WALNUT GENOME ANALYSIS

Jan Dvorak, Ming-Cheng Luo, Mallikarjuna Aradhya, Dianne Velasco, Charles A. Leslie, Sandie L. Uratsu, Monica T. Britton, Russell L. Reagan, Jiajie Wu, Yong Q. Gu, Yuqin Hu, Frank M. You, Jirui Wang, Gale H. McGranahan, and Abhaya M. Dandekar

ABSTRACT

The goal of this project is to build a set of comprehensive genomic tools to facilitate a more precise evaluation of breeding populations as well as to access unique germplasm for the walnut improvement program. A remaining deliverable was a genetic map that can be used to map phenotypic traits of significance for walnut quality and production. This map can be used to align any phenotypic trait to linkage groups that represent discreet regions on walnut chromosomes. Single nucleotide polymorphisms (SNPs) were used to create the map. These SNPs were identified by comparing three different sources of DNA sequence data, BAC end sequences, SOLiD shotgun genome sequence and RNAseq data obtained from 20 different cDNA libraries representing various walnut tissues. This comparison revealed ~6000 SNPs that were then processed by Illumina to create an Infinium array that was then used to create the linkage map by analyzing 352 progeny from a Chandler x Idaho cross. The data was processed and generated a very dense map with 35 linkage groups by assigning 1600 SNPs. The genetic map will accelerate development of improved walnut cultivars to address the needs of both California growers and the consumers of this important agricultural commodity. A total of 20 cDNA libraries were sequenced using next generation DNA sequencing to define the walnut transcriptome and over a billion reads were processed to obtain 85,045 consensus sequences or genes that define the gene space. These genomic tools will significantly strengthen ongoing California walnut breeding efforts by facilitating marker-assisted selection strategies. The use of well-defined markers will significantly increase selection efficiency, the discovery of new genes, and rapid integration of these genes into genetic backgrounds adapted to California environmental conditions, thus accelerating the development of improved walnut cultivars.

OBJECTIVES

Development of a linkage map to facilitate the mapping of economically important walnut traits

PROCEDURES

Discovery, Validation and Processing of Walnut SNPs

DNA sequence data is required for the discovery of SNPs so we used all of the available DNA sequence information available to us and this included the following:

1. Walnut BAC End sequences (BES): 48, 661. Total bases: 35,084,098 bp. Average read length: 721 bp.
2. Walnut (cv Chandlar) shotgun SOLiD sequence data: Total qualified reads (low quality reads were removed): 395,528,231. Total bases: 19.776 Gbp. Read length: 50 bp.

3. Walnut cNDA contig sequences: 85,045 contigs. Total bases: 136,806,538 bp. Average contig length: 1,608 bp.
4. Walnut FPC contig map: 916 contigs. Total clones: 113,063. Clones in contigs: 108,233.

SOLiD reads were mapped to all walnut BES using the AGSNP pipeline. The average ($\overline{X}$) mapping depth of SOLiD reads and standard deviation (*s*) were estimated based on extreme valuation distribution. All SNPs with extremely high mapping depth were considered to be present in repeats or in multi-gene families, and were thus removed. The remaining reads were used for SNP identification. Putative SNPs were aligned against BES, cDNAs and physical contigs. SNP selection for genotyping was based on these criteria: one best SNP for each BES; at least one SNP marker in each of 649 FPC contigs which have putative SNPs; SNP markers more or less evenly distributed along a FPC contig; if multiple BES hit the same gene, only one BES was chosen; only Infinium II type SNPs.

Single Nucleotide Polymorphism (SNP) Genotyping

The *Juglans regia* genetic mapping population genotyped on the Illumina Infinium SNP array consisted of three hundred and fifty-two F_1 progeny resulting from the cross of cultivars „Chandler" x „Idaho". These progeny were previously confirmed by the simple sequence repeat (SSR) genotyping as true F1s hybrids. The two parents „Chandler" and „Idaho" and 30 other cultivars from the UC Davis Walnut Breeding Program (WBP) germplasm collection with either diverse origins or varying relationship to Chandler (Table 1, Figure 1) were also included in the first round of SNP genotyping.

The remaining approximately 690 samples to be genotyped using the Infinium array include seventy-four remaining F_1s, 41 remaining WBP germplasm, five elite named selections, 137 WBP selections to augment association mapping, approximately 430 USDA germplasm accessions which includes the association mapping population. A set of ~25 species of wild *Juglans* from the Davis NCGR collection and ~40 *Carya* spp. from the National Clonal Germplasm Repository for Pecans and Hickories, Somerville, TX 77879 will also be included to check for cross compatibility of walnut SNPs developed in this project across wild *Juglans* and *Carya*.

