Memorandum of Understanding

Multinational Effort to Sequence the Arabidopsis Genome (Sept. '96)

On August 20-21, 1996, representatives of six research groups committed to sequencing the Arabidopsis genome met in Washington DC to discuss strategies for facilitating international cooperation in completing the genome project. All six groups have secured major funding to pursue large-scale genomic sequencing of Arabidopsis, and the EU and Japanese groups have been engaged in large-scale sequencing for some time. The primary objectives of this meeting were to establish Arabidopsis as a model for international coordination of sequencing efforts and to develop guidelines for rapid and efficient completion of the sequencing project by the year 2004. Representatives from Japan, France, the EU, and the USA were present at the meeting. A complete list of participants and observers is attached as Appendix I.

A remarkable degree of consensus was reached by the end of the meeting on the general strategy for the Arabidopsis sequencing project. All parties agreed to follow several practices that were seen as facilitating international cooperation. This document was drafted to serve as a modus operandi for the participating groups until such time as it is modified by mutual agreement of representatives of the participating groups. All signatories to this document have agreed to the following:

1. The Arabidopsis Genome Initiative (AGI) is intended to be an inclusive international collaboration. Any group that intends to engage in the sequencing of hundreds of kilobases of contiguous Arabidopsis genomic DNA will be invited to participate as a coequal collaborator in the AGI and will be expected to follow the guidelines outlined in this document.

2. A coordinating committee with representation from each of the participating groups was formed. This committee will be responsible for making all decisions that affect the overall goals and operations of the AGI. In particular, it is anticipated that the AGI coordinating committee will be a planning and brokering system for establishing efficient ways of completing the genome. The committee will endeavor to apportion regions of the genome to the various groups in such a way as to minimize needless duplication of effort while maximizing progress toward complete sequencing of the genome. The committee will also be responsible for keeping the Arabidopsis community informed of continuing advances in the sequencing project.

Members of the committee for 1996-97 are Mike Bevan (Chair; EU consortium), Satoshi Tabata (Kazusa DNA Research Institute), Joe Ecker (Stanford-University of Pennsylvania-Plant Gene Expression Consortium {the SPP consortium}), Dick McCombie (Cold Spring Harbor-Washington University -Applied Biosystems Consortium {CSH-WU-ABI}), Steve Rounsley (The Institute for Genomic Research {TIGR}), Francis Quertier (French Genome Center) and David Meinke (Multinational Arabidopsis Steering Committee). Each member of the committee will be responsible for arranging a temporary or permanent replacement from the represented group when appropriate. New members will be invited to join the committee based on a nomination from one member of the committee and an affirmative vote by a majority. It is anticipated that the committee will maintain regular communication and will meet annually.

Mike Cherry (Curator of TAIR) will develop an email server to facilitate correspondence between members of the committee.

3. The six research groups are expected to complete different amounts of finished sequence because they have different capabilities and levels of funding devoted to this project. In order to prevent duplication of effort, it was considered useful to have the various groups initiate sequencing in different well-defined regions of the genome. It was agreed that each group should begin by nucleating sites over a contiguous region of a size that could be completed with the funding available. It was recognized that it may not be possible to define such a region with high accuracy because of variation in the ratio of genetic distance to physical distance. The goal in this respect should be to avoid situations where one group obtains scattered regions of sequence that must eventually be finished (i.e., linked up) by other groups. Exceptions to this strategy are noted elsewhere in this document.

