

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, Gale H. McGranahan, and Abhaya M. Dandekar

ABSTRACT

The goal of this project is to build a set of comprehensive genomic tools for walnut. These will facilitate a more precise evaluation of breeding populations and will accelerate development of improved walnut cultivars to address the needs of both California growers and the consumers of this important agricultural commodity. Development of these tools includes (1) construction of a physical map of the walnut genome, (2) a detailed survey of walnut gene expression, and (3) fine-scale genetic and association mapping of economically important traits. bacterial artificial chromosome (BAC) libraries were constructed from in vitro grown shoots of Persian walnut (*Juglans regia* cv. Chandler), using cloning enzymes *Hind*III and *Mbo*I. There are total of 129,024 clones, 64,512 per BAC library. 48,218 high-quality BAC end sequences (BESs) were generated, with an accumulated sequence length of 31.2 Mb, representing ~3.9% of the walnut genome. Analysis of repeat DNA content in BESs revealed that approximately 15.42% of the genome consists of known repetitive DNA, while walnut unique repetitive DNA identified in this study constitutes 13.5 % of the genome.

Variability, genetic structure and differentiation based on morphological traits contribute significantly to accuracy and reliability of associations between molecular and economic traits in the genetic and association mapping of economic traits. A morphometric analysis of walnut germplasm based on nineteen economic traits comprising mostly nut characteristics indicated significant levels of variability. The coefficient of variability (CV) reflects the variation around the mean for different traits ranged from 7.8% for the number of pistillate flowers per cluster to 40.6% for nut shape and the variability is fairly high for all the nut traits. Kernel weight and shell thickness showed considerable variation with CV of 24.01 and 23.68 percent, respectively. The frequency distribution for most traits followed the normal curve suggesting the suitability of data for various univariate and multivariate analyses. The pattern of distribution of variability across the germplasm collection suggested a mild genetic structure, which is highly desirable for association genetic analysis.

A total of 20 cDNA libraries were sequenced using next generation DNA sequencing to define the walnut transcriptome, 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

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 was 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 was fragmented with different restriction enzymes and ordered into contiguous sequences based on the overlap of fragment patterns. Ends of the BAC clones were 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 was identified using bioinformatic tools and will be confirmed through their expression in walnut tissues (Objective 3). The GSTs will be used to obtain SNPs for the genetic mapping. 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. Genotypic data using microsatellite polymorphisms was used to validate the full-sib nature of the mapping population and to assess the genetic structure of the association mapping population. Phenotypic data has been collected every year to facilitate the genetic mapping once the SNP platform to perform the mapping is completed

Objective 3: Functional mapping of the walnut genome

Genetic and physical maps describe the structure of the genome, however, what needs to be defined is the gene space. This will be analyzed by deep sequencing of mRNA transcripts from different tissues to serve as a link to specific traits (Objective 2) and GSTs (Objective 1). A key step toward this is gene transcript sequencing to map the reads to expressed genes. Twenty tissue-specific gene transcript libraries have been constructed and sequenced to generate slightly more than one billion reads representing nearly all expressed genes (with the exception of a few rarely expressed genes). These reads have been assembled to build a library of 85,045 consensus sequences that represent and describe the walnut gene space. These sequences will be mapped to BES to identify high quality SNPs. They also will be used to derive functional information using

bioinformatics software to identify genes involved in important metabolic pathways. These represent key steps in progress from the genetic map to the complete walnut genome sequence.

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

A web-based browser has been developed for the walnut research community to access genomics resources. When complete, the database will contain all physical and linkage mapping information as well as all EST sequences and their integration with the walnut physical and genetic maps.

