Journal of Heredity Advance Access published September 8, 2008

Journal of Heredity
doi:10.1093/jhered/esn068

Tracing the Geographic Origins of Major Avocado Cultivars

HAOFENG CHEN, PETER L. MORRELL, VANESSA E. T. M. ASHWORTH, MARLENE DE LA CRUZ, AND MICHAEL T. CLEGG

From the Department of Ecology and Evolutionary Biology, University of California, Irvine, CA 92697 (Chen, Morrell, Ashworth, and Clegg); and the School of Biological Sciences, University of California, Irvine, CA 92697 (de la Cruz). P. L. Morrell is now at the Monsanto Co, 700 Chesterfield Parkway North, Mail Stop: GG5B, Chesterfield, MO 63017.

Address correspondence to Michael T. Clegg at the address above, or e-mail: mclegg@uci.edu.

Abstract

It has been difficult to infer the genetic history of avocado breeding, owing to the role of hybridization in the origin of contemporary avocado cultivars. To address this difficulty, we used the model-based clustering program, STRUCTURE, and nucleotide polymorphism in 5960 bp of sequence from 4 nuclear loci to examine population structure in 21 wild avocado accessions. The origins of 33 cultivars were inferred relative to the wild sample. Nucleotide sequence diversity in domesticated avocados ranged between 80% and 90% of that observed for the same loci in wild avocado, depending on the diversity statistic used for comparison. Substantial genetic differentiation among 3 geographic groups of wild germplasm corresponded to the classically defined horticultural races of avocado. Previously undetected genetic differentiation was revealed in wild populations from Central Mexico, where 2 subpopulations were distinguished based on elevation and latitude.

Key words: *assignment testing, avocado, domestication, genetic resources, haplotype phasing, SNPs*

The domestication of plants in the Americas began at least 10 000 years ago with squash in Mexico (Smith 1997), followed by a rich variety of plants in Mesoamerica including maize, beans, and somewhat later tree crops such as cacao and avocado. Our knowledge of these early events is fragmentary and relies on scant archeological remains. The origins of avocado present a particularly challenging case, having been domesticated at least 3 times from geographically distinct populations of the progenitor species (summarized in Davis et al. 1998 and Ashworth and Clegg 2003). Even the names of the contemporary races of domesticated avocado are misleading, as one race has the appellation "West Indian," despite the fact that wild avocado does not occur in the West Indies. Moreover, many modern avocado cultivars are thought to be hybrids of 2 or more geographic races, although actual hybrid origins are a matter of speculation. We employ resequencing data, coupled with statistical analyses, to unravel the genealogical history of domestication of avocado and to identify the hybrid origins of various cultivars.

Avocado (*Persea americana* Mill) was domesticated in Mesoamerica where archeological sites in Coxcatlán (in the region of Tehuacán, Puebla State, Mexico) document human consumption as far back as approximately 8000–7000 BC (Smith 1966). According to Galindo-Tovar et al. (2007), avocado was cultivated and domesticated by the first Mesoamerican cultures (*The Mokayas*) who must have transmitted this practice to later cultures such as the Mayas and Olmecs. Larger seed sizes in more recent archeological strata at both the Tehuacán site and sites in the Oaxaca Valley (Oaxaca State, Mexico) suggest that human selection may have begun between 4000 and 2800 BC (Smith 1966, 1969). Linguistic evidence further supports the notion that indigenous cultures in Mesoamerica had used the avocado as food for a considerable length of time (Gama-Campillo and Gomez-Pompa 1991).

The putative wild progenitor of cultivated avocado was a polymorphic tree species that spanned a broad geographic area from the eastern and central highlands of Mexico through Guatemala to the Pacific coast of Central America (Smith 1966). Neolithic peoples selected primitive domesticated descendants (henceforth referred to as "wild" forms) from these populations, replacing the wild *P. americana* ancestor completely by 3 well-demarcated ecotypes of avocado known by their putative centers of origin as the Guatemalan (*Persea americana* var. *guatemalensis* Williams), Mexican (*Persea americana* var. *drymifolia* Blake), and West Indian (*Persea americana* var. *americana* Mill) horticultural races

(Bergh and Ellstrand 1986). Coastal Guatemala is now held to be the actual center of origin for the "West Indian" race (Scora et al. 2002). Ethnobotanical data (Williams 1976; Storey et al. 1986) and genetic marker studies (Furnier et al. 1990; Ashworth and Clegg 2003) suggest that these 3 races underwent separate domestication and did not come into contact until after European contact in the 16th century.

