

University of California  
Division of Agricultural Sciences

## PROJECT PLAN/RESEARCH GRANT PROPOSAL

Project Year 2008-09 Anticipated Duration of Project 4 years

Project Leader Abhaya M. Dandekar Location UC Davis

Cooperating Personnel Jan Dvorak, Malli Aradhya, Ming Ching Luo, Gale McGranahan, Chuck Leslie

Project Title Walnut genome Analysis

Keywords Physical mapping, DNA genotyping, association mapping, genetic mapping, EST sequencing, genome database, walnut, genomics

Commodity(s) Walnut, Juglans regia Relevant AES/CE Project No. 5271-H

### **Problem and its significance:**

The California walnut industry is chiefly concerned with producing high yields of superior quality walnuts, while minimizing production costs and controlling pests and diseases. The ideal walnut cultivar would be late leafing yet harvest in early autumn, and would combine high production capacity with reduced chemical input. The nutshell would be relatively smooth, well sealed, and comprise no more than 50% of the nut weight. The kernel would be plump and light colored, weigh over seven grams, and come out easily in halves. In addition, walnuts contain high levels of polyunsaturated fats and are an excellent source of omega-3 fatty acids, which have been shown in clinical studies to be beneficial to cardiovascular disease, diabetes, and other medical conditions. Therefore, maintenance and improvement of these nutritional traits is very desirable.

It is also very desirable to be able to predict these traits in seedlings and saplings and identify potentially superior genotypes. Traditional breeding is time-consuming and resource-demanding due to the long juvenile period before progenies can be evaluated. In addition, traits change over time and do not stabilize until the tree is about 15 years old. Molecular tags tightly linked to economically valuable traits could identify seedlings with high potential, and could also be used to eliminate the majority of undesirable seedlings within a short time after germination.

Rootstocks are primarily interspecific hybrids, such as 'Paradox' (*J. hindsii* x *J. regia*), which are difficult to breed due to male sterility in the progeny. Desirable characters in rootstocks are resistance to pests and diseases (e.g., *Phytophthora*, crown gall disease, nematodes, blackline, etc.) and traits which enhance desirable characteristics of the scion, such as improved tree architecture and vigorous growth.

Genomic technology is an integral component of meeting the constraints facing today's agriculture industry. Applications of this technology include: 1) rapid and efficient production of new rootstocks and scion cultivars to meet growers' needs and consumer preferences; 2) modification of orchard management to reduce the costs of land, labor, and chemical inputs; 3) accurate and rapid assessment and amelioration of plant health problems using global genomic expression of genes, proteins, and important metabolites; and 4) closely associated methods to rapidly identify pests and diseases.

During the 1990s, genomic initiatives in model organisms, including plants such as *Arabidopsis* and rice, led to improvement of major agricultural crop sectors including grasses (rice, wheat, and maize) and legumes (soybean). The infrastructure and innovation developed in large genomics projects have increased the throughput and reduced

the cost of similar endeavors that benefit other crops. Major genomics initiatives are already underway for many fruit, nut, and timber crops such as citrus, prunus (e.g., peach and almond), apple, grape, poplar, olive, and chestnut. The genome of poplar has been sequenced and sequencing of the apple, grape, peach, and citrus genomes will be completed by 2008.

Data from these ongoing projects can be used, in turn, to improve the efficiency of similar initiatives for other plants. Therefore, now is the ideal time to leverage our fundamental understanding of model systems to address the needs and desires of the walnut community. We present here an integrated genomics project that begins with a combination of gene expression analysis and development of genetic and physical maps of the walnut genome, and will ultimately lead to complete genome sequence analysis.

### **Objectives:**

The primary goal of this proposal is to build a set of comprehensive genomic tools, integrate them with available genetic and molecular resources to provide greater precision in evaluation of breeding populations, and thus facilitate rapid development of improved walnut cultivars addressing the needs of growers and consumers at large. To accomplish this goal, three major components must be developed and integrated into the ongoing walnut genetic improvement program. These three key components include: (1) structural genomics, or construction of a physical map of the walnut genome; (2) functional genomics, which is a detailed survey of walnut gene expression; and (3) fine-scale genetic mapping to construct pedigree-based high density linkage maps and association analysis. Together, these will build a knowledge base that will support and significantly strengthen ongoing California walnut breeding efforts. This synergy will allow development of genetically sound marker-assisted selection strategies to significantly increase selection efficiency, trait identification and integration, and rapid genetic gains in accelerated development of improved walnut cultivars.