RESULTS AND DISCUSSION

Objective 1: Physical mapping of the walnut genome

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 *Hind*III and *Mbo*I. A total of 129,024 clones, 64,512 per BAC library, were arrayed in 336 plates (384-well). The average insert size is estimated to be 135 kb and 120 kb for the *Hind*III and *Mbo*I libraries, respectively. Assuming approximate walnut genome size of 800 Mb, these two BAC libraries represent ca. 20x genome equivalents. A total of 52,840 BAC clones from *Hind*III and *Mbo*I libraries of Persian walnut (*Juglans regia* cv. Chandler) were sequenced from one end that rendered 50,625 successful sequence reads. Following trimming and vector sequence removal, 48,557 BESs were obtained with quality score above 20 and sequence length longer than 100bp. From those, 339 BESs were removed due to high similarity to *Arabidopsis* organelle sequences. The resulted 48,218 walnut BESs will be deposited at GenBank. Among them, 25,948 BESs were from *Hind*III library, and 22,270 were from *Mbo*I library. In total, 31,181,448 bp of walnut genomic sequences were generated with average read length of 647 bp per BES. Assuming the walnut genome of approximately 800 Mb, these sequences represented 3.9% of the walnut genome. The average GC content of the BES was 37.7%, which is slightly higher than those sequenced dicot genomes such as papaya, poplar and *Arabidopsis* with the GC contents ranging from 33-35%, but lower than that of rice which is 43% (Ming et al., 2008). Blast2go analysis was performed to assign GO terms according to the blastx hits. Totally 1,330 unique GO terms were assigned to 6,396 BESs. Distribution of GO terms in categories of biological process, molecular function and cellular component was presented in Figure 1. The walnut genes are shown to cover a broad range of functional categories and biological processes. In molecular function categories, binding and catalytic activity were two most abundant categories, followed by transporter activity, electron carrier activity, and transcription regulator activity. The most represented biological processes were metabolic process, cellular process, localization, biological regulation, and response to stimulus. The abundance of each category for walnut was comparable to previous reports on other trees (Cheung and Town, 2007; Han and Korban, 2008; Barakat et al., 2009). Close to 4,000 SNPs have been discovered by aligning of BES with ESTs and whole genome shotgun sequences. These SNPs will be genetically mapped in the population of “Chandler x Idaho”, which will ultimately lead to anchoring BAC contigs onto linkage map. More SNPs from gene-containing region will be discovered by aligning ESTs and whole genome shotgun sequences (in progress). The linear order of the genes with SNPs will be placed on linkage map and then be integrated

into physical maps. The collection of SNPs will also serve as base of association mapping described below.

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

Nineteen morphological traits of economic importance were analyzed to assess phenotypic variability (Table 1) and to classify walnut germplasm using a cluster analysis (CA; Fig. 2) and principal component analysis (PCA; Fig. 3). Phenotypic variability and population structure are important for performing genetic and association mapping of economic traits. Here we estimate and describe the variability parameters for important economic traits in walnut germplasm and deduce the pattern of distribution of variability within and among the groups produced by the CA showing the level of population substructure within the collection (Figs 2 and 3). While assessment of morphological variability is important, the nature of its distribution in the germplasm collection, as reflected in the genetic structure, contributes to the nature of associations between the molecular markers and economic traits.

There is a significant amount of variability for all the economic traits analyzed as reflected in the descriptive statistics (Table 2 & Fig. 4). The coefficient of variability (CV) reflecting the variation around the mean for different traits ranged from 7.8% for the number of pistillate flowers per cluster to 40.6% for nut shape and the variability is fairly high for all the nut traits. Kernel weight and shell thickness showed considerable variation with CV of 24.01 and 23.68 percent, respectively. The frequency distribution for most traits followed the normal curve suggesting the suitability of data for various univariate and multivariate analyses (Fig. 4).

The CA and PCA of walnut germplasm indicated that the genetic structure based on important morphological traits is mild and therefore suitable for association genetic analysis. Six clusters were evident in the CA at the chi-square distance of 0.25 (Fig. 2). However, these clusters were not obvious in the PCA (Fig. 3). These results further confirm that the genetic structure is weak and excellent for association analysis.

Morphological analysis of mapping population ('Chandler' x 'Idaho') is currently being performed to understand the nature of segregation for the important quantitative traits. These traits will be mapped along with molecular markers (single nucleotide polymorphisms) in fine scale genetic analysis to develop a high density genetic map.

Objective 3: Functional mapping of the walnut genome

Functional mapping represents a composite profile of the transcriptome of walnut. This was accomplished by isolating mRNA from 20 different walnut tissues outlined in Table 3 that was used to construct cDNA libraries. The libraries were then sequenced using the next generation Illumina Genome Analyzer II sequencer. The cDNA were fragmented into 300-500 bp lengths and each was sequenced from both ends. The first few libraries were sequenced 125 x 125 bases from each end, but all subsequent cDNA libraries were sequenced 85x85. The number of reads obtained from each library is indicated in Table 3. In total over a billion reads were obtained. The raw cDNA sequence data from 23 lanes of Illumina GAI paired-end sequencing and 1 lane of single-end sequencing were adaptively trimmed with a custom script. The trimmed Illumina reads derived from each sample were assembled using velvet v. 0.1.12 / oases v. 0.1.15 (Zerbino and Birney, 2008; Zerbino et al., 2009). Contigs and singletons of at least 125 bp from the velvet

assemblies were then assembled using tgiel/CAP3 (Huang and Madan, 1999). The CAP3 output contigs and singletons of at least 200 bp (128,286 sequences) were consolidated to remove redundancy by mapping all assembly input sequences with bwa (Li and Durban 2009), and setting a threshold of 10 reads per kilobase (rpk) mapped to arrive at a final set of transcriptome contigs (85,045 sequences) (Table 3).