By the late 1800s, avocado improvement gained momentum via interracial hybridization between Guatemalan cultivars and Mexican germplasm in California and between Guatemalan cultivars and Cuban (West Indian race) germplasm in Florida (Robinson 1926; Davenport 1986). A long period of open-pollination and interracial hybridization has resulted in modern cultivars that are complex and often inaccurately characterized mixtures of the 3 horticultural races (Davis et al. 1998; Scora et al. 2002; Ashworth and Clegg 2003; Schnell et al. 2003).

Previous attempts at characterizing the diversity of cultivars and germplasm resources using microsatellite markers in conjunction with genetic distance and principal components analyses (Ashworth and Clegg 2003; Schnell et al. 2003) have suffered from the necessity of a priori (inter)racial assignments when attempting to draw conclusions on racial composition and from the difficulty of dealing with hybrids in genetic distance analyses. We address these problems by delimiting the racial boundaries de novo using haplotype data from a resequencing study of 4 nuclear loci in a sample of 21 wild avocado accessions (Chen et al. 2008). We use data on the same loci from a panel of 33 contemporary cultivars to revisit their presumed racial assignments and estimate the geographic sources that contributed to the genomes of modern cultivars.

Materials and Methods

Sampling strategy and data collection

Samples of 29 cultivars were obtained from the avocado germplasm collection maintained at the South Coast Research and Extension Center, Irvine, California. M. L. Arpaia of the University of California, Riverside, provided 4 cultivars: Andes 3, Andes 4, Leaven's Hass, and Puebla. Cultivar names and salient characteristics are shown in Table 1. Wild avocado sampling was described in Chen et al. (2008); the accessions represent wild avocados from central and southern states of Mexico, Costa Rica, Ecuador, and the Dominican Republic. In this study, we sequenced the same 4 loci as in the previous study (Chen et al. 2008), they are endo-1,4-D-glucanase (*Cellulase*), chalcone synthase (*CHS*), flavanone-3-hydroxylase (*F3H*), and serine-threonine-kinase (*STK*). Avocado genomic DNA was extracted from fresh leaves of cultivars using the DNAeasy plant mini-prep kit (Qiagen, Valencia, CA). Genomic DNA from dried wild materials was extracted using the cetyltrimethylammonium bromide method (Ausubel et al. 1994). For each sample, we used leaves from one tree for DNA extraction. Genes were amplified using primers listed in Chen et al. (2008), DNA sequencing and sequence assembly followed Morrell et al. (2003), using BigDye version 3.1 chemistry mix (Applied Biosystems, Foster City, CA) and Better Buffer sequencing buffer (The Gel Company, San Francisco, CA) according to the manufacturer's protocols. Sequence reads were assembled using PHRED/PHRAP/CONSED (Ewing and Green 1998; Gordon et al. 1998). POLYPHRED 5.04 (Nickerson et al. 1997; Bhangale et al. 2006; Stephens et al. 2006) was used to detect heterozygous single-nucleotide polymorphism (SNP) sites and insertions/deletions (indels). The sequences were aligned using CLUSTALW (Thompson et al. 1994) and alignments were manually adjusted. Haplotypes were phased experimentally using allele-specific polymerase chain reaction (PCR) (Chen HF 2006, unpublished dissertation, University of California, Riverside, CA). The accuracy of haplotype data was inferred with the error detection program EDUT (Toleno et al. 2007).