The SPP group will begin nucleating on chromosome 1. The EU group will nucleate the bottom arm of chromosome 4. The CSH-WU- ABI group will nucleate a 4 Mb region on the top arm of chromosome 4 and a 2 Mb region on the top arm of chromosome 5 (the latter in collaboration with the EU group and the Kazusa group). The TIGR group will nucleate chromosome 2. The Kazusa group will nucleate the lower part of chromosome 5. The region at the top of chromosome 5 of mutual interest to the EU, CSH-WU-ABI and Kazusa group will be sequenced collaboratively. The Kazusa group, which anticipates a monthly sequencing rate of approximately 500 Kb, expects to begin nucleating a region of chromosome 3 in 1997. The EU, TIGR, SPP and CSH-WU-ABI groups anticipate an average monthly rate of approximately 200, 220, 150 and 150 Kb per month, respectively. Thus, when all the groups are operating at full capacity, the average monthly rate for the entire AGI collaboration is expected to exceed 1.2 Mb per month. The philosophy of the AGI collaboration is that as the assigned regions near completion, the coordinating committee will designate new regions of unfinished sequence to the groups in proportion to their sequencing capabilities. For example, the French genome Center is tentatively interested in sequencing BAC ends during the first year or two of operation but after that time it is anticipated that they will engage in sequencing a contiguous region of genomic DNA that will be decided at a later date.

Several of the participants had differing views about the relative merits of sequencing unique sequences versus regions of repetitive sequence such as centromeres and telomeres. On the one hand, it may be expected that the maximum number of coding sequences will be found by sequencing the regions of low copy number. On the other hand, it will be interesting to know the structure of the centromeric and telomeric regions. The majority view appeared to be that it was not necessary at this time to resolve this issue. However, the majority view was that renewals of existing grants should take into account the fact that some regions of sequence will be more difficult to complete than others and large stretches of contiguous sequence are more difficult to achieve than small scattered regions.

Sequencing efficiency should be the sole criterion for choosing which clone to sequence. It was agreed by all parties that none of the groups should perform service sequencing for outside groups interested in particular clones. The reason for this is that the sequencing groups should not be seen to be favoring certain colleagues.

4. The most efficient strategy for sequencing the Arabidopsis genome is to shotgun sequence large clones such as BACS, YACS or inserts from P1 clones. Most of the groups have had preliminary experience with BACS and YACS and preferred BACS. The fact that most of the groups are currently satisfied with the available public BAC libraries will facilitate coordination and exchange of information. In particular, in order to minimize the requirement for additional physical mapping, it is desirable to obtain several hundred base pairs from the ends of a large number of BAC clones so that the minimum tiling path from a region of sequence to an overlapping clone can be determined by database analysis. The groups led by Craig Venter (TIGR) and Francis Quertier (French Genome Center) agreed to sequence the ends of approximately 14,000 BACS from public BAC libraries during the next two years and to make the information freely available to the community.

All of the groups will use public BAC, YAC or P1 libraries constructed from the Columbia ecotype that will be freely available to the world community. A suitable BAC library to begin with is the TAMU BAC library constructed by Choi et al (http://probe.nalusda.gov:8000/otherdocs/ww/vol2/choi.html) that is currently available at the Ohio Stock Center. The other BAC library was constructed by Thomas Altmann and collaborators (altmann@mpimp-golm.mpg.de) and is also publicly available (http://www.rzpd.de). A P1 library (the 'M library') developed by Bob Whittier and colleagues at Mitsui is also available at the Ohio Stock Center and a second library (the 'K library') is being tested at the Kazusa Institute.

5. The objective of the AGI is to obtain high accuracy sequence of the entire genome. There was general agreement that it was not possible to set a standard for exactly what high accuracy means or for mechanisms to enforce high accuracy. However, it was generally agreed that a minimal standard would be that >97% of all sequence would be obtained on both strands or by two chemistries. It was the opinion of the group that these criteria were of similar importance and that with most clones, about seven-fold redundancy of sequencing would be required for shotgun sequencing.