These contigs were compared to the Arabidopsis TAIR10 protein sequences with blastx (and then assigned Arabidopsis gene identifiers to the walnut assembly contigs that had blast hits below a particular e-value cutoff). The walnut contigs are also being annotated using Blast2GO software (Conesa and Gotz, 2008; Gotz et al. 2008). The GO categories and Enzyme Commission identifiers (Bairoch, 2000) can be used to map the associated proteins onto metabolic pathways to reveal which metabolic pathways are active in each tissue and at various time points in the growing season. From this, a functional map of the gene activities in walnut can be constructed. Using unsupervised clustering, gene set enrichment and functional visualizations, biomarkers will be identified: the genes which determine quality traits, and enable pathogen resistance and other stress responses.

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 at <http://walnutgenome.ucdavis.edu> will have two distinct components when completed: one for visualizing the physical map (now available for browsing) and one for visualizing the genetic map. The database will provide access to all of the fingerprinting data along with the BAC end sequences. 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.

REFERENCES

- Bairoch, A. (2000). The ENZYME database in 2000. *Nucleic Acids Res.* 28, 304.
- Barakat, A., DiLoreto, D.S., Zhang, Y., Smith, C., Baier, K., Powell, W.A., Wheeler, N., Sederoff, R., and Carlson, J.E. (2009). Comparison of the transcriptomes of American chestnut (*Castanea dentata*) and Chinese chestnut (*Castanea mollissima*) in response to the chestnut blight infection. *BMC Plant Biol* 9, 51.
- Cheung, F., and Town, C.D. (2007). A BAC end view of the *Musa acuminata* genome. *BMC Plant Biol* 7, 29.
- Conesa A and Gotz S (2008) Blast2GO: A Comprehensive Suite for Functional Analysis in Plant Genomics. *Int J Plant Genomics*. 2008: 619832.
- Götz S, García-Gómez J.M., Terol J., Williams T.D., Nagaraj S.H., Nueda M.J., Robles M., Talón M., Dopazo J., Conesa A. (2008). High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic Acids Res.* 36:3420-3435.
- Han, Y., and Korban, S.S. (2008). An overview of the apple genome through BAC end sequence analysis. *Plant Mol Biol* 67, 581-588.
- Huang X and Madan A (1999) CAP3: A DNA Sequence Assembly Program. *Genome Research*, 9: 868-877.

- 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.
- Ming, R., Hou, S., Feng, Y., Yu, Q., Dionne-Laporte, A., Saw, J.H., Senin, P., Wang, W., Ly, B.V., Lewis, K.L., Salzberg, S.L., Feng, L., Jones, M.R., Skelton, R.L., Murray, J.E., Chen, C., Qian, W., Shen, J., Du, P., Eustice, M., Tong, E., Tang, H., Lyons, E., Paull, R.E., Michael, T.P., Wall, K., Rice, D.W., Albert, H., Wang, M.L., Zhu, Y.J., Schatz, M., Nagarajan, N., Acob, R.A., Guan, P., Blas, A., Wai, C.M., Ackerman, C.M., Ren, Y., Liu, C., Wang, J., Na, J.K., Shakirov, E.V., Haas, B., Thimmapuram, J., Nelson, D., Wang, X., Bowers, J.E., Gschwend, A.R., Delcher, A.L., Singh, R., Suzuki, J.Y., Tripathi, S., Neupane, K., Wei, H., Irikura, B., Paidi, M., Jiang, N., Zhang, W., Presting, G., Windsor, A., Navajas-Perez, R., Torres, M.J., Feltus, F.A., Porter, B., Li, Y., Burroughs, A.M., Luo, M.C., Liu, L., Christopher, D.A., Mount, S.M., Moore, P.H., Sugimura, T., Jiang, J., Schuler, M.A., Friedman, V., Mitchell-Olds, T., Shippen, D.E., dePamphilis, C.W., Palmer, J.D., Freeling, M., Paterson, A.H., Gonsalves, D., Wang, L., and Alam, M. (2008). The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya* Linnaeus). *Nature* 452, 991-996.
- 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.
- Zerbino DR and Birney E (2008) Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.* 2008. 18: 821-829. doi: 10.1101/gr.074492.107
- Zerbino DR, McEwen GK, Margulies EH, Birney E (2009) Pebble and Rock Band: Heuristic Resolution of Repeats and Scaffolding in the Velvet Short-Read *de Novo* Assembler. *PLoS ONE* 4(12): e8407. doi:10.1371/journal.pone.0008407
- 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. Descriptors used in morphological evaluation of walnut germplasm