Data analysis

Descriptive statistics for sequence diversity and frequency spectrum were computed using the program COMPUTE and DESCPOLY from the Libsequence library of software (Thornton 2003). Reported statistics include the number of segregating sites, haplotype number, θ_W (Watterson 1975), π (Tajima 1983), and Tajima's *D* (Tajima 1989). The impact of recombination (R_m) and extent of linkage disequilibrium (LD) were estimated using the 4-gamete test (Hudson and Kaplan 1985) and Wall's *B* (Wall 1999).

The genetic clustering and assignment programs STRUCTURE (Pritchard et al. 2000) and STRUCTURAMA (Huelsenbeck and Andolfatto 2007) were used to investigate population structure in a sample of 21 wild avocado accessions. STRUCTURE was used to assign 33 popular avocado cultivars to one or more of the wild-inferred clusters according to their probability of membership in each wild cluster. STRUCTURE uses a Bayesian model-based clustering method that infers population structure based on multilocus genotype data. The clustering method assumes a model in which there are *K* populations (where *K* may be unknown). Each population is characterized by a set of distinct allele frequencies across loci.

In this study, genotype data consist of alleles or haplotypes at a series of "haplotype segments" at each of 4 loci (see Morrell and Clegg 2007). A frequency filter was applied to haplotype data, such that SNPs that were singletons within the full resequencing panel (wild and cultivated avocados) were removed (singleton SNPs can not be geographically informative) (Chen et al. 2008). Haplotype segments were inferred on the basis of direct evidence of recombination among 4-gamete intervals identified by DnaSP4.0 (Rozas et al. 2003). Haplotype segments were defined to contain all SNPs up to the boundary of a 4-gamete interval, thus delimiting a series of quasi-recombinationally independent chromosomal segments along each locus. STRUCTURE analysis was performed both with a linkage model (treating blocks from each of the 4 loci as linked with distances based on physical distance

Table 1. Key characteristics and breeding history of avocado cultivars in this study

Name	Characteristics and breeding history	Previous assignment	Results reported in this study
Anaheim	Originated in Anaheim, CA, 1910; seed size medium to small; skin green, medium, rough, or glossy; fruit 18–32 oz, shape ellipsoid to obovate; cold sensitive, tree tall and slender	G	G (99%)
Andes 3	A new variety from Hijuelas, Chile, 1999	N/A	G (95%)
Andes 4	A new variety from Hijuelas, Chile, 1999	N/A	G (95%)
Arue	From the Society Islands, 1932; fruit 20–30 oz, skin rough; seed large	WI	WI (99%)
Bacon	Originated 1928 in Bueno Park, CA; fruit ovoid, 7–12 oz; skin green, thick, and smooth; flesh very pale yellow-green; excellent frost tolerance; tree tall and slender	M × G	G (94%)
Daily 11	Originated near Camarillo, CA, 1941; fruit shape oblong, 30 oz; skin smooth, thick, color green; seed size small	Not stated	G (96%)
Duke 6	Rootstock; probably progeny of 'Duke'	M	M (100%)
Duke 7	Rootstock; probably progeny of 'Duke', more vigorous with greener foliage than 'Duke 6'	M	M (99%)
Esther	Seed size medium; fruit spheroid, 14–28 oz; skin green, dark green when soft, thick	G	G (99%)
Fuerte	From Atlixco, Mexico, 1911; fruit pyriform, 16 oz; skin dark green with small raised pale spots, thin; seed medium; alternate bearing; tree open, spreading, tall	M × G	M (99%)
Ganter	Originated as seedling in 1905 in Whittier, CA; fruit spheroid, 4–9 oz; skin smooth, green; seed size medium	M	M (98%)
Gwen	Presumed seeding of 'Thille', UCRBP; fruit pyriform to round, 10 oz; skin green, moderately thick, rough; seed small	M × G	G (99%)
H287	Fruit narrowly obovate; skin green; seed size small; 'Hass' progeny	Hybrid	M × G (57%, 43%)
H670	Similar to 'Hass', UCRBP; fruit narrowly obovate to spheroid, 7–12 oz; skin color green, black when ripe	Hybrid	M × G (50%, 50%)
Hass	Originating in La Habra Heights, CA, 1926; fruit pyriform, 7–10 oz; skin turning dark on tree, black when soft, pebbled, leathery; seed small, tight in cavity; flesh creamy; tree starts bearing second year	G	M × G (42%, 58%)
HX48	Fruit pyriform, 7–11 oz; skin color green, black when ripe; seed size medium; grandchild of 'Hass'	Hybrid	M × G (53%, 47%)
Khan	Also known by the name 'Toro Canyon'	N/A	M (100%)
Leaven's Hass	Detected in Ventura, CA, 2005; produces "early" 'Hass'-like fruit	N/A	G (95%)
Linda	Originating in Antigua, Guatemala, 1914; fruit round to oblong, 16–48 oz; skin dull purple, smooth, medium thick; seed small, tight in cavity; flesh yellow; tree low and spreading; regular bearing	G	G (99%)
Lyon	Originated in Hollywood, CA, 1911; fruit shape narrowly obovate, 11–23 oz; skin color green, medium thickness	Hybrid	M (99%)
Mexicola	Rootstock; fruit obovate, 4–6.5 oz; skin color black; very susceptible to root rot	M	M (99%)
Nabal	Originating in Antigua, Guatemala, 1917; fruit nearly spherical, 12–17 oz; skin green, smooth; seed medium to small, tight in cavity; flesh yellow; marked alternate bearing	G	G (99%)
Nimlioh	Originating in Antigua, Guatemala, 1917; fruit round, 28–40 oz; seed medium, tight in cavity; skin thick, black, and rough; alternate bearing	G	G (100%)
Noga	Fruit shape obovate, 12–14 oz; skin color green, black when ripe; skin thickness medium; seed size large	Hybrid	M × G (71%, 29%)
Pinkerton	Seedling of 'Hass' × 'Rincon', 1974; fruit pyriform, 8–14 oz, skin green, leathery; seed small, separates well from flesh; tree habit low and spreading	M × G	G (98%)
Puebla	Introduced in 1911 from Atlixco, Mexico; fruit shape obovate, 6–16 oz; skin color black, seed is tight in cavity	M	M × G × WI (6%, 82%, 12%)

