Plant Genome Research Program Grant: DBI9975718/0196098
"Global Expression Studies of the Arabidopsis Genome"

PI, Joseph R. Ecker, The Salk Institute for Biological Studies, 10010 N. Torrey Pines Road, La Jolla, CA 92037, Email: ecker@salk.edu
Co-PI, Ronald W. Davis, Stanford Genome Technology Center, Stanford University, 855 California Ave, Palo Alto, CA 94304, Email: dbowe@sequence.stanford.edu
Co-PI, Athanasios Theologis, Plant Gene Expression Center, U.C. Berkeley, 800 Buchanan Street, Albany, CA 94022, Email theo@nature.berkeley.edu

Funding Period: October 1, 1999 – September 30, 2002
URL for the Salk, Stanford, PGEC, (SSP) Consortium Project: http://signal.salk.edu/SSP/index.html

1. Project Summary. In order to carry out functional genomic and proteomic studies using the recently completed *Arabidopsis* genomic sequence, we must be able to readily manipulate and express all of the genes. Unfortunately, current computational approaches for Arabidopsis gene prediction are not able to precisely predict or, in some cases, even recognize many of the genes. These limitations prohibit the use of new emerging technologies for global gene functional analysis genomes. The aim of our program is to experimentally define the transcription units for all Arabidopsis genes. This will provide an accurate determination of the gene structures and allow the construct full-length cDNAs for each gene. Determining the sequences of the transcription units will resolve ambiguities in the annotated genomic sequence and allow precise position of introns/exons and 5' transcription start and 3' polyadenyation addition sites,. The identification of full-length cDNAs for all *Arabidopsis* genes is of primary importance for the entire plant biology community as these clones will be essential for many future global functional genomic and proteomic studies.

2. Responsibilities and Deliverables of the SSP Consortium.

- Isolation and complete sequencing of full length cDNAs for 8,000 genes with immediate deposit of cDNA sequences in GenBank.

- Construction of 8,000 open reading frame (ORF) clones into a universal recombination plasmid vector (pUNI). The ORF clones, which are fully sequence validated and terror-free, are deposited in the *Arabidopsis* Biological Resource Center (ABRC) at Ohio State University (no MTA) Among the 8,000 ORF clones, 7,000 will be constructed by PCR from full length cDNAs and the last 1,000 are being identified from non-expressed annotated genes (hypothetical).

- Identification novel Arabidopsis transcription units using custom Affymetrix genome tiling arrays and mRNA samples prepared from various plant tissues and conditions.

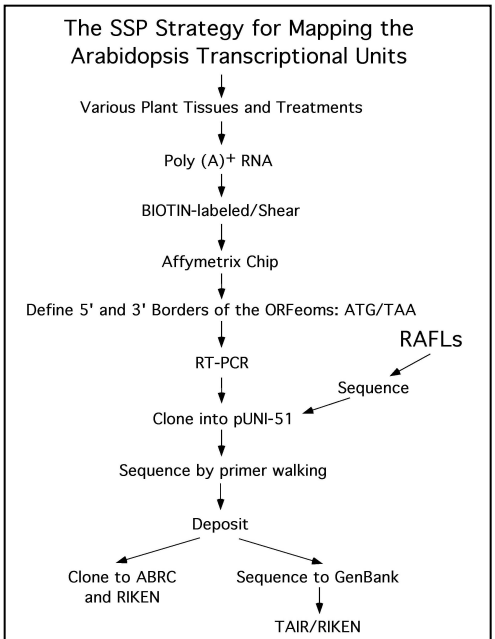

Figure 1. SSPC Strategy Diagram

3. Methodology.

The strategy for isolating full-length/ORF cDNA clones for 8,000 *Arabidopsis* genes is shown in Figure 1. The strategy utilizes three complementary approaches for achieving of our goals:

Approach 1: Construction of ORF clones by RT-PCR (see below for the terminology of various clones). Sixty percent of the *Arabidopsis* genes have an identified ESTs and the source of mRNA for this clone is known. Using RT-PCR and gene specific primers at the ATG and TAA full-length cDNA can be isolated for a larger number of these genes. The annotated ATG/TAA can be tested experimentally to determine whether it is correct by designing RT-PCR primers for potential upstream ATG(s) using the genome sequence.

Approach 2: The second approach utilizes the RIKEN *Arabidopsis* full length (RAFL) clones constructed by Dr Kazuo Shinozaki at the RIKEN Genome Center (http://www.gsc.riken.go.jp/Plant/index.html). This collection was made available to SSP by an agreement between the RIKEN Genome Science Center and the Salk Institute, Stanford University and UC Berkeley. The RAFL collection consists of ~15,000 clones representing ~10,500 unique *Arabidopsis* genes. The RAFL cDNAs (R clones) have been sequenced by the SSP (Table 1). Subsequently the ORF of each RAFL clone is transferred into a pUNI vector by PCR/subcloning and each of the ORF clones (U clones) is then fully sequenced. The sequences of the error-free U clones are deposited in GenBank while the clones themselves are deposited with the Arabidopsis Biological Resource Center at Ohio State University.

Approach 3: Finally, we have developed a novel strategy for identifying the "missing genes" that utilizes custom high density genome tiling arrays constructed by Affymetrix. Use of several different types of custom high density oligonucleotide arrays has allowed the identification of numerous transcriptional units that, thus far, have not been found in any of the deep EST or cDNA collections. We have developed protocols for labeling mRNA and calibrating the hybridization conditions for the transcript mapping chip. Importantly, we have also developed a first generation software tool for scanning of the genome tiling arrays that allows interpreting this massive amount of expression data (see http://signal.salk.edu/msample.html).