Code	Score	Full Characteristic Name	Description
LFDA	Date	50% Leafing Date	Date at which approximately 50% of leaf buds have opened to the point of reflexing (one or more leaves with petioles arching outwards)
FPINF	1, 2, >3	Flowers Per Inflorescence	Most frequent number of [pistillate] flowers recorded in 10 terminal inflorescences [generally 1-3 in <i>J. regia</i> ; 2 is most common]
HULDEH	1-3	Hull Dehiscence	[2008 Hulls and nuts dried together]
NUTSHAPE	1-9	Nut Shape	Longitudinal and perpendicular to suture
STEXTURE	1-9	Shell Texture	[cv. Chandler score is approximately 4-5]
SCOLOR	1-9	Shell Color	[cv. Chandler score is approximately 4]
SINTEGR	1-3	Shell Integrity	[Integrity of shell, not inclusive of insect damage]
NDIA		Nut Diameter	Face to face [face=cheek, or non-sutured side] in mm
NLENGTH		Nut Length	length in mm
NWT		Nut Weight	Inshell nut weight (g) [average of 10 nuts]
KWT		Kernel Weight	Kernel weight [average of 10 nuts]
K%		Kernel Percentage	Kernel weight/nut weight * 100
SSTRENGTH	1-7	Shell Strength	
SSEAL	1-9	Shell Seal	Suture integrity/strength [whether the suture had any openings, or separated easily during cracking]
PKTISSUE	1-9	Packing Tissue Thickness	[working with C. Leslie, most values were between 4 and 6, with 5 being most common (Chandler ~4)]
KFILL	1-7	Kernel Fill	[How tightly does the kernel fill the shell]
KREMOVL	1-9	Kernel Removal	Ease of removal of kernel halves
KPLUMP	1-7	Kernel Plumpness	
STHICK		Shell Thickness at cheek	Near center of half shell [scored from one randomly selected typical shell]

Table 2 Variability measures for different economic traits in walnut germplasm

Code	Mean±SE	Range	SD	CV
LFDA	79.072±0.555	63-113	7.752	9.8
FPINF	2.0049±0.011	1-3	0.1569	7.82
HULDEH	2.9138±0.0173	2-3	0.2462	8.45
NUTSHAPE	5.324±0.151	1-8	2.159	40.56
STEXTURE	5.049±0.058	3-7	0.829	16.42
SCOLOR	5.1373±0.0492	4-7	0.7024	13.67
SINTEGR	2.9265±0.0208	2-4	0.2969	10.15
NDIA	35.431±0.252	26.908-47.683	3.596	10.15
NLENGTH	40.905±0.349	29.956-58.318	4.981	12.18
NWT	153.25±2.3	82.9-264.5	32.91	21.47
KWT	72.36±1.22	25.5-120.7	17.38	24.01
K%	0.47399±0.0051	0.25994-0.68807	0.0729	15.38
SSTRENGTH	5.2549±0.0638	3-8	0.9117	17.35
SSEAL	4.7549±0.0422	3-7	0.6027	12.68
PKTISSUE	5.1471±0.0285	4-7	0.4068	7.9
KFILL	5.1225±0.0637	3-7	0.9095	17.76
KREMOVL	5±0.0573	3-8	0.8185	16.37
KPLUMP	5.3137±0.0466	4-7	0.6656	12.53
STHICK	1.5173±0.0252	0.62-2.48	0.3592	23.68

Table 3: Functional mapping of the walnut genome (Objective 3), walnut tissues gathered April to Nov 2008 for cDNA library construction and deep sequencing using Illumina GAI

