperation in completing the genome project. In order to avoid duplication of efforts, the six groups of the Arabidopsis Genome Initiative (AGI) agreed to focus on different regions of the genome (Bevan et al., 1997, Plant Cell 9:476-487). In July 1998, the members of the AGI met again in Arlington, VA to discuss progress to date, to anticipate barriers to timely completion, and to establish an oversight committee for the U.S.-based labs (see Appendix).

At present, the major sequencing domains of the AGI groups have been assigned as follows:

Chromosome 1 (30 Mb)	SPP group (Stanford, PennU, PGEC)
Chromosome 2 (14 Mb)	TIGR group
Chromosome 3 (top - 13.5 Mb)	Kazusa group
Chromosome 3 (top - 5 Mb)	TIGR group*
Chromosome 3 (bottom - 9 Mb)	EU project chrom3 (coordinated by Genoscope)
Chromosome 4 (top - 4 Mb)	CSHSC (CSH-WU-ABI group)
Chromosome 4 (bottom - 13 Mb)	EU group (ESSA I, II, III)
Chromosome 5 (top - 9 Mb)	EU group (ESSA III)
Chromosome 5 (top + middle - 4 Mb)	CSHSC (CSH-WU-ABI group)
Chromosome 5 (top + bottom - 17 Mb)	Kazusa group

* TIGR will begin sequencing this region in spring 1999.

Sequencing is being done on BAC and P1 clones. Two different strategies are pursued. Both the SPP group and the TIGR group have selected nucleating sites ("seed BACs") around which BAC contigs have been established by using BAC end sequences to select adjacent clones with minimum overlap. This sequential sequencing procedures involves 32 and 16 starting points on chromosomes 1 and 2, respectively. The other sequencing strategy adopted by CSHSC, ESSA and Kazusa involves building of BAC or P1/TAC tiling paths with minimum overlap of adjacent clones ("sequence ready maps"). This procedure requires more preparative work but once established, large regions can be sequenced in parallel, e.g. by the several sequencing groups within the ESSA group.

Lists of clones selected for sequencing can be found on the web sites of the sequencing groups. Start dates for sequencing are indicated and it is agreed that the finished sequences will be released within 4-6 month after the start of sequencing (for details, see Appendix). The current state of genome sequencing is as follows (for overview by chromosome region, see TAIR / Arabidopsis Sequencing View and the homepages of the AGI groups):

AGI Progress at the end of 1998

Chr.	Est. Size	Completed		In Progress		In Preparation		Total Sequence	
	(Target)	Clones	Mb	Clones	Mb	Clones	Mb	Clones	Mb
1	30 Mb	52	5.52	16	2.02	16	1.7	84	9.25
2	17 Mb	107	10.29	45	4.32	27	2.64	179	17.25
3	22.2 Mb	13	1.09	29	2.5	27	1.2	69	5.8
4	18.5 Mb	153	11.71	51	5.2	28	2.6	231	19.4
5	29.2 Mb	208	14.62	15	1.2	38	2.7	261	18.54
Total	120 Mb	534	43.32	156	15.2	136	11.8	826	70.3

Note that the total sequence entered into TAIR and summarised above includes overlaps between adjacent clones (except for those submitted by ESSA and WashU, which have overlaps almost all removed). For this reason the total number of clones sequenced is a better estimate of progress. With 10% overlap, 120 Mb will require 1,390 BAC clones. On 31 December 1998 the following finished clones had been deposited in Genbank:

SPP 50 BAC clones
TIGR 105 BAC clones
CSHSC 60 BAC clones
ESSA 117 BAC and cosmid clones
Kazusa 202 P1, TAC and BAC clones
Total 534 clones (approx. 39% of the total genome)

As of 31 December 1998, the TAIR Sequencing View displays 46 Mb (39% of estimated 120 Mb genome size) of complete sequence. This figure is 17 Mb higher than that given at the end of June 1998, indicating that the current rate of sequencing is close to 3 Mb per month for the entire AGI project. Taking into account the sequences that have not been released, the actual amount of sequence information is close to 55 Mb (almost 50% of the unique sequences). It is thus a realistic goal to finish the sequence of the Arabidopsis genome (excluding telomeric and centromeric regions as well as NORs) by the end of the year 2000.

Completion of the sequence is defined as each chromosome arm between subtelomeric repeats and centromeric repeats consisting of a single fully sequenced contig. This excludes the rDNA repeats (NORs on chromosomes 2 and 4 each of which accounts for ca. 3.5 Mb) and other internal tandem repeat regions. For these regions, it will be sufficient to sequence one repeat unit and to estimate the repeat number at each site. By these criteria, sequencing of chromosomes 2 (14 Mb) and 4 (17 Mb) can be expected to be complete before the end of 1999.

As sequencing is reaching the closing phase, boundaries between sequencing domains have to be defined precisely to avoid duplication of efforts by different sequencing groups. This difficulty has already been encountered by all the sequencing groups, resulting in duplication of sequences and mismapped clones (see table). For example, on chromosome 4 both CSHSC and ESSA sequenced two different but overlapping clones and had to reassign remaining projects in a common region of ca. 900 kb. TIGR and SPP have abandoned or mismapped at least 4 BACs and a chimaeric YAC, while Kazusa has sequenced several duplicate clones on chromosome 5. Depending on different rates of progress, it may seem advisable, in the interest of the Arabidopsis community, to reallocate genomic regions between the sequencing groups (see Appendix 1 and 2). The fingerprint map constructed at Washington University and the hybridisation-based map constructed by T. Altmann have the potential for delineating these regions before they are sequenced, and will probably be used for this purpose.

Useful Web Sites

Information on sequencing (lists of BAC and P1 clones, maps)

chromosome 1:	http://sequence-www.stanford.edu/ara/by_locus.html http://pgec-genome.pw.usda.gov/ http://cbil.humgen.upenn.edu/~atgc/ATGCUP.html
chromosome 2:	http://www.tigr.org/tigr_home/tdb/at/atgenome/atgenome.html
chromosome 3:	http://www.genoscope.cns.fr/ http://www.inra.fr/Versailles/BIOCEL/CHR3-INRA/chromosome3.html
chromosome 4&5:	http://nucleus.cshl.org/protarab/
chromosome 3,4&5:	http://mips.gsf.de/proj/thal/
chromosome 3&5:	http://www.kazusa.or.jp/arabi/
entire genome:	http://genome-www3.stanford.edu/cgi-bin/AtDB/Schrom

Stock Center Resources and Data Bases

The Arabidopsis Biological Resource Center (ABRC), Ohio State University, USA and the Nottingham Arabidopsis Stock Centre (NASC), University of Nottingham, UK have been providing stocks and information to the Arabidopsis community, since 1991. They have endeavored to accumulate the broadest possible range of stocks to provide the best platform of genetic diversity and genetic tools for Arabidopsis researchers and the genome project. Combined, they have sent 35,000 samples of seeds, 12,000 individual DNA clones and additional thousands of clones as gridded large-insert libraries to researchers world-wide during the last year.

The seed stocks currently available from the two centers include mutant lines (600), T-DNA lines and pools (30,000+), mapping strains, the G. P. Rédei collection of mutants and research lines (300+), the A. R. Kranz collection of mutants and ecotypes (700+), transposon/transposase lines (100+), RI lines (3 populations), ecotypes (400+), transgene lines and related species. The genetic mapping resources of the centers and the T-DNA and transposon resources complement the AGI sequencing efforts and the current research focus on functional genomics.

DNA stocks of ABRC include cloned genes (200), RFLP mapping clones (300+), expressed sequence tagged (EST) clones (30,000+), cDNA libraries (7), a phage genomic library, YAC libraries (6), BAC and P1 libraries used in genome sequencing (3) and two-hybrid libraries (2). In addition, filters of BACs, P1s and YACs for hybridization and isolated DNA from T-DNA populations (12,000 lines) are available.

The EST collection has been organized so that a set of 11,000, non-redundant based on the sequences available to TIGR, is being used by AGI. The 3' sequences of these clones are being analyzed by the MSU EST project to further eliminate redundancy. Copies of BAC and P1 clones, for which sequences have been published, are being sent to many research laboratories. In this connection, ABRC requests that all sequencing projects adhere, if at all possible, to the agreed clone-naming conventions when publishing sequences so that researchers can identify, without confusion, the proper clones to obtain.

NASC and ABRC are working to enlarge the collections of characterized mutants and clones. In addition, it is expected that large numbers of T-DNA lines will be received so that, within the next year, the available T-DNA lines will represent essential saturation of the genome. In connection with the accumulating genomic and cDNA sequence information, these resources will prove invaluable to the research community. In addition, new transposon-tagged populations, recombinant inbred mapping populations, a tetrad mapping populations and GFP lines are being incorporated into the collections.

RI mapping

The Nottingham Arabidopsis Stock Centre (NASC) curates the Lister and Dean RI maps that were originally developed and maintained by Clare Lister and Caroline Dean (JIC, Norwich). NASC also offers a weekly community mapping service. Anyone can submit data to NASC for mapping using the specially designed data submission form. The positions of all markers mapped at NASC are made publicly available through the NASC WWW server, the Arabidopsis Genome Resource and TAIR. For private mapping, all the marker scores are available from NASC. However, the aim for the community is to have as many markers as possible placed on the canonical map and so the submission of mapping data for inclusion on the RI map is appreciated.

Linking maps and sequence for comparative analysis in the Arabidopsis Genome Resource

The Arabidopsis node of the BBSRC funded UK-Crop Plant Bioinformatics Network (UK-CropNet) based at NASC has established the Arabidopsis Genome Resource (AGR). AGR is being developed as a repository of Arabidopsis data of value in the comparative analysis of plant genomes and as an essential tool to aid in the cloning of homeologous genes of agronomic importance.

Comparative analysis in plants relies upon genetic and physical mapping of common probes between species. To this end AGR has made available the YAC physical maps of chromosomes IV and V (from C.Dean, R.Schmidt, M. Stammers). AGR also includes the Recombinant Inbred Maps from NASC integrated with the AGI sequence template clones (locations provided through TAIR). Arabidopsis nucleotide sequences are also included within AGR.

Integrating these data sets is the next key step in the development of AGR. Sequence overlap between completed AGI clones define contigs of BACs and P1s. These contigs will be fixed to the YAC physical maps using the results of BAC-YAC hybridisations. Contigs may be anchored on the RI maps through the nearest marker information from individual clones. RI maps and YAC physical maps are to some extent integrated through the use of some RI markers as probes in YAC physical mapping.

In collaboration with Martin Trick (John Innes Center), these data will be used to generate comparative map displays between Arabidopsis and the Brassicas.

Contact Persons

Randy Scholl, ABRC email: scholl.1@osu.edu
Mary Anderson, NASC email: arabidopsis@nottingham.ac.uk

Web sites

ABRC:	http://aims.cps.msu.edu/aims/
NASC:	http://nasc.nott.ac.uk/
AtDB:	/

Database Issues

The Arabidopsis Data Base (TAIR) is, at this time, located at Stanford University, Mike Cherry, P.I. The explosion of data, both genomic and biological, makes it clear that the data base as it now exists is operating at a minimal, not an optimal, level. The recognition that the community had to express its needs in a more concrete way resulted in two workshops addressing the issues of database composition and management. One was held in 1993 in Dallas, TX and that report can be accessed at /db/dallas.report.html.

However, a more recent workshop on the same topic was held at the international meeting at Madison, WI in 1998 and that report is attached as an appendix. The needs are for a central database with links to other useful databases and information which is organized in a user-friendly fashion. Recognition of the needs of the Arabiopsis community as well as other interested communities has resulted in a call for proposals to the NSF titled "Arabidopsis thaliana Information Resource Project (AtIR)" The deadline date is March 22, 1999 and a copy of that announcement is attached to this report as an appendix.

Recommendation on information management

Large-scale genomic sequencing has reached a critical stage, with about half the genome in hand. Although the AGI sequencing groups provide information for specific regions of chromosomes, it is difficult and time-consuming for the Arabidopsis community to retrieve the relevant information. To take full advantage of all the progress that has been made in the analysis of the Arabidopsis genome, it will be necessary to establish a well-funded unified genome database that displays sequence and related features together with biological information in a user-friendly way.

National and Transnational Research Activities

Australia

Arabidopsis research in Australia is focused on building an understanding of fundamental aspects of plant biology. There is no direct commitment to large scale genome sequencing at this stage.

Among recent highlights, Liz Dennis, Jim Peacock and colleagues from CSIRO Division of Plant Industry in Canberra have discovered a second nonsymbiotic leghemoglobin gene from Arabidopsis (Proc. Nat. Acad. Sci. US 94, 12230-12234, 1997). They propose that all plants have two classes of leghemoglobins, as exemplified by the two genes in Arabidopsis. In the evolution of symbiosis, the product of one or other of the genes has been recruited on different occasions to play a new role in association with the symbiont. In most cases class 1 gene products have been involved, but the newly discovered class 2 proteins are also potentially symbiotic.

Another highlight has been the discovery of a gene encoding the catalytic subunit of cellulose synthase (Science 279, 717-720, 1998). Tony Arioli and colleagues in Richard Williamson's research group in the Research School of Biological Sciences at ANU in Canberra have walked to the locus of a temperature sensitive mutant that leads to root swelling (RSW1). The gene that complements the mutant phenotype is related to a cellulose synthase subunit gene from cotton. In the mutant there is widespread accumulation of beta-1,4-glucan but it is not crystallised into microfibrils, suggesting such assembly is a role of the RSW1 gene product.

Other active programs include studies of various aspects of flowering, from induction through floral organ morphogenesis to fertilisation and seed development. Also topics as diverse as aspects of photosynthesis, analysis of effects of abiotic stresses including heavy metals and UV, epigenetic effects of cytosine methylation, and the roles of the MYB gene family are being actively investigated.

A major commitment is being made to host the 10th International Conference on Arabidopsis Research in Melbourne from 4-8 July 1999. A Regional Advisory Committee, with colleagues from Japan, South Korea, Singapore and New Zealand, has been set up to give the meeeting a Western Pacific focus. This will be the first time the Arabidopsis community has met outside Europe and North America, and we look forward to welcoming scientists and students to Australia where plant science continues to thrive.

Contact Person: David Smyth, Monash University, Melbourne

E-mail Address: David.Smyth@sci.monash.edu.au

Belgium

As Belgium is a federal country we have both federal and Flemish initiatives to support research using *Arabidopsis thaliana* as the experimental organism.

A Flemisch project is running on the isolation and characterization of new ethylene mutants in *Arabidopsis thaliana.* This project aims at the isolation of a new series of mutants in the ethylene signal transduction pathway. A combined morphological, physiological and molecular-genetical approach will elucidate a number of previously unknown elements and will provide a better insight in the control of plant development by this hormone.

Belgian governement also stimulates interactions between the different universities. In this frame a project is running between the universities of Gent, Antwerp, Brussels and Liège on the growth and development of higher plants. Many external factors such as light intensity, light quality, temperature, the availability of nutrients and the interaction with pathogenic organisms influence to a great extent, growth and development of higher plants. The current knowledge on the molecular processes that control growth and development is still very limited. The national network aims at making a contribution to developmental biology by studying a limited number of aspects of plant development. Wherever possible, *Arabidopsis thaliana* will be used as a model plant. Keyprojects include the identification and cloning of key regulatory genes involved in leaf morphogenesis, the molecular analysis of the formation of syncytia (=large feeding cells) in nematode infected Arabidopsis roots. The Flemish community also supports these projects.

Contact Person: Nancy Terryn /Marc Van Montagu, University of Ghent

E-mail Address: nater@gengenp.rug.ac.be

China

Research using *Arabidopsis* as a model system was further established in China at national research institutes and universities in the past year. The research areas mainly include biosynthesis of amino acids, signal transduction and metabolism of plant hormones, cell wall formation, seed storage proteins, response to environmental stresses, isolation of various mutants affecting growth and development, and characterization of transposable elements. Interests in reverse genetics and functional genomics are also greatly increased with the focuses on genetargeting, constructing a large transgenic population with mapped Ds randomly distributed at a high density, developing an expression library to transform in planta and establishing cDNA array to monitor gene expression and identify functional genes. Grants to support the research projects mentioned above are mainly from National Natural Science Foundation of China, Chinese Academy of Sciences and Hong Kong Research Grant Council/UPGC Grant HKU.