Table 1. Continued

Name	Characteristics and breeding history	Previous assignment	Results reported in this study
Reed	Originated from Carlsbad, CA, 1948; probably derived from 'Anaheim' × 'Nabal'; tree slender; fruit round, 10–24 oz; skin medium to thick, green, slightly pebbled, easy to peel	G	G (100%)
Teague	Fruit shape obovate, 7–16 oz; skin thin, color green	M	M × G (82%, 16%)
Thille	Originated in Santa Paula, CA, 1954; fruit shape spheroid to obovate, 8 oz; skin thick, color green	Hybrid	G (100%)
Thomas	Rootstock; fruit ovate, low quality; skin black, thin, and smooth; seed large; survivor tree showing resistance to *Phytophthora cinnamomi* root rot	M	M (100%)
Topa Topa	Rootstock; originated in Ojai, CA, 1907; fruit oblique pyriform, 6–10 oz, poor eating quality; skin black, glossy, and smooth	M	M (100%)
Whitsell	Fruit shape obovate, 10–18 oz; skin thick, color green; seed size small	Hybrid	G (98%)
Zutano	Originating in Fallbrook, CA, 1926; fruit pyriform, 8–12 oz; skin very thin, pale green; flesh watery; seed medium; cold tolerant; tree upright	M × G	M × G (43%, 56%)

Previously hypothesized botanical race origins are compared with assignments inferred in this study.

[a]Data are from the University of California, Riverside (UCR) Web site (http://ucavo.ucr.edu). (UCRBP = UCR Breeding Program).

among the midpoints of the segments) and without linkage (treating each segment as unlinked). Independence among haplotype segments was tested using a Fisher's Exact test for linkage as implemented in GDA 1.1 (Lewis and Zaykin 2001).

