WALNUT GENOME ANALYSIS

Jan Dvorak, Ming- Cheng Luo, Mallikarjuna Aradhaya, Charles A. Leslie, Gale H. McGranahan, and 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).

PROJECT 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. To accomplish the construction of a physical map, walnut genomic DNA fragments were cloned in a bacterial artificial chromosome (BAC) vector. Each BAC clone is fragmented with different restriction enzymes and ordered into contiguous sequences based on the overlap of fragment patterns. Each end of each of the BAC clones is then sequenced using Sanger DNA sequencing technology. Each of the BAC end sequences (BES) generated by this process is collinear with the BAC segments and thus corresponds to the sequence of nucleotides along a walnut chromosome. The presence of gene sequence tags (GSTs) within the BES will be confirmed through their expression in walnut tissues (Objective 3, Fig 1). The physical map also provides a scaffold upon which to assemble the complete walnut genomic sequence when such sequencing is performed.

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

Two different approaches were proposed 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. Extensive DNA libraries of both mapping and association mapping populations have been developed. We are currently developing genotypic data using microsatellite polymorphisms to validate the full-sib nature of the mapping population and assessment of genetic structure of association mapping population.

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 early 2009, from tissue samples collected over the 2008 growing season. 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, and compiled and annotated with computer analysis to 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 is being developed as data from Objectives 1, 2 and 3 begins to accumulate. 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) to better inform the walnut research community and to provide access genomic resources.

RESULTS AND DISCUSSION

Objective 1: Physical mapping of the walnut genome

Physical mapping consists of cloning large genomic DNA fragments in a suitable vector such as a BAC and ordering the fragments so that their sequence reflects the order of nucleotides in a chromosome. The first step in the construction of physical maps is the construction of BAC libraries. Two bacterial artificial chromosome (BAC) libraries were constructed from DNA isolated from *in vitro* grown shoots of Persian walnut (*Juglans regia* cv. Chandler). Walnut genomic DNA was fragmented with either *Hind*III or *Mbo*I restriction endonucleases. 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. Assuming the walnut genome is approximately 800 Mb, these two BAC libraries represent approximately 20x genome equivalents. Each BAC library has been stored in triplicate.

BAC fingerprinting and BAC end sequencing was initiated in 2008. We are using 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. The "SNaPshot" fragment pattern for each BAC clone, using a different restriction endonuclease, is analyzed using a 96-capillary ABI3730XL robotic DNA sequence analyzer. These BAC fingerprints are rapidly edited using a previously developed computer software (GenoProfiler; You et al. 2006). Another computer program, FPC (Soderlund et al. 2000), searches for overlaps between BAC fingerprints, creating contiguous sequences of BAC clones (contigs). These contigs reflect the sequences of nucleotides along individual chromosomes. As of November 30, 2008, 65,280 BAC clones have been fingerprinted with the 5-dye SNaPshot fingerprinting technique, and 63,000 BAC fingerprints were obtained (Fig. 2). The fingerprints of those clones have been edited. Approximately 92% of fingerprinted clones are suitable for contig assembly. A trial assembly was performed, which generated over 2000 contigs (Fig. 3) and about 24% of clones remained as singletons. A total of 3,840 BAC end sequences (BES) have also been produced.

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

1. Genotyping of mapping populations to confirm their full-sib origin.

As a first step, the mapping population from the cross ('Chandler' x 'Idaho') belonging to three age groups (3-year, 2-year and 1-year old trees) have been genotyped using 15 microsatellite loci to verify their full-sib status and to check whether or not they are true hybrids. Only seven individuals out of a total of 265 F1 individuals were found to be of half-sib origin and therefore were recommended for removal. The DNA library of this mapping population is currently being archived until SNPs are identified and genotyping platform developed. The phenotyping of these trees in the mapping population is being conducted by the walnut breeding group as the traits appear on these trees. Currently only one part of the mapping population is in bearing stage (about 140 trees) and as the remaining populations attain bearing stage they will be evaluated annually for at least two years. As suggested in the proposal, the mapping population is being evaluated for the following traits of breeding value: (1) Lateral vs. terminal bearing; (2) Leafing and harvesting dates; (3) Nut size; (4) Nut shell thickness; (5) Nut shell seal; (6) Plumpiness or fill (nut/kernel ratio); (6) Kernel color; (7) Nut yield. Oil content and fatty acid composition (especially omega-3 acids) will also be evaluated for representative year.