DNAs were extracted with a CTAB protocol (Doyle and Doyle, 1987), modified with the addition of PVP-40 and PVPP, then treated with RNase A. Samples were evaluated for quality and concentration prior to genotyping using a NanoDrop 1000 spectrophotometer (Thermo Fisher Scientific, Wilmington, Delaware, USA) and VICTOR^{2}D fluorometer (PerkinElmer, Waltham, Massachusetts) readings of double strand DNA specific Quant-iT PicoGreen (Invitrogen, Molecular Probes, Eugene, Oregon, USA) assay. The NanoDrop provided A260/A280 DNA quality values to assess whether they contained contaminants that would interfere with the genotyping process as well as initial DNA concentration readings. The DNA samples were then diluted to give a solution that contained ~10 ng/uL based on the NanoDrop concentrations, and then assessed for concentration by PicoGreen assay readings from VICTOR^2. Based on the PicoGreen and VICTOR^2 concentration readings, 750 ng of each sample were aliquoted to a 96-well PCR plate and then dehydrated using a vacuum centrifuge, and reconstituted with 15 uL of 1X TE to provide a final concentration of 50 ng/uL. Samples were submitted for genotyping to the DNA Technologies Core at the UC Davis Genome Center.

Genetic Mapping of Walnut SNPs.
The 6000 SNPs along with their flanking sequences were submitted to the Illumina for development of an Infinium assay design and to direct the bead manufacture. Typing was performed in the UC Davis Genome Center. The raw typing data were analyzed and visualized with the GenomeStudio software package. The SNPs for which Chandler showed polymorphism but monomorphic in Idaho behave as "test cross" in the Chandler x Idaho population and therefore were used for linkage map construction. The linkage groups were generated using the MultiPoint software package.

RESULTS AND DISCUSSION

Discovery, Validation and Processing of Walnut SNPs

SOLiD reads were mapped to all walnut BES using the AGSNP pipeline. Average ($\overline{X}$) of mapping depth of SOLiD reads and standard deviation (s) were estimated based on extreme valuation distribution (Figure 1), being 15.9 reads and 19.1 reads, separately. That is, the estimated average sequencing coverage of SOLiD reads is ~16X. A cutoff of + 2s for SNP discovery was calculated to be 54 reads. All SNPs with a mapping depth > 54 were considered to be repeats or present in a gene family, and were thus removed.

A BLAST search of 48,661 BES against walnut cDNA contig sequences was performed at 1e-10. A total of 29,223 BES were found to have hits to cNDA sequences. Thus, only those gene-related BES were retained for further SNP discovery.

A total of 29, 180 BES out of 29,223 gene-related BES were found to contain at least one putative SNP and were used for further SNP filtering.

4. A total of 14, 829 putative SNPs were identified, which were located in 9,672 BES.

5. There were a total 916 FPC contigs present in the walnut physical map. Among them, 649 of the contigs including 7,948 clones with putative Infinium II SNPs.

Single Nucleotide Polymorphism (SNP) Genotyping

A few of the SNPs, such as SNP loci JH009I21r_570 and JH016K09r_31, were homozygous in both Chandler and Idaho, but were found to be either heterozygous or alternately homozygous in the WBP. This may be due to the nature of a few samples collected for expression data. However, maternal tissue samples collected from a few of the F1s may have provided somewhat more diverse genetic material. This is because a few of the F1s were not true to type, i.e. not the progeny of Chandler x Idaho, but had not been removed from the field thus providing a few SNPs that are homozygous in both Chandler, Idaho, and their progeny. While a few others, such as SNP locus JH003A24r_235, were likely divergent enough in the flanking sequence to not hybridize properly and prevent genotyping. Among fourteen individuals from the WBP germplasm samples unable to be genotyped at JH003A24r_235, one was Conway Mayette, a grandparent of Chandler whereas Chandler is heterozygous at this locus. Otherwise, over 50% of the 5421 SNP loci were scoreable. Forthcoming genotype data for the association mapping population will be combined with phenotypic data an analyzed to elucidate trait-genotype associations for further investigation.

Genetic Mapping of Walnut SNPs.
Out of the 6000 potential SNPs selected for Infinium assay, 5421 passed the Illumina‟s manufacturing pipeline. 1639 SNPs showed haploid type polymorphism in Chandler with Idaho being monomorphic; these SNPs were used for linkage map construction (Fig 3) The linkage groups were generated using the MultiPoint software package. There are currently 35 linkage groups, significantly more than the expected 16 linkage groups, which might be caused by the homozygous segments existing in the Chandler genome (Fig 4).