Contact Person: Jiayang Li, Institute of Genetics, Chinese Academy of Sciences

E-mail Address: jyli@ss10.igtp.ac.cn

France

Genome sequencing

During the last year, three French laboratories (M. Delseny/Perpignan, M. Kreis/Orsay and R. Mache/Grenoble) have systematically sequenced three BACS (300 kb) as part of the EU-ESSA II Program. Delseny's group has also continued to sequence cDNA clones corresponding to the 60kbp locus, Em1, on chromosome 3. A French sequencing center, Genoscope CNS has been created and part of its activity is devoted to sequencing the Arabidopsis genome. In collaboration with TIGR and Upenn, Genoscope is generating end sequences from all 23,000 BAC clones from the TAMU and IGF libraries to expedite the selection of clones with minimal overlap with those already sequenced. They are also coordinating a new EU project aimed at sequencing the lower arm of chromosome 3 (9Mb). This project involves 16 sequencing groups. The goal for Genoscope and three academic French laboratories is about 2 Mb.

Synteny with other genomes

A program was developed between INRA Rennes and Versailles groups to identify consensus markers between rapeseed and Arabidopsis for a number of agronomically relevant genes. A collaboration between laboratories in Perpignan, Davis and Poznan has found synteny between five adjacent genes in the chromosome 3 Em locus of Arabidopsis and genes in B. oleracea, B. nigra and B. rapa. The EU program EuDicotMap has started to select highly conserved ESTs of rice and Arabidopsis and to map them in Arabidopsis as well as important European crops in order to identify synteny blocks between different families.

Generation of insertion lines and reverse genetics screenings

INRA-Versailles has now generated more than 38,000 T-DNA mutant lines. Screening of the collection is being done via a coordinated effort between INRA, CNRS and various European laboratories. Out of approximately a hundred target genes selected for the screen insertions were identified in 50% of them. The systematic characterization of flanking sequences tags of insertions in over a thousand mutants has now begun. About 11,000 lines will be donated to NASC by the beginning of next year.

A summary of *Arabidopsis* genes under study

Research in many areas of plant genetics and biology is being actively pursued in French laboratories. Plant hormone and signal transduction, cell wall, secreted and membrane proteins, metabolism, development, and plant pathogen interactions are being investigated in laboratories throughout France.

Contact Person: Michel Caboche, INRA Versailles

E-mail Address: caboche@tournesol.versailles.inra.fr

Germany

Arabidopsis research is still increasing in scope at universities and research institutions. The national research program on "Arabidopsis as a model for analysing plant development" is in its final two-year funding period. Because its tremendous success, an initiative has been made by Arabidopsis researchers to establish a new program focusing on plant cell biology. Another six-year national research program on plant hormones to start in 1999 includes several groups working on *Arabidopsis*. Beside these programs, *Arabidopsis* research is funded within European projects and by DFG grants on an individual basis or as part of local research programs.

Several *Arabidopsis* projects are related to genome research. ZIGIA, a program operated at the Max-Planck-Institut in Cologne, aims at the functional analysis through gene inactivation by transposon insertion. High throughput endprobe hybridization of BAC clones from the IGF library was done at the Max-Planck-Institut in Golm. These data were integrated with information made available by other groups to assemble a complete BAC-based physical map of the *Arabidopsis* genome**.** Projects on transcript profiling have been initiated at the DKFZ in Heidelberg, the MPI in Golm and the IPK in Gatersleben. The Federal Ministery of Education and Science (BMBF) has made a call for proposals within a newly-established Plant Genome Analysis program (GABI). A joint Arabidopsis proposal involving 32 projects from 27 different institutions has been submitted, aiming at a functional analysis of the genome.

An EMBO (European Molecular Biology Organisation) Course held at the Max-Planck-Institut in Cologne in May 1998 entitled "Molecular and Biochemical Analysis of Arabidopsis" was attended by 16 participants representing 13 European countries. The course covered the theoretical and practical aspects of forward and reverse genetics, genetic and physical mapping, transformation, transient gene expression, in situ hybridisation, cell biology, physiology, the yeast two-hybrid system, complementation of yeast mutants and bioinformatics over an eleven-day period. EMBO Course seminars from ten invited speakers were integrated with a two-day meeting of the national Arabidopsis research program.

Contact Person: Gerd Jürgens, Universität Tübingen

E-mail Address: gerd.juergens@uni-tuebingen.de

Italy

Research in Italy with *Arabidopsis* is growing. About twenty laboratories are presently attending to researches regarding this model system. Investigations cover: plant pathogen relationships, expression of PG and PGIF genes, role of rolB and rolD in plant differentiation, HD-ZIP transcription factors in plant morphogenesis, complementation of yeast by Arabidopsis genes, selection of Ca2+ and K+ transport mutants, genes involved in heat and cold resistance, myb transcription factors, genes of the polyamine pathway, induction of noduline genes in plants by Rhizobium, use of antisense RNA to inhibit nitrogen transport, study of agravitropic mutants in earth and micro g conditions (ESA-ASI projects). Financial support for the researches is coming from different sources, e.g. the National Research Council, the Ministry of Agriculture, the European IV Frame Programs, the ESA-ASI Space Programs, and a few other National Agencies. Research groups are located both in universities and in National Institutes (National Research Council, ENEA, National Institute of Nutrition). The Italian association of researchers interested in Arabidopsis (ARABITALIA) met for the first time in September 1997 in Abbadia di Fiastra (Macerata, central Italy). In this occasion the scientists present to the meeting furnished a report of their Arabidopsis investigations and projects, and a booklet carrying the information about research on Arabidopsis in Italy was also distributed. In this occasion some young Italian researchers, who are working in foreign countries (USA, and UK) also reported about their recent investigations. The 1998 annual Meeting was held at the end of September in Viterbo (central Italy) in the occasion of the EUCARPIA Symposium on plant breeding. A document is in preparation about the state of Arabidopsis research in Italy, and about the actions that can be started to obtain the financial support that is needed to foster it.

Contact Person: Fernando Migliaccio, CNR (Monterotondo)

E-mail Address: miglia@nserv.icmat.mlib.cnr.it

Japan

Arabidopsis research is well-established in Japan. The number of laboratories using the model plant for research and education is still increasing gradually in universities, national institutes, and private companies. Areas of research are widely spread from developmental biology, metabolic regulation, gene expression, environmental stress signaling, and DNA methylation, to large scale DNA sequencing. The results of the researches were reported in international meetings such as the " International Congress of Arabidopsis Research" in Madison, WI, the "Joint Meeting of Japanese and American Societies of Plant Physiologists" in Vancouver, BC and in national meetings, especially in the "Workshop on Arabidopsis Studies", an annual meeting. The 8th workshop was organized by Kazuo Shinozaki, Minami Matsui, Yuji Kamiya, and Richard E. Kendrick from October 11 to 13, 1997, at Riken Institute at Wako city, Saitama. The workshop was joined with Frontier Research Forum, "Recent Progress of Plant Hormone Research in Arabidopsis". We had nearly 250 participants, 20 poster presentations, and 37 speakers including 7 guest speakers from abroad. The 9th workshop was held in Kazusa Academia Center from Nov. 19 to 20, 1998. The workshop organized by Satoshi Tabata had nearly 300 participants, 40 poster presentations and 11 presentations. Topics of the presentations included systemic genome analyses, patent, and postgenome tactics, as well as mutant analyses, gene cloning, and newly-developed techniques.

The Japanese Arabidopsis communication network, nazuna-net, started in January 1995, now includes 442 members (Sept. 1998) from 99 organizations including 17 private companies (contact: Dr. Takayuki Kohchi: kouchi@bs.aist-nara.ac.jp). A large-scale genome sequencing project showed extensive progress at Kazusa DNA Research Institute in coordination with the Multinational Arabidopsis Genome Initiative (contact: Dr. Satoshi Tabata: tabata@kazusa.or.jp). Nearly 12.5 Mb covering 174 P1 clones have been sequenced and reported in the journal "DNA Research" (contact: http://www.uap.co.jp), on a homepage (http://www.kazusa.or.jp/arabi/). The Sendai Seed Stock Center (SASSC) is operated by Dr. Nobuharu Goto (n-goto@ipc.miyakyo-u.ac.jp) since 1993.

Contact Person: Kiyotaka Okada, Kyoto University

E-mail Address: kiyo@ok-lab.bot.kyoto-u.ac.jp

The Netherlands

The Dutch Arabidopsis groups organized their annual meeting in Utrecht on February 19, which was attended by approximately 80 participants. Arabidopsis groups are located at the Universities of Leiden, Utrecht and Wageningen and at CPRO-DLO in Wageningen. Important research topics are in Leiden (Hooykaas) recombination, auxin action and apoptosis, in Utrecht sugar sensing (Smeekens), root development (Scheres) and acquired resistance (van Loon), in Wageningen embryogenesis (de Vries, van Lammeren) and flowering and seed- development (Koornneef), transposons, genome sequencing, plant disease resistance genes and developmental biology (Stiekema, Pereira, Angenent, Groot all CPRO-DLO). The groups collaborate through their involvement in graduate schools and EU programs.

Contact Person: Maarten Koornneef, Agricultural University Wageningen

E-mail Address: Maarten.Koornneef@BOTGEN.EL.WAU.NL

Spain

No special funding programme supports *Arabidopsis* research in Spain. However, more than 20 research groups are currently active in research with this organism, mainly funded by the National Biotechnology Programme, Basic Research Programmes, and the European Union BIOTECH Programme. Some of these groups are involved in large-scale genome sequencing and function search, specially in the case of the Myb family of transcriptional factors. Spanish groups interested in *Arabidopsis* development are mainly focused on seed, leaf and flower development, and flowering induction. This area is seing the incorporation of new groups of Arabidopsis users, some of them also interested in cell differentiation. In the area of plant physiology and metabolism some topics that have seen significant contributions during the year are the study of secondary metabolism, the identification of new elements in the signal transduction pathways involved in different environmental stress responses, and the analysis of sulfur and phosphate assimilation. *Arabidopsis* has also being increasingly used for studies in plant pathogen interactions to identify new elements in the response signal transduction pathways.

The Spanish Arabidopsis network, funded by the National Biotechnology Programme, generated a collection of 10000 T-DNA lines that is being actively used in mutant screenings at both the phenotypic and DNA levels, in many laboratories. This network that includes all the Spanish laboratories working with *Arabidopsis* is now discussing future join activities. Many more Spanish scientists are currently involved in *Arabidopsis* research in other laboratories around the world. Their succesful integration in the Spanish R&D system would strongly contribute to steer the field and increase the contribution of our country.

Contact Person: José Martinez Zapater, Centro Nacional de Biotecnología (Madrid)

E-mail Address: zapater@cnb.uam.es

United Kingdom

There are over 190 projects at present in the UK involving *Arabidopsis*. The European Commission continues to be a major source of funding and the newly announced Framework V programme is due to begin calls for proposals. Although there are no longer any special initiatives aimed specifically at Arabidopsis research, The Biotechnology and Biological Sciences Research Council (BBSRC) funds projects through competitive grants and special initiatives, contributing approximately £ 6.8m to *Arabidopsis* research in the UK.

An Arabidopsis Gene Function Search Network is currently under development by Mike Bevan at the John Innes Centre. This is a network of consortia, groups of labs with a common goal, being brought together with the aim of doing large scale screening programmes to reveal the functions of very large numbers of genes being revealed by the genome project.

The Genetical Society of Great Britain chose Arabidopsis as the subject area for their annual autumn meeting in 1997. The Mendel Lecture was given by Elliott Meyerowitz who was preceded during the day by Mike Bevan, Rob Martienssen, Joe Ecker, Ben Scheres, Caroline Dean, Gerd Jurgens and Brain Staskawicz."*Arabidopsis thaliana*: Big Ideas from a Small Plant" was such a success that the Society has decided to host a biennial conference on *Arabidopsis.*

An EMBO (European Molecular Biology Organisation) Course held at the John Innes Centre in May 1997 entitled "*Arabidopsis* as an Experimental Organism" was attended by 12 participants representing seven European countries. The course covered the theoretical and practical aspects of mutant screening, genetic and physical mapping, plant pathology, microscopy, biolistics, the yeast two-hybrid system, and sequence fragment and data analysis over a ten day period which also included seminars from ten invited speakers.

The Chelsea Flower Show judges awarded a prestigious Silver Medal to the John Innes Centre Science Communication and Education Department exhibit, entitled "Arabidopsis - a Wonderful Weed". The exhibit demonstrated how *Arabidopsis* is used to recognise genes of agronomic importance in agricultural crops. The public exposure and media coverage the display attracted in the UK and abroad has helped to increase awareness of the importance of plant molecular biology.

In the last year the Nottingham Arabidopsis Stock Centre (NASC) in collaboration with the Arabidopsis Biological Resource Center (ABRC) has continued to accumulate the broadest possible range of stocks to provide the best platform of genetic diversity and genetic tools for the investigation of this model system. Currently NASC maintains and distributes over 20,000 accessions of Arabidopsis to the research community. New stocks generated within the UK and shortly to be made available include the first 10,000 of the Sainsbury Laboratory *Arabidopsis* transposants (SLAT) lines (Jonathan Jones, Sainsbury Lab, UK), 100 GFP lines (Jim Haseloff, Cambridge, UK) and a Recombinant Inbred population of Nd (Niederzenz) x Columbia generated by Eric Holub, Jim Beynon and Ian Crute (HRI Wellsbourne, UK).

Contact Person: Caroline Dean, John Innes Centre, Norwich

E-mail Address: caroline.dean@bbsrc.ac.uk

United States

Arabidopsis research continues to flourish in both academic and corporate laboratories in the United States. One of the most obvious indicators of the value of information that can be gleaned from *Arabidopsis* research has been the establishment of several genomics companies that are exploiting *Arabidopsis* genetics. Thanks to continued support from the National Science Foundation (NSF), the Department of Energy (DOE) and the U.S. Department of Agriculture (USDA), the *Arabidopsis* genome is on track for being completely sequenced by the end of 2000. A total of 46 Mb of finished sequence had been deposited in public databases as of January 1999, of which the US sequencing groups contributed more than 24 Mb. Importantly, the groups in the US Arabidopsis Genome Initiative (AGI) finished the first phase of their sequencing effort in less than the original 3 year time allowed, and could thus begin during 1998 with the second phase of sequencing ahead of time. In addition to its value for database mining and other more traditional genomic approaches, the availability of large amounts of genome sequence together with physical maps that cover almost the entire genome have begun to eliminate positional cloning as a bottleneck in *Arabidopsis* genetics. Much of this information is conveniently accessed through the *Arabidopsis thaliana* database (TAIR) at Stanford University. The growing importance of *Arabidopsis* research has also been evident in the increasing number of participants at the Eight and Ninth International Conferences on *Arabidopsis* Research, which were held in Madison, WI, and drew 817 and 998 participants, respectively.

Apart from the genome sequencing efforts, important tools are being developed for reverse genetics and functional genomics. A significant advance in this area has been an $8.7M award from the NSF Plant Genome Research Program for a cooperative effort to provide high-throughput gene expression profiling as well as gene knock out services to the Arabidopsis community. The identification of gene knock outs has been made possible through the availability of large numbers of T-DNA insertion lines, of which 48,500 have already been deposited with the Arabidopsis Biological Resource Center (ABRC) at Ohio State University. This number can be expected to at least double in 1999. The ABRC continues to be an important resource for the Arabidopsis community. It shipped 29,500 seed and 13,000 DNA stocks in 1997; and 46,500 seed and 16,000 DNA stocks in 1998.

As a direct consequence of the improvements in scientific infrastructure, significant scientific advances have been made in every area of *Arabidopsis* research, including hormone and light signaling, circadian clock, responses to biotic and abiotic stress and developmental biology. Some of the most noteworthy discoveries in 1998 included the discovery of master regulatory genes that protect *Arabidopsis* from cold damage and the identification of proteins that transport auxins.

Contact Person: Detlef Weigel

E-mail Address: detlef_weigel@gm.salk.edu

Contact Person: Jeff Dangle

E-mail Address:dangle@email.unc.edu

Appendix 1

NSF *ARABIDOPSIS* GENOME MEETING REPORT

INTRODUCTION

In 1990, a report entitled "A Long-range Plan for the Multinational Coordinated *Arabidopsis thaliana* Genome Research Project" was published by the National Science Foundation (NSF 90-80). The report detailed plans made by members of the *Arabidopsis* research community in the U.S. and abroad, to collaborate in the sequencing of the genome of this model plant, and to characterize the structure, function and regulation of all *Arabidopsis* genes. In 1998 it became possible to set a realistic goal of finishing the sequence by the end of the year 2000.