2. Association mapping using the walnut germplasm collection:

In the association mapping approach, we have attempted to analyze the genetic structure and differentiation within the English walnut germplasm collection. This analysis not only provides information on the amount and organization of genetic variation, which is fundamental to the association genetic analysis, but also permits identification of diverse genotypes for a re-sequencing panel to identify SNPs based on sequence comparisons.

The cultivated walnut (*J. regia*) germplasm collection of 399 trees from 204 diverse accessions representing the Eurasian distribution maintained at the USDA germplasm repository and 62 elite germplasm frequently used in walnut breeding have been analyzed for genetic structure and differentiation using 14 polymorphic microsatellite loci. A cluster analysis (CA) using the neighbor-joining method revealed a mild genetic structure with Chinese accessions forming a major group with a number of subgroups (Figs 4 & 5). The genotypes originating from South Asia formed a cluster and there were two other clusters containing mixtures of accessions of the South and East Asian origin along with some genotypes from the Carpathian region. The elite germplasm including some of the selections and cultivars from the California breeding program and some introduced from northeastern Europe formed one cluster, which showed considerable differentiation as compared to other groups. Observed heterozygosity was consistently lower than the expected for all loci, which ranged from 0.33 for the locus WGA106 to 0.64 for WGA178 with an average of 0.52 while the expected levels ranged from 0.41 for WGA106 to 0.85 for WGA276 with an average of 0.69 (Table 1). The deficiency of heterozygotes across loci is probably due to sub-structuring observed in the CA or the species has not fully recovered from the historical bottlenecks during the Pleistocene glaciations in cryptic Caucuses and Carpathian refugia and subsequent expansion and human selection during domestication. The inbreeding coefficient for the total population (F_{IT} = 0.2466) is apportioned into inbreeding within groups (F_{IS} = 0.1522) and due to differentiation among groups (F_{ST} = 0.1108), suggesting significant genetic differentiation among groups (Table 2). The analysis of molecular variance indicated ~87% of the variation was found within populations with only 13% accounting for differentiation among groups (Table 3) suggesting significant differentiation within cultivated walnut. On the whole, the genetic structure within the cultivated walnut reflects historical bottlenecks during the glaciations, founder effect on the post-glaciations population expansion, introgression and human selection during domestication.

We are currently working on enlarging genotypic data by adding more microsatellite loci to reassess genetic structure using both distance-based and model-based Bayesian approaches. Both of these analyses will be repeated on a much larger SNP genotype data, when once the SNP platform is ready to be used for genotyping.

Objective 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 represents 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. Each complimentary DNA (cDNA) library represents 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. 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/).

New ultra-high-throughput (UHT) DNA sequencing is an approach that can profile an entire transcriptome in a single run using technology such as the Roche 454 Life Sciences GS FLX Titanium system (http://www.454.com/applications/transcriptome-sequencing.asp) and massively parallel signature sequencing (MPSS) with Solexa technology on an Illumina Genome Analyzer (http://genomecenter.ucdavis.edu/dna_technologies/uhtsequencing.html). These new technologies are able to profile an entire transcriptome in a single run by generating hundreds of thousands of sequences corresponding to a greater diversity of the mRNA population at a lower cost than traditional Sanger sequencing. The resulting sequences are then compiled to generate a Unigene set, representing a high percentage of the walnut genespace.

For this project, 15 samples of walnut tissue were gathered from Chandler trees in the Stuke block at UC Davis between April and October 2008 (Table 4). Three additional samples were taken from Chandler plant material maintained in the lab of Gale McGranahan. RNA isolation and cDNA library construction will be initiated in early 2009 (Table 4). Sequencing will primarily be conducted using a Roche 454 FLX Titanium Genome Sequencer. The cDNA generated from each sample will be sequenced separately so that the profile of genes expressed in each tissue can be determined. In total, approximately 2 million sequences of 400-500 bp in length will be generated. Bioinformatics for sequencing processing will be conducted at the UC Davis Genome Center Bioinformatics Core, where these sequences will be compiled using Newbler and GenomeQuest assembler software (http://www.genomequest.com/) to produce a profile of the transcriptome across the growing season.

