WALNUT GENOME ANALYSIS

Jan Dvorak, Ming- Cheng Luo, Mallikarjuna Aradhaya, Charles A. Leslie, Gale H. McGranahan, Abhaya M. Dandekar

ABSTRACT

The goal of this project is to build a set of comprehensive genomic tools for integration with currently available genetic and molecular walnut resources. These tools will facilitate more precise evaluation of breeding populations and accelerate development of improved walnut cultivars to address the needs of California growers and consumers of this important agricultural commodity. These tools include (1) construction of a physical map of the walnut genome, (2) a detailed survey of walnut gene expression, and (3) fine-scale genetic mapping of economically important traits. Accomplishing this goal will significantly strengthen ongoing California walnut breeding efforts by facilitating marker-assisted selection strategies, which significantly increase selection efficiency, discovery of new genes, and their rapid integration into genetic backgrounds adapted to California environmental conditions, thus accelerating development of improved walnut cultivars (Fig 1).

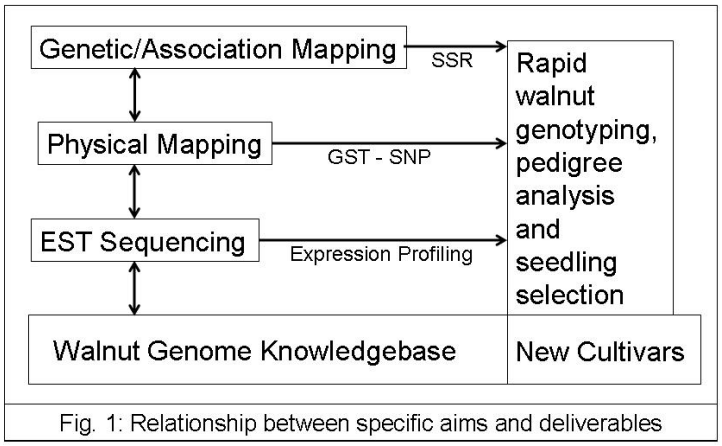

Fig. 1: Relationship between specific aims and deliverables

OBJECTIVES

1. Physical mapping of the walnut genome
2. Genetic and association mapping of economically important walnut traits
3. Functional mapping of the walnut genome
4. Development of a 'Walnut Genome Resource (WGR)', a web-based knowledge base of walnut genomic information

PROCEDURES

Objective 1: Physical mapping of the walnut genome

A physical map of the walnut genome will be built concurrent with development of the genetic map. During physical map construction, DNA fragments cloned in a bacterial artificial chromosome (BAC) vector are ordered on the basis of overlapping sequences. The sequence of DNA fragments generated by this process corresponds to the sequence of nucleotides along a chromosome. The presence of gene sequence tags (GSTs) within these sequences will be confirmed through their expression in walnut tissues (Objective 3, Fig 1). The physical map

provides a scaffold upon which to assemble the complete walnut genomic sequence when such sequencing is performed.

Construction of two walnut BAC libraries: Two bacterial artificial chromosome (BAC) libraries were constructed from *in vitro* grown shoots of Persian walnut (*Juglans regia* cv. Chandler) using cloning enzymes *Hin*dIII and *Mbo*I. A total of 129,024 clones, 64,512 per BAC library, were arrayed in 336 384-well plates. The average insert size was around 135 kb and 120 kb for the *Hin*dIII and *Mbo*I libraries, respectively (Fig 2). Assuming the walnut genome is approximately 800 Mb, these two BAC libraries represent ca. 20x genome equivalents. Each BAC library was stored in triplicate. BAC fingerprinting and BAC end sequencing will be initiated in the middle of January, 2008.

