



COLLEGE OF AGRICULTURE AND  
ENVIRONMENTAL SCIENCES  
AGRICULTURAL EXPERIMENT STATION  
DEPARTMENT OF PLANT PATHOLOGY  
TELEPHONE: (530)752-0300  
FAX: (530)752-5674

ONE SHIELDS AVENUE  
354 HUTCHISON HALL  
DAVIS, CA 95616-8680

May 9, 2007

Dr. Anita Klein  
National Science Foundation  
Plant Genome Program

Dear Dr. Klein:

As chair of the Scientific Advisory Board (SAB) for Dr. Ronald Sederoff's Fagaceae Genome Project, I am writing to provide our assessment of project activities to date. The annual meeting for the Fagaceae Genome Project was held at the Penn State Mount Alto campus on April 30-May 1, 2007. SAB members in attendance were myself, Dr. Jennifer Koch and Dr. Jeanne Romero-Severson. Dr. Jeff Dangl, Dr. Susan McCouch and Dr. Antoine Kremer were unable to attend, while Dr. Gerald Tuskan has resigned due to a heavy professional load.

The meeting consisted of presentations from each project coPI, describing progress to date and near term plans. In addition to ad hoc questions and dialogue that followed each presentation, the SAB met in closed session on the morning of May 1 to evaluate progress in the Fagaceae Genome Project and to formulate our recommendations. A summary of these deliberations is presented below.

Please let me know if you require clarification on any of these items.

Best regards,

A handwritten signature in black ink that reads "Doug Cook".

Doug Cook

---

#### **Evaluation and comments on progress to date**

The SAB recognizes that the research team has made progress on essentially all project objectives, in most cases providing a solid foundation for research in the coming year. Perhaps the two most significant advances are the production of an apparently high quality 11X HindIII Bacterial Artificial Chromosome (BAC) clone library of Chinese chestnut and the generation of initial transcript sequence data from the 454 platform. Physical map construction using the High Information Content Fingerprinting (HICF) approach of Dvorak et al has recently begun, suggesting that significant

progress on the physical map will be made in the coming project year. Physical map development will be augmented by the production of a second BAC library for Chinese chestnut based on a different restriction enzyme. Progress on the physical map during the coming year will be an important milestone for the project.

All 454 data analyzed to date was obtained using the original 454 platform, which suffers from short reads. The project's instrument was recently upgraded to the 454 FLX model and 2 ¼-plate sequencing runs have been conducted. Despite the relatively short (average 100 bp) reads in the initial 454 data set, the project has produced several million base pairs of non-redundant (i.e., contiged) transcript data from infected tissue cDNA libraries of American and Chinese chestnut genotypes. Preliminary analyses of all data include contig assembly and BLAST vs Arabidopsis and Populus, with more thorough analyses of gene structure and content, rates of potential polymorphism, and comparisons between the two genotypes to be conducted. A cursory analysis for sequence variation suggests that SNP and SSR polymorphisms will be relatively abundant. With the upgraded instrument in hand and several cDNA sources now ready for analysis it is reasonable to expect that the majority of polymorphisms required for subsequent molecular-genetic analysis will be identified in the coming months.

Less progress has been made in sequencing of ESTs by means of Sanger sequencing using the ABI 3730xl. However this is not viewed as a problem, as the relatively modest sequencing goal of 20K paired-end reads can be completed quickly, once high quality cDNA libraries are produced. Thus, the production of high quality cDNA libraries is an important priority in the coming months. The tradeoff between redundancy in non-normalized cDNA libraries versus shorter clones in subtracted/normalized cDNA libraries has been a topic of some discussion among project investigators. The SAB endorses a model where the emphasis is placed on cDNA insert size, dispensing with the idea of subtractive hybridization for normalization. Sequencing a modest number of clones (e.g., 1,000 to 2,000) from each of several high quality libraries that represent distinct tissues and/or conditions is likely to offer a high rate of novel transcript discovery while retaining cDNA integrity. Prior to a given cDNA library entering bulk sequencing, paired end sequencing of 48 random clones (5' and 3' reads for each of 48 clones = 96-well test plate) could provide an initial check of transcript redundancy and cDNA quality. It will be important to calculate more than raw insert size; the extent to which individual cDNAs are near full-length should be estimated based on comparison to established protein databases and gene structures that are known or predicted from well-characterized genomes, such as Poplar, Medicago and Arabidopsis.