These new sequences will be combined with the 18,000 existing walnut ESTs to generate a Unigene set, which is expected to describe most of the expressed walnut genome. This set of genes will be annotated using Blast2GO (http://blast2go.bioinfo.cipf.es/). This software package was recently used to annotate the set of 8622 walnut genes compiled in our walnut-nematode project. There, Gene Ontology (GO) categories were assigned to 57% of the sequences, including 18% which were also matched to Enzyme Commission (EC) identifiers. The GO categories and EC identifiers can be used to map the associated proteins onto metabolic pathways. These data can then be used to determine which metabolic pathways are active in each tissue and at various time points in the growing season.

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 database that provides access to genetic, physical, and functional mapping data generated in this project. The resource, which is now web accessible (http://walnutgenome.ucdavis.edu/) will have two distinct components: one for visualizing the genetic map and one for visualizing the physical map. The database provides access to all of the fingerprinting data that will be generated this year along with the BAC end sequences as they too are generated. Tools are available to integrate and represent this information as a physical map showing individual contigs. The physical map is a scaffold on which we will integrate genetic mapping data of walnut phenotypic information, molecular markers, and expressed genes as that information becomes available.

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 walnut using the Pathway Tools Omics Viewer (http://www.plantcyc.org; Zhang et al. 2005), or MapMan (http://gabi.rzpd.de/projects/MapMan; Thimm et al. 2004). Both of these visualization platforms organize plant genes, primarily from model species such as *Arabidopsis*, into ontologies, or hierarchical classifications corresponding to the more than 2000 known structural, regulatory and enzymatic functions. Both the Pathway Tools and MapMan visualization platforms can display high throughput expression data of individual genes or whole pathways in combination with standard pathway maps and graphical classifications of gene functions. The Omics Viewer is useful for integrating data from other types of “omics” experiments into the visualizations, and for generating animated graphics from timecourse experiments. We have used MapMan to visualize expression data from experiments on tomato, apple, citrus, and grape. MapMan is more versatile for incorporating custom graphics, including our own pathway figures more relevant to our particular study.

REFERENCES

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.

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

Thimm, O., Blasing, O., Gibon, Y., Nagel, A., Meyer, S., Kruger, P., Selbig, J., Muller, L.A., Rhee, S.Y. Stitt, M. (2004). MAPMAN: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. The Plant Journal 37, 914-939.

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.

Zhang, P., Foerster, H., Tissier, C., Mueller, L., Paley, S., Karp, P., Rhee, S.Y. (2005). MetaCyc and AraCyc. Metabolic Pathway Databases for Plant Research. *Plant Physiology* 138: 27-37.

Table 1. Within cluster genetic variability – locus-wise heterozygosity

Locus	N	Obs_Het	Exp_Het*	Nei**	Ave_Het
WGA001	882	0.6032	0.8038	0.8029	0.7145
WGA004	896	0.5826	0.6779	0.6771	0.6154
WGA069	904	0.5597	0.8124	0.8115	0.7341
WGA089	912	0.5000	0.6674	0.6667	0.5664
WGA106	908	0.3304	0.4128	0.4123	0.3822
WGA118	846	0.6052	0.7893	0.7883	0.6602
WGA178	864	0.6435	0.7472	0.7463	0.7232
WGA202	840	0.6143	0.8307	0.8297	0.7629
WGA237	900	0.3489	0.5999	0.5993	0.4652
WGA276	750	0.4213	0.8517	0.8505	0.8104
WGA321	860	0.5814	0.7186	0.7178	0.6493
WGA331	826	0.5666	0.6583	0.6575	0.5955
WGA338	912	0.5110	0.5555	0.5548	0.5336
WGA384	822	0.4647	0.5808	0.5801	0.4420
Mean	866	0.5238	0.6933	0.6925	0.6182
St. Dev		0.0993	0.1254	0.1252	0.1286

* Expected heterozygosity were computed using Levene (1949)
** Nei's (1973) expected heterozygosity

Table 2 Genetic differentiation within the cultivated walnut

Locus	N	F_{IS}	F_{IT}	F_{ST}
WGA001	882	0.1559	0.2534	0.1155
WGA004	896	0.0664	0.1461	0.0854
WGA069	904	0.2383	0.3112	0.0957
WGA089	912	0.0940	0.2365	0.1573
WGA106	908	0.1107	0.2048	0.1059
WGA118	846	0.0969	0.2439	0.1628
WGA178	864	0.1061	0.1336	0.0308
WGA202	840	0.1967	0.2673	0.0880
WGA237	900	0.2850	0.4435	0.2217
WGA276	750	0.4483	0.4791	0.0558
WGA321	860	0.0939	0.1899	0.1059
WGA331	826	0.0537	0.1396	0.0907
WGA338	912	0.0542	0.0879	0.0356
WGA384	822	-0.0158	0.2242	0.2362
Mean	866	0.1528	0.2466	0.1108