An unknown factor affecting the accuracy of the sequence concerns the fidelity of the BAC clones. Preliminary experience suggests that the BACS are generally faithful clones of the genome. However, it will be essential to verify the integrity of each BAC. A minimum criterion is that both ends of the BAC should map to the same region of the genome, typically to the same YAC. When 14,000 BAC ends are sequenced, it is expected that, on average, we will have 500 bp of sequence every 5 kb on average throughout the genome. The resulting library of end-sequenced BACS will represent a check on BAC integrity that will assist in revealing any major rearrangements, deletions or additions. No standard was agreed upon for BAC (or P1) integrity checking. However, most groups indicated that comparing fingerprints of tiled BACS would be the most appropriate criterion for integrity.

After some discussion, it was agreed that a single-pass shotgun sequence of the entire genome would not be worthwhile because the combination of available ESTs and the high output rate of the AGI collaboration would obviate much of the value of single-pass shotgun sequencing for gene discovery. However, because chromosome 3 will be sequenced later than the other regions, the group endorsed a proposal by the SPP consortium to do a feasibility study involving shotgun

sequencing of clones from chromosome 3. After the meeting was concluded, the SPP consortium decided that this was not a good idea and reverted to their original plan in which "limited testing of some of the new instrumentation being developed at Stanford will utilize clones from a whole-genome shotgun library."

6. All of the participating laboratories are committed to early data release via the internet. One approach discussed at the meeting involved daily release of preliminary sequence information (ie., sequences that have been edited to remove vector and regions of high ambiguity and condensed into >1 kb contigs). The C. elegans sequencing groups follow this approach and the community has found it very useful. Two of the US groups, the SPP consortium and the CSH-WU-ABI consortium intend to release data in this way. Both groups anticipate release of finished, annotated sequence within six months of beginning to sequence a clone. The EU group does not consider it feasible, at the moment, to do daily releases because the consortium is composed of seventeen relatively small sequencing groups with varying levels of technical capabilities. The EU anticipates release of finished annotated sequence within one month of completion. The TIGR and Kazusa groups do not wish to release unfinished sequence because they believe that carefully edited sequence will be most useful to the community. Both groups promised release of information on a given clone to public databases within three to six months after sequencing began. The TIGR group will release finished, annotated sequence within three months of beginning to sequence a BAC. The Kazusa group estimates that they will release finished, annotated sequence within four to six months of beginning to sequence a clone. In all cases, the start date for sequencing a specific clone will be announced on linked WWW sites so that members of the community will know when to expect the finished sequence. In summary, all of the groups agreed to establish linked WWW pages for posting complete lists of all clones that have been sequenced to date, along with the start dates of those clones that are still in progress, and the anticipated start dates for the next set of clones to be sequenced in the future. Each clone will therefore have a start date that will be widely advertised to the community. All of the groups anticipate that it will take less than six months to completely sequence and annotate a BAC, YAC or P1 clone and that they will deposit the complete annotated sequences in a public database (eg.,GenBank, EMBL, JDB). No sequence information will be withheld from the community for the sole purpose of benefitting selected individuals, groups, or private companies.

7. There was consensus that the value of the sequence obtained is proportional to the quality of annotation. Thus, each group will attempt to achieve a common standard of annotation. Each group will perform BLAST (or FASTA) searches to align ESTs and known genes and gene products to the genomic sequence. In addition, each group will use programs such as GRAIL and GeneFinder to identify ORFs. Annotation should be presented to the community in a format that can be readily accessed and understood by plant biologists worldwide.

It was agreed that all unassigned ORFs would be named according to the *C. elegans* system. A provisional agreement was reached that the following rules of nomenclature will apply: The first letter is the library name. T=TAMU BAC, F=IGF BAC, M=Mitsui P1 clone, K=Kazusa P1 clone, C=cosmid clone from Goodman library. The first letter is followed by the microtiter plate number, then the row and column numbers followed by a dot and the number of the ORF (numbered sequentially from one side of the clone to the other). Thus, a typical ORF might be called t23a11.12 (i.e., a TAMU BAC from plate 23, well a11, the 12th ORF from one end). It

was agreed that zeros will not be included (ie., t23a11.12 but not t23a11.012). It was also suggested that the names be all lowercase for consistency. Sometimes it will happen that after all the ORFs have been named, a new one will be found by some functional test or other criteria. In this case the two ORFs will be named with an extension to the name (eg., t23a11.12.1 and t23a11.12.2). When two ORFs are found to belong to the same gene or an ORF is found not to be expressed, the name will be deleted. When one ORF spans two or more clones, the entire ORF will be given the name of the 5' region of the ORF.