Since then, a multinational genome sequencing project involving laboratories in the United States, in Europe, and in Japan, has been engaged in achieving this goal. This report is the proceedings of a meeting held to discuss progress to date, to anticipate barriers to timely completion, and to establish an oversight committee for the U.S. -based labs. The meeting was held at the National Science Foundation in Arlington, Virginia on July 9 and 10, 1998.

Participants Representing

Elliot Meyerowitz, California Institute of Technology Chair

Ian Bancroft, John Innes Centre ESSA

Michael Bevan, John Innes Centre ESSA

Ellson Chen, Perkin-Elmer Applied Biosystems CSHSC

Ronald Davis, Stanford University SPP

Nancy Federspiel, Stanford University SPP

Gerd Jürgens, University of Tübingen MSC

Richard McCombie, Cold Spring Harbor Laboratory CSHSC

Rob Martienssen, Cold Spring Harbor Laboratory CSHSC

David Meinke *Arabidopsis* community

Xiaoying Lin, TIGR TIGR

Curtis Palm, Stanford University SPP

Daphne Preuss, University of Chicago *Arabidopsis* community

Francis Quetier, Genoscope Genoscope

Steven Rounsley, TIGR TIGR

Marcel Salanoubat, Genoscope Genoscope

Satoshi Tabata, Kazusa Kazusa

Athanasios Theologis, USDA Plant Gene Expression Ctr. SPP

Richard Wilson, Washington University CSHSC

Observers

Mary Clutter NSF

Machi Dilworth NSF

DeLill Nasser NSF

James Tavares DOE

Jane Peterson NIH

Adam Felsenfeld NIH

Peter Bretting USDA

Liang-Shiou Lin USDA

STRUCTURE AND PROGRESS

There are six different sequencing consortia participating in the sequencing phase of the *Arabidopsis* genome project, three from the United States, two from the European Community, and one from Japan. Each is sequencing a different region of the genome, and each has its own model for distribution of the necessary work among consortium members. The progress of each follows, taking them in turn.

TIGR (The Institute for Genome Research, http://www.tigr.org/tdb/at/at.html)

TIGR has taken on two aspects of the sequencing project. The first is BAC end sequencing (along with SPP and Genoscope), to provide one-pass sequences of both ends of the 22,000 BAC clones that are one type of clone being used for sequencing in the genome project. The purpose of this is to allow sequential progression from a single sequenced BAC to the two adjacent genomic regions with minimal overlap. TIGR has sequenced 16,392 BAC ends from a total of 9,572 BAC clones, providing a total of 7.34 Mb of BAC end sequence. The total BAC end sequence from all three groups is 36,574 BAC ends from 18,746 clones, representing 13.64 Mb.

The second TIGR project is the sequencing of chromosome 2. They have chosen 16 well-spaced starting points (by use of the Goodman lab chromosome 2 contig map), and are sequencing BAC clones in parallel, starting with the original clone in each location, and proceeding by use of BAC end sequences to adjacent clones with minimal overlap. The average overlap between adjacent BAC clones has been 8.2 kb, with a range from 150 bp to 30 kb. At present 4.83 Mb is complete and annotated, 3.25 Mb has shotgun sequencing or annotation in progress, and 1.38 Mb of BAC clones are in preparation for sequencing, for a total of 9.46 Mb.

The only problem encountered so far is a gap with no clones to cross it in present BAC collections, in the m336 large contig. Fiber FISH done at the University of Wisconsin indicates a gap size of 500 kb, and the sequence at either side of the gap shows no special features. There has also been a BAC difficult to close due to long tandem dinucleotide repeats, but there is no theoretical barrier to completion of such clones.

The total estimated length of chromosome 2 is less than 14 Mb, not including an estimated 3.5 Mb of ribosomal DNA tandem repeats at one end of the chromosome. The current rate of sequencing in this phase of the project at TIGR is presently 8 Mb per year, and there is an existing proposal to increase that to 12 Mb per year. It is estimated that, barring unforeseen problems, chromosome two, excluding highly repetitive centromeric regions and the rDNA repeats, will be completed by the end of 1999; if the full capacity is to be used, clones on other chromosomes will have to be started by the end of 1998.

SPP (Stanford University, Plant Gene Expression Center, University of Pennsylvania; http://pgec-genome.pw.usda.gov; http://cbil.humgen.upenn.edu/~atgc/ATGCUP.html; http://sequence-www.stanford.edu/ara/ArabidopsisSeqStanford.html)

These three groups have as a goal completing the sequence of chromosome 1. They have divided some of the preparative tasks, with Stanford providing automated template preparation, Penn mapping chromosome 1 BACs and providing BAC end sequences to the project, and PGEC making the sequencing libraries. All groups are involved in sequencing. The strategy is similar to that of TIGR, whereby seed BAC clones chosen by the Penn laboratory are used a sequencing origins, and progress made by use both of BAC end sequences and BAC fingerprints, to provide minimal overlap. Initially 20 starting points were used, there are plans to add an additional 20 soon.

SPP has provided 8,936 BAC end sequences to the 36,574 BAC end total.

The chromosome 1 sequencing done or in progress has so far totaled 5.64 Mb, which is the sequence of 55 BACs and 1 YAC clone. Excluding overlap between adjacent clones leaves a total unique sequence in progress or finished of 5.36 Mb. Of this 4.02 Mb are complete, 0.65 Mb in finishing and 0.97 Mb in shotgun phase. Overlap between adjacent clones has been 2 to 38 kb, with an average less than 7 kb; there has as yet been no failure to find the adjacent clone from any sequenced BAC.

The total estimated length of chromosome 1 is 30 Mb. Capacity exists to finish it by the end of 2000, given sufficient funding - completion will require sequencing approximately 300 BAC clones in the next 3 years, or 33 BACs per year per participating site.

CSHSC (Cold Spring Harbor Sequencing Consortium; http://www.cshl.org/arabweb/; http://genome.wustl.edu/gsc/)

This consortium includes Cold Spring Harbor Laboratories, Washington University and Perkin-Elmer Applied Biosystems. They are taking a different approach to choosing the BAC clones to sequence, which involves HindIII and EcoRI fingerprinting of BAC clones, and from the clone overlaps inferred from fingerprint identity, producing deep contigs of overlapping clones. Each contig is then to be anchored to known chromosomal positions by use of the abundant public information on BAC clone map positions, or by cross-hybridization with the YAC contigs already established for chromosomes 4 and 5 at the John Innes Centre in the U.K. Once a genome-wide set of BAC contigs is available, a minimal tiling path can be calculated and many clones can be sequenced in parallel. This approach requires the same degree of preparative work as BAC end sequencing for a comparable cost, but has the advantages of providing a physical map to the *Arabidopsis* community prior to the completion of the genomic sequence, and also will allow parallel sequencing of clones rather than the necessarily sequential sequencing using BAC end sequences. In addition, this method will allow gaps to be identified in advance of sequencing in the gapped region, and thus may allow a longer time to close gaps before they become a critical problem with sequence completion.

So far an estimated 71 MB of the perhaps 120 Mb nuclear genome is contained in 66 BAC contigs, which contain 10,840 BAC clones. The chromosome totals are:

Chromosome	Mb	Contigs
1	22.5	13
2	>4	7
3	17.0	11
4	15.3	8
5	13.4	8

The current rate of BAC clone fingerprinting and editing is 15 Mb per month. It is expected that all 22,000 available BAC clone will be added to this map by the end of 1998. Concentration at present is on chromosome 5, where the CSHSC is sequencing, and chromosome 3, where Genoscope plans to sequence using the CSHSC contigs.

The CSHSC is committed to sequencing the top of chromosome 4 and a region of approximately 4 Mb around the centromere and on the north arm of chromosome 5. Sequence data has been contributed by all three collaborating partners. Totals finished so far are 690 kb from ABI, 1.22 Mb from CSH and 1.64 Mb from Washington University, adding up to 3.54 Mb (with overlap subtracted). In addition to this, approximately 3 Mb of sequencing is in progress, making a total of more than 6.0 Mb in 61 BAC clones and 1 YAC. If this rate were to be continued, the proposed chromosome 4 region could be completed by the end of 1998, with chromosome 5 region completion either 1998 or early 1999.

ESSA (European Scientists Sequencing *Arabidopsis*;
http://mips.gsf.de/proj/thal/db/index.html)

The ESSA project is in three phases. Phase I, which is complete, was to sequence two contiguous regions on chromosome 4. One, surrounding the FCA genetic marker, is 1.92 Mb (Bevan et al. 1998, Nature 391:485), the other, around the genetic marker AP2, is 0.41 Mb, for a total completed ESSA I sequence of 2.33 Mb. ESSA II, which is to be completed in October 1998, has the goal of completing a 5 Mb region on the long arm of chromosome 4. So far 3.16 Mb is completed and annotated, an additional 1.73 Mb completed and in annotation phase, for a total of 4.89 Mb sequenced. Another 0.24 Mb is nearly complete, for an overall total of ESSA II complete and nearly complete contiguous sequence of 5.13 Mb. The ESSA I and ESSA II total of completed and nearly completed sequence is thus 7.46 Mb.

The two-year ESSA III project begins in August, 1998. Its goal is to complete the sequence of the long arm of chromosome 4 (estimated to total 13 to 13.5 Mb) and to sequence two regions of the north arm of chromosome 5 (with others to be done by CSHSC and Kazusa), with a total goal of sequencing 9 Mb.

The ESSA procedure is to use the existing YAC contig maps of chromosomes 4 and 5 to group BAC clones in bins according to their YAC cross-hybridization, then to use SalI digestions and pulsed-field gel electrophoresis followed by blotting and iterative hybridization with BAC clones to establish both BAC contigs and an overall SalI restriction map of both chromosomes. A minimal BAC tiling path is then defined and called the "sequence ready map,", the clones from this map are then sent to one of 9 collaborating sequencing laboratories for nucleotide sequencing. The data are collected and annotated at MIPS, the Munich Information Center for Protein Sequences.

The only problems encountered so far have been two difficult clones, one with a large hairpin and the other with a large region of tandem repeats. Both have been nearly completed, with the tandem repeats solved by long PCR as a supplement to the shotgun sequencing.

Kazusa DNA Research Institute (http://www.kazusa.or.jp/arabi/)

The Kazusa Institute is engaged in sequencing the long arm of chromosome 5 and along with ESSA and CSHSC, portions of the short arm of this chromosome (totaling 17.2 Mb when complete), and they are beginning the sequencing of the long (13.2 Mb) arm of chromosome 3.

The clone libraries used are from the Mitsui Plant Biotechnology Research Institute, and consist of P1 and TAC clones. Clones from these libraries are initially selected by cross-hybridization to mapped clone markers. The clones are then anchored on the YAC contig (for chromosome 5 clones), and fingerprinted as an integrity check. They are then shotgun sequenced, assembled, and annotated. A collection of YAC, TAC and P1 clone end sequences has been made for tiling the chromosome 5 clones, it includes 1254 sequences from 690 CIC YAC clones and 706 sequences from 389 P1 or TAC clones on chromosome 5. Similar methods for chromosome 3 are starting, using the YAC contig map of that chromosome produced by D. Bouchez and collaborators at INRA. At present, two large contigs for chromosome 3 exist, one of 13.6 Mb for the long arm, and one of 9.2 Mb for the bottom arm.

Progress to date has been the release of 8.89 Mb of completed, annotated sequence, with release of an additional 1.60 Mb scheduled by August 1. Thus by August 1, 1998, 10.49 Mb will have been completed and released. 10.15 Mb of this is on chromosome 5, 0.34 Mb on chromosome 3. An additional 2 Mb of chromosome 5 sequencing is in progress. At current rates of 700 to 800 kb per month, it is expected that 27 months will be required for completion of this part of the project, which is estimated to include (in addition to the 10.49 Mb to be completed by August 1) 7.05 Mb of chromosome 5 and 13.3 Mb of chromosome 3. Genoscope has proposed to do 5 Mb of the long arm of chromosome 3 (see below), if they are able to take this on (a matter now being considered there, and dependent upon the demand for their resources by human genome sequencing) the total sequence proposed by Kazusa will be reduced, and completion will be expected within 2 years.

Genoscope (Centre Nationale de Sequencage;
http://www.genoscope.cns.fr/externe/arabidopsis/Arabidopsis.html)

Genoscope is involved in the second European project. They have already provided BAC end sequences totaling approximately 11,500 completed end sequences, with plans to provide 2,000 more. Once this is complete 91% of the 22,000 BAC clones used in the sequencing project (from the IGF and the TAMU collections) will have available end sequences.

Their sequencing plan is to use the Bouchez chromosome 3 YAC contigs to make a minimal BAC tiling path by use of fingerprints done at Genoscope and at CSHSC, then to sequence the bottom (9 Mb) arm of chromosome 3. Complete contigs for this region have been supplied by CSHSC. 16 different European sequencing groups are receiving the BAC clones from Genoscope, and the data are returned to MIPS for annotation and entry into a public database. The sending out of clones is to begin within weeks, and completion of the 9 Mb region is expected by the end of 2000.

Genoscope has in addition explored with Kazusa the possibility of sequencing an additional 5 Mb on the top arm of chromosome 3; their ability to do this will depend upon the amount of their sequencing capacity that will be required to do their part of human chromosome 14, and their ability to generate extra sequencing capacity. A decision on whether Genoscope or Kazusa will sequence this 5 Mb is planned for September, 1998.

Summary of Progress

Chromosome	Est. Size (Mb)	Complete (Mb)	Group
1	~30	4.02	SPP
2	14 (+rDNA)	4.83	TIGR
3	23	0.34	Kazusa & Genoscope
4	17 (+rDNA)	9.02	ESSA & CSHSC
5	~30	10.15	Kazusa, CSHSC, ESSA
TOTAL	~114 Mb +rDNA	28.36	

In addition, shotgun sequencing libraries are in preparation for an additional 2.80 Mb, and sequencing is in progress but not yet complete for an additional 2.98 Mb. Furthermore, 36,574 BAC ends from 18,746 clones, representing 13.64 Mb, provided by TIGR, SPP and Genoscope are completed, as are 1254 end sequences from 690 CIC YAC clones and 706 sequences from 389 P1 or TAC clones, provided by Kazusa.

COMPLETING THE SEQUENCE

Defining completion

In addition to the gene-rich and highly informative regions of the genome (with one gene every 4-5 kb), there are regions of repetitive DNA, and perhaps of lower gene density.

One instance is the ribosomal DNA repeats, which are arranged in two uninterrupted tandem arrays. Each repeat unit contains a gene for 18S, 5.8S and 25S structural ribosomal RNAs and is 10-10.5 kb in length. The large tandem arrays of repeat units are found at the top arms of chromosomes 2 (NOR2) and 4 (NOR4). Each is on the order of 3-3.5 Mb, or 300-350 repeat units.

Centromeric regions are only beginning to be defined at the molecular level in *Arabidopsis*, but cloning and chromosome *in situ* hybridization studies have shown that these regions contain multiple tandem repeats of short sequences, a major element of which is 180 bp repeats and related repeats. In one case (chromosome 1) an estimate of the repeat length is 950 kb. For chromosome 4 the functional centromere is probably on one side of a 180 bp repeat region, and so far does not seem to be unclonable. There is some indication that BAC clones from this region may have a higher amount of repetitive sequence in tandem arrays than other BAC clones sequenced to date, and one BAC clone from the chromosome 2 centromere region has only 3 genes, a much lower density than the typical 1 gene per 4-5 kb found elsewhere. Another BAC from the centromere region of chromosome 4 has a more typical density.

Telomeres and subtelomeric regions in *Arabidopsis* have been characterized and appear to be small (totaling perhaps 100 to 200 kb in the genome) and not difficult to sequence so far.

There are also small regions of simple tandem repeats, as for example as described above in the ESSA project progress report. This clone, BAC F9F13, contained 10 tandem copies of a 3.5 kb repeat, as well as 2 additional copies of the same repeat.

Because the exact sequence and number of tandem repeats is not thought to be consequential for any functional analysis, and in fact is quite polymorphic between ecotypes, it was decided that a sufficient characterization of these repeats would be a sequence of one subunit, and an estimation from blotting or long-range PCR of the number of tandem copies at each site.