Table 3 Analysis of molecular variance

Source of Variation	Sum of squares	Variance components	Percentage variation
Among Populations	445.339	0.62242**	12.49557
Within Populations	3734.725	4.35873	87.50443
Total	4180.064	4.98115	

F_{ST} = 0.12496**

Table 4: Walnut tissues gathered April to Nov 2008 for cDNA library construction and EST analysis

Sample No	Tissue Source	Developmental Stage	Source	Harvest Date
1	Vegetative Bud	Vegetative	Tree	4/1/08
2	Leaf - early	Vegetative	Tree	4/15/08
3	Root	Vegetative	Pot	8/27/08
4	Callus	Vegetative	In Vitro	10/15/08
5	Pistillate Flower	Vegetative	Tree	4/17/08
6	Catkins	Immature	Tree	4/1/08
7	Somatic Embryo	Immature	In Vitro	10/14/008
8	Leaf- mature	Vegetative	Tree	6/5/08
9	Fruit	Immature	Tree	6/5/08
10	Hull	Immature	Tree	6/5/08
11	Packing Tissue	Immature	Tree	6/5/08
12	Hull Peel	Mature	Tree	9/4/08
13	Hull Cortex	Mature	Tree	9/4/08
14	Packing Tissue	Mature	Tree	9/4/08
15	Pellicle	Mature	Tree	9/4/08
16	Embryo	Mature	Tree	9/4/08
17	Leaf – late	Senescent	Tree	10/15/08
18	Hull – dehiscing	Senescent	Tree	10/15/08