The informativeness for assignment of the 21 haplotype segments was calculated using INFOCALC with I_n (informativeness for assignment), I_a (informative for ancestry coefficients), and ORCA (optimal rate of correct assignment) from both an 1-allele and 2-allele estimate (Rosenberg et al. 2003; Rosenberg 2005) computed relative to the 3 clusters of accessions identified in the STRUCTURE analysis (Table 2). Average informativeness of the markers is much higher than previously reported in similar studies that included haplotype segments from resequencing data, for example, in wild barley (Morrell and Clegg 2007).

In addition to the linkage model, we compare data models in STRUCTURE that included correlated or uncorrelated allele frequencies, models with and without population admixture, and combinations of these models (the linkage model requires an admixture model). Each round of analysis used 100 000 iterations for the burn-in period and a run length of 100 000 for assignment estimation. Each of the combinations of models with the data was run through 5 replicates; we report the model and assignment matrix only for the replicate with the highest log likelihood.

Initial STRUCTURE analysis included only wild samples, which were clustered into $K = 2$–5 clusters without the use of geographic location of origin. The number of distinct clusters in wild avocado was also estimated using STRUCTURAMA, which can explicitly estimate the number of distinct populations under a Dirichlet process prior (Huelsenbeck and Andolfatto 2007). For a second round of STRUCTURE analyses, geographic location was used for wild accessions, with wild accessions treated as a learning sample. STRUCTURE was used to infer the probabilities of origin of each cultivar from all given wild avocado clusters, with an admixture model permitting the inference of origin from more than one inferred (wild) cluster.

Results

Nucleotide Sequence Polymorphism and Diversity

Resequencing of 4 loci in 33 cultivated avocado accessions resulted in 5960 bp of aligned sequence. Resequencing data from the same 4 loci in 21 wild avocado accessions was reported by Chen et al. (2008). Because of difficulties with amplification and sequencing of some cultivated samples, not all individuals could be sequenced at all loci, resulting in sample sizes of 31 for *Cellulase*, 26 for *CHS*, 25 for *F3H*, and 33 for *STK*. All individuals were experimentally phased using allele-specific PCR. Thus, the number of chromosomes sampled is twice the number of sampled accessions. Observed heterozygosity in cultivars is overall higher than in wild accessions reflecting the importance of hybridization in cultivar origins. For *Cellulase*, *CHS*, *F3H,* and *STK*, observed heterozygosity in cultivars was 48.4%, 69.2%, 68%, and 60.6%, respectively, versus 40%, 70.6%, 50%, and 50% for the same loci in wild accessions (Chen et al. 2008).

Nucleotide diversity for both wild and domesticated accessions is reported in Table 3. Cultivars maintain a large proportion of the diversity observed in wild accessions. Mean θ_W for cultivated accessions is 5.65×10^{-3} versus 7.09×10^{-3} for the same loci in wild accessions. The mean value for π is 6.02×10^{-3} for cultivated accessions versus $\pi = 6.58 \times 10^{-3}$ for the same loci in wild accessions. The

Table 2. Four measures of marker information content are shown for 21 haplotype segments from 4 loci in wild avocados

Locus/segment	I_n[a]	I_a[a]	ORCA[a] (1-allele)	ORCA[a] (2-allele)
*Cellulase*_01	0.739	0.166	0.792	0.870
*Cellulase*_02	0.619	0.132	0.708	0.865
*CHS*_01	0.628	0.139	0.759	0.880
*CHS*_02	0.506	0.122	0.741	0.891
*CHS*_03	0.731	0.170	0.852	0.965
*CHS*_04	0.283	NA	0.537	0.616
*F3H*_01	0.680	0.153	0.757	0.851
*F3H*_02	0.510	0.105	0.688	0.822
*F3H*_03	0.492	0.100	0.667	0.792
*F3H*_04	0.711	0.163	0.833	0.958
*F3H*_05	0.357	0.074	0.604	0.741
*F3H*_06	0.644	0.143	0.785	0.930
*F3H*_07	0.023	NA	0.354	0.374
*F3H*_08	0.458	0.102	0.668	0.794
*F3H*_09	0.673	0.151	0.806	0.950
*F3H*_10	0.615	0.143	0.764	0.916
*F3H*_11	0.439	0.103	0.722	0.880
*STK*_01	0.759	0.172	0.806	0.942
*STK*_02	0.295	0.068	0.589	0.697
*STK*_03	0.447	0.097	0.653	0.834
*STK*_04	0.327	NA	0.583	0.738

The wild avocado samples were divided into Central Mexican, Southern Mexican/Guatemalan, and "West Indian" subsamples. "NA" indicates values that could not be calculated.