Given this, the complete sequence of the nuclear genome will be considered to be in hand when each chromosome arm is fully sequenced as a single contig from subtelomeric repeat to "centromeric" tandem repeats, with internal tandem repeat regions (including rDNA repeats) characterized only as far as demonstrating that they are pure tandem repeats, with the sequence of one repeat unit determined, and an estimate of repeat number at each site provided. This characterization already exists for the rDNA repeats (Copenhaver et al. (1995) Plant J. 7:273-286). This definition may have to change if unclonable regions are found, or if non-tandemly organized but nonetheless impossible to sequence (with available relevant technology) clones are found. To date there is no indication of either unclonable regions or of clones impossible to sequence for reasons other than large numbers of small tandem repeats.

Other sequence parameters

Accuracy

All of the participants have agreed before, and continue to agree, that the standard for sequence accuracy should be one error in 10,000 nucleotides or better, and the projects so far seem to be achieving this goal. The U.S. groups agreed to a common pair of tests to monitor sequence accuracy. The first would be using base calling programs such as Phred (Ewing et al. (1998) Genome Res. 8:175-185) or TIGR Assembler to assess sequence accuracy in each sequencing run. The second is to independently determine the sequence of all regions of overlap between adjacent clones, and only after sequence finishing to compare them for mismatches. This serves as an independent method to determine sequence accuracy, and since all mismatches are to be resolved by further analysis, this test will in addition indicate the degree of sequence change due to mutation in the clones being used for sequencing.

The European and Japanese groups have different methods to measure sequence accuracy, but have the same goal of less than one error in 10,000 bases.

Annotation

Proper annotation of sequences to indicate the position, structure and nature of each of the coded genes is a critical component, and in fact the primary product, of the genome project. It is clear, though, that initial annotation of sequences is not fully (or even very) accurate, as the software and algorithms used for gene recognition can miss exons and introns, and can also indicate the presence of exons or introns where there are none. This is as true in animal genome projects as in plant projects. Thus, annotation will have to be done in stages, with initial annotations that can be useful, but that must be acknowledged to be flawed.

Each of the sequence groups performs its own annotation, as this is not only an interesting part of the work, but also helps with continued sequencing. It was agreed that, to provide the highest quality initial annotation, each group would use multiple software programs for gene recognition, and would indicate in its output the product of each of the programs (something that GenBank cannot do; thus this requires output to be in a form other than that sent to GenBank or equivalent public databases). It should be emphasized that doing this does not remove the requirement for inclusion of the output in public databases like GenBank or DDBJ. In addition, experimental means of annotation are to be used by each group - that is, sequences must be compared with the EST sequences that are available and that indicate actual RNA sequences, and must be compared with the genes of known structure that have been individually studied. Furthermore, feedback from the community of *Arabidopsis* researchers should be invited by each group, to allow correction or improvement of each group's annotations.

As the genome project proceeds, it is important to consider additional experimental methods for gene recognition, and the application of such methods should be considered important goals for the project. Among the experimental methods to be considered is sequencing of related genomes (such as those of *Arabis lyrata* or *Cardaminopsis petraea*, see http://www.arabis.net/wild.htm). Because exonic sequences change more slowly than intronic or intergenic sequences, this could serve as a very useful indicator of gene location and exon boundaries. Additional experimental means for improving annotations include RNA blots and RT-PCR to find if suggested genic sequences in fact correspond to RNAs, and full-length sequencing of large numbers of cDNA clones for comparison to genomic sequences.

Maintenance of summary lists of identified genes according to the type of protein coded (see Bevan et al. 1998, Nature 391:485) is also an important aspect of annotation.

Because annotation methods and the experimental information on which they are based is subject to continual improvement, frequent reannotation is worthwhile. Both the Kazusa and TIGR groups have plans for systematic reannotation of sequences from all groups. To facilitate this and, especially, to facilitate community access to annotations, it was agreed that all groups would work toward a standardized format for data presentation, and that groups doing large-scale reannotation would make their data freely available for mirroring on the web sites of all groups that wish to display them.

Data release

Each of the U.S. groups sends sequence out unannotated and in small fragments as soon as it reaches either approximate 2 kb contigs or 7x average coverage. The sequences from two of the three groups are sent at this stage to the high throughput genome sequence (HTGS) part of GenBank, the third group has agreed to start doing this as well. The sequences are now sent to each group's own web page, each of which supports BLAST searches, and are also sent at short intervals to TAIR, the public *Arabidopsis* database, where they are also BLAST searchable (http://www.arabidopsis.org/Blast).

The structure of the European projects, where sequence-ready clones are allocated to many groups, and each group has some discretion (and rules from their own national government) in how to sequence and when to submit completed sequence, does not lend itself to identical release methods or policies. Nonetheless, the groups agree to collect and distribute through MIPS and TAIR all sequences as soon as practicable, at latest after completion and before annotation.

The Japanese group also has its own policies and level of funding for informatics, which so far have dictated that sequence be released only after both completion and annotation, and then posted to DDBJ (DNA Database of Japan) and GenBank. This entails a delay in public access relative to other groups, as the time from completion to annotation is about a month, and the time from acquisition of the earliest data to completion is also appreciable. The Japanese group will consider mechanisms for earlier release, within the constraints of policy and of funding for this aspect of the project.

Clone registration (intention to sequence)

One critical aspect of the project is coordination between groups on the clones to be sequenced, as without tight coordination, duplication of effort will occur, especially in the closing phases of the project. In addition, as different groups complete their assigned regions, reallocation of regions may become necessary so that groups ahead of their predicted rate can help by sequencing clones originally assigned to other groups. At present this coordination has been supplied by direct communication between the groups, and by the function of an international coordinating committee of the *Arabidopsis* Genome Initiative (AGI: see http://genome-www3.stanford.edu/cgi-bin/Webdriver?MIval=atdb_registry_info.html). This committee will remain the arbitrator of international sequencing efforts, but will be supplemented with a new committee that will allow for closer coordination of the U.S. groups. This new committee has been mandated by the U.S. funding agencies, as a replacement for the three separate advisory groups that now exist, one for each group.

One of the tasks of the U.S. committee will be clone reallocation, and in addition frequent communication with the members of the international AGI committee, as a way of stimulating continued discussion among all groups. As representatives of all groups will be invited to the meetings of the U.S. committee, these meetings may also be able to serve as a forum for discussion and decisions of the AGI committee. This may help the AGI by increasing the frequency of its considerations.

NEW U.S. STEERING COMMITTEE

Given the important new role of the mandated U.S. Steering Committee as arbitrator and communication facilitator between the U.S. groups, and as aid to the AGI committee on the international front, the role a responsibilities of the committee were discussed and agreed upon.

The U.S. Steering Committee will have the following responsibilities:

1) Setting boundaries between the U.S. sequencing groups (ideally, to be defined by sequenced clones) to avoid duplication of effort in chromosomes where more than one group is working

2) Reallocation of clones or chromosome regions from one group to another to fit sequencing capabilities to the remaining work.

3) Monitoring and enforcement of the common agreements described earlier in this report, namely the agreement to work toward a common annotation format, to provide quality control information both from base calling programs and from clone overlap regions, and to monitor sequence release compliance.

4) Providing annual progress reports to the *Arabidopsis* community and to the U.S. funding agencies, separate from the progress reports of each of the individual sequencing groups. These reports will include a careful consideration not only of amount of sequence provided by each group, but of progress in all respects, balanced so that groups taking on difficult clones to sequence, or who are in closing phase and thus must devote time to closing gaps, are given full credit for such efforts. In addition, these reports are to detail progress in the informatics aspects of the project, including a summary of the progress and needs of the *Arabidopsis* database - as an interface between the database and its advisory committee, the sequencing groups, and the *Arabidopsis* community.

5) Provide an interface between the U.S. groups and the international AGI committee, and act to facilitate the setting of boundaries and clone reallocation at an international level.

6) The committee should endeavor to meet in person at least once a year, and have regularly scheduled meetings by electronic mail or conference call.

The composition of the committee is as follows:

Members:

- 3 members of the U.S. *Arabidopsis* community, initially appointed; with rules for succession and for input from the North American *Arabidopsis* Steering Committee. Chairmanship will rotate among these members annually.
- 1 non-U.S. member of the international *Arabidopsis* research community, to be appointed by the chair of the Multinational Steering Committee
- 2 genome sequencing experts from projects other than the *Arabidopsis* project
- 1 expert in genome databases
- 1 principal investigator of the *Arabidopsis* genome database TAIR

Ex officio:

- representatives of each of the six genome sequencing groups
- representatives of U.S. and international funding agencies

The actual members of the committee who have so far agreed to serve:

Elliot Meyerowitz, chair (U.S. *Arabidopsis* community)

Daphne Preuss (U.S. *Arabidopsis* community)

Gerd Jürgens (international *Arabidopsis* community)

Ex officio:

Joe Ecker, SPP

Dick McCombie, CSHSC

Steve Rounsley, TIGR

Ian Bancroft, ESSA III

Francis Quetier, Genoscope

Satoshi Tabata, Kazusa

Recommendations for the other members were:

Joanne Chory, Pam Green or Detlef Weigel (U.S. *Arabidopsis* community)

Mark Johnson, Richard Gibbs, John Sulston, Maynard Olsen (sequencing experts)

Mark Boguski (database expert)

Mike Cherry (TAIR representative)

FINAL PROSPECT

Given sufficient funding, which seems very likely, there is no technical obstacle to the completion of the *Arabidopsis* nuclear genome sequence by December 31, 2000. Although the efforts of the project members must be focused tightly on finishing the sequencing, it is not too early to begin considering the next steps, among them experimental methods for annotation, and functional analyses of genes and gene families.

submitted by:

Elliot M. Meyerowitz July 15, 1998

Appendix 2

Summary of December 1998 AGI Meeting at CSHL

1. Daphne Preuss summarized her work on centromeric regions and presented detailed information on approximate map locations of BAC contigs and sequenced BACS based on hybridization (Altmann) and fingerprint (WashU) data. She agreed to make this information available to the community. Rob Martienssen stressed that individual clones would need to be compared closely with fingerprint contigs constructed at WashU because some hybridization data were unreliable.

2. Each group discussed their estimated sequencing capacity and assigned chromosomal regions for the coming year. Kazusa expects to finish their assigned regions on III and V by the end of 1999. ESSA and CSHL/WashU may also complete their assignments on IV and V at about the same time. SPP is continuing with chromosome I and was encouraged to avoid starting many additional nucleation points in order to focus on the same closure issues being addressed by the other groups. Genoscope has begun sequencing the bottom arm of III and will continue with this region through 2000. TIGR expects to finish chromosome II by summer 1999 and will therefore be the first funded group to run out of an assigned region to sequence.

3. AGI members discussed the importance of finishing difficult areas within assigned regions of the genome while also continuing to make rapid progress on other regions to maximize release of information to the community.

4. Both TIGR and Kazusa proposed to begin sequencing the "unassigned" top 5-6 Mb of chromosome III during 1999. After considerable discussion, both at the AGI meeting and later in the conference when Satoshi Tabata arrived, a consensus was reached to have TIGR begin

sequencing this region of chromosome III during the spring of 1999 with the aim of finishing this region by January 2000.

5. Starting in January 2000, TIGR, Kazusa, CSHL, and ESSA will likely have residual sequencing capacity ready to shift to centromeric regions and portions of chromosome 1 that have not yet been completed. By this time a minimal tiling path based on fingerprint data should be available to facilitate assignment of remaining BACs to AGI members. SPP has funding to complete most or all of chromosome I but recognizes that the entire genome

may be completed more rapidly if other groups contribute in the year 2000 to sequencing portions of this chromosome (or possibly part of the bottom of chromosome III depending on progress made by Genoscope) after their own assigned regions have been essentially completed.

6. Marcel Salanoubat and Francis Quetier led a discussion of the Genoscope policy for sequence release. While it was clear that the informatics capabilities of the individual laboratories in their program varied significantly, there was a general agreement that the group should strive for immediate release of sequences (at least for the bigger laboratories within their program).

7 . Rob Martienssen and David Meinke discussed the status of the CSHL/WashU consortium plans to continue sequencing and fingerprinting efforts. NSF has now received all of the necessary paperwork for continued funding of this consortium and expects to make an award at a level sufficient to enable sequencing another 2.4 Mb per year starting early in 1999. In addition, NSF has recommended funding an informatics person at WashU to finish editing of fingerprinted contigs and establishment of an interactive version of the BAC physical map that can be accessed via the Internet. This person will work closely with TAIR to avoid duplication of effort.

8. The CSHL/WashU group has agreed to release to other sequencing groups all of their edited contig information and fingerprint database through their ftp site no later than the end of January, 1999. The SPP and TIGR groups are particularly anxious to make use of this information in order to avoid repeating the contig-building steps that have already been completed elsewhere. Rob Martienssen agreed to provide as soon as possible a minimal BAC tiling path for regions of the genome that may require coordination during the final year of the project..

9. Joe Ecker and David Meinke discussed a proposal by Hiroaki Shizuya at Caltech to fingerprint and end-sequence a new BAC library with large inserts (180 kb average). The general consensus was that although this library might be very useful in regions of the genome with minimal coverage and could reduce the overall cost of sequencing other regions by reducing overlaps, it was unlikely that many AGI participants would immediate move away from using TAMU and IGF clones for the bulk of their sequencing efforts. NSF is willing to discuss further the potential value of this library with interested AGI members.

10. Rob Martienssen agreed to serve as the next AGI chairperson. There was general agreement that AGI members should meet again in summer 1999, perhaps at the next Arabidopsis meeting in Australia, to assess progress and make specific plans for the future.

Joe Ecker, AGI chairperson

Appendix 3

DATABASE NEEDS OF THE ARABIDOPSIS COMMUNITY

I. VENUE AND PARTICIPANTS

To assess the current and future database needs of the Arabidopsis community, an NSF-supported workshop on this topic was convened in Madison Wisconsin on June 28, 1998. The workshop participants included the following individuals:

Rick Amasino, University of Wisconsin
Mary Anderson, Nottingham University
Mike Cherry, Stanford University
Joanne Chory, Salk Institute
Maarten Chrispeels, University of California San Diego
Jeff Dangl, University of North Carolina
Keith Davis, Ohio State University
Allan Dickerman, National Center for Genome Research
David Flanders, Stanford University
Pam Green, Michigan State University
Bertrand Lemieux, University of Delaware
David Meinke, Oklahoma State University
Larry Parnell, Cold Spring Harbor Laboratory
Daphne Preuss, University of Chicago
Ralph Quatrano, Washington University
Ernie Retzel, University of Minnesota
Steve Rounsley, The Institute for Genomic Research
Randy Scholl, Ohio State University
Chris Somerville, Carnegie Institution of Washington and Stanford University (chair)
Desh Pal Verma, Ohio State University

The following individuals provided valuable written comments prior to the meeting (Appendix I):

Jean Greenberg, University of Chicago
Katie Krolikowski, Harvard University
Russell Malmberg, University of Georgia
Jose Martinez-Zapater, Biology Molecular y Virologia Vegetal, CIT-INIA
Natasha Raikhel, Michigan State University
Pierre Rouze, Flanders Institute of Biotechnology
Chris Town, Case Western Reserve University
Desh Pal S Verma, The Ohio State University

In addition, the workshop was attended by the following observers:

Peter Bretting, USDA/ARS National Program Staff
Greg Dilworth, Department of Energy
Machi Dilworth, National Science Foundation
Margarita Garcia, Stanford University
Paul Gilna, National Science Foundation
Xiaoying Lin, The Institute for Genomic Research
Bob MacDonald, US Department of Agriculture
DeLill Nasser, National Science Foundation

II. GOALS

The general goals of the workshop were to examine the present and future database needs of the Arabidopsis community and to outline in general terms the main issues which should be addressed in any future proposals concerning the development of new or expanded Arabidopsis databases. The discussions were intentionally focused on biological and community issues and there was no attempt to define or specify issues which are related to specific computer hardware or specific database programs. In particular, no assumptions were made concerning continued government funding of any current Arabidopsis database activities.