QTL mapping
The SNP loci selected for developing high density marker maps will be combined with the economic traits and analyzed for QTL mapping to identify chromosomal regions that affects the economic traits. Markers linked to these regions controlling the traits (co-segregate with QTL) detected by the analysis of variance will be used in developing marker assisted selection schemes. Further presence of QTL and QTL-marker linkages will have to be validated in diverse genetic backgrounds before implementing marker-assisted selection.

REFERENCES

Doyle JJ, Doyle JL (1987). A rapid DNA isolation procedure form small quantities of fresh leaf tissue. *Phytochem Bull* 19: 11-15.

Tulecke, W., and G. McGranahan. 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 Harold I. Forde. Report No. 13. University of California Genetic Resources Conservation Program, Davis, CA.

Table 1. Individuals in first 384 samples genotyped using high throughput Infinium SNP assay.

Sample	ID	Origin	Parentage	Relation to Chandler
Chandler (CR)	64-172	WBP	Pedro x 56-224	Self
Idaho (ID)	DJUG 171	Idaho, USA		
CR x ID seedlings	various	WBP	Chandler x Idaho	F1 progeny (352 individuals)
Howard	64-182	WBP	Pedro x 56-224	Full sibling
Pedro	53-113	WBP	Conway Mayette x Payne	Parent
Conway Mayette	031	Calif., USA	sel. of Mayette (1915)	Grandparent (parent of Pedro)
Sharkey	053	China[1]	discovery (1925)	Grandparent (parent of 56-224)
Payne	001	Calif., USA	French[1] x Chinese[1]	Grandparent (parent of Pedro) and Great grandparent[1] (parent of Marchetti)
Eureka	007	Iran	discovery (~1903)	Great grandparent[1] (parent[1] of Marchetti)
Abbotbad #1	85-040	Pakistan		
Alsoszentivani-117	85-042	Hungary		
Badajoz	87-002	Spain		
Cascade	87-018	Wash., USA	Russian, Carpathian, Manchurian parentage	
Concha	005	Chile	selection (1979)	
Franquette	96-002	France		
Lara	86-016	France		
Lozeronne	92-002	France		
Manregian	049	China		
Meylan	042	France		
O-20-1072	025	Iran		
PI 159568	048	Afghanistan		
Poe	91-136	Calif., USA		
Red Zinger	86-011	Germany		
Ronde de Montignac	96-003	France		
Rouge de la Donan	93-004	France		
Serr	59-129	California	Payne x PI 159568	
Sheinovo sd	85-043-1	Bulgaria		
Sinensis #5	054	Japan		
Sir Bon	071			
Waterloo	056	Calif., USA	sel. of Eureka (1934)	
Weeper	84-006	USA		
Xinjiang 6	85-008	China		
XXX Mayette	058	Calif., USA	sel. of Mayette	