It was recognized that annotation of a clone at the time of deposit in public databases will rapidly be rendered obsolete because of information about genes being discovered by the community at large. Thus, there will be an ongoing need for annotation of previously sequenced clones. Because most of the groups are funded to produce new sequence, it will be difficult for the groups producing sequence to also take responsibility for revising the annotation of previously completed sequence. There was broad agreement that the task of annotation revision should be institutionalized by assigning responsibility for revision to the curators of the Arabidopsis database (TAIR). The group expressed its strong enthusiasm and support for the continued funding of TAIR to make certain that essential informatics components of the Arabidopsis genome project are not overlooked. Mike Cherry agreed that it was a suitable responsibility for TAIR and agreed to accept the task to the extent that resources permit.

8. Because the US groups associated with the Arabidopsis Genome Initiative will need to reapply for funding within 2.5 years, there was concern about the criteria that will be used to evaluate success. It was agreed that each of the groups will be evaluated based on their overall contribution to the AGI collaboration and that the criteria will not simply be dollars per kb.

9. It is considered essential to keep the entire community well informed of technical advances and practical applications of the genome project. Each group will mount a WWW page that will report the contribution of the group to the multinational sequencing effort. Each group will also work through Mike Cherry (TAIR) and the coordinating committee to make certain that community members receive the training required to make efficient use of the extensive sequence data that will be generated over the next several years. In addition, the coordinating committee will evaluate the feasibility of appointing a part-time public relations specialist to produce user-friendly documentation about the progress of the Arabidopsis Genome Initiative. These efforts should help to advertise the dramatic impact that sequencing the Arabidopsis genome will have on basic and applied research in plant biology.

Signed: Mike Bevan, Ian Bancroft (EU consortium) Satoshi Tabata, Kiyotaka Okada (Kazusa DNA Research Institute) Joe Ecker, Sakis Theologis, Nancy Federspiel (SPP consortium) Dick McCombie, Rob Martienssen, Rick Wilson, Ellson Chen (CSH- WU-ABI) Craig Venter, Steve Rounsley, Owen White, Chris Somerville (TIGR) Francis Quertier (French Genome Center) David Meinke (Multinational Arabidopsis Steering Committee). September 14, 1996

List of Participants

Multinational Arabidopsis Seering Committee

David Meinke Department of Botany Oklahoma State University Stillwater, OK 74078 Phone:405-744-6549 fax 405-744-7673 E-mail:meinke@osuunx.ucc.okstate.edu

EU Group

Mike W. Bevan The Cambridge Laboratory AFRC Institute of Plant Science Research John Innes Center Norwich Research Park Colney, Norwich NR4 7UJ England phone 4460352571 fax 44-603-502 270 michael.bevan@bbsrc.ac.uk

Ian Bancroft The Cambridge Laboratory AFRC Institute of Plant Science Research John Innes Center Norwich Research Park Colney, Norwich NR4 7UJ England phone 4460352571 bancroft@bbsrc.ac.uk

Kazusa DNA Research Institute

Satoshi Tabata Kazusa DNA Research Institute Laboratory of Gene Structure 2 1532-3 Yana, Kisarazu Chiba 292, Japan e-mail tabata@kazusa.or.jp Tel +81-438-52-3933 Fax +81-438-52-3934

Kiyotaka Okada Department of Botany, Graduate School of Science, Kyoto University Kitashirakawa-Oiwake-Cho, Sakyo-Ku, Kyoto 606, Japan Phone: 81-75-753-4247, 4249, 4147 FAX: 81-75-753-4257 E-mail: kiyo@ok-lab.bot.kyoto-u.ac.jp