A previous workshop with these goals was held on June 5th and 6th, 1993. A copy of the published summary that workshop was provided to all participants and served as a reference to earlier views and objectives of the Arabidopsis community. [1993 Dallas Workshop Report] In addition, participants were provided with a draft summary of a BBSRC-USDA bilateral plant bioinformatics and coordination meeting held at Llangollen Wales, March 22-24, 1998. A copy of a memorandum, dated February 26, 1998, from the North American Arabidopsis Steering Committee to the curators of TAIR, concerning the current Arabidopsis community database needs was also provided. [NAASC Memorandum] Finally, in preparation for the meeting, written comments solicited from the community on the Arabidopsis electronic newsgroup were provided to the participants before the meeting. A copy of the solicitation and written comments are appended as Appendix I.

III. RATIONALE FOR AN ARABIDOPSIS DATABASE

The genomes of higher plants, such as Arabidopsis, contain approximately 25,000 genes. During the next several years, the sequence of the Arabidopsis genome will be completed and extensive sequence information will become available for many other species, including many plants. Most or all of the Arabidopsis genes will be used to develop gene chips or microarrays that permit simultaneous measurements of the expression (mRNA levels) of all of the genes. These will be used to generate information about the expression of all the genes in the organism in response to a wide variety of treatments and genetic backgrounds. Each experiment could have as many as 25,000 data points for each time point or treatment of each genotype! Comprehensive libraries of insertional mutations will permit the isolation, by reverse genetics, of null mutations in any Arabidopsis gene. Extensive collections of enhancer-trap or promoter-trap lines are being developed that permit sensitive analyses of the spatial patterns of gene expression down to the single-cell level. Thousands of new classes of mutants will be isolated by selecting for suppressors or enhancers of existing mutations. The corresponding genes will be cloned by very high resolution mapping of the mutations so that a limited number of candidate genes which are evident in the delimited region of genomic sequence can be directly tested for complementation. This will depend on the development of very high resolution maps. It seems likely that high resolution proteomics methods will become important for identifying the substrates of the thousands of kinase genes that form many of the regulatory networks in Arabidopsis and other plants. Additionally, extensive genomic-based work in other plant species will produce a flood of sequence information. The value of much of that information will be greatly enhanced by comparison with the aggregate information available in Arabidopsis. Thus, we are entering an era of explosive growth of knowledge about Arabidopsis in particular, and plants in general. Most of the data generated by the projects described above will never appear in printed journals and will only be available to the community through electronic databases.

Because Arabidopsis is one of the most intensively studied organisms, and is a direct model for 250,000 closely related species, we believe that it is appropriate to undertake a major investment in developing new information retrieval tools (IRTs) for Arabidopsis in particular and plants in general. By this we mean that because we will know everything about Arabidopsis, it is a suitable object on which to focus the building of a comprehensive database or set of linked databases. However, because the value of Arabidopsis derives from its utility in understanding other plants, it would be desirable to build a structure that permits facile high resolution linking of specific information about Arabidopsis to all other plants.

Looking into the future more generally, it is apparent that scientific publishing is undergoing a much needed revolution. All of the major journals will be electronic within a few years and once that transition is complete, scientists will develop new tools for interacting with data. The complexity of biological knowledge in many fields is such that new mechanisms for integrating data are required. The development of computer programs that calculate genetic maps "on the fly" from currently available data is an early example of what will become a more general mechanism for integrating data. Integrated graphical representations of patterns of gene expression in individual cells of three dimensional models of organisms at various developmental stages is another example that is under development. With such a model it will be possible to find relationships between objects (eg., genes) and processes that would be difficult or impossible with current information retrieval technologies.

Because of the changes taking place in publishing, there may be an opportunity to develop databases that will eventually be self supporting in the same way that journals are self supporting. As the distinction between the format blurs, the concept of paying for a database subscription will become commonplace. However, there are many complex issues associated with imposing charges for database use and the question is largely academic at present.

There are many challenges in developing a new generation database. Perhaps the foremost is the difficulty in collecting information from the thousands of scientists who produce primary information for conventional publication in journals.

IV. CURRENT PUBLICLY SUPPORTED DATABASE ACTIVITIES

The principal publicly supported Arabidopsis database activities are the TAIR database at Stanford University and the stock center databases maintained by the Arabidopsis resource centers at Ohio State University and the University of Nottingham. In addition, the University of Minnesota supports an EST database for all plants, and each of the Arabidopsis genome sequencing groups provides database access to genomic sequences, including BAC end sequences.

The TAIR goal is to provide the plant-biology research community with convenient and correlated access to the publicly available results of Arabidopsis research. This includes published and otherwise freely available information about the genome, the genes it contains, the gene products, their positions on genetic and physical maps, as well as DNA sequences. The users of the database are very diverse, ranging from Arabidopsis molecular biologists to biologists focusing on any other organism. The members of the TAIR project are currently shared with the Saccharomyces Genome Database, and the database administrator is shared with the Expression Microarray database and Genetic Footprinting database projects, all located at the Department of Genetics at Stanford University. In an effort to minimize wasteful duplication of effort, the TAIR project uses much of the same software and staffing structure as the Saccharomyces Genome Database (SGD). The combined SGD and TAIR groups thus benefit from an economy of scale by sharing computing and human resources.

At a meeting of the Arabidopsis genome community in 1992 at the Cold Spring Harbor Banbury Center, a consensus was reached that TAIR should take responsibility for providing centralized access to Arabidopsis databases, a recommendation that has been repeatedly endorsed by the North American Arabidopsis Steering Committee. Since that time TAIR has been supported by a grant from the National Science Foundation. However, the annual level of support for TAIR has been only a small fraction of the support provided for database activities for similarly advanced models such as Drosophila, yeast and mouse.

V. SUMMARY OF CONCLUSIONS AND RECOMMENDATIONS

- The main conclusions from the workshop were consistent with the conclusions of the 1993 workshop.
- The Arabidopsis community has a large number of unmet needs for database services that are required to make efficient use of existing information.

The highest priorities for database content are:

- Integration of the physical map and genetic map
- Detailed and consistent annotation of the genomic sequence
- Gene chip and DNA microarray data
- Information about forward and reverse genetics
- Spatial and temporal information about gene expression and tools for visualizing such information.
- Protein localization and proteomics
- Phenotypic information about mutants

- Federal government support for increased Arabidopsis database capabilities is of crucial importance to continuing progress in understanding all aspects of plant biology. This knowledge is a vital component of the mechanisms that support continuing agricultural productivity, environmental stewardship and development of a robust agricultural biotechnology industry. Support for such databases should be long-term and contingent upon guarantees that all information in the databases will be freely available to the international scientific community.
- The main focus of Arabidopsis database activities should be service components. However, in those cases where publicly available computer programs do not meet the needs of the Arabidopsis community, adequate resources should be made available to support the development of components that are required to fully implement the service functions of Arabidopsis databases.
- Arabidopsis databases should provide an intellectual focus for the interpretation, synthesis and integration of biological data. The value of such a resource will be proportional to the ability of the databases to acquire all relevant data. Since past experience has indicated that it is not feasible to rely on the community to submit all useful information to databases, the Arabidopsis databases must be professionally curated by paid curators. In addition, mechanisms must be explored for obtaining data directly from authors in conjunction with publication in journals. The development of user friendly internet-accessible data entry forms that would allow direct deposit of information by members of the community into public databases must also be pursued.
- US Arabidopsis databases activities must be linked to the community through an oversight committee that includes representation from, and is approved by, the North American Steering Committee.
- Arabidopsis databases must be accessible internationally via the internet using commonly available internet browsers.
- Although all Arabidopsis information need not be archived in a single database, it is essential that all Arabidopsis databases be able to seamlessly communicate with each other and with other major databases, such as the nucleic acid databases.
- Because the databases are a world resource, an effort should be made to coordinate the development of Arabidopsis database activities internationally with a view to sharing the costs of curation and development of new tools.
- It is not desirable or appropriate to attempt to implement partial or complete cost recovery of Arabidopsis database services in the forseeble future.
- An encyclopedic database which interrelates all aspects of Arabidopsis biology remains an attractive long-term goal.

VI. WHAT SHOULD BE IN THE DATABASES?

The long-term goal is to provide interconnected access to all information about Arabidopsis. However, certain classes of information should have a higher priority for immediate inclusion and also require a high degree of curation in order to be most useful to the community.

A. Map-Based Information

At present, many laboratories are engaged in cloning genes by map-based cloning methods. The use of map-based cloning is expected to continue indefinitely and to become the most widely used method of cloning genes in the future. The ease with which this can be accomplished is directly proportional to the availability of information about genetic and physical maps, polymorphisms, and large clones. Thus, the greatest current need is a unified genetic and physical map that incorporates all available information about polymorphic markers (eg. CAPS, SSLPs, RFLPs), mutations, BAC and YAC clones, mapped clones and insertions or other modifications of the genome.

Because of the pending completion of the genomic sequence, the state of the genetic map is expected to change dramatically during the next several years as sequence-based markers become anchored on the genomic sequence. The availability of the sequence information will enhance the value of the integrated map because it will stimulate map-based cloning efforts which will remain dependent on a high density of polymorphic markers. The integration of the genetic and physical maps should be undertaken by a group with appropriate expertise in both genetic and physical maps and database management and curation.

Ready, access to primary mapping data should be given highest priority in database development. Map information should be collected and presented in a manner that allows the user to determine what is known, plus what remains questionable or unresolved with respect to map locations of genetic and molecular markers in combination with a complete physical map anchored to the complete nucleotide sequence. In constructing the database, it should be remembered that recombination data generally provide only rough estimates of map location, and that mapping data may differ widely in quality and reliability. Therefore, some database users may prefer direct access to primary mapping data in order to compare their results with those obtained in other laboratories. A database that provides options for visualizing several different maps constructed with different mapping functions or subsets of markers and primary mapping data would be particularly valuable to the Arabidopsis community.

Any proposal for database development should also discuss in some detail how the integrity of these maps would be verified and maintained. Some mutations and cloned genes are likely to be known by several different names. It will therefore be important to establish a database that will accommodate multiple changes in nomenclature. Other plant databases are moving toward the use of standard gene names as described in the Mendel database. The Arabidopsis databases should also adopt this policy to ensure compatibility with other databases.

Provisions should also be made to add new types of information to genetic and physical maps as they become available (break points of chromosomal aberrations; regions of extensive heterochromatin; regions with a high/low degree of sequence homology to related plants; etc.).

B. Sequence information

The value of the genomic sequence will depend on the quality of the annotation. The goal for the quality of annotation should be similar or identical to that of other higher organisms. It should be possible to arrive at an integrated map of a gene by various routes. A user should be able to begin a query with a sequence, a gene name, a keyword or a genetic map location. A user should be able to highlight a region of the genome on a graphical display and move to increasingly higher levels of resolution with the click of a mouse. For example, one might start with a whole chromosome, then move to a ~10 cM region which shows the contigs of BACs and YACs, the mapped mutations, the sites of insertional mutations or launching pads for transposons. Next the user should be able to visualize a ~1 cm region showing all of the above features plus the locations of open reading frames (theoretical and verified), ESTs, polymorphic markers, potentially polymorphic markers (ie,. SSLPs). Finally, at the next level of resolution the user should be able to visualize the DNA sequence, the various putative open reading frames indicated by gene finding programs, experimentally verified genes, ESTs, BAC and YAC end sequences, polymorphisms, mutations and other known aberrations. The open reading frames should be linked to information about gene expression, experimentally verified information about gene function, mutant phenotypes associated with classical mutations or over or under expression, theoretical information about gene function based on inference from other organisms, subcellular localization of the gene product, known or predicted modifications of the gene product. If there are other genes of similar structure in the genome, the presence of these genes should be indicated. Similarity to genes from other plants should be indicated with a link to the appropriate databases. The control regions of the genes should be annotated with known or predicted motifs and with information about the identity of other genes with similar motifs.

The sequence information should not simply be a link to raw sequence in GenBank because the level of annotation and tools to manipulate that sequence do not directly support the kinds of queries made by most biologists. Thus, the sequence should be directly available from a specialized database which provides useful tools for manipulating the sequence. It should be possible to retrieve from the database sequence information based on map position, type of sequence, or other specific requirements. All information should be linked to publications describing the data when possible.

Because the sequencing groups are not expected to have the resources to provide continued annotation, there will be a need for a group to take responsibility for continued upgrading of the annotation of the genomic sequence as information about the sequence becomes available from direct experimentation and from computational analyses based on experimental results obtained with other organisms.

C. Expression information

The use of microarrays and gene chips are expected to provide a massive amount of new information. Most or all of the Arabidopsis genes will be used to develop gene chips or microarrays that permit simultaneous measurements of the expression (mRNA levels) of all of the genes. These will be used to generate information about the expression of all the genes in the organism in response to a wide variety of treatments and genetic backgrounds. Each time point or treatment could have as many as 25,000 data points. Because the experiments are technically straightforward, it seems likely that a common type of experiment will be to prepare mRNA from a mutant and a wild type and to compare the consequences of the mutation on the expression of all the genes in the organism. In addition to simply archiving the raw data it should be possible to query the data in various ways. For instance, as data from different treatment accumulates, it will become possible to search for genes that are coregulated with a gene. This kind of query may provide insights into the identity of otherwise anonymous genes or reveal the existence of networks. It should also be possible to identify all the factors that cause altered expression of a gene, to identify all genes that specifically respond to certain treatments, to identify mutations that cause similar effects on gene expression. For these kinds of queries it will be necessary to have software that can identify data sets that are most similar from among hundreds or thousands of different data sets produced by different treatments.

There is also a large need for a repository for information about spatial aspects of gene expression. There are now many transgenic lines which exhibit specific spatial patterns of reporter gene expression, and cloned genes which confer such patterns. In the short term a database with a controlled vocabulary for the various cell and tissue types and linked images of the patterns of gene expression would meet immediate needs. In the longer term, it would be useful to have graphical tools that would integrate the patterns of gene expression into an organismic model.

D. Phenotypic Information

Because of the diversity of processes that are being analyzed by a mutational approach in Arabidopsis, there is a need for facile access to information about gene function as it relates to the organism. One aspect of the problem involves determining the genetic basis for a phenotype. In this case it should be possible to enter a description of a phenotype and obtain a ranked list of probable genetic alteration that could give rise to the phenotype. Conversely, it would be very helpful to be able to enter a gene name and obtain a description of the corresponding mutant. This capability will greatly enhance the efficiency with which new mutations will be studied as the number of known mutations begins to plateau. It is expected that we will soon have saturating collections of transposon mutants, so having ways of describing these phenotypes, and making them accessible, will be important. No capability of this kind currently exists.

One strategy may be to use organizational schemes as entry points (phenotypic indexes, so to speak). One such index is the genetic map position. Knowledge of this provides an entry point to other mutants and papers. Another possible organizing scheme could be based on the EcoCyc database format of metabolic pathways, so that biochemical phenotypes could be correlated, or the knowledge of existing pathways could be queried. The user would click on a pathway and learn what was known about this. Another way of indexing and accessing the data for development might be to have a standardized Arabidopsis growth animation - at appropriate times during the growth animation, a user could click on a graphic representation of an organ or other feature, and then this would lead to additional information. Clicking on a rosette leaf might lead to various types of leaf cells or indexed leaf morphologies.

E. Stock-Based Information

The databases maintained by the two Arabidopsis resource centers at Ohio State University and the University of Nottingham provide excellent access to information on the availability of biological and chemical materials related to Arabidopsis research. These databases have implemented many of the recommendations of the 1993 workshop report and should continue to assume responsibility for descriptive information concerning seed stocks, clones, vectors, libraries, cDNAs, oligonucleotides, and any other materials that may require distribution to the Arabidopsis community. Emphasis should be placed on careful documentation of biological materials, controlled vocabularies, and maximal utilization of sophisticated graphics to display plant phenotypes, molecular hybridization patterns, and other data where appropriate.

With respect to seed stocks, it should be possible to search the database by general phenotype, not just by gene symbol, in order to obtain a broad listing of ecotypes and mutant lines with similar features. Information on phenotypes, screening methods, growth conditions, and differences between alleles should be included for all mutants available through the stock centers. It should also be possible to obtain information on additional mutants or alleles that have been isolated in specific laboratories but are not available from the stock centers.

Individuals should be able to search for specialized libraries, vectors, transgenic lines, and molecular reagents (antibodies, purified proteins, unusual compounds, and biochemical standards) required for Arabidopsis research.

The stock center databases should be directly linked to a central Arabidopsis database so that queries about the properties of a gene or mutant can lead directly to a query about the availability of the resources used to study these or related aspects of the biology.

