NSF 2010 MCB MCB0115103
Title: A Sequence-Indexed Library of Insertion Mutations in the Arabidopsis Genome
PI: Joseph R. Ecker, The Salk Institute for Biological Studies, EMAIL: ecker@salk.edu

Overall Goals: With the availability of the entire Arabidopsis genome sequence ([, 2000 #37]), the next challenge is to begin to uncover the functions of the more than 25,500 genes in this reference plant [Chory, 2000 #30]. Given the scope of the National Science Foundation 2010 program, "to identify the function of all Arabidopsis genes in the next decade", more efficient and cost effective (systematic) approaches will be necessary to identify mutations in as many genes as possible. The goal of our NSF funded 2010 project grant is to create a sequence-indexed library of mutations in the Arabidopsis genome.

Method: The Salk Institute Genome Analysis Laboratory (SIGnAL) has established high-throughput genome sequencing methods to identify the sites of insertion of *Agrobacterium tumefaciens* T-DNA insertions in the Arabidopsis genome. Individual T-DNA transformed plants from the Alonso/Crosby/Ecker collection (Col-0 strain) are grown in a 96 well format, genomic DNA is prepared, flanking plant DNA is recovered by adapter ligation/suppression PCR amplification of the T-DNA insertion site and DNA sequence of the products are determined. As is typical for T-DNA transformation, ~50% of the transformed plants contain more than one T-DNA integration event. However, no attempt is made to physically separate the products prior to sequencing, as this would create unmanageable tracking issues. In most cases, where two or more plant flanking sequences are amplified from a single plant line, a single high quality DNA sequence is obtained from the longest insertion site PCR product. Each T-DNA sequence is aligned with the latest version of the annotated Arabidopsis genome in GenBank (current version: January 10, 2002). A single best location (based on E value) for each insertion sequence is determined, and annotation of a best approximation of insertion site is added (5'UTR,exon, intron, 3' UTR) (see FAQ page for more details). The sequence data is made available via a web accessible graphical interface- SIGnAL Arabidopsis Gene Mapping Tool (http://signal.salk.edu/cgi-bin/tdnaexpress) that provides both text and DNA searches of the insertion mutant database. All T-DNA insertion site sequences with genome homology are deposited into GenBank (GSS Division) and also provided to The Arabidopsis Information Resource (TAIR- http://www.arabidospis.org).

Resources made available: Each month, seeds (~100 µl vol.) from each Salk T-DNA insertion line is deposited with the Arabidopsis Biological Resource Center (ABRC) at Ohio State University. The ABRC distributes seed to the community and to the Nottingham Arabidopsis Stock Centre; the SIGnAL laboratory does not distribute seeds to individual investigators. ABRC is propagating a subset of the Salk T-DNA insertions mutants. Each month we provide ABRC with ~ 6,250 insertion lines (~63 boxes of 100 individual T-3 generation seeds) and a corresponding gene "hit" list. This allows the ABRC to prioritize their seed propagation program to initially focus on amplification of plant lines containing insertions within genes (vs. lines with T-DNA insertions between genes). Importantly, no attempt is being made to identify lines that are homozygous for the insertion. Investigators are cautioned to confirm the presence of the expected T-DNA insertion using PCR (see FAQ page). We have made every attempt to reduce tracking and contamination problems. However, like other high throughput operations, it is inevitable that due to mechanical or human error such events will occur. Therefore, the Salk

insertion lines are provided to the ABRC "as is". Users are expected to confirm our results before initiating their experiments. Please check our FAQ page (http://signal.salk.edu/tdna_FAQs.html) for experimental details regarding the confirmation of insertion targets before contacting the PI with questions.

Progress: During the 6 months since initiation of funding (September 1,2001), we have identified ~32,500 sequence indexed insertion lines and the seeds for each corresponding mutant have been made available through the ABRC. This number corresponds to ~ 5,000 insertion mutants/month and translates to ~ 9,000 unique gene mutations. To put these results in perspective, the total number of Arabidopsis gene mutations identified by this program is greater than the entire accumulated total number of community identified gene mutations available in the public domain.

Cost: The full cost of the project (direct and indirect costs) for propagation of individual lines, preparation of genomic DNA, ligation/PCR amplification of plant flanking sequences, DNA sequencing reaction/product separation, sequence analysis, insertion site gene annotation, database development, cleaning/packaging of seed, bar coding/shipment of individual mutant lines is $20.00/individual Salk insertion line, easy to use graphical web interface (SIGnAL Arabidopsis Gene Mapping Tool) in conjunction with and the availability of the corresponding mutant lines in public stock centers provides researchers with ready access to complete or partial loss-of-function mutants in most Arabidopsis genes. This resource will allow investigators to begin to test hypotheses about plant gene function at an unprecedented rate and an unprecedented scale- thousands of genes in parallel.

Acknowledgement. This project is supported by a grant (MCB0115103) from the National Science Foundation Eukaryotics Genetics Program.