#### **Specific Aim 1: Physical mapping of genetic traits in the walnut genome**

A physical map of the walnut genome will be built concurrent with development of the genetic map. Unlike genetic maps, where the position and arrangement of genetic markers are determined based on recombination frequencies, the physical map will allow specific genes and genetic markers to be precisely positioned and arranged along the chromosomes based on physical distances measured in base pairs. Physical maps are also rich sources of genetic markers which can then be used to enrich the genetic map. The development of a physical map provides a scaffold representing the chromosomes and is a necessary precursor to facilitate assembly of the complete walnut genome when such sequencing is performed.

#### **Specific Aim 2: Genetic and association mapping of economic traits in walnut.**

Genetic maps are constructed by analyzing genetic markers in segregating populations derived from crosses between cultivars selected for progeny that will segregate for traits of interest. Genetic recombination between markers during meiosis permits estimation of map distances, which in turn allows ordering markers within and between linkage groups. Association analysis/mapping, in contrast, exploits correlations (linkage disequilibrium; LD) between markers and economic traits (QTL) that exist in natural populations as a consequence of mutation, recombination, selection, and drift on an evolutionary time scale. The interplay of evolutionary forces, especially genetic recombination, in natural populations gradually erases loose linkages, leaving behind only genes that are tightly linked. As a result, association analysis presents an excellent opportunity for higher-resolution mapping of molecular markers with genes controlling simple and complex traits than pedigree-based linkage analysis, which accounts for a limited recombination among genetic loci. Overall, the main goal of the genetic mapping is to link genetic markers with specific economic traits of interest so that this information can then be used to design an effective marker-assisted juvenile selection strategy to improve selection efficiency, traits integration, and rapid cultivar development addressing the needs of growers and consumers.

#### **Specific Aim 3: Functional mapping of the walnut genome**

While genetic and physical maps describe the structure of the genome, it is also essential to precisely document gene

expression, allowing specific traits to be linked to underlying metabolic and biochemical processes. This is accomplished through sequencing of gene transcripts which, in turn, allows identification of expressed genes. Several tissue-specific gene transcript libraries will be constructed and sequenced to generate thousands of Expressed Sequence Tags (ESTs). The ESTs will be deposited in public databases where the information can be processed through computer programs to identify genes involved in important metabolic pathways. These sequences are used to generate experimental tools such as microarrays that investigators can use to analyze expression levels of thousands of genes in parallel, which allows comparisons to be made between different cultivars and conditions (e.g., diseased and healthy plants). ESTs are also important for structural analysis of the walnut genome by facilitating both detection of genes on the physical maps and map-based cloning of genes.

#### **Specific Aim 4: Development of a ‘Walnut Genome Resource (WGR)’, a web-based knowledgebase of walnut genomic information.**

A web-based browser will be developed for the walnut research community to access genomics resources. The database will contain all association and physical mapping information as well as all ESTs and their integration with the walnut physical and genetic maps.

#### **Plans and Procedures:**

The genus *Juglans* resides within the order *Fagales*, which contains several genera of important nut and timber trees including birch, beech, oak, chestnut, butternut, and pecan. Very little genomic data currently exists for any of these plants, although there is a genomics initiative for chestnut that is primarily targeted towards resistance to chestnut blight. Data developed for any of these trees, including walnut, can be used to improve the knowledge base for all related tree species. Outside *Fagales*, the model plants most closely related to walnut are legumes (i.e., *Medicago* sp., *Lotus* sp., soybean) and poplar. Slightly more distant relatives include grape and *Arabidopsis*.

Within *Juglans*, desirable timber and rootstock traits are found in black walnut species (e.g., *J. hindsii*, *J. californica*, and *J. nigra*) and yield and nut quality traits in *J. regia* germplasm. Through development of genomics tools, the germplasm available at Davis in the UC Germplasm collection and the National Clonal Germplasm Repository can be screened to discover desirable traits that can be brought into the breeding program to improve both scion cultivars and rootstocks.

In defining the objectives of a walnut genomics program, it is important to include American black walnut (*J. nigra*), which is grown as a forest tree in both natural forests and in plantations in the eastern United States, where it is an important timber and nut crop. Inclusion of other *Juglans* species and strategic alliances with other closely related industries, including pecan and chestnut, should attract nationwide interest in funding a comprehensive genomics program for *Fagales* species.