3. SSP deliverables: Funding from our first year award was used to develop protocols for all the steps in the strategy in order to carry out the experiments using the Affymetrix arrays. The second year of funding was primarily used for large-scale cDNA sequencing and construction/sequencing of the ORF clones. Essential to the entire enterprise was the development of a cDNA sequence and mapping database, software for automating the

sequencing and annotation procedures for full length cDNA sequencing and ORF production and software for analysis of high density genome tiling arrays.

i. Full-length cDNA and ORF clones construction and sequencing production: Below are definition of the various cDNA clone types being generated by the SSP and the total number of clones constructed, sequenced and submitted to GenBank from each class as of March 15, 2002 (Table 1).

TABLE.
SSP Consortium Full–Length cDNAs and ORF Clone Submissions
(March 15,2002)

Clone	Completed sequences deposited in GenBank
R cDNA Clones	8,729
C pUNI Clones	236
U pUNI Clones	2,257
S cDNA Clones	238
TOTAL:	11,452

R Clone: It is a fully sequenced RIKEN Arabidopsis full-length (RAFL) cDNA clone (including 5' and 3' UTRs) that is clone in Bluescript.
U Clone: It is a cDNA that contains only the ORF (ATG to STOP) of an R clone. This subclone is fully sequence verified and cloned into the universal vector pUNI 51.
S Clone: It is a fully sequenced error-free cDNA product (including 5' and 3' UTRs) generated by RT-PCR using mRNAs prepared from various plant tissues (see below).
C Clone: It is a fully sequenced, error-free pUNI ORF clone generated by RT-PCR using mRNAs prepared from various plant tissues (see below).

Additional details about the DNA sequences and clone/vector information can be found at the SSPC web site, http://signal.salk.edu/SSP/index.html. This site contains all of the SSPC data in one location for ease of access to the community with links to each of the three participants web sites.

ii. Preparation of mRNAs for transcription unit discovery: We have prepared 107 distinct mRNA population from a varity of plant tissues and treatments
iii. Hybridization data: All data from Affymetrix pilot tiling chip and whole genome chip hybridizations experiments used for transcription unit discovery will be available at the end of the project.

Overall assessment of cost. With a 3yr budget of $7.5 million (direct/indirect cost), our NSF funded Arabidopsis full-length cDNA sequencing and ORF clone construction project is the largest publicly funded program of its type. This amount of funding translates to ~$500 sequenced validated cDNA clone. An equivalent project called the

Mammalian Gene Collection (http://mgc.nci.nih.gov/Info/ProjectSummary) is being carried under the sponsorship of 19 NIH and NCI Institutes and involving 22 academic laboratories and companies. The current total unique full-length cDNAs (as of 23- Mar-02) are 7,646 (human) and 4,416 (mouse) for a cost of $25 million. This amount of funding translates to ~$2000 cDNA (with no ORF clone). Therefore, the SSP Consortium project compares favorably with other similar pubic projects, confirming that our approach is very cost effective.

Material distribution. DNA Sequences: All completed cDNA sequences are immediately deposited in Genbank. A variety of cDNA seach tools are available on our web site (http://signal.salk.edu/cgi-bin/sspsearch). **cDNA clones:** Sequence validated, error-free ORF clones in pUNI51 are deposited and available through the ABRC (http://godot.ncgr.org/abrc). Beginning at the end of April 2002, all of the RIKEN Arabidopsis Full-Length (RAFL/R clone) cDNA clones whose full-length cDNA sequences have been determined by the Salk, Stanford, PGEC (SSP) Consortium will be available from RIKEN Bioresource Center. Contact the Bioresource Center (RIKEN BRC) (PI: Dr. Masatomo Kobayasi, Email: kobayasi@rtc.riken.go.jp) for any of the Arabidopsis RAFL cDNA clone. These clones will also become available through the ABRC. See our "where to order from" web page for further details (http://signal.salk.edu/SSP/ssporder.html).

4. Summary. The creation of an easy to use graphical web interface (SIGnAL Arabidopsis Gene Mapping Tool) to our cDNA database and the availability of the corresponding full-length cDNAs and ORF clones in public stock centers provides researchers with ready access to their genes of interest. Full-length cDNAs and ORF clones are prerequisite for the construction of whole proteome arrays, for high throughput protein structural studies and for the rapid creation of protein fusion (GFP, TAP-tagged, etc.) for all proteins. For example, the ability to rapidly create translational fusions for any protein tag to any Arabidopsis protein will allow for large scale in vivo protein complex/mass spectrometry studies. These resources will allow investigators to begin to test hypotheses about plant gene function at an unprecedented rate and an unprecedented scale (i.e. thousands of genes in parallel).

5. Citation of the project. Since we plan to submit the results of this study for publication, we request that you do not cite this project summary as a reference to our project. Instead, until publication, we suggest the following acknowledgement: "We thank the Salk, Stanford, PGEC (SSP) Consortium and the RIKEN Genome Science Center for providing the sequence validated full-length cDNAs." Finally, we request that investigators include the Genbank accession numbers for RAFL cDNAs and SSP ORF clones in all publications that describe cDNAs produced by our consortium.

Acknowledgement. This project is supported by the National Science Foundation Plant Genome Research Program Grant: DBI9975718/0196098