F. Community-Based Information

During the past several years there has been a proliferation of electronic resources that provide easy access to information on a wide range of community issues. For instance, it is now relatively easy to retrieve contact information for colleagues or previous postings on the Arabidopsis newsgroup, the abstracts for meetings are available on line and there is an electronic Arabidopsis journal, Weeds World, which provides a forum for discussion of methods and problems and publication of short papers. Many laboratories have mounted web pages that provide detailed information about specialized methods, specialized databases or collections of genetic materials. The curators of TAIR have provided convenient access to these diverse resources by providing a web page that facilitates connection to these resources.

While it is desirable to continue having one group take responsibility for maintaining a centralized launcher or "data warehouse" for Arabidopsis-related web sites, this should be a relatively inexpensive activity and should not require significant public financial support. The distinction between this activity and a database does not seem to be fully appreciated by the community. The result is that, because of the proliferation of sites which are all superficially similar, the users do not know how to efficiently find information. Therefore, it may be desirable to maintain a clear distinction between a centralized internet launcher and any future attempts to develop a unified Arabidopsis database.

G. Biology-Based Information

The focus of research with Arabidopsis is likely to change in the future from the immediate emphasis on mapping, sequencing, and gene identification, to the long-term questions of general biology and gene function during plant growth and development. Thus, there is a long-term need to develop Arabidopsis database(s) that provide facile access to information that may be of critical importance during this second phase. In proposing a vision of the future requirements one correspondent wrote the following (Appendix I):

"I envision a data base organized by levels of organization that can be addressed at different levels. This database should contain both structural and functional data organized at different levels. In this way starting, for example, with the keyword root, one can access information about root structure, root cell components, root development, nutrients uptake, etc. and end up in the interactive pathways and proteins responsible for these processes and the corresponding genes. It should also be possible to address the database by processes - for example elongation or flowering or pollination. Of course this is likely far away from real possibilities. Going down to the specifics, the information in the database could be implemented with information on pathways and networks, protein interaction maps, protein structures, subcellular organelles, cell structure, etc. This will be a way to reach to a database as described above."

Examples of topics that might be included in this category, include: information on plant pathogens that infect Arabidopsis and details on the molecular interactions that take place between host and pathogen; information on the chemical composition of specific plant parts (sugars, lipids, proteins, polysaccharides, specialized compounds, etc.); physiological data on the normal life cycle and the response of mutant and wild-type plants to various environmental and experimental treatments; protein profiles of different plant parts revealed through 2-D gel electrophoresis; information on the natural distribution and ecology of Arabidopsis and closely related species; detailed comparisons of the different ecotypes with respect to morphology, physiology, and molecular biology; information on the taxonomy of Arabidopsis with particular attention to related plants used in agriculture; light and electron micrographs of different types of cells in wild-type plants; records of expression patterns of specific genes during growth and development; and computer-enhanced reconstructions of serial sections through various plant structures.

At present it appears that the development of these resources will be best accomplished by the individual initiative of members of the community with specific knowledge and interests in specialized information of the kinds described above. The eventual integration of specialized databases of this type into a unified Arabidopsis database will be facilitated by encouraging the open exchange of schema between database developers. Therefore, public support for Arabidopsis databases should be contingent on unrestricted access to all schema and source code used in Arabidopsis databases.

VII. STANDARDS FOR QUALITY OF DATA

All data that is acquired by the databases should be available to users. However, where data is suspect or in conflict with other data, it may be desirable or necessary to provide various views of data. Thus, it may be desirable to provide a user with a curated version of a certain kind of data and an uncurated version. A specific example might be in the interpretation of open reading frames. Since the various gene finder programs do not always make the same prediction, it should be possible to provide the curators best guess as one view and the various alternatives as another view. A simple tab associated with each view would provide a convenient tool for meeting this need. It is also desirable to provide access to the primary mapping data used to position mutations and genes on the genetic map.

Publication of data should not be a prerequisite for inclusion in the databases. Indeed, the vast majority of data is unlikely to ever be available via traditional publishing methods.

VIII. HOW WILL THE DATABASE BE USED? WHAT LINKS SHOULD BE MADE BETWEEN CATEGORIES OF INFORMATION?

In addition to the specific ability to perform searches as described in the previous sections, the categories of information must be linked with user friendly interfaces. To facilitate maximal utility of Arabidopsis databases, there is a need to develop a standard interface for access to Arabidopsis genomic sequence information. All information must have the name of the individuals that provided the data. Attention should be paid to tight coordination between the genetic map and related genes, clones, and sequences, so that selection of any of these will lead transparently to accession of the others. Also, it is highly desirable for the database to have simple links for comparative sequence and mutant analysis with other plants and beyond that, with all organisms. The interface should allow viewing in a variety of ways.

As examples of the types of links desired, we list below a series of questions that the system should be able to answer.

(1) If a user enters two cloned markers, the system should return a list of all markers of a specified type that map between them.

(2) If a user points to a location on a genetic map, the genes, clones, and sequences should appear. Likewise, a user should be able to derive map position if a DNA sequence is used as the starting point.

(3) For any gene, the expression pattern of the RNA encoded, by the clone should be readily accessed. Since information may be available about how the expression of the gene changes with different treatments, or in different mutants, it will be necessary to allow the user to define a set of comparator genes that can serve as standards. If available, links to spatially resolved information should be available as images.

(4) If a user finds a mutant that is altered in a particular way, the system should retrieve all mutants altered in a similar manner. A cross-species accession to similar mutants in other plants might be useful.

(5) It should be possible for a user to rapidly determine the map positions for all genes in a given biochemical or developmental pathway.

(6) If a user has new mapping information, the system should have the ability to download archived data in that region for manipulation.

IX. COMMUNITY ISSUES THAT MUST BE CONSIDERED IN THE DESIGN AND OPERATION OF THE DATABASE

A. Advisory Committee

All Arabidopsis database proposals should include a provision for an oversight committee that will represent the community of Arabidopsis researchers and will advise database investigators on priorities and data to be included. The oversight committee should also include individuals with technical expertise in database design and management. It may also be desirable to have representation from individuals involved in development and operation of other plant databases. In order to maximize accountability, it would be desirable to have the oversight committee formally approved by the North American Arabidopsis Steering Committee (NASC).

B. Curation, Entry, Correction, and Long-Term storage of Data

One of the major problems associated with developing a database is collecting data. Because database deposits do not currently generate a citation for inclusion in an individuals vita, there is no incentive to make the effort to deposit data. One mechanisms for encouraging deposits may be to implement a citation system for database deposits which would resemble those currently used for journal publications (ie., Author, title, date, accession number).

The task of data acquisition would be greatly facilitated if the journals would require authors to make deposits of data directly into appropriate databases at the time of publication in much the same way that all journals now require GenBank accession numbers. There was unanimous agreement that this would be a desirable development and there are indications that at least some of the plant journals are willing to implement such a change. Future proposals should include a plan for creating user-friendly interfaces that can be used by the authors of journal articles to enter data directly into an internet accessible form. Such forms could also be used by members of the community to enter unpublished data into the database. There was broad enthusiasm for a requirement that anyone receiving public research support be obliged by the funding agencies to describe how the data and research materials from previous supported research have been made available to the stock centers and databases.

Previous attempts to acquire data by soliciting input from the community have been generally unsuccessful and curation of data by the community is not considered feasible. Thus, the Arabidopsis databases must be curated by professional curators. Professional curators of Arabidopsis databases should make every effort to leverage the database activities undertaken elsewhere and to adapt existing software when appropriate for use in the Arabidopsis research community. Thus, the major activity of Arabidopsis databases should be the collection, entry, and correction of data rather than writing software for storing, retrieval, and presentation of data. It is clear from past experience that full time professional curators are required for the development and operation of an adequate database. In order to recruit and retain highly skilled personnel to develop and operate the Arabidopsis databases, it is essential that there be a reasonable expectation of stable long- term funding.

C. Relation to Other Databases and Programs

All Arabidopsis databases should use industry-standard hardware and software, so that they are both compatible with and can communicate transparently with other data bases. However, as stated elsewhere in this report, the primary goal should be to collect and store data using currently accepted database models rather than to develop new database software. The most important principle, therefore, in the design of next generation databases is that the data be entered in a form that makes it possible to interface easily with other databases and which makes the data portable to future generation database software. Any software that is written specifically for an Arabidopsis database (display of genetic maps, for example) should be layered and use industry standard interfaces so that the software, as well as the underlying data, is also compatible with and portable to future generation databases. In adition, consideration should be given to production of generic database structures that can be used for a variety of different organisms.

Databases are currently being developed for most plants of economic significance. Because all higher plants are very closely related and are thought to contain a similar basic gene set, the information in these databases can be readily interrelated by biological criteria. However, because of the various concerns of the groups developing other plant databases, and because of the different kinds and amount of information available, it is not feasible at this time to consider a common database structure that would accommodate Arabidopsis and other plants. Therefore, in order to facilitate future interconnectivity between the Arabidopsis databases and other plant databases, a concerted effort must be made to adopt common standards whenever possible. The use of the Mendel gene nomenclature conventions is a case in point. The developers of Arabidopsis databases should be informed about major activities with other plants and wherever possible should endeavor to share software.

D. Access of Databases

Data accumulated by a publicly funded database should be community property. There should be no restrictions on the availability of the data in the databases and they must be accessible internationally by the internet.

E. One or Several Databases?

It is desirable to facilitate full expression of the collective genius of the world Arabidopsis community. Because talent in bioinformatics and enthusiasm for Arabidopsis is distributed around the world, and because of the ease with which databases can communicate via the internet, a distributed database should be the goal. However, based on past experience, the users experience difficulty if information is fragmented or presented in a variety of different interfaces. Thus, the current situation in which users must navigate six separate databases to view genome sequence information is unacceptable. Bringing all genome sequence annotation into a common format should have a top priority. Thus, if there are several databases, each should have a clear and defined subset of the database task, and appropriate links to the others. It is imperative that they be integrated and that the staff operating the different databases be committed to cooperating with each other. Unrestricted access to all schema and source codes should be a requirement for public support. The goal should be to have a single user interface for a specific class of information. Proposals requesting support for database development must address this issue.

F. Education

The ADB investigators should be provided funding for the provision of community education and training. This would include the development of on-line help, training manuals, workshops, and short courses. The ADB developers should maintain complete documentation and source code. This information should be in the public domain. Because educators and students in higher education (including high-school students) may make use of ADB, sufficient documentation for non-sophisticated users should be made available.

G. Financial Support for Arabidopsis Databases

Although there is a general willingness of most members of the community to pay directly for database services in much the same way that journal subscriptions are currently purchased, it was concluded that the disadvantages of imposing charges outweigh the likely benefits for the foreseeable future. Thus, at present, it would be inappropriate to impose charges for the use of publicly supported databases. As with other organism-specific databases, the burden of funding must be borne by government agencies. In order to retain highly qualified database curators and developers, there must be reasonable assurance of continuing support.

H. Ownership of Databases

Because of the convergence of electronic publishing and database activities, potential liability issues, and because of the intrinsic value of established databases, consideration needs to be given to the legal ownership of databases. At present, databases developed with US federal grants are the property of the institutions that administer the grants. Because of the importance of ensuring unrestricted public access to Arabidopsis databases, proposals for funding of future database activities should provide assurances that institutional policies are consistent with the continuing need for free unrestricted access.

X. WHAT DESIGN-FEATURE ISSUES NEED TO BE CONSIDERED?

The design considerations for Arabidopsis databases are essentially unchanged from the 1993 workshop report. One of the most pressing needs reported was for improved graphical visualization tools for various forms of data.

A. Design Considerations that Should be Discussed in any Proposal:

- Any field upon which a user might be expected to initiate a search should contain controlled-vocabulary entries.
- Data must always be in a form that will be portable to new database systems, and new computers and computer types. Current industry trends are towards layered software systems and client-server databases.
- All data should be available for bulk access or bulk downloading, as discussed above.

- The database I should have update information on its own contents: date/time coding should be considered for all data and links between data, so that update data can be obtained by users on a regular basis.
- The database should contain cross-references to all other relevant databases (eg., GenBank Nucleotide Sequence Database; Arabidopsis thaliana stock center database; Cell and/or probe repository catalogue number(s); and Genetic map databases for species showing significant synteny with Arabidopsis thaliana).
- To ensure the development of a robust and stable production quality system, the database should be based upon readily available, proven software.
- To facilitate inter-database linking, and referencing of items in the database in an unambiguous manner for publications and other reports, and to ensure the long- term ready availability of the data in the database, primary entities ("unit records") should be identified by public, unchanging unique identifiers (accession numbers).
- The database structure should be described in a data dictionary or repository which would be available to database users
- To facilitate user and developer access, the database should be maintained on a computer (or network of computers) which is connected to the internet.

B. Research Goals

Developers should consider and propose to carry out some short-term research relevant to improving the quality of the Arabidopsis thaliana database. Some possibilities for short-term research would be:

- Develop methods for defining and controlling differential access to the data.
- Develop a means for providing an audit trail, or other historical record, of all changes to the database.
- Investigate methods for, facilitating inter-database interactions and connections.
- Develop a stable, documented application program interface (APT) to the database.
- Develop a method for representing variations in data quality and for recording uncertainty.
- Develop means for integration of physical mapping data with genetic and cytogenetic maps.
- Develop means for providing ready user access to underlying supporting data (maintained in remote laboratory databases) through -the database on-line user interface.
- Develop improvements in data presentation, including graphical representation of maps.

C. Possible Long-Term Research Goals

- Investigate new database systems and new data models.
- Monitor advances in hardware improvement and develop plans for using new hardware to improve the quality of the database.

Appendix 4

DIRECTORATE FOR BIOLOGICAL SCIENCES

Arabidopsis thaliana Information Resource Project (AtIR)

NATIONAL SCIENCE FOUNDATION

Division of Biological Infrastructure

DEADLINE: MARCH 22, 1999

MATRIX OF PROGRAM REQUIREMENTS

General Information

- **Program Name:** *Arabidopsis thaliana* Information Resource Project (AtIR)
- **Short Description/Synopsis of Program:**

The Directorate for Biological Sciences (BIO) of the National Science Foundation (NSF), through the Biological Database Activities Program in the Division of Biological Infrastructure, has identified as a priority support for the design, development, and implementation of biological information resources for the Multinational Coordinated *Arabidopsis thaliana* Genome Research project. Therefore, the Biological Database Activities Program announces a special competition for an on-line resource to extend, maintain and distribute a user focused, on-line resource for biological information on *Arabidopsis thaliana,* termed here the ***Arabidopsis thaliana*** Information Resource (AtIR). The successful awardee of this competition will be required to incorporate and build on the existing *Arabidopsis thaliana* Database (AtDB, /), which continues to be an unique resource in its role as a primary repository of *Arabidopsis* information.

- **Cognizant Program Officer(s):**
- Paul Gilna, by phone (703) 306-1469 or by e-mail *pgilna@nsf.gov*
- Applicable Catalog of Federal Domestic Assistance (CFDA) No.: **47.074-Biological Sciences**

Eligibility

- Limitation on the categories of organizations that are eligible to submit proposals: **None**
- PI eligibility limitations: **Limited to categories 1 and 2 of the Grant Proposal Guide *(GPG)* NSF 99-2, Chapter I, Section D**
- Limitation on the number of proposals that may be submitted by an organization: **None**

Award Information

- *Type* of award anticipated? **Grant or Cooperative Agreement**

- Number of awards anticipated in FY 1999: **1**
- Amount of funds available: The total award size is expected to range up to $1 million per year for 5 years.
- Anticipated date of awards: August 1999

Proposal Preparation Instructions

Proposal preparation instructions: **Standard Grant Proposal Guide (*GPG)* plus supplementary guidance**

Deviations from standard *GPG* proposal preparation instructions: **PIs must complete the BIO Proposal Classification Form (PCF)**

Budgetary Information

Cost sharing/matching requirements: **None**

Indirect cost (F&A) limitations: **None**

Other budgetary limitations**: Funds may not be requested or used for construction or renovation of facilities.**

FastLane Requirements

Use of FastLane in Proposal Preparation & Submission: **Entire Proposal Required**

FastLane point of contact for this program: **E-mail biofl@nsf.gov**.