#### **Specific Aim 1: Physical mapping of genetic traits in the walnut genome**

Unlike genetic maps, which order markers and genes based on linkage strength, physical maps order markers and genes based on physical location along the chromosome. Distances between markers and genes on genetic maps are in recombination units, which do not reflect actual distances between markers and genes on the chromosome, but distances between markers and genes are in nucleotides on physical maps, and do reflect physical distances between markers and genes. For this reason, physical maps are an essential prerequisite for gene isolation schemes. Physical maps are also rich sources of markers for genetic map construction and development of marker-assisted selection schemes. Ultimately, the genomic scaffold generated by a physical map facilitates assembly of shotgun genome sequences.

Construction of physical maps is entirely lab-based and highly automated. Hence, the rate of progress is the same as it is for any model plant species as it would be for a long-lived tree such as walnut. Physical mapping of a genome uses the following basic strategy: The genome is segmented with restriction enzymes that cut DNA into large fragments. These fragments are inserted into a Bacterial Artificial Chromosome (BAC) vector, creating a BAC library. Usually, several libraries are developed using different restriction enzyme cut sites to improve the diversity

and distribution of DNA segments in each BAC library. BAC clones are then randomly picked to 15X or more genome equivalents and then fingerprinted. We developed a fluorescence-based, high-throughput BAC DNA fingerprinting technique (Luo et al. 2003) which sizes DNA fragments from each BAC clone by capillary electrophoresis creating a fragment profile that is unique for each BAC clone, a fingerprint. With a single robotic DNA sequence analyzer (96-capillary ABI3730XL), about 1,000 BAC clones can be fingerprinted daily. To match the high-throughput of the BAC fingerprinting technique we also developed computer software for rapid editing (GenoProfiler; You et al. 2006) and matching (FPC computer program; Soderlund et al. 2000) of BAC fingerprints. These computer assisted fingerprint matching programs are used to identify overlapping BAC clones. Contiguous sequences contained within the overlapping regions of BAC clones (contigs) are then assembled, identifying specific sequences of nucleotides along the chromosome.

The identification and extraction of DNA sequence based markers located on these BACs is the next step which then integrates the physical mapping information with that generated with genetic and functional mapping explained below. For this at least one end of each BAC clone (BAC-ends) is sequenced on the same robotic DNA sequence analyzer explained above with essentially the same DNA preparation created for the fingerprint analysis. Each BAC-end DNA sequence (BES) contains DNA sequence information for a discreet portion of the walnut genome. Since this sequence represents randomly distributed segments of the walnut genome and some of these segments may overlap with regions that contain a gene, those regions are referred to as Gene Sequence Tags (GSTs). The GSTs can be generated by comparing the BES to the Expressed Sequence Tags (ESTs) generated in the functional mapping explained below to rapidly identify the GSTs. Recently we were able to confirm this methodology with *Brachypodium distachyon* (genome size 380 Mb) where 20% of the BES matched ESTs, since walnut is double the size of this plant we would expect at least 10% to match ESTs. Hence, the 120,000 BAC-end sequences planned here we could expect to detect as many as 12,000 different walnut genes. GSTs also provide means for rapid identification of Single Nucleotide Polymorphisms (SNPs) present in the walnut genome. SNPs are essential for rapid and low-cost genetic map construction utilizing the Illumina Golden Gate™ high-throughput genotyping platform. With this strategy, we have recently constructed a genetic map of *Aegilops tauschii*, a close relative of wheat, containing nearly 1,500 SNPs, in less than two months (M.C. Luo, Y-Q Ma, and J. Dvorak, unpublished).

### **Specific Aim 2: Genetic and association mapping of economic traits in walnut.**

We propose two different approaches for walnut genome mapping: (1) Linkage analysis in a pedigree context using a conventional mapping population derived from a bi-parental cross between diverse parental lines differing for traits under consideration; and (2) Association genetic analysis using a natural population such as a germplasm collection consisting of genotypes of unknown or mixed ancestry that represent a common gene pool.