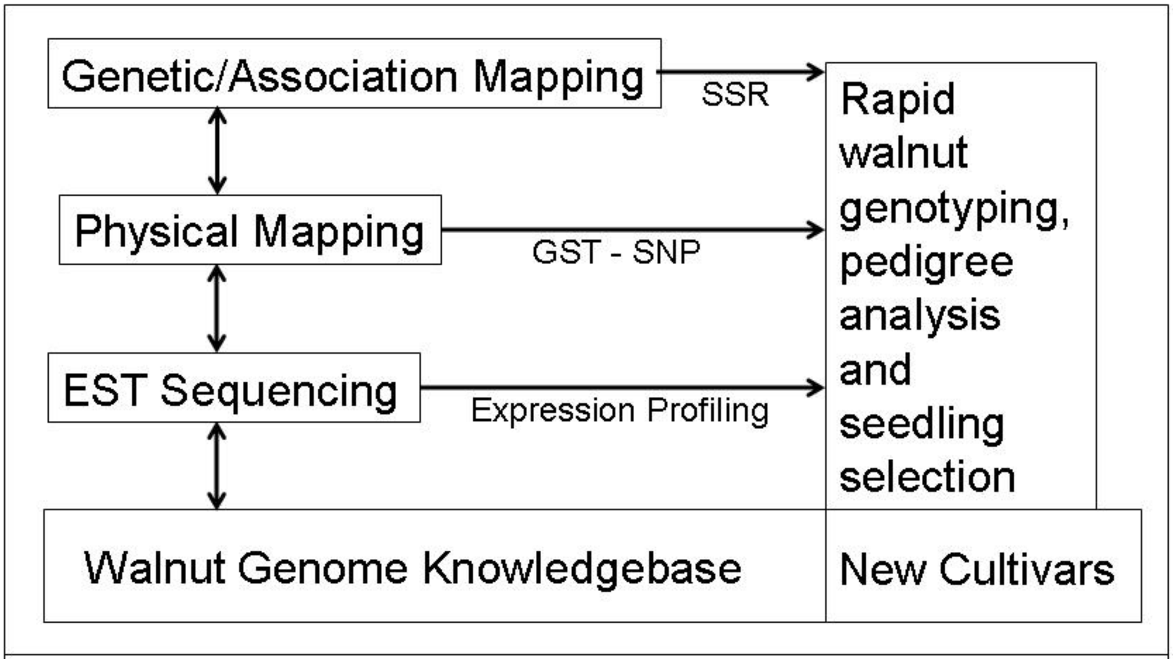
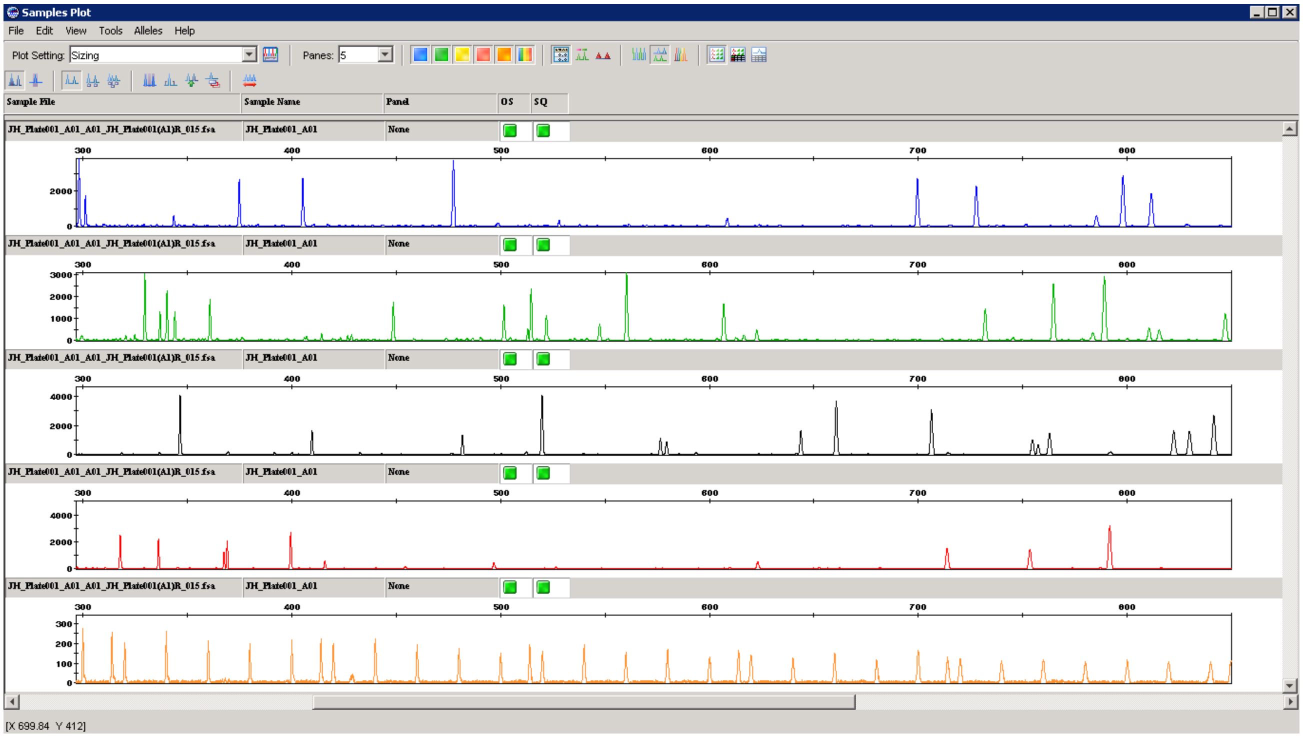

Genetic/Association Mapping
SSR
Physical Mapping
GST - SNP
EST Sequencing
Expression Profiling
Rapid walnut genotyping, pedigree analysis and seedling selection
Walnut Genome Knowledgebase
New Cultivars

Fig. 1: Relationship between specific aims and deliverables

Fig. 2: Example of a single finger print for a walnut BAC clone. To date 65,280 BACs have been fingerprinted with 92% being of high quality suitable for contig assembly.