1- speculated; WBP = Walnut Breeding Program

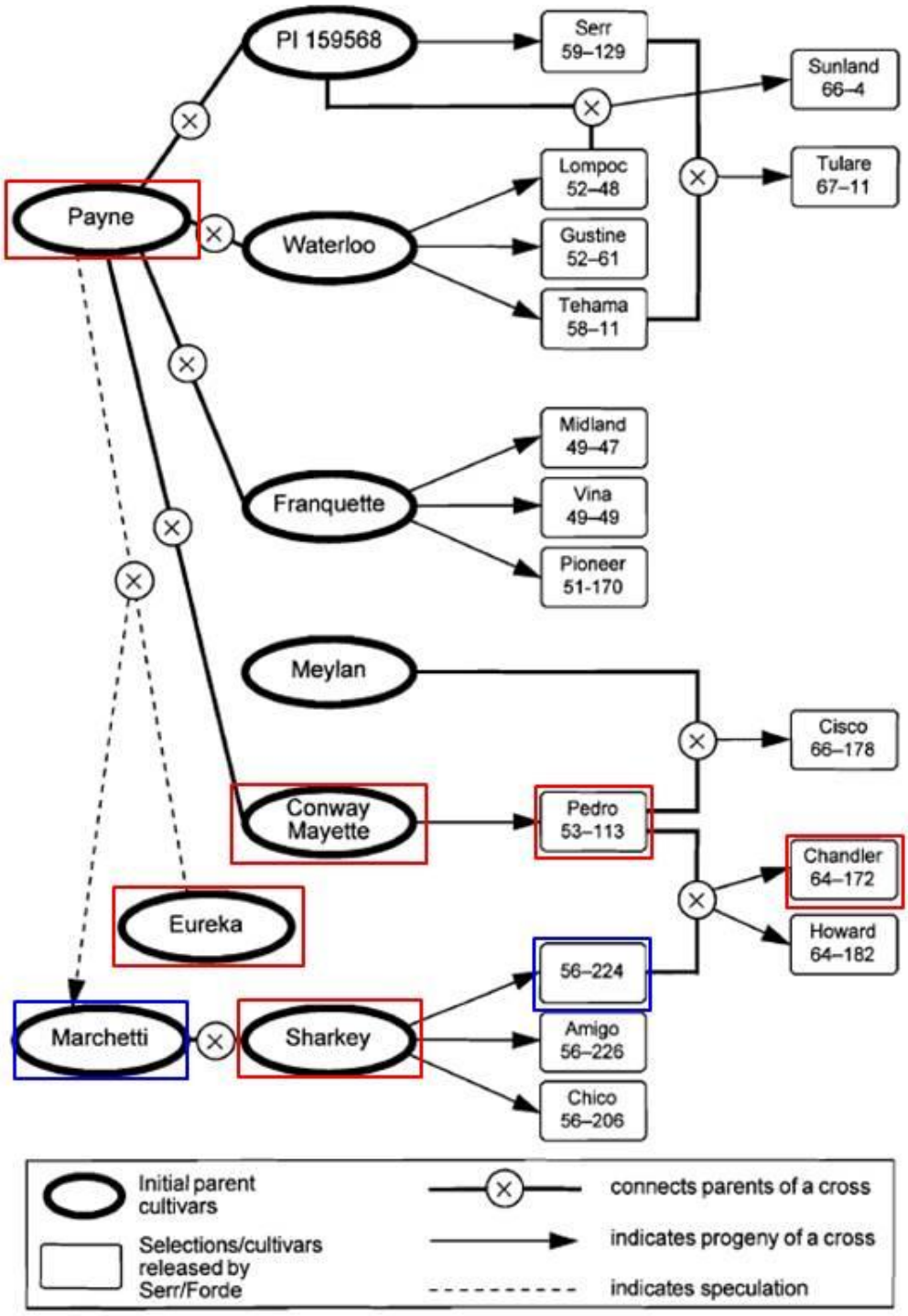

Figure 1. Pedigree chart of major cultivars in the Serr and Forde walnut breeding program (Tulecke and McGranahan, 1994). Red boxes indicate cultivars in Chandler pedigree that have been SNP genotyped and blue boxes indicate those still needing to be SNP genotyped.

Fig 2: Principal component analysis biplot of RNA-Seq data for 20 libraries derived from different tissues and developmental stages of walnut, listed in upper-right (PK = mature packing tissue, PT = immature packing tissue). Two principal components are derived from similarities and differences between abundances of specific transcripts found in each library. Some, but not all, tissue types cluster together in groups.