SNP data will be mined based on comparison of 454 to Sanger data, and also by comparison between 454 data sets. It will be important to establish criteria for distinguishing sequence error from genotype variation; for example, based on the independent occurrence of particular nucleotide variants (alleles) or by means of experimental validation such as re-sequencing of a subset of putative alleles. Having a firm grasp of this issue will be relatively important as the project moves to the genotyping phase.

On the bioinformatics side, a project web site has been produced by the Clemson group. Similarly, the project LIMS system is functional, or nearly so, and an initial but cursory analysis of sequence data has been conducted. Developing a mature project web site, with data analyses and views that support the project's molecular-genetic objectives and that will also enable activities by Fagaceae researchers

external to the project will become increasingly important in the coming year. This issue is discussed in some detail, below.

The production and curation of mapping populations appears to be progressing well. Current estimates are that roughly 300 full sib advanced backcross progeny are available for both American X American (GHorn X GMBig) and Chinese X Chinese (Mahogany X Nanking). The major development activity in the coming year is to increase the size of the American X Chinese F2 population.

Because a fraction of the mapping progeny are expected to be the product of uncontrolled outcrossing, it will be important to test all progeny for correct parentage by means of PCR. This analysis deserves a degree of priority, as the results will dictate whether additional progeny need to be developed. As a prelude to this analysis (and for genetic mapping), SSR markers should be tested across all parents of both Chinese and American populations to verify their utility across populations and to test for unbiased amplification in heterozygous backgrounds.

Various international collaborations and outreach activities are ongoing or planned. The project has awarded their first Kenan Fellow, a high school teacher from the Raleigh area who is now working with the project to develop a K-12 teaching module. Agreements for data sharing and research collaborations with the European EvolTree project are progressing; already SSR data has been provided to US investigators and a plan is in place to for the Penn State group to provide 454 sequencing related to parallel EvolTree objectives. A formal material transfer agreement is currently being negotiated.

## **Recommendations and suggestions from the SAB to project investigators**

### **1. Public web site for genes and markers:**

To maximize value of the project's data for Fagaceae researchers external to the project, it will be important to develop organized, web-accessible views of project data. For the first 2 years of the project the primary data type will be transcript sequences, derived from a modest level of Sanger sequencing in a single genotype and relatively deep 454 sequencing of transcripts from multiple genotypes and species.

Project investigators need to develop a scheme to organize these data in ways that benefit both the project and the broader Fagaceae community.

Primary data types internal to the project will be near full-length cDNAs from a reference genotype of Chinese chestnut (this will provide a reference transcript set), 454 data (some fraction of which will have identity to the FL-cDNA reference set and much that will not), and eventually BAC end data. Additional dimensionality will be the presence of 454 data from multiple genotypes within individual species, as well as data from distinct species. Important data types external to the project include whole genome sequences and deduced proteomes from model plant species, and the larger collection of all NR proteins.

Analyses of the Fagaceae transcript set will add additional dimensionality to the data, and this will need to be incorporated into the database structure. Examples include the use of sequence alignments to relate 454 data to the reference FL-cDNA set, as well as to the genomes of model species;

polymorphisms (SNPs and SSRs) can be discovered by sequence alignments or motif finding programs; minimum predicted gene sets for each species can be derived from programs such as MegaBLAST and CAP3; rationale annotations for deduced proteins can be obtained by comparison to public resources such as the annotated Arabidopsis genome and the GeneOntology initiative; gene structure information (including intron prediction, and coding and non-coding transcript regions) can be deduced by comparison to the sequenced genomes of related model plants; sets of potential oligonucleotide primers that span SSRs and/or introns, or that target SNPs for genetic analysis can be developed; transcript abundance in 454 data can provide estimates of gene expression in specific tissues and/or conditions; etc.

A key question is how should this information be organized to suit the needs of the project and external investigators? Data for individual species will need to be organized independently to derive contigs, transcript frequency information, and organize metadata (e.g., tissue of origin, genotype, etc). Comparisons will then need to be made between individual species to produce a Fagaceae-level view of the transcriptome. When homologs are available, the reference FL-cDNA collection from Chinese chestnut can provide an organizing principle for closely related sequence contigs and singletons from the 454 data of all Fagaceae species. In many cases, however, the breadth of 454 data will exceed that of the Sanger FL-cDNA data; in such cases, a closely related model genome, i.e., Medicago and/or Populus, can serve as a surrogate for the FL-cDNA reference set.