[a] Values reported are I_n, I_a, and ORCA from both a 1-allele and 2-allele estimate. For all measures, larger values indicate higher information content.

difference in diversity is attributable to the *Cellulase* and *STK* loci, where the number of SNPs observed in cultivated accessions is reduced relative to the wild accessions despite the larger cultivar sample size (Table 3, Figure 1). Three of the 4 loci have larger values of Tajima's *D* in domesticated accessions than in wild accessions, as expected, if the selection of cultigens has resulted in the loss of some rare alleles. However, haplotype number for cultivars is relatively large for all loci (Table 3) with many observed haplotypes sampled in both wild and cultivar accessions (Figure 1). Observed values of Wall's *B* indicate relatively low levels of intralocus LD, with similar levels of intralocus LD between wild and cultivar accessions (Table 3). All loci show evidence of recombination, as indicated by R_m (Table 3), in both wild and cultivated accessions.

Assignment Analyses

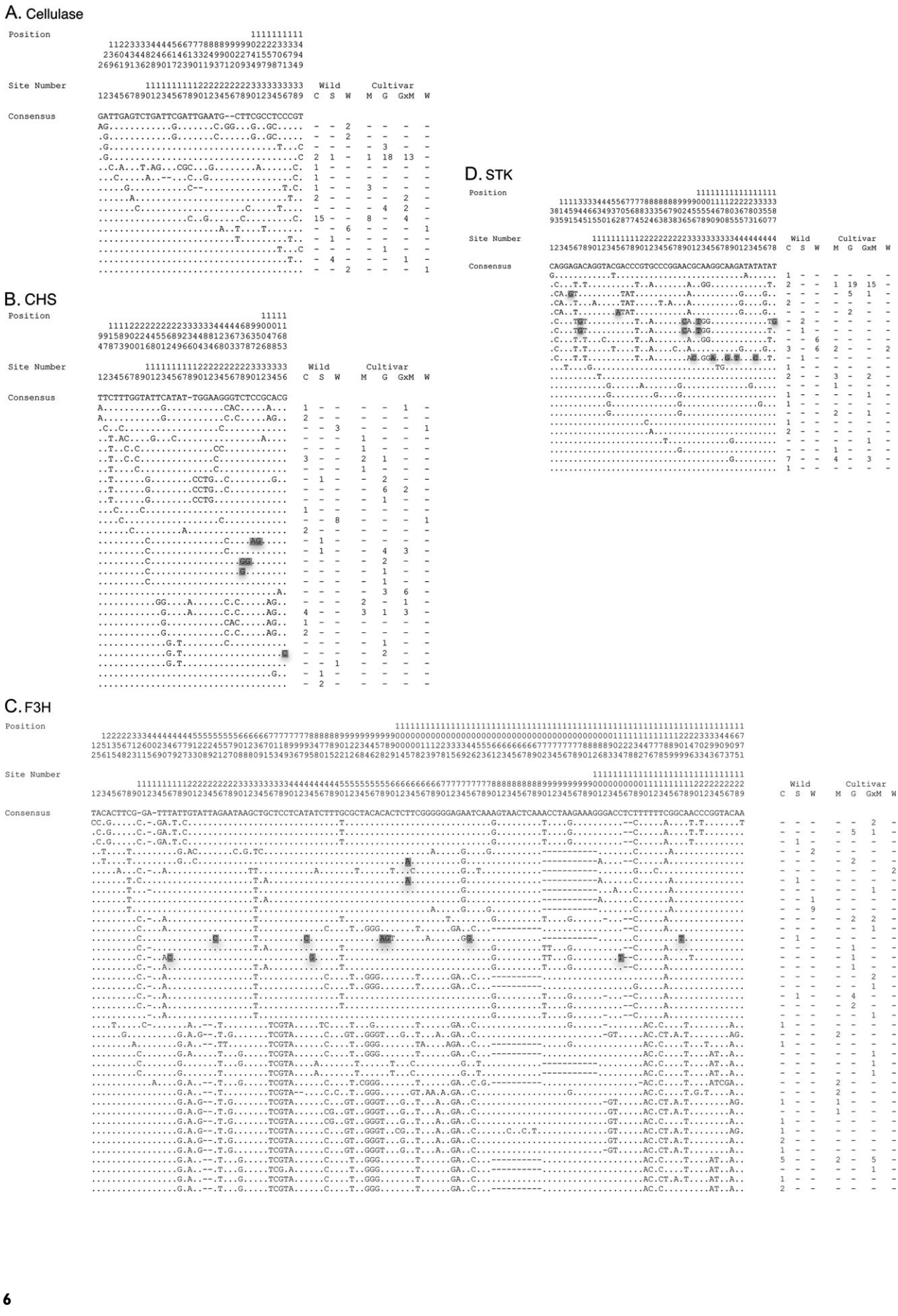
Using the genetic assignment methods implemented in STRUCTURE (Pritchard et al. 2000) and STRUCTURAMA (Huelsenbeck and Andolfatto 2007), we explore population structure in wild avocados by allowing the number of inferred clusters (K) to vary from 2 to 5. In both analyses, $K = 3$ resulted in the highest likelihood value for the dataset. The STRUCTURE analysis applied a model with no admixture and uncorrelated allele frequencies. The data model did not use linkage among markers. This choice is supported by a Fisher's Exact test for LD that did not identify significant LD (within the largest single sample from Central Mexico) between haplotype segments, including those from the same sequenced locus, at $P \leq 0.05$, particularly after correcting for multiple tests.