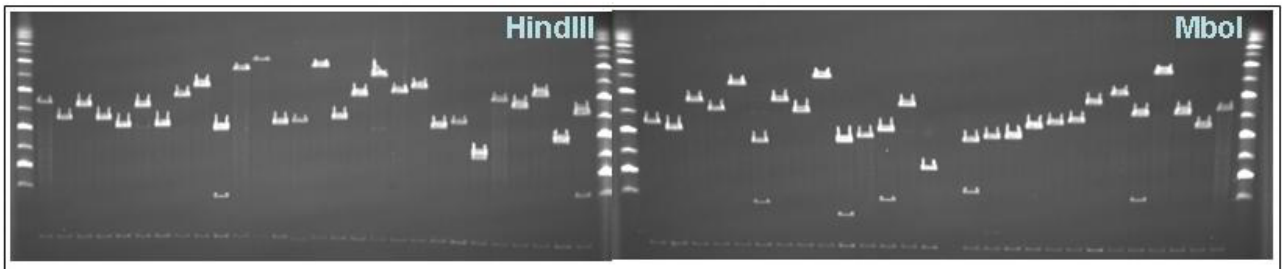

Fig 2: Size distribution of BAC inserts obtained by pulse field gel electrophoresis, *Hin*dIII BACs (left) and *Mbo*I BACs (right). The 25 kb ladder markers are shown on both sides of each gel.

Objective 2: Genetic and association mapping of economic traits in walnut.

Molecular fingerprinting of walnut mapping population: One hundred thirty-five F_1 individuals from a cross between 'Chandler' x 'Idaho' were fingerprinted using eighteen microsatellite loci to confirm the hybridity and parentage. The test resulted in the identification of six out-crossed individuals possessing alleles other than the ones present in the parents. The results are being analyzed to verify the pattern of inheritance of molecular markers. About half of this F_1 population is already in bearing stage and phenological data, bearing habit, and fruit and nut traits are being collected. The remaining half will come into bearing in the next two years.

DNA isolation of walnut (*J. regia*) germplasm collection at the Davis repository: There are about 800 trees of *J. regia* in the germplasm collection of the National Clonal Germplasm Respository. So far, DNA has been isolated from 100 trees and it will continue during the spring of 2008. We plan to complete (1) DNA isolation of entire population, (2) genotyping using 24 microsatellite loci, and (3) genetic structure analysis and identification of association mapping population during 2008 crop season. Shown below in Fig 3 is the scheme for development of marker assisted selection strategy for walnut.

Association analysis using elite walnut germplasm: Fifty five elite walnut cultivars actively used in the breeding program have been fingerprinted using 15 microsatellite loci and this data will be merged with 39 phenotypic traits collected from the same set of cultivars and analyzed for associations.