Deadline/Target Dates

Full Proposal Deadline: **March 22, 1999**

Proposal Review Information

- Standard NSB Approved Merit Review Criteria plus supplementary criteria

Description of supplementary criteria: In addition, reviewers will focus on the following issues:

- responsiveness to the expected scope
- potential to advance international *Arabidopsis* genome and plant research;
- effectiveness of the project's organizational plan to reflect technical advances and new scientific discoveries;
- extent to which operation is focused on research community's needs;
- soundness and openness of the information-sharing plan and management of intellectual property rights;

- quality of the training environment for junior scientists; and,
- commitment to promote participation of members of under-represented groups.

Where appropriate, reviewers will also consider:

- cohesiveness and soundness of the planned coordination for a multi-investigator project; and,
- efficiency and cost-effectiveness of the proposed approach for infrastructure development.

Award Administration Information

Special grant conditions anticipated: **None**

INTRODUCTION

The Directorate for Biological Sciences (BIO) of the National Science Foundation (NSF), through the Biological Database Activities Program in the Division of Biological Infrastructure, has identified as a priority support for the design, development, and implementation of biological information resources for the Multinational Coordinated *Arabidopsis thaliana* Genome Research project.

The Multinational Coordinated *Arabidopsis thaliana* Genome Research project was established in 1990 to develop *Arabidopsis thaliana* as an experimental model system for flowering plants. During the next several years, the sequence of the *Arabidopsis* genome will be completed and extensive sequence and mapping information will become available for this and many other plant species. New technologies such as microarrays and gene chips now present the capacity to study the functional expression of thousands of genes at a time, while new capabilities in creating libraries of insertional mutations will allow detailed studies and ultimately manipulation of specific gene function. Drawing on the original goals of embarking on model organism genomes, the value of the *Arabidopsis* project lies in the utility of the information gathered in seeking to understand the biology of flowering plants.

Therefore, the Biological Database Activities Program announces a special competition for an on-line resource to extend, maintain and distribute a user focused, on-line resource for biological information on *Arabidopsis thaliana,* termed here the ***Arabidopsis thaliana*** Information Resource (AtIR). The successful awardee of this competition will be required to incorporate and build on the existing *Arabidopsis thaliana* Database (AtDB, /), which continues to be an unique resource in its role as a primary repository of *Arabidopsis* information

PROGRAM DESCRIPTION

The ***Arabidopsis thaliana*** Information Resource (AtIR) is expected to serve as a repository for data and information generated from multiple genomic studies on *Arabidopsis*. Operational priorities for this project will be predominantly needs-driven as defined by the *Arabidopsis* (and related) research communities, and as gathered through mechanisms established by the awardee. While it is understood that some software development will be required to meet these needs, the major mission of AtIR should be viewed as the collection, entry, and updating of data and information.

The project will be expected to focus on specific needs that have been defined by the Arabidopsis research community during the course of meetings held in Dallas, Texas in 1993 (/db/dallas.report.html) and updated in a meeting in Madison, Wisconsin, in 1998 (/db/database.needs.html).

Integration of Arabidopsis physical and genetic map data.

The greatest current need is a unified genetic and physical map that incorporates all available information about polymorphic markers (*e.g.,* CAPS, SSLPs, RFLPs), mutations, BAC and YAC clones, mapped clones and insertions or other modifications of the genome. This should be viewed as a critical component of the AtIR service.

Phenotypic and Genotypic Information.

Because of the diversity of processes that are being analyzed by a mutational approach in *Arabidopsis*, there is a need for the entire scientific community to have facile access to information about gene function as it relates to the organism. This capability will greatly enhance the efficiency with which new mutations will be studied as the number of known mutations begins to plateau. AtIR will be expected to incorporate this capability.

Interoperation with databases of related information.

AtIR should contain cross-references to all other relevant databases (e.g., GenBank nucleotide sequence database; *Arabidopsis thaliana* stock center databases; cell and/or probe repository catalogue number(s); and genetic map databases for other species showing significant synteny with *Arabidopsis thaliana*).

Storage and dissemination of expression data. Most or all of the *Arabidopsis* genes will be used to develop gene chips or microarrays that permit simultaneous measurements of the expression (mRNA levels) of all of the genes. The use of microarrays and gene chips are expected to provide a massive amount of new information. The ability to query this information may provide insights into the identity of otherwise anonymous genes, reveal the existence of networks or identify factors that cause altered expression of a gene. While it is not necessarily expected that the AtIR will serve as a primary repository for such data, it is expected that user access to such resources will be enabled through the use of appropriate links to other such databases.

Links to stock-based information. The databases maintained by the two *Arabidopsis* resource centers at Ohio State University and the University of Nottingham provide excellent access to information on the availability of biological and chemical materials related to *Arabidopsis* research. These databases will continue to assume responsibility for descriptive information concerning seed stocks, clones, vectors, libraries, cDNAs, oligonucleotides, and any other materials that may require distribution to the *Arabidopsis* community. The AtIR should be directly linked to the stock center databases so that queries about the properties of a gene or mutant can lead in turn to information about the availability of, and ordering procedures for, associated reagents.

Mechanisms for data acquisition by direct submission.

The task of data acquisition would be greatly facilitated if members of the *Arabidopsis* research community could deposit data directly. The AtIR should include a plan for creating user-friendly interfaces that can be used by scientists to deposit data directly to the AtIR via the internet, and address approaches to be taken to encourage direct submission of data from the research community.

Curation and maintenance of data.

Curation and maintenance refers to the need to add, validate and update the biological attributes of repository data. Approaches to this task have ranged from an "in-house" staff of curators or annotators to dependency on community-based methods of data correction, maintenance and updating, to, conceivably, a highly automated suite of computational tools. Curation of data in an *Arabidopsis* data resource has been and will continue to be an important community need and will be an important facet of the AtIR operation. Proposors will be expected to outline approaches to this task and address the utility of automated or community-based approaches to data curation.

Extensibility of database architecture to other plant genome information management projects.

The *Arabidopsis* database should use industry-standard hardware and software, so that it is both compatible, and can communicate transparently with, other databases. An important principle in designing the resource will be that the storage architecture is structured in a form that makes it possible to interface easily with other databases. Some consideration should be given to production of generic database structures that can potentially be adopted for use in a variety of different organisms and particularly in related mapping and/or sequencing activities in the Plant Genome Research community.

Summary

Proposals submitted in response to this announcement must discuss the structure of the proposed database with these goals and scope in mind, and provide detailed plans for long-term management and distribution of the database. The data should be structured and maintained in a way that permits the development and use of complex queries by knowledgeable users or by third party software developers. The AtIR will be expected to collaborate with other efforts relevant to plant databases, both nationally and internationally. Plans detailing how such collaborations might work should be provided. However, formal arrangements for the collaborations need not be made prior to an award. The proposals must also provide plans for the incorporation into the AtIR of information currently found in the *Arabidopsis thaliana* Database

(TAIR) and for the timely assumption of responsibility for data entry, repository maintenance and database distribution, all of which are now provided by TAIR.

ELIGIBILITY

The *Arabidopsis thaliana* Information Resource Project competition, will accept applications from eligible institutions as described in the NSF *"Grant Proposal Guide" (GPG)*, NSF 99-2, Chapter I, Section D, in categories 1 and 2 only. The *GPG* is available on the NSF web site at the URL (*http://www.nsf.gov/cgi-bin/getpub?nsf992*). Paper copies of the *GPG* may be purchased from the NSF Publication Clearinghouse, P.O. Box 218 Jessup, Maryland 20794-0218, telephone (301) 947-2722, or by e-mail from *pubs@nsf.gov*.

Consortia of eligible individuals or organizations may also apply, but a single individual or organization must accept overall management responsibility. International collaboration is encouraged; however, financial support for any non-U.S. participant organization must be provided from within the participant's country or other non-U.S. sources.

PRINCIPAL INVESTIGATOR AND OTHER SENIOR STAFF

The Principal Investigator (PI) and other senior staff responsible for the project must have the necessary skills to successfully carry out the tasks covered in this announcement, or the proposal must present convincing plans to hire such staff. The PI should have demonstrated the leadership necessary to meet the challenges of managing a large community database in a rapidly changing technological and scientific environment. The PI and other members of the senior staff should, in the aggregate, have experience with aspects of plant biology research relevant to the database, have current knowledge about computerized databases and their management, and have a demonstrated ability to interact with the members of the various scientific disciplines and other groups important for the successful operation of the database. Experience with the successful management of a database effort of comparable scope and complexity will be considered an important asset.

AWARD INFORMATION

The NSF expects to make one five year award in Fiscal Year 1999 depending on the quality of submissions and the availability of funds. The total award size is expected to range up to $1 million per year. The exact amount will depend on the advice of reviewers and on the availability of funds. It is anticipated that the award will be administered as a grant or cooperative agreement.

Note, while the term "award" and "awardee" used herein imply a single entity, NSF is not necessarily constrained by this model and is open to proposals of innovative models involving more than one entity by which the primary functions of AtIR might be administered (*e.g.*, a "virtual resource"). Again, a single individual or organization must accept overall management responsibility.

INSTRUCTIONS FOR PROPOSAL SUBMISSION

A. Proposal Preparation Instructions

Proposals to *Arabidopsis thaliana* Information Resource (AtIR) Project competition require electronic submission via the NSF FastLane system in accordance with the guidelines provided in the "Instructions for Proposal Preparation" found in the *GPG*, Chapter II. The *GPG* is available on the NSF Web Site at the URL *http://www.nsf.gov/cgi-bin/getpub?nsf992*. Paper copies of the *GPG* may be purchased from the NSF Publication Clearinghouse, P.O. Box 218 Jessup, Maryland 20794-0218, telephone (301) 947-2722, or by e-mail from *pubs@nsf.gov*.

Include in proposals to AtIR the components listed in *GPG*, Chapter II, Section D. State information in each component as clearly and concisely as possible for merit review. Take special care in adhering to the requirements for page limitations, font size, and margins (see *GPG*, Chapter II, Section C). **Proposals not strictly adhering to the requirements of the *GPG* and these guidelines are returned without review.** Instructions and guidelines for the FastLane submission of proposals are detailed in *Instructions for Preparing and Submitting a Standard Proposal via FastLane* located at *http://www.fastlane.nsf.gov/a1/newstan.htm*. Also, see the "FastLane Submission" section below.

Guidelines are provided for specific sections of the proposal as follows:

- **Proposal Cover Sheet (NSF Form 1207)**

In the NSF FastLane system follow instructions on proposal preparation. When completing the Cover Sheet click on the "Add Org Unit" button. Highlight "DIRECT FOR BIOLOGICAL SCIENCES" and click "OK." Highlight "Database Activities" and click "OK." Clicking "OK" designates this program as the NSF organizational unit of consideration. In the box labeled "Program Announcement/Solicitation No." enter "NSF 99-50" with no additional characters.

Begin the title of the proposal with "AtIR: . . . ."

The first-listed Principal Investigator (PI) is designated as the primary PI and is responsible for coordinating the entire proposed project.

- **Project Summary**

Provide a brief (200 words or less) description of the project.

- **Project Description (maximum length 25 pages)**

Particular attention must be paid to the following major aspects in preparing a description of the proposed project. Although some relevant technical issues are mentioned below, these details are intended only as guidelines. This section must not exceed 25 pages inclusive of tables, diagrams or other visual material. Clearly label sections and major subdivisions of the project description.

Long-Term Vision

Describe your vision for the long-term future of such a database as the AtIR and the role this operation should play in the overall plant genome research forum. Address issues such as long-term economic sustainability of the database, potential economic models that invoke alternative sources of support, and possible transition plans to such models.

Repository Structure

The proposal should provide a description of (1) the logical or conceptual model for the data, and (2) a general outline of the physical implementation schema for the repository. The general features and overall design of both must be justified in the context of efficient data management and researcher support functions. Extensibility of the design to the maintenance of data and information from other databases of plant research information may be discussed here.

Data Acquisition

Proposals should describe the manner in which the data to be placed in the resource will be acquired. Specifically, if it is intended that data be acquired from investigators as the original source of the data, procedures for the handling of such submissions should be described, including any standard or proprietary data exchange formats or tools to be used.

Because it is anticipated that the volume and rate of data generation will continue to increase in the future, an important technical issue to be considered is the development and use of approaches which are capable of scaling to anticipated increases in the volume of data.

Database Content

Proposals should describe precisely the expected content of the database. The description should include some definition of what constitutes a minimum dataset, as well as a description of what might constitute a fully annotated dataset.

Minimum criteria for insuring the completeness and consistency of entries at the time they are placed in the database should be described, as should procedures for assuring that the criteria have been met. It is expected that the utility of the criteria and procedures will be periodically reviewed and approved using the formal external advisory mechanism.

Database Maintenance

Proposals should address the technical issues involved in the maintenance of a highly automated information repository, with convenient public access and off-site backup or other provision for protection from software or hardware failure. Provisions for maintenance of internal and external links should be discussed. The focus of the proposal should be the operation of a basic repository.

Database Distribution

Proposals should also describe the distribution methods envisioned, for example network access to the complete collection using the WWW or other means, and periodic production of tapes, CD-ROM or other media containing current entries.

If mirror sites are to be used, describe how the central and mirror sites will interact, estimate the time and effort required to operate a typical mirror and provide the criteria to be used in selecting mirror sites.

Any planned charges for copies on tape or other media, or for permission to provide such copies, should be discussed briefly in the proposal. All such charges will be subject to approval by the NSF. Periodic assessment of the utility of the distribution methods will be expected as part of the management and oversight of the AtIR.

NSF expects that Principal Investigators agree to complete and open sharing of data and material in an expeditious manner. By submitting a proposal, it is understood that the submitting institution and all participants agree to these guidelines (see the NSF *GPG*, NSF 99-2, Chapter VII, Section H).

Direct Access

Describe how users will be able to develop and use direct queries of the database. The interaction with the repository and the means to insure stability and security should be specified.

Assumption of Responsibility for Database Operation

Provide a timetable for the assumption of responsibility for new data entry and distribution of the database, including any efforts necessary for incorporation of entries now found in the TAIR into the new database. It is anticipated that the time required for complete assumption of the responsibility will not exceed one year from the date of the award.

Quality Control

Describe provisions for insuring the quality of the database and its operation, including procedures for obtaining and responding to user feedback on issues related to quality.

Management

A sound management plan will be a crucial aspect of the proposal. The responsibilities of the various senior personnel must be clearly described, as must the time and effort to be committed by each. A mechanism for replacing key personnel who leave the project must also be described. In the event senior personnel will participate in multiple activities related to the database (e.g., outreach, data acquisition, etc.), estimate the anticipated effort with respect to each activity.

External Input/Advice

The awardee will be expected to establish a formal mechanism for insuring ongoing external input from relevant groups and interested individuals regarding AtIR policies and practices. An appropriate mechanism could, for example, consist of a standing external advisory board with relevant technical and managerial expertise. The function of such an advisory board could be to advise senior management of the AtIR and the awardee institution(s) on policies such as those regarding operational priorities, format, content and validation of entries and reports, those related to other aspects of use or distribution of the database, etc. Periodic review and approval of the utility and appropriateness of any such criteria will be expected.

Implementation of the mechanism should insure that the views of relevant research communities are represented as part of this advice. In general, the mechanism should provide an opportunity for input from the international *Arabidopsis* research community. The appropriateness and adequacy of the mechanism, as implemented, will be subject to approval by the NSF.

Outreach and Training

Describe provisions for timely and widespread communication of activities of the AtIR, in particular procedures for alerting user/developer communities to impending changes in software/formats/policies, etc. Describe any activities planned to train new or experienced users in use of the resource. Activities supported by this award may provide an ideal environment to train young scientists in cutting-edge research technologies and to expose them to new paradigms in plant biology informatics. In addition, these activities should promote increased participation by members of under-represented groups. Proposers should describe plans to increase diversity whenever feasible.

Results From Prior NSF Support (maximum length 5 pages)

If the PI or any Co-PI has received federal support for the establishment or operation of a publicly available database within the last five years, provide a brief description of the relevant features of the database together with the name of the agency providing support, the award number and title, and the amount and duration of the award. This section should include a general description of the type of database, number of users, means of distribution, etc. If the database is available electronically, provide the relevant URL. If awards for more than one project have been received, describe the project most relevant to the current proposal. **This section is limited to a maximum of 5 pages, including any references and is included as part of the Project Description 25 page limit.**

- **Biographical Sketches**

For each of the key personnel, including senior staff and any other staff whose participation is critical to the success of the project, provide a curriculum vitae or short biographical sketch. Briefly describe relevant experience and list up to 10 publications (to include the individual's 5 most important and up to 5 other relevant publications). Include an alphabetical list of current and past collaborators of all key personnel whose biosketches are included, and of any other staff or collaborators mentioned by name in the proposal. Additionally, include names of all graduate students and postdoctoral fellows who have trained with these individuals, as well as anyone with whom these individuals have co-authored a paper within the last 4 years. The information may not exceed 2 pages for each individual. Applicants may include letters of support in the

FastLane submission by scanning the documents and adding them at the end of the Project Description file, clearly labeled.