1) Pedigree-based molecular linkage map development: A mapping population consisting of an F<sub>1</sub> trees of a diverse cross ('Chandler' x 'Idaho') is being developed. The parental cultivars were selected so that the progeny segregate for a number of economic traits. A mapping population of 100 individuals from this cross is just beginning to bear. An additional 500 nuts from this cross were produced in 2005 and have been germinated. The cross was repeated in 2006 and the resulting nuts will be germinated in the spring of '07. The bearing F<sub>1</sub> trees will be checked to confirm the parentage and for normal segregation of traits of interest. Altogether, about 300 true F<sub>1</sub> trees will be genotyped with about 300 microsatellite markers and about 1,500 SNPs. Genetic map containing both SSR and SNP markers will be assembled. For phenotypic evaluation of economic traits, seedlings will be grafted on to 'Paradox' rootstocks and field experiments will be established in two locations by following an augmented block design with standard check cultivars. All standard horticultural practices will be followed to establish mapping populations and data collection will be begin as the traits appear at appropriate times after planting as recommended by the walnut breeder.

2) Association analysis (Measurement of linkage disequilibrium): The walnut germplasm collection maintained at the USDA repository representing a wide spectrum of diversity for economic traits within *J. regia* will be used for association analysis. All the genotypic data (300 SSR and 1500 SNP loci) from this collection will be subjected to a population structure (Q-matrix; Pritchard et al., 2000) and relative kinship (K-matrix; Hardy and Vekemans, 2002)

analyses. The significance of LD between molecular markers will be examined using Fisher's exact test (Lewontin, 1995; Weir, 1996). General linear, step-wise, and mixed regression models will be used to examine the association of molecular markers with the economic traits. In order to account for structured association, the marker-inferred population structure (Q-matrix) and relative kinship (K-matrix) will be incorporated in the mixed model (Yu et al. 2006) as covariates in all the association analyses. Because the walnut F<sub>1</sub> mapping population is still under development, the project will be started with the association mapping approach using the walnut germplasm collection maintained at the USDA Germplasm Repository in Davis followed by linkage map development using pedigree approach.

### **Specific Aim 3: Functional mapping of the walnut genome**

Functional mapping profiles the transcriptome, or the transcribed RNA, of any particular organism. In plants, gene expression is both temporally and spatially regulated in different tissues. A profile of the transcriptome in a plant organ such as fruit, for example, would sample all mRNA expressed in this organ, and thus represent all genes expressed in fruit. Since only the functional part of the genome is observed, this provides a rapid and direct way to analyze genes that regulate fruit traits. Single stranded mRNAs are easily converted to double-stranded cDNA via *in vitro* cDNA synthesis and cloned into a bacterial vector system. Complimentary DNA (cDNA) libraries represent the mRNA population of a particular plant and tissue at a particular time during development. Randomly sequenced individual cDNA clones can be used to create an EST database (a repository of all EST sequences for a particular commodity) to catalog the sequences. An EST represents a single or paired DNA 'long' sequence run of the 5' and/or 3' ends of an individual cDNA clone and corresponds to an individual mRNA. Random analysis of a cDNA library provides a random sampling of corresponding tissue mRNA. The key is to sample, effectively and efficiently, the less abundant or unique mRNAs that represent the greatest diversity of genes expressed in a plant organ like fruit. This requires extensive sequencing of 'short' stretches of individual cDNAs. New ultra-high-throughput (UHT) DNA sequencing approaches using the Life Sciences 454 machine (<http://www.454.com/enabling-technology/index.asp>) and massively parallel signature sequencing (MPSS) using Solexa technology (<http://www.solexa.com/wt/page/mpss>) are strategies that can profile an entire transcriptome in a single run by compiling many short sequences corresponding to a greater diversity of the mRNA population. Here, mRNA from different tissues can be pooled to increase the diversity of the mRNA profile.

In 2001, we constructed the first cDNA libraries of walnut embryo tissues and deposited ~4000 ESTs in GenBank. In 2004, we prepared the first cDNA library of *P. vulnus*, a key walnut pest, and deposited ~2500 sequences representing the first known protein-coding sequences for this nematode pest. Last year, we constructed and sequenced five new cDNA libraries, four from walnut and one from *P. vulnus*, and deposited an additional 16,000 EST sequences. Our lab has produced over 95% of the more than 18,000 walnut sequences currently in GenBank. These sequences are available via the NCBI GenBank (<http://www.ncbi.nlm.nih.gov/>) and the UC Davis CAES Genomics Facility website (<http://cgf.ucdavis.edu/>). Our laboratory has extensive experience building cDNA libraries for EST sequencing in several plant species including citrus, prunus, apple, and walnut. All of our sequences are publicly available through NCBI and the UC Davis CGF websites. We also have experience using microarrays under several different platforms (e.g., Affymetrix, Combimatrix) to analyze differential gene expression in tomato, apple, and citrus tissues.