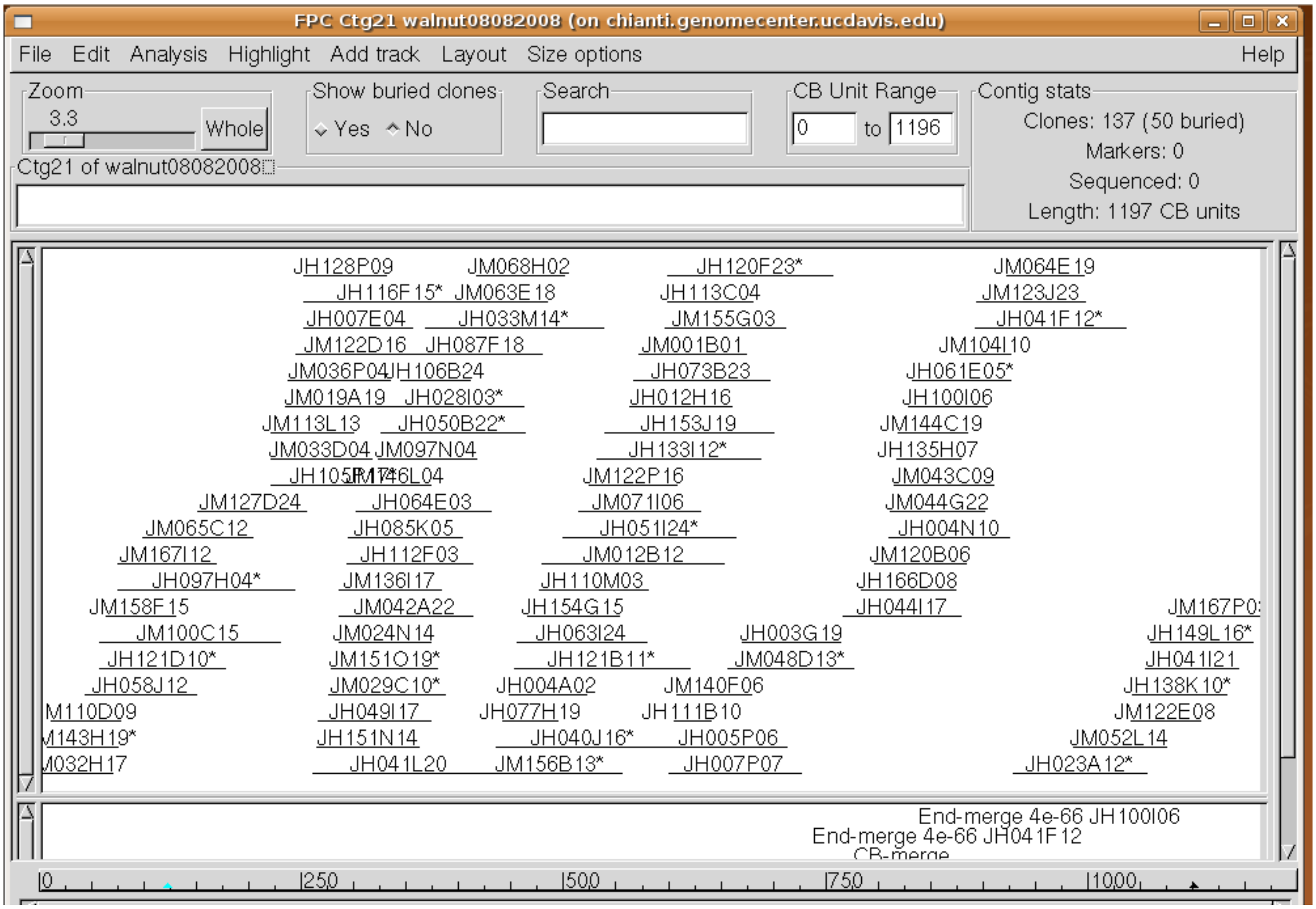
Plot Setting: Sizing

Panes: 5

Sample File	Sample Name	Panel	OS	SQ
JH_Plate001_A01_A01_JH_Plate001(A1)R_015.fsa	JH_Plate001_A01	None	🟩	🟩
JH_Plate001_A01_A01_JH_Plate001(A1)R_015.fsa	JH_Plate001_A01	None	🟩	🟩
JH_Plate001_A01_A01_JH_Plate001(A1)R_015.fsa	JH_Plate001_A01	None	🟩	🟩
JH_Plate001_A01_A01_JH_Plate001(A1)R_015.fsa	JH_Plate001_A01	None	🟩	🟩
JH_Plate001_A01_A01_JH_Plate001(A1)R_015.fsa	JH_Plate001_A01	None	🟩	🟩

[X 699.84 Y 412]

Fig 3: Example of a single contig assembly. To date 2000 contigs have been assembled, 24% of BACs are singletons. 3840 BES have been completed