Cross genome comparisons should be a central focus of the database. Sets of highly similar homologous sequences (potential orthologs, perhaps identified by means of a reciprocal best hit approach) will need to be identified and organized into a searchable database and web interface. The goal should be to allow researchers to mine data within and between species, identifying polymorphisms, accessing suggested primer pairs, comparing transcript frequencies between species, making key word queries for specific gene functions, etc..

The above discussion is intended as food for thought for the investigators, not as a strong recommendation for any specific strategy. It will be valuable if a prototype database schema, including proposed functionalities and possible screenshots of a user interface can be developed in the coming months. If available by the time of the September project meeting in Virginia, there would be an opportunity to present the concept in a public forum and solicit input from the community.

## **2. Prioritize transcript sequencing:**

The SAB expects that maximum value from transcript sequencing will be delivered by a two-tier strategy: (1) Sanger sequencing is focused on near-full length cDNA libraries using a paired end approach, with the goal of producing a partial but high quality reference transcript set for the Fagaceae; (2) 454 sequencing across genotypes and species, with the goal of gene and polymorphism discovery. The 454 data will also provide data for a rudimentary expression analysis based on EST frequency.

This recommendation is consistent with the current project strategy, but the SAB wants to emphasize the importance of producing high quality cDNA libraries before sequencing on the ABI 3730xl is scaled up.

### **3. Prioritize selection of ESTs for mapping and BAC hybridization:**

The stated goal of the project is to map ~500 ESTs by means of EST-SSRs. Given that the number of EST-based SSRs is likely to be large compared to the stated objective (525 SSRs were reported from the initial 454 survey), and because not all ESTs will have equal value as genetic markers, we recommend that the project develop a rubric for prioritizing the selection of ESTs for markers.

Possible issues to consider are:

*a. Select SSRs from the 3' untranslated regions of genes, where rates of polymorphism are highest.*

*b. Focus on SSRs with total length >20 nucleotides and motifs of 2-7 bp.*

*c. Prefer low copy genes over gene families.* Low copy genes will have greater utility for construction of comparative maps within the Fagaceae, as well as for relating genome structure to *Populus* and *Medicago*. The *Populus* and *Medicago* genomes are likely to be useful references to eliminate moderate sized gene families.

*d. Prefer transcripts that overlap intron regions.* Some EST-SSRs will not be conserved across species. In such cases, mapping the homologs in other Fagaceae species (for example, oak or beech) may require the use of EST-associated SNPs, which will be most abundant in introns. Intron position can be predicted with reasonable confidence by alignment of Fagaceae ESTs to highly conserved genome regions in *Populus* and *Medicago*.

*e. Other criteria?*

It should be possible to automate the identification of these features in the transcript dataset, and thus to rank ESTs by priority without the need for manual intervention.

### **4. Using the Illumina platform for SNP detection:**

It seems likely that a relatively large number of SNP polymorphisms will be evident in the 454 sequence data. Project investigators are considering the use of recent technical innovations for SNP analysis that will allow mapping of larger numbers of genes and at lower cost per data point. SNP analysis on the Illumina platform using the Golden Gate assay is one such technology to consider, *and the SAB is very supportive of this approach.* Depending on the number of individuals analyzed in a biparental mapping population, and the nature of variation in the heterozygous parental genotypes, it may be possible to bring the cost per data point into the range of \$0.06.

Using such an approach would permit the project to move directly from sequence analysis to genetic mapping, without the need for PCR or fragment analysis.

### **5. Provide a plan for the genetic mapping and phenotyping of disease resistance in hybrid populations.**

It would be helpful for the SAB members to have access to a general scheme for the proposed mapping of disease resistance phenotypes, including which populations will be used for interval mapping, how phenotyping will be conducted, how the efficiency of the assay (i.e., rate of escapes) will be

determined, etc. This could be in the form of a diagram and provided to the SAB prior to the September project meeting.

#### **6. Project management.**

The SAB would like to have a concise summary of the major responsibilities and the anticipated timeline for the MAJOR deliverables from each participating laboratory. This might take the form of a schematic or short paragraph for each research site.

The SAB recommends having a regular (every other week, for example) teleconference. This would provide a venue for individual investigators to present recent progress, hear feedback from other researchers, and would generally enhance interaction among the project sites.