### **Specific Aim 4: Development of a 'Walnut Genome Resource (WGR)' a web based knowledgebase of walnut genomic information.**

A genome resource or knowledgebase is a database that stores and visualizes genetic, physical, and functional mapping data. This resource will have two distinct components: one for visualizing the genetic map and one for visualizing physical maps. Tools are available to integrate and represent this information. The physical map is a scaffold on which to integrate phenotypic traits, molecular markers, and DNA sequence data. The main objective of functional mapping is gene annotation, with emphasis on functional categorization of ESTs. For example, we can transform gene expression data and visualize expression changes in entire pathways and categories of genes in apple using the MapMan tool developed by Mark Stitt at the German Resource Center for Genome Research (<http://gabi.rzpd.de/projects/MapMan/>) (Thimm et al. 2004). MapMan project collaborators have developed an ontology which classifies *Arabidopsis* genes into 35 broad categories and nearly 2000 sub-categories corresponding

to all known functions in *Arabidopsis*. The Image Annotator, MapMan's data visualization tool, includes several pathway diagrams that show expression levels and functional classifications of many individual genes. The software tool can be customized for conditions that match more closely with our particular study, and we can add our own visualization diagrams. To date, the standard ontology and pathway figures of the MapMan package have enabled us to visualize expression data from experiments on tomato, apple, citrus, and grape based on orthologs to *Arabidopsis* genes in each species.

## References:

- Hardy, O.J. and Vekemans, X. (2002). SPAGeDi: a versatile computer program to analyze spatial genetic structure at the individual or population levels. *Mol. Ecol. Notes* 2: 618-620.
- Lewis P.O., Zaykin D. (1997) Genetic data analysis: Computer program for the analysis of allelic data. Version 1.0. A free program distributed by the authors over the internet from the GDA Home Page at <http://chee.unm.edu/gda>
- Lewontin, R.C. (1995). The detection of linkage disequilibrium in molecular sequence data. *Genetics* 140: 377-388.
- Luo M.C., Thomas C., You F.M., Hsiao J., Ouyang S., Buell C.R., Malandro M., McGuire P.E., Anderson O.D., Dvorak J. (2003) High-throughput fingerprinting of bacterial artificial chromosomes using the SNaPshot labeling kit and sizing of restriction fragments by capillary electrophoresis. *Genomics* 82:378-389
- Pritchard, J.K. Stephens, M. and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics* 155: 945-959.
- Soderlund C., Humphray S., Dunham A., French L. (2000) Contigs built with fingerprints, markers, and FPCV4.7. *Genome Research* 10:1772-1787
- Weir, B.S. (1996). Genetic data analysis II. Sunderland, MA: Sinauer Associates.
- You F.M., Luo M.C., Gu J.Q., G.R. L., Dvorak J., Anderson O.D. (2006) GenoProfiler: Batch Processing of High Throughput Capillary Fingerprinting Data. *Bioinformatics Advance Access* published on October 2, 2006; doi:10.1093/bioinformatics/btl494
- Yu, J., Pressoir, G., Briggs, W.H., Bi, I.V., Yamasaki, M., Doebley, J.F., McMullen, M.D., Gaut, B.S., Nielsen, D.M., Holland, J.B., Kresovich, S. and Buckler, E.S. (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat. Genet.*, 38: 203-208.

**BUDGET REQUEST**

Budget Year 2008-2009

Funding Source

Salaries and Benefits

Postdocs/RA's

SRAs

Lab/Field Assistance

\$ 35,041

\$ 45,503

\$ \_\_\_\_\_

Subtotal

Sub 2

\$ 80,544

Employee benefits

Sub 6

\$ 21,557

SUBTOTAL

\$ 102,101

Supplies and Expenses

Sub 3

\$ 46,399

Equipment

Sub 4

\$ \_\_\_\_\_

Travel

Sub 5

\$ 1,500

TOTAL

\$ 150,000

Department account number \_\_\_\_\_

WMB Total \$150,000

UC Discovery \$150,000

\_\_\_\_\_ Date Nov. 28, 2007

Originator's Signature

COOPERATIVE EXTENSION

County Director \_\_\_\_\_

Date \_\_\_\_\_

Program Director \_\_\_\_\_

Date \_\_\_\_\_

AGRICULTURAL EXPERIMENT

Department Chair \_\_\_\_\_

Date Nov. 28, 2007

STATION

LIAISON OFFICER

\_\_\_\_\_ Date \_\_\_\_\_

D2454-2(1/84)

(Rev. 9/96)