The clusters identified correspond closely to the known botanical races. The major clusters of wild avocados are associated with large differences in haplotype composition, as illustrated for the 4 loci in Figure 1. A number of sampled SNPs and haplotypes are unique to single clusters; for example, SNPs marked in gray in Figure 1 are private to Southern Mexican/Guatemalan wild and cultivar accessions. This is consistent with the substantial morphological differences between botanical races, suggesting that clusters identified in the STRUCTURE analysis represent genetically distinct populations.

When we assign the genomic composition of cultivars relative to the 3 groups ($K = 3$) identified in the wild avocado sample, all cultivars thought to have Guatemalan ancestry based on existing breeding records cluster with the Southern Mexican subpopulation. This strongly suggests

Table 3. Summary of measures of nucleotide diversity at 4 loci in wild and cultivated avocado samples

Gene	Aligned length (bp)	Sample	Sample size	Segregating sites	Haplotype number	$\theta_W \times 10^3$	$\pi \times 10^3$	Tajima's *D*	Wall's *B*	R_m
Cellulase	1540	Total	51	33	15	4.14	2.93	−0.896	0.094	2
		Wild	20	30	12	4.60	3.46	−0.850	0.103	1
		Cultivar	31	16	10	2.22	2.24	0.0346	0.067	2
CHS	1210	Total	43	35	27	6.09	4.92	−0.608	0.125	3
		Wild	16	27	15	5.66	4.36	−0.812	0.200	2
		Cultivar	27	29	19	5.67	5.02	−0.384	0.115	3
F3H	1812	Total	41	97	36	11.6	12.4	0.214	0.067	10
		Wild	16	76	16	10.9	12.3	0.503	0.151	5
		Cultivar	25	81	23	10.7	11.9	0.398	0.118	5
STK	1398	Total	53	48	22	6.56	5.51	−0.502	0.085	3
		Wild	20	43	16	7.23	6.18	−0.514	0.119	3
		Cultivar	33	27	11	4.06	4.98	0.717	0.115	1
Average	1490	Total	47	53.3	25	7.10	6.43	−0.448	0.093	—
		Wild	18	44	14.75	7.09	6.58	−0.418	0.143	—
		Cultivar	29	38.3	15.75	5.65	6.02	−0.200	0.104	—