Sample No	Tissue Source	Genotype	Developm'l Stage	Source	Code	cDNA Library	No of reads x10 ⁶
1	Vegetative Bud	Chandler	Vegetative	Tree	VB	Yes	39.9
2	Leaf – Young	Chandler	Vegetative	Tree	LY	Yes	63.3
3	Root	Chandler	Vegetative	Pot	RT	Yes	38.8
4	Callus Interior	Chandler	Vegetative	In Vitro	CI	Yes	59.3
5	Callus Exterior	Chandler	Vegetative	In Vitro	CE	Yes	29.8
6	Pistillate Flower	Chandler	Vegetative	Tree	FL	Yes	69.8
7	Catkins	Chandler	Immature	Tree	CK	Yes	56.4
8	Somatic Embryo	Chandler	Immature	In Vitro	SE	Yes	27.8
9	Leaf – mature	Chandler	Vegetative	Tree	LM	Yes	50.4
10	Leaves	Chandler	Vegetative	Tree	LE	Yes	60.1
11	Fruit immature	Mixed	Immature	Tree	IF	Yes	57.0
12	Hull immature	Chandler	Immature	Tree	HL	Yes	115.8
13	Packing Tissue	Chandler	Immature	Tree	PT	Yes	62.8
14	Hull Peel	Chandler	Mature	Tree	HP	Yes	43.3
15	Hull Cortex	Chandler	Mature	Tree	HC	Yes	62.8
16	Packing Tissue	Chandler	Mature	Tree	PK	Yes	56.7
17	Pellicle	Chandler	Mature	Tree	PL	Yes	42.7
18	Embryo	Mixed	Mature	Tree	EM	Yes	35.5
19	Hull – dehiscing	Chandler	Senescent	Tree	HU	Yes	59.5
20	Transition wood	<i>J.nigra</i>	Transition Zone	Tree	KW	Yes	48.4
21	Leaf – late	Chandler	Senescent	Tree	LS	No	N/A
Total number of reads x10 ⁶ :							1080
Total number of genes:							85,045

Fig. 2. Cluster analysis showing the population structure in walnut (based on nineteen morphological traits)