Copies of letters indicating agreement to participate should be provided by all senior personnel who do not endorse the cover page as PI or Co-PI. Such letters should include a brief description of the individual's expected role in the project and an estimate of the time and effort to be required. Scan the letters and add them at the end of the Project Description file, clearly labeled as Appendix A. **This information *is not*** counted as part of the 25 page limit of the Project Description.

- **Budget (NSF Form 1030)**

Provide a budget and budget justification for each year of support requested as well as a separate, cumulative budget for all years. If funds for subcontracts are requested, then a separate budget and budget justification must be prepared by each subcontractor to show the distribution of subcontract funds across categories. Funds for facility construction or renovation may not be requested.

A brief justification for funds in each budget category should be provided. For major equipment or software materials, a particular model or source and the current or expected price should be specified whenever possible. A brief explanation of the need for each item whose cost exceeds $10,000 should be provided. This section should also include details of institutional cost sharing, if any, and of other sources of support for the project, such as government, industry, or private foundations. Although cost sharing is not required, any such commitment specified in the proposal will be referenced and included as a condition of an award resulting from this solicitation.

Appropriate documentation of any such commitments should be provided in an appendix (Appendix B). Scan the documents and add them at the end of the Project Description file, clearly labeled as Appendix B. **This information *is not*** counted as part of the 25 page limit of the Project Description.

- **Current Support (NSF Form 1239)**

Provide a complete list of current and pending support for all PIs and Co-PIs

- **Facilities, Equipment, & Other Resources (NSF Form 1363)**

Include a brief description of available facilities, including space and computational equipment available for the project. Where requested equipment or materials duplicate existing items, explain the need for duplication. This section is limited to 2 pages.

- **BIO Proposal Classification Form (PCF)**

Complete the BIO PCF, available on the NSF FastLane system. The PCF is an on-line coding system that allows the Principal Investigator to characterize his/her project when submitting proposals to the Directorate for Biological Sciences. Once a PI begins preparation of his/her proposal in the NSF FastLane system and selects a division, cluster, or program within the Directorate for Biological Sciences as the first or only organizational unit to review the proposal, the PCF will be generated and available through the Form Selector screen. Additional information about the BIO PCF is available in FastLane at *http://www.fastlane.nsf.gov/a1/BioInstr.htm*.

- **Special Information and Supplementary Documentation**

Plans requiring collaborative effort by an individual not employed at the submitting institution(s) should be supported by a signed letter from the individual. Besides indicating a willingness to collaborate, the letter should provide a brief outline of the goals of the collaboration and estimate the time and effort the individual expects to devote to the collaboration. Biographical sketches should not be provided for such individuals, unless requested by NSF. A collaborator whose primary purpose is advisory (e.g., service on a committee that will provide policy advice) does not need to provide/submit such a letter.

Scan the letters and other relevant Special Information and Supplementary Documentation, as specifically described in the *GPG,* Chapter II, Section D.12, and add them at the end of the Project Description file after Appendices A and B, clearly labeled as "Special Information and Supplementary Documentation." Only documentation as described in the *GPG*, Chapter II, Section D.12 and detailed above is allowed. **This information *is not*** counted as part of the 25 page limit of the Project Description.

- **Appendices**

Only the appendices described in the "Budget Justification", and "Biographical Sketches", are allowed. Other letters of endorsement may not be included.

B. Proposal Due Dates

Proposals must be sent by 5:00 p.m., submitter's local time, March 22, 1999 via the NSF FastLane system.

Mail the following materials directly to the Biological Database Activities Program:

- a paper copy of the cover sheet, including the completed certification page (page 2 of 2) signed by the PI and all Co-PIs and by an institutional representative; and
- the BIO classification form.

Do not mail copies of the full proposal. NSF will make the appropriate number of copies of the proposal.

The grantee is responsible for ensuring that the materials are **received** by March 26, 1999. Send materials to:

Arabidopsis thaliana Information Resource Project-NSF 99-50

Division of Biological Infrastructure
National Science Foundation
4201 Wilson Boulevard
Room 615
Arlington, VA 22230

Unless requested by NSF, additional information may not be sent following proposal submission.

C. FastLane Submission

In order to use NSF FastLane to prepare and submit a proposal, you must have the following software: Netscape Navigator 3.0 or above, or Microsoft Internet Explorer 4.01 or above; Adobe Acrobat Reader 3.0 or above for viewing PDF files; and Adobe Acrobat 3.X or Aladdin Ghostscript 5.10 or above for converting files to PDF.

To use FastLane to prepare the proposal your institution needs to be a registered FastLane institution. A list of registered institutions and the FastLane registration form are located on the FastLane Home Page. To register an organization, authorized organizational representatives must complete the registration form. Once an organization is registered, PIN for individual staff are available from the organization's sponsored projects office.

To access FastLane, go to the NSF Web site at *http://www.nsf.gov*, then select "FastLane," or go directly to the FastLane home page at *http://www.fastlane.nsf.gov/*. Please see "Instructions for Preparing and Submitting a Proposal to the NSF Directorate for Biological Sciences" located at *http://www.fastlane.nsf.gov/a1/BioInstr.htm*. Additionally, read the "PI Tipsheet for Proposal Preparation" and the "Frequently Asked Questions about FastLane Proposal Preparation," accessible at *https://www.fastlane.nsf.gov/a1/A1Prep.htm*.

IMPORTANT NOTE: For technical assistance with FastLane, please send an e-mail message to *biofl@nsf.gov*. If you have inquiries regarding other aspects of proposal preparation or submission, please contact the cognizant program officer, preferably *at least three weeks before the competition deadline*.

MERIT REVIEW

A. NSF Proposal Review Process

Reviews of proposals submitted to NSF are solicited from peers with expertise in the substantive area of the proposed research or education project. These reviewers are selected by Program Officers charged with the oversight of the review process. NSF invites the proposer to suggest, at the time of submission, the names of appropriate or inappropriate reviewers. Special care is taken to ensure that reviewers have no immediate and obvious conflicts with the proposer. Special efforts are made to recruit reviewers from non-academic institutions, minority serving institutions, adjacent disciplines to that principally addressed in the proposal, first time NSF reviewers, etc.

Proposals will be reviewed against the following general merit review criteria established by the National Science Board. Following each criterion are potential considerations that the reviewer may employ in the evaluation. These are suggestions and not all will apply to any given proposal. Each reviewer will be asked to address only those that are relevant to the proposal and for which he/she is qualified to make judgments.

1. What is the intellectual merit of the proposed activity?

How important is the proposed activity to advancing knowledge and understanding within its own field and across different fields? How well qualified is the proposer (individual or team) to conduct the project? To what extent does the proposed activity suggest and explore creative and original concepts? How well conceived and organized is the proposed activity? Is there sufficient access to resources?

2. What are the broader impacts of the proposed activity?

How well does the activity advance discovery and understanding while promoting teaching, training, and learning? How well does the proposed activity broaden the participation of underrepresented groups (e.g., gender, ethnicity, geographic, etc.)? To what extent will it enhance the infrastructure for research and education, such as facilities, instrumentation, networks, and partnerships? Will the results be disseminated broadly to enhance scientific and technological understanding? What may be the benefits of the proposed activity to society?

In addition, reviewers will focus on the following issues:

- responsiveness to the expected scope;
- potential to advance international *Arabidopsis* genome and plant research;
- effectiveness of the project's organizational plan to reflect technical advances and new scientific discoveries;
- extent to which operation is focused on research community's needs;
- soundness and openness of the information-sharing plan and management of intellectual property rights;
- quality of the training environment for junior scientists; and,
- commitment to promote participation of members of under-represented groups.

Where appropriate, reviewers will also consider:

- cohesiveness and soundness of the planned coordination for a multi-investigator project; and,
- efficiency and cost-effectiveness of the proposed approach for infrastructure development.

Integration of Research and Education

One of the principal strategies in support of NSF's goals is to foster integration of research and education through the programs, projects and activities it supports at academic and research institutions. These institutions provide abundant opportunities where individuals may concurrently assume responsibilities as researchers, educators, and students and where all can engage in joint efforts that infuse education with the excitement of discovery and enrich research through the diversity of learner perspectives. PIs should address this issue in their proposal to provide reviewers with the information necessary to respond fully to both NSF merit review criteria. NSF staff will give this careful consideration in making funding decisions.

Integrating Diversity into NSF Programs, Projects, and Activities

Broadening opportunities and enabling the participation of all citizens-women and men, underrepresented minorities, and persons with disabilities-is essential to the health and vitality of science and engineering. NSF is committed to this principle of diversity and deems it central to the programs, projects, and activities it considers and supports. PIs should address this issue in their proposal to provide reviewers with the information necessary to respond fully to both NSF merit review criteria. NSF staff will give this careful consideration in making funding decisions.

B. Review Protocol and Associated Customer Service

Most proposals submitted to the NSF are reviewed by mail review, panel review, or some combination of mail and panel review.

Proposals submitted to this activity will be evaluated by a special emphasis panel formed to review the applications and mail reviewers. Site visits may be conducted as needed. NSF will be able to tell applicants whether their proposals have been declined or recommended for funding within six months for 95 percent of proposals in this category.

GRANT AWARD AND ADMINISTRATION INFORMATION

A. Notification of the Award

Notification of the award is made ***to the submitting organization*** by a Grants Officer in the Division of Grants and Agreements. Organizations whose proposals are declined will be advised as promptly as possible by the cognizant NSF Program Division administering the program. Verbatim copies of reviews, not including the identity of the reviewer, will be provided automatically to the lead Principal Investigator.

B. Grant Award Conditions

Grants awarded as a result of this announcement are administered in accordance with the terms and conditions of NSF GC-1 (10/98), "Grant General Conditions" (10/98), or FDP-III (7/97), "Federal Demonstration Partnership General Terms and Conditions," or CA-1 "Cooperative Agreement General Terms and Conditions" (2/98), depending on the grantee organization. Copies of these documents are available at no cost from the NSF Publications Clearinghouse, P.O. Box 218, Jessup, Maryland 20794-0218, telephone (301) 947-2722, or via e-mail to pubs@nsf.gov. More comprehensive information is contained in the NSF *Grant Policy Manual* (NSF 95-26), available on the NSF OnLine Document System located at http://www.nsf.gov/, or for sale through the Superintendent of Documents, Government Printing Office, Washington, D.C. 20402.

C. Reporting Requirements

For all multi-year grants (including both standard and continuing grants), the PI must submit an annual project report to the cognizant Program Officer at least 90 days before the end of the current budget period.

Within 90 days after expiration of a grant, the PI also is required to submit a final project report. Approximately 30 days before expiration, NSF will send a notice to remind the PI of the requirement to file the final project report. Failure to provide final technical reports delays NSF review and processing of pending proposals for the PI. PIs should examine the formats of the required reports in advance to assure availability of required data.

NSF has implemented a new electronic project reporting system, available through FastLane, which permits electronic submission and updating of project reports, including information on: project participants (individual and organizational); activities and findings; publications; and other specific products and contributions. Reports will continue to be required annually and after the expiration of the grant, but PIs will not need to re-enter information previously provided, either with the proposal or in earlier updates using the electronic system.

Effective October 1, 1998, PIs are required to use the new reporting format for annual and final project reports. PIs are strongly encouraged to submit reports electronically via FastLane. For those PIs who cannot access FastLane, paper copies of the new report formats may be obtained from the NSF Clearinghouse as specified above. NSF expects to require electronic submission of all annual and final project reports via FastLane beginning in October, 1999.

D. New Awardee Information

If the submitting organization has never received an NSF award, it is recommended that the organization's appropriate administrative officials become familiar with the policies and procedures in the NSF *Grant Policy Manual* which are applicable to most NSF awards. The "Prospective New Awardee Guide" (NSF 97-100) includes information on: Administration and Management Information; Accounting System Requirements and Auditing Information; and Payments to Organizations with Awards. This information will assist an organization in preparing documents that NSF requires to conduct administrative and financial reviews of an organization. The guide also serves as a means of highlighting the accountability requirements associated with Federal awards. This document is available electronically on NSF's Web site at http://www.nsf.gov/cgi-bin/getpub?nsf97100.

CONTACTS FOR ADDITIONAL INFORMATION

Inquiries regarding the announcement should be directed to the cognizant NSF official: Dr. Paul Gilna, Division of Biological Infrastructure, National Science Foundation, 4201 Wilson Boulevard, Room 615, Arlington, VA 22230. Telephone: (703) 306-1469; FAX: (703) 306-0356; E-mail: pgilna@nsf.gov

GENERAL INFORMATION

The National Science Foundation (NSF) funds research and education in most fields of science and engineering. Grantees are wholly responsible for conducting their project activities and preparing the results for publication. Thus, the Foundation does not assume responsibility for such findings or their interpretation.

NSF welcomes proposals from all qualified scientists, engineers, and educators. The Foundation strongly encourages women, minorities, and persons with disabilities to compete fully in its programs. In accordance with federal statutes, regulations, and NSF policies, no person on grounds of race, color, age, sex, national origin, or disability shall be excluded from participation in, be denied the benefits of, or be subjected to discrimination under any program or activity receiving financial assistance from NSF. Some programs may have special requirements that limit eligibility.

Facilitation Awards for Scientists and Engineers with Disabilities (NSF 91-54) provide funding for special assistance or equipment to enable persons with disabilities (investigators and other staff, including student research assistants) to work on NSF-supported projects.

The National Science Foundation has Telephonic Device for the Deaf (TDD) and Federal Information Relay Service (FIRS) capabilities that enable individuals with hearing impairments to communicate with the Foundation regarding NSF programs, employment, or general information. TDD may be accessed at (703) 306-0090; FIRS at 1-800-877-8339.

PRIVACY ACT AND PUBLIC BURDEN STATEMENTS

The information requested on proposal forms and project reports is solicited under the authority of the National Science Foundation Act of 1950, as amended. The information on proposal forms will be used in connection with the selection of qualified proposals; project reports submitted by awardees will be used for program evaluation and reporting within the Executive Branch and to Congress. The information requested may be disclosed to qualified reviewers and staff assistants as part of the review process; to applicant institutions/grantees to provide or obtain data regarding the proposal-review process, award decisions, or the administration of awards; to government contractors, experts, volunteers, and researchers and educators as necessary to complete assigned work; to other government agencies needing information as part of the review process or in order to coordinate programs; and to another Federal agency, court or party in a court or Federal administrative proceeding if the government is a party. Information about Principal Investigators may be added to the Reviewer file and used to select potential candidates to serve as peer reviewers or advisory committee members. See Systems of Records, NSF-50, "Principal Investigator/Proposal File and Associated Records," 63 *Federal Register* 267 (January 5, 1998), and NSF-51, "Reviewer/Proposal File and Associated Records," 63 *Federal Register*

268 (January 5, 1998). Submission of the information is voluntary. Failure to provide full and complete information, however, may reduce the possibility of receiving an award.

Public reporting burden for this collection of information is estimated to average 120 hours per response, including the time for reviewing instructions. Send comments regarding this burden estimate and any other aspect of this collection of information, including suggestions for reducing this burden, to: Reports Clearance Officer; Information Dissemination Branch, DAS; National Science Foundation; Arlington, VA 22230.

The program described in this announcement is in the category 47.074 (BIO) of the Catalog of Federal Domestic Assistance.

YEAR 2000 REMINDER

In accordance with NSF Important Notice No. 120 dated June 27, 1997, Subject: Year 2000 Computer Problem, NSF awardees are reminded of their responsibility to take appropriate actions to ensure that the NSF activity being supported is not adversely affected by the Year 2000 problem. Potentially affected items include computer systems, databases, and equipment. The National Science Foundation should be notified if an awardee concludes that the Year 2000 will have a significant impact on its ability to carry out an NSF-funded activity. Information concerning Year 2000 activities can be found on the NSF Web site at *http://www.nsf.gov/oirm/y2k/start.htm*.

OMB NO. 3145-0058
P.T.: 34
K.W.: 1002037

NSF 99-50 Electronic Dissemination Only