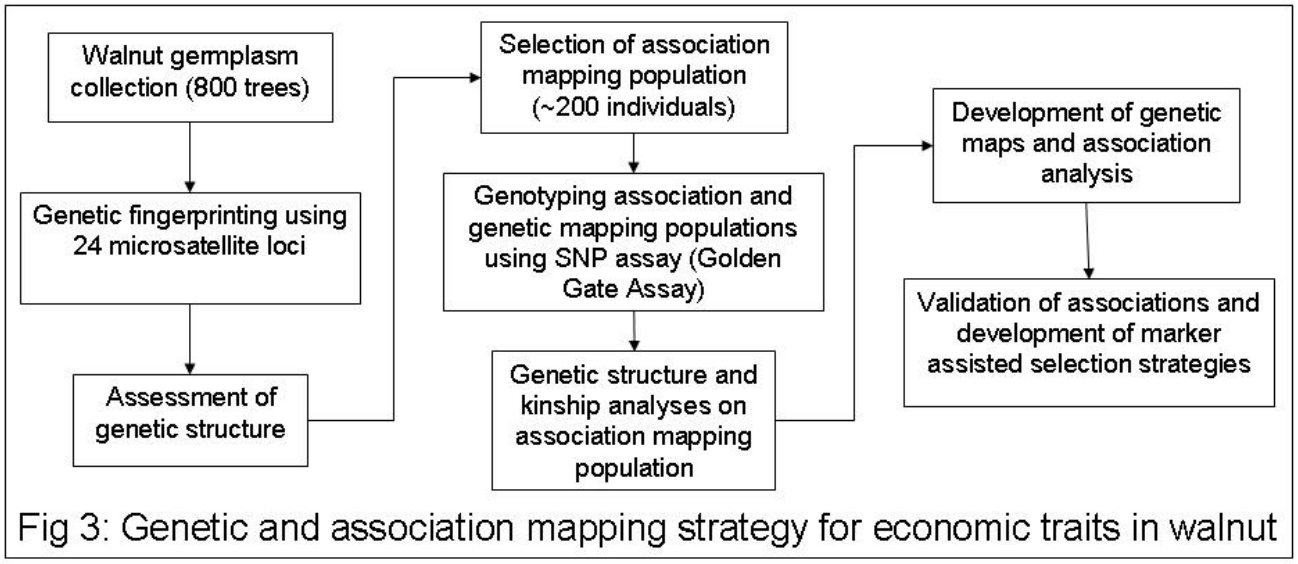

Fig 3: Genetic and association mapping strategy for economic traits in walnut

Objective 3: Functional mapping of the walnut genome

Genetic and physical maps describe the structure of the genome, but it is also essential to precisely document gene expression and to link specific traits (Objective 2) and GSTs (Objective 1) to underlying metabolic and biochemical processes (Fig 1). A key step toward this is gene transcript sequencing to identify expressed genes. Several tissue-specific gene transcript libraries will be constructed starting in Spring 2008 with the onset of walnut fruit development. These transcripts will be copied into DNA, cloned and sequenced to generate thousands of Expressed Sequence Tags (ESTs). The ESTs will be deposited in public databases where computer analysis can identify genes involved in important metabolic pathways. The ESTs will also be used to generate microarrays later in the program to validate GSTs through their expression pattern in different walnut tissues.

Objective 4: Development of a 'Walnut Genome Resource (WGR)', a web-based knowledge base of walnut genomic information.

A web-based browser will be developed in 2008 once data begins to accumulate to better inform the walnut research community and to provide access genomic resources. The database will contain all physical and linkage mapping information as well as all ESTs and their integration with the walnut physical and genetic maps (Fig 1).

RESULTS AND DISCUSSION

Objective 1: Physical mapping of the walnut genome

Physical mapping consists of cloning large genomic DNA fragments in a suitable cloning vector such as a BAC vector and ordering the fragments so that their sequence reflects the order of nucleotides in a chromosome. The construction of physical maps is entirely lab-based and highly automated. The basic strategy involves cutting genomic DNA into large fragments with restriction enzymes. These fragments are inserted into a BAC vector creating a BAC library a process that has been successfully completed. We have constructed two libraries one using restriction enzyme HindII and the other using the restriction endonuclease MboI (Fig 2). These

restriction endonucleases are DNA sequence specific and cut DNA at different sites to ensure coverage of the entire genome. BAC clones have been successfully picked to a 20X genome coverage and these will be fingerprinted. We use a previously developed fluorescence-based, high-throughput BAC DNA fingerprinting technique (Luo et al. 2003) that sizes DNA fragments from each BAC clone by capillary electrophoresis, creating a unique fragment profile or BAC clone fingerprint for each BAC clone. With a single robotic DNA sequence analyzer (96-capillary ABI3730XL), about 1,000 BAC clones can be fingerprinted daily. To fully exploit the high-throughput BAC fingerprinting technique, we also developed computer software for rapid editing of fingerprints (GenoProfiler; You et al. 2006). The computer program FPC (Soderlund et al. 2000) searches for overlaps between BAC fingerprints. Contiguous sequences of BAC clones (contigs), reflecting the sequences of nucleotides along chromosomes, are then assembled.