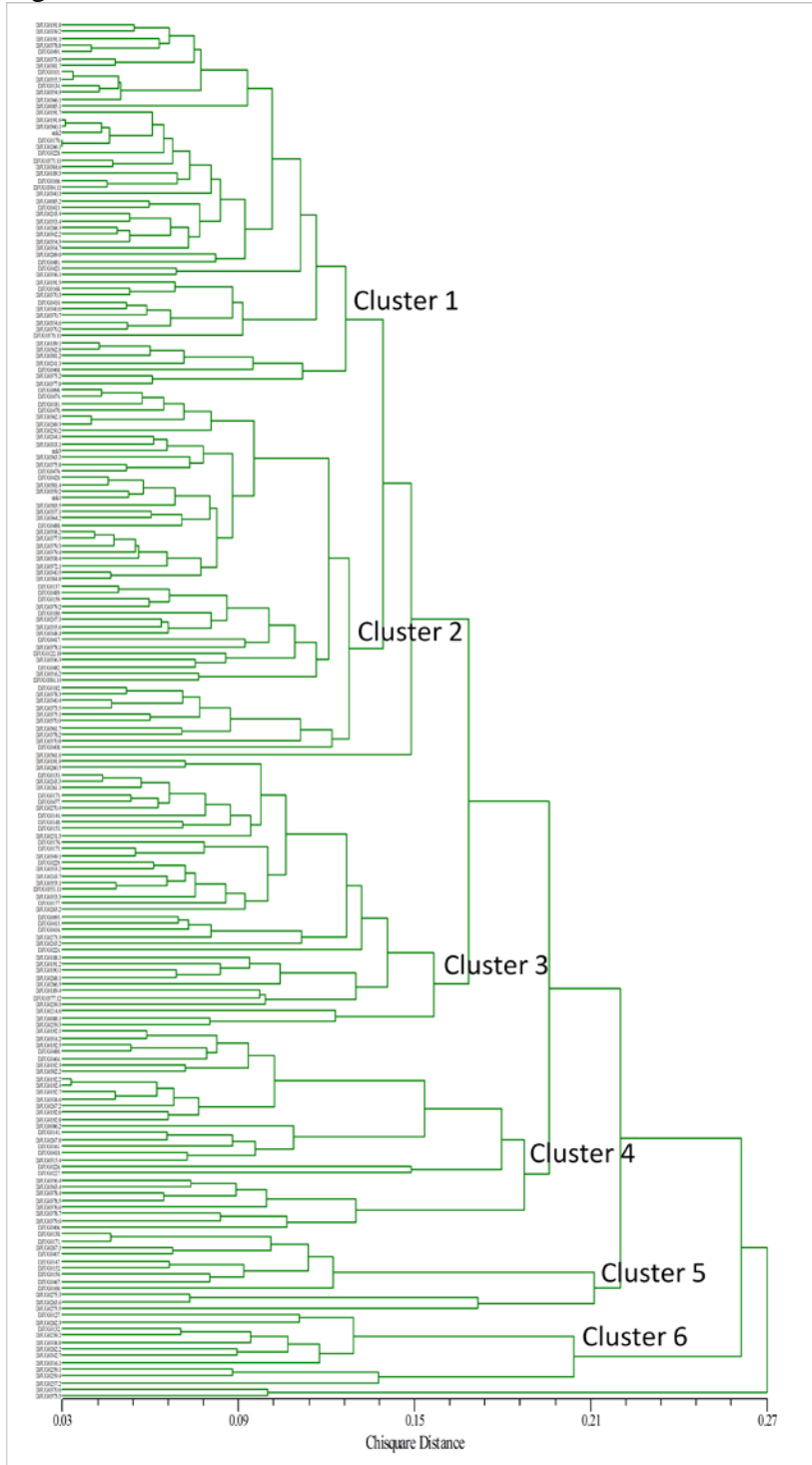


Fig. 3. 2D projection of walnut accessions along the first two axes (PCA based on nineteen morphological traits)

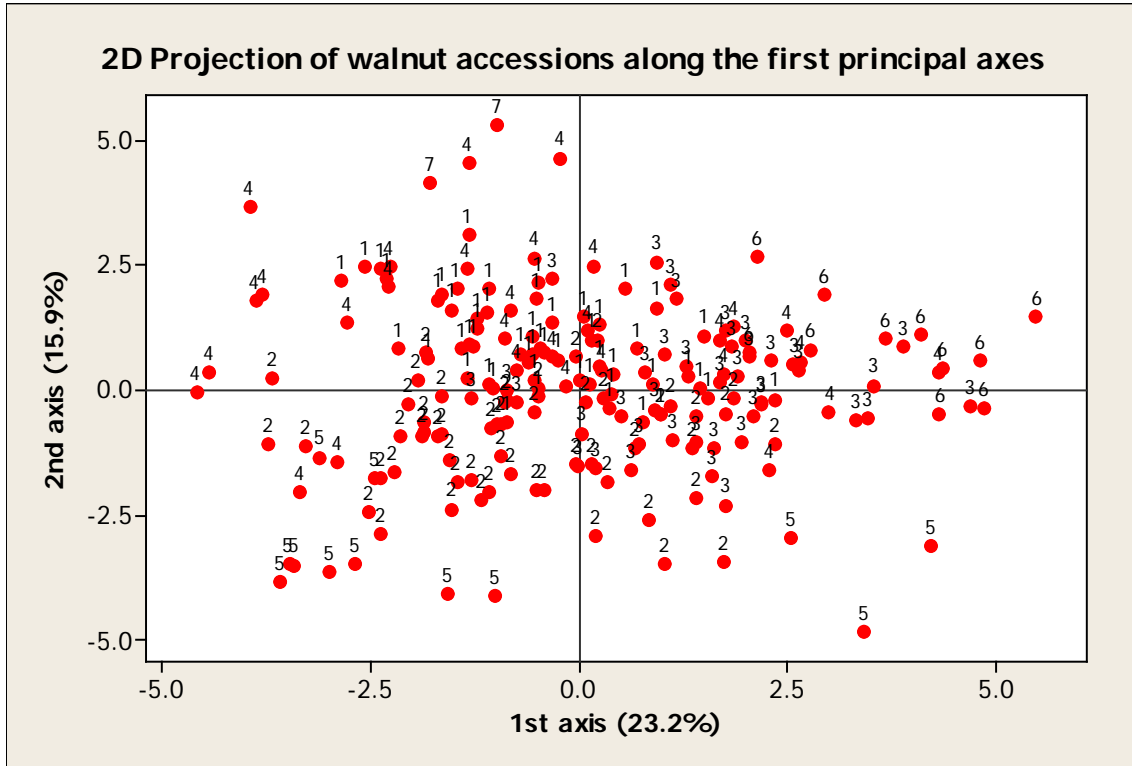


Fig. 4. Graphical depiction of variability of important morphological traits in walnut