FPC Ctg21 walnut08082008 (on chianti.genomecenter.ucdavis.edu)

Zoom 3.3

Show buried clones ◇ Yes ◆ No

Search

CB Unit Range 0 to 1196

Ctg21 of walnut08082008

Contig stats
Clones: 137 (50 buried)
Markers: 0
Sequenced: 0
Length: 1197 CB units

End-merge 4e-66 JH100I06
End-merge 4e-66 JH041F12
CB-merge

Fig 4: A midpoint rooted tree depicting an analysis of the genetic structure and differentiation within the walnut germplasm collection and within cultivated walnut.

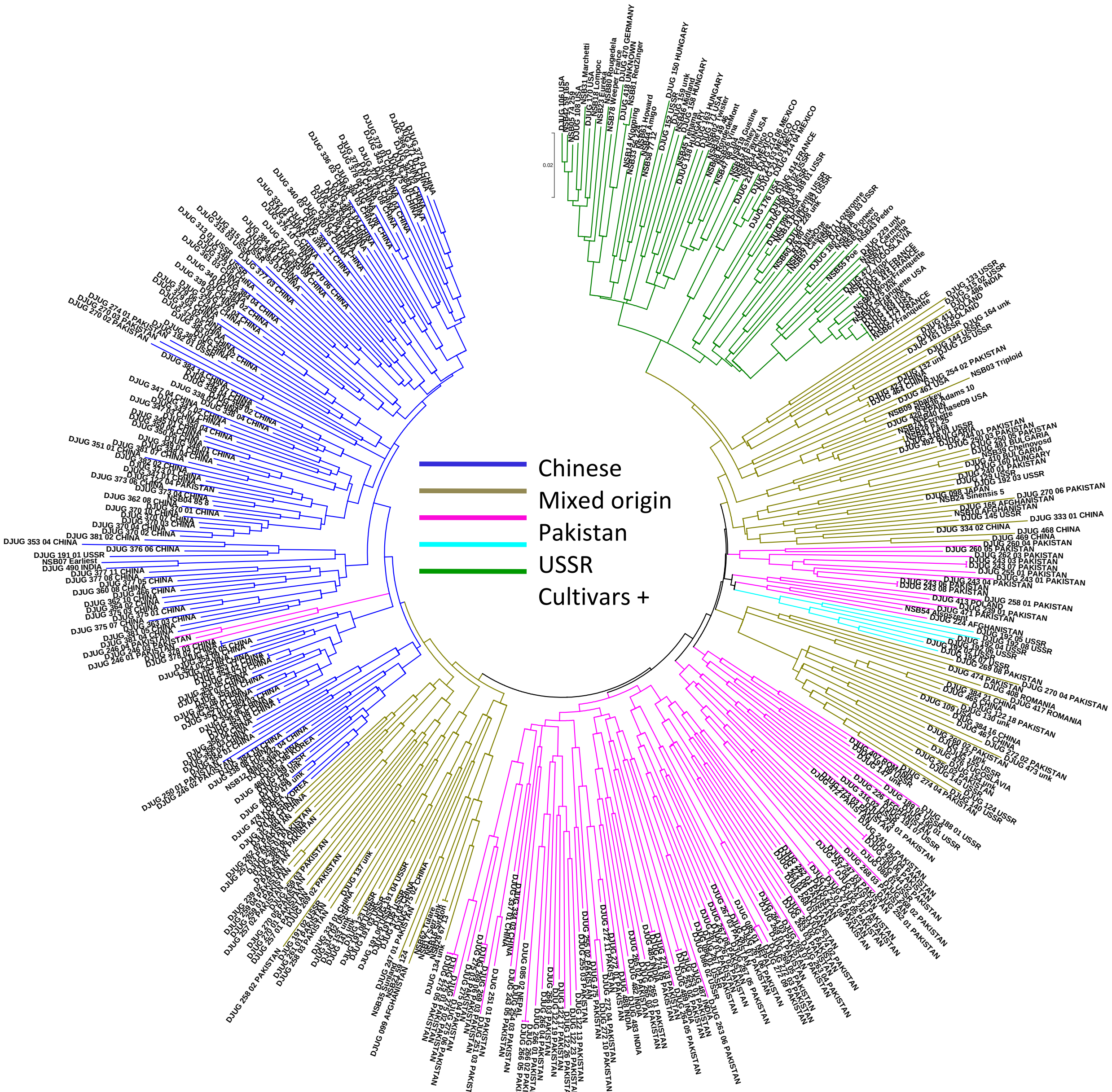
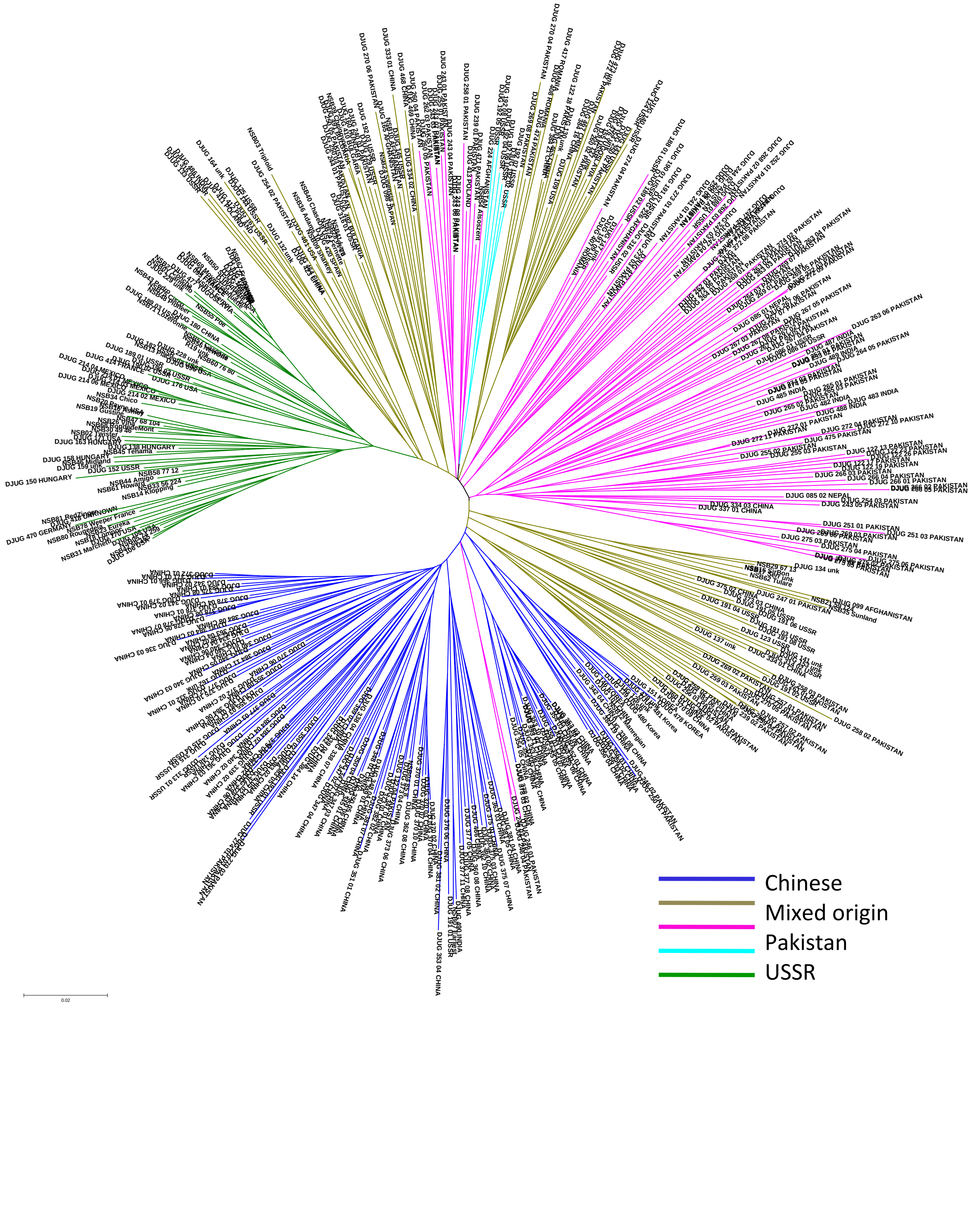


Fig 5: An unrooted tree depicting an analysis of the genetic structure and differentiation within the English walnut germplasm collection and within cultivated walnut.


0.02
Chinese
Mixed origin
Pakistan
USSR