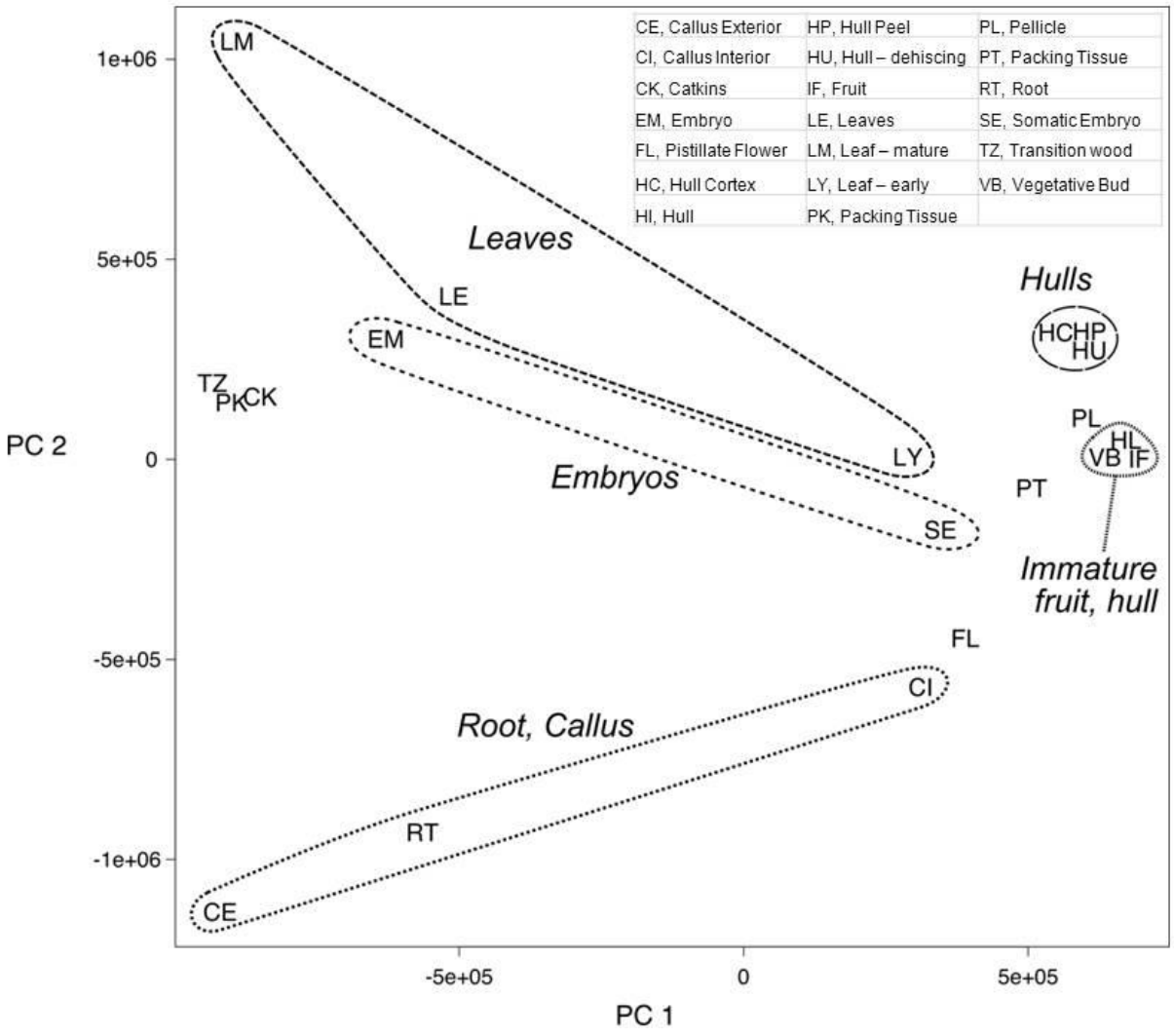

Fig. 3: Cluster of genotypes for SNP polymorphism. The left group shows the homozygous allele A; the right group shows the homozygous allele B, and the middle group shows heterozygous state (A+B).

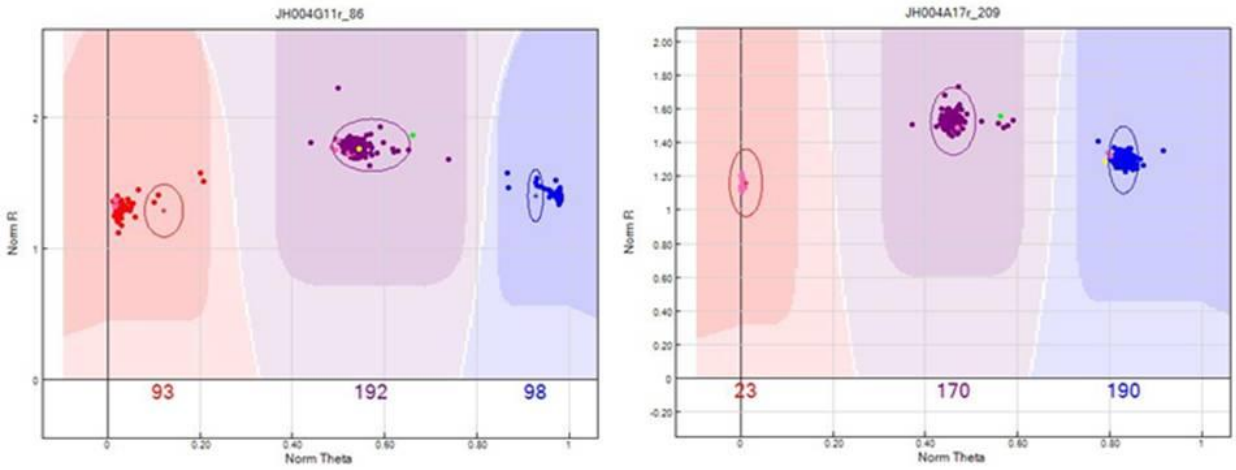

Fig 4: The linkage maps based on SNP Infinium assay.