A. Cellulase
B. CHS
C. F3H
D. STK
Position
Site Number
Consensus
Wild
Cultivar

that Southern Mexican populations are quite similar to the Guatemalan group and probably share similar haplotypes; thus, we can treat Southern Mexico as a proxy for Guatemalan wild populations. When we infer population structure of wild avocados with $K = 2$, the West Indian group clusters with the Southern Mexican/Guatemalan, suggesting that these 2 populations are more closely related than either is to populations from Central Mexico (data not shown). This result appears to corroborate suggestions that "West Indian" avocados originated in lowland Guatemala and is consistent with the proposed origin of the avocado races (Sanchez-Colín et al. 1998).

The largest cluster identified in the STRUCTURE analysis was from Central Mexico. We then further explored population structure within Central Mexican wild avocado with $K = 2$. The resulting analysis identified clusters distinguished by elevation and latitude of origin (Table 4 shows the original locations and elevations of Central Mexican accessions), where 1 cluster of 5 accessions is found below 2000 m above sea level, and largely in the southern portion of the region, and a second cluster of 5 accessions is found above 2000 m and largely in the northern portion of the region (Figure 2).

In a second round of analyses, we used the clusters inferred from wild avocado to assign cultivars to regions of origin with wild avocado treated as a learning sample. We assigned cultivars based on haplotype data, but without use of prior information on geographic location of origin, to one or more of 4 clusters, representing high elevation Central Mexican, low elevation Central Mexican, Guatemalan, and West Indian wild populations. Figure 3 depicts the inferred contribution of each of the 4 wild clusters to individual avocado cultivars. As expected, some cultivars are derived from interracial hybrids. The admixture depicted for individual cultivars in Figure 3 accords well with anecdotal information on hybrid origins (see Table 1) and also provides new information on cultivar ancestries. Generally, high elevation Mexican wild avocados have left a greater genetic footprint in hybrid cultivars than their low elevation counterparts. But it is noteworthy that both the high and low elevation clusters appear to have contributed to modern cultivars. For example, the Mexican parent of the leading commercial cultivar, 'Hass' assigns to the high elevation and northern cluster.

Most cultivars included in this study were also part of previous studies of restriction fragment length polymorphisms (Davis et al. 1998) and microsatellite diversity (Ashworth and Clegg 2003). The present SNP-based results are similar to those of the previous studies but provide a much clearer image of cultivar ancestry. For example, these results confirm that the cultivar Hass is a hybrid between Guatemalan (G) and Mexican (M) races, with inferred probability of assignment of 42% Mexican and 58% Guatemalan. This accords well with anecdotal information that 'Hass' is a G × (M × G) backcross. However, cultivars Gwen and Fuerte, also considered to be G × M hybrids, have a very high assignment to Guatemalan (99.4%) and Mexican (99.3%) ancestry, respectively. Only one cultivar, Puebla, combines elements from all 3 races (Mexican, Guatemalan, and West Indian). The 5 rootstocks sampled showed a 99–100% inferred probability of assignment to the Mexican race. However, previously unsuspected differentiation emerged, with 'Duke 7' and Thomas representing the high elevation subpopulation, 'Duke 6' and 'Topa Topa' the low elevation subpopulation, and 'Mexicola' combining both but with the high elevation component dominating.

Table 5 reports diversity within cultivar groups assigned to Guatemalan, Mexican, or mixed ancestry (West Indian cultivars are not reported due to limited sample size). As expected, diversity within individual groups is generally lower than in the full cultivar sample (Table 3) but remains relatively large for all partitions of the data. LD as measured by Wall's B is slightly larger among the mixed group as might be expected given the role of admixture, but it is noteworthy that the mixed group does not show higher diversity levels despite its hybrid origins.

Table 4. Collection localities and elevations of 10 wild avocado accessions from Central Mexico

Wild avocado designation	Locality, State	Elevation (m)
46	Tepetl, México	2660
63	Comonfort, México	2286