SPP Consortium

Joe Ecker University of Pennsylvania Department of Biology Plant Sciences Institute Philadelphia, PA 19104 fax 215 898 8780 phone 215 898 9384 jecker@atgenome.bio.upenn.edu

Sakis Theologis USDA Plant Gene Expression Center 800 Buchanan Street Albany, CA 94710-1198 phone 510 559 5910 fax 415 559 5678 theo@mendel.berkeley.edu

Nancy Federspiel Department of Biochemistry Stanford University School of Medicine Stanford, CA 94305 fax 650 723 6783 nfeder@genome.stanford.edu

CSH-WU-ABI Consortium

Rob Martienssen Cold Spring Harbor Laboratory 1 Bungtown Road PO Box 100 Cold Spring Harbor, NY 11724-2213 fax 516 367 8369 phone 516 367 8884 martiens@cshl.org

Dick McCombie Cold Spring Harbor Laboratory 1 Bungtown Road PO Box 100 Cold Spring Harbor, NY 11724-2213 fax 516 367 8369 phone 516 367 8884 mccombie@cshl.org

Rick Wilson Washington University School of medicine Department of Genetics St Louis MO 63108 rwilson@watson.wustl.edu

Ellson Chen ACGT ABD-Perkin Elmer 850 Lincoln Centre Drive Foster City, CA 94404 phone 650 638 5107 fax 650-638 6177 cheney@perkin-elmer.com

TIGR group

Craig Venter, The Institute for Genomic Research 9712 Medical Center Drive Rockville, MD 20850 phone 301-838-3500 fax 301 838 02208 jcventer@tigr.org lholland@tigr.org (secretary)

Steve Rounsley The Institute for Genomic Research 9712 Medical Center Drive Rockville, MD 20850 phone 301-838-3500 fax 301 838 02208 rounsley@tigr.org

Owen White The Institute for Genomic Research 9712 Medical Center Drive Rockville, MD 20850 phone 301-838-3500 fax 301 838 02208 owhite@tigr.org

Chris Somerville Carnegie Institution 290 Panama Street Stanford, CA 94305 phone 650-325-1521 extn 203 fax 650-325-6857 crs@andrew.stanford.edu

French Genome Center

Francis Quetier GENETHON 1 rue de l'Internationale 91002 EVRY Cedex France Tel ; (33) 1 69 47 69 89 Fax (33)1 60 77 09 51 email quetier@genethon.fr

Observers

Mike Cherry Department of Genetics Stanford University School of Medicine Stanford CA 94305-5120 phone 650-723-7541 fax 650-723-7016 cherry@genome.stanford.edu

Greg Dilworth Department of Energy Biosciences, ER-17 US Department of Energy 19901 Germantown Road Germantown, MD 20874-1290 fax 301-903-1003 Phone 301-903-2873 greg.dilworth@oer.doe.gov

Machi Dilworth DIR/BBS National Science Foundation fax 703 306 0356 phone 703 306 1471 mdilwort@note.nsf.gov

Edward Kaleikau NRI Competitive Grants Program CSREES/USDA AGBox2241,WashingtonDC20250-2241 Phone:2024011901 E-mail:ekaleikau@reeusda.gov

Delill Nasser BIO/MCB National Science Foundation fax 703-306-0356 phone 703-306-1439 dnasser@note.nsf.gov

Henry Shands National Genetics Resource Program ARS/USDA Bldg 005, BARC-West Beltsville, MD 20705 Phone:3015045059 E-mail:shands@ars-grin.gov

Jim Tavares Department of Energy Biosciences, ER-17 US Department of Energy 19901 Germantown Road Germantown, MD 20874-1290 fax 301-903-1003 Phone 301-903-2873 james.tavares@oer.doe.gov