Objective 2: Genetic and association mapping of economically important traits in walnut

The existing California walnut breeding program develops cultivars through conventional breeding following quantitative genetic approaches and phenotypic selection (Tulecke and McGranahan, 1994). Assessment of progeny is resource-intensive, time consuming, and influenced by genotype x environment interactions, seriously limiting the number of trees that can be evaluated. Indirect selection using molecular markers tightly linked to economically important traits (marker-assisted selection; MAS) could significantly increase selection efficiency and integration of traits to address complex breeding objectives while reducing time and cost. Further, marker-assisted selection permits identification of genotypes with great economic potential as juveniles, allowing large segregating populations to be screened for potentially useful genotypes at minimum cost.

We propose two different approaches for walnut genome mapping: (1) Linkage analysis of a conventional mapping population derived from a cross between parents that differ for traits under consideration; and (2) Association genetic analysis of a natural population such as a germplasm collection with genotypes of unknown or mixed ancestry that represent a common gene pool.

A mapping population of F_1 trees from a diverse cross ('Chandler' x 'Idaho') is being developed. The parent cultivars were selected to produce progeny that segregate for economically important traits. A mapping population of 100 individuals from this cross is just beginning to bear and an additional 500 seedlings are currently being propagated. The bearing F_1 trees will be checked to confirm parentage and for segregation of traits of interest. About 400 true F_1 trees will be included in the final mapping population. These will be genotyped using ~300 microsatellite and ~1,500 SNP markers. A genetic map of both SSR and SNP markers will be assembled. For phenotypic evaluation, seedlings will be grafted onto 'Paradox' rootstocks and established in two locations following an augmented block design with standard check cultivars. Standard horticultural practices will be followed to establish mapping populations. Data collection will be initiated as traits appear and continued over three years.

The walnut germplasm collection maintained at the USDA repository, representing a diversity of economically important traits within *J. regia,* will be used for association analysis. Generally populations of outcrossing species such as walnut possess mild genetic structure approximating

panmixia. LD decays relatively rapidly over short physical distances within linkage groups; an ideal situation for association mapping. A preliminary analysis of genetic structure in walnut revealed a weak structure among 47 diverse genotypes representing a profile of the germplasm collection (Dangl et al., 2005). We will reevaluate the genetic structure of the entire germplasm collection (~800 individuals) using a set of 50 unlinked or distantly linked markers before a final association mapping population is selected. A model-based approach (Pritchard et al., 2000) and distance-based clustering strategies such as the neighbor-joining method and principal component analysis will be used to determine genetic affinities within the germplasm collection. The results will aid in selecting a suitable population for association mapping. Association of alleles from different marker loci will be examined using 2 x 2 contingency tables for χ^2 test. Probability (*P*) of allele independence is calculated using Fisher's exact test. General linear, step-wise, and mixed linear regression models will be used to examine the association of molecular markers with economically important traits. To account for structured associations, the marker-inferred ancestry (Q-matrix), computed from the estimated number of subdivisions in the final mapping population, will be incorporated into the mixed model (Yu et al.2006) as a covariate in the association analyses.

Objective 3: Functional mapping of the walnut genome

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* and deposited ~2500 sequences representing the first known protein-coding sequences for this key 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 efforts have produced over 95% of 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/). Starting in Spring 2008 we will start building cDNA libraries for EST sequencing corresponding the mRNA in different walnut tissues to provide a comprehensive view of the expressed portion of the walnut genome.

Objective 4: Development of a 'Walnut Genome Resource (WGR)', a web-based knowledge base of walnut genomic information

A genome resource, or knowledgebase, is a data portal for viewing genetic, physical, and functional mapping data. This resource will have two distinct components: one for visualizing the genetic map and one for the 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 database which we have developed for other crops (http://wheat.pw.usda.gov/cgi-bin/westsql/map_locus.cgi) will be expanded to accommodate the walnut physical mapping information. To display the function information, i.e., the gene expression data we will use 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 and functional classes 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

Aradhya, K.M., Potter, D. Gao, F. and Simon, C.J. (2005). Cladistic biogeography of Juglans (Juglandaceae) based on chloroplast DNA intergenic spacer sequences. P. 143-170. In: Motley, T.J., Zerega, N. and Cross, H. (eds). Darwin's harvest – New approaches to the origin, evolution, and conservation of crops. Columbia University Press, New York.

Aradhya, M.K., Potter, D., Gao, F. and Simon, C.J. (2007). Molecular phylogeny of Juglans (Juglandaceae): A biogeographic perspective. Tree Genetics and Genomes (in press).

Aradhya, M.K., Potter, D., Woeste, K.E. and Simon, C.J. (2001). Development of a high-density map of walnut (Juglans regia L.) using AFLP and SSR markers. Abstract, Plant and Animal Genome IX Conference, January 13-17, 2001, San Diego, CA. p. 195.

Dangl, G.S., Woeste, K., Aradhya, K.M., Koehmsted, A., Simon, C.J., Potter, D., Leslie, C.A. and McGranahan, G. (2005). Characterization of 14 microsatellite markers for genetic analysis and cultivar identification of walnut. J. Amer. Soc. Hort. Sci., 130: 348-354.

Fjellstrom, R.G., Parfitt, D.E. and McGranahan, G.H. (1994). Genetic relationships and characterization of persian walnut (Juglans regia L.) cultivars using restriction fragment length polymorphisms. J. Amer. Soc. Hort. Sci., 119:833-839.

Hardy, O.J. and Vekemans, X. (2002). SPAGeDi: a verstile 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 buy 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

Nei, M., Li, W., 1979. Mathematical model for studying genetic variation in terms of restriction endonucleases. Proc. Natl. Acad. Sci., USA. 76, 5269-5273.

Nelson W.M., Dvorak J., Luo M.C., Messing J., Wing R.A., Soderlund C. (2006) Efficacy of clone fingerprinting methodologies. Genomics 89:160-165

Pritchard, J.K. Stephens, M. and Donnelly, P. (2000a). Inference of population structure using multilocus genotype data. Genetics 155: 945-959.

Pritchard, J.K. Stephens, M., Rosenberg, N.A. and Donnelly, P. (2000b) Association mapping in structured populations. Am. J. Hum. Genet., 67: 170-181.

Royal Laboratories. 2006. Costs for fatty acid profile. http://www.royallab.net/services/cassays.htm (personal communication)

Soderlund C., Humphray S., Dunham A., French L. (2000) Contigs built with fingerprints, markers, and FPCV4.7. Genome Research 10:1772-1787

Soderlund C., Nelson W., Shoemaker A., Paterson A. (2006) SyMAP: A system for discovering and viewing syntenic regions of FPC maps. Genome Research 16:1159-1168

Swofford, D.L. and Selander, R.B. (1989). BIOSYS-1 A computer program for the analysis of allelic variation in population genetics and biochemical systematics. Illinois Natural History Survey, Urbana, Illinois.

Thornsberry, J.M., Goodman, M.M., Doebley, J., Kresovich, S., Nielsen, D. and Buckler, E.S. (2001). Dwarf8 polymorphisms associated with variation in flowering time. Nature Genet., 28: 286-289.

Tulecke, W. and McGranahan, G. (1994). The walnut germplasm collection of the University of California, Davis: A description of the collection and a history of the breeding program of Eugene F. Serr and Helord J. Forde. Rpt. 13. Univ. Calif. Resources Conservation Program, Davis.

Weir, B.S. (1996). Genetic data analysis II. Sunderland, MA: Sinauer Associates.

Woeste, K., Burns, R., Rhodes, O. and Michler, C. (2002). Thirty polymorphic nuclear microsatellite loci from black walnut. J. Hered., 93: 58-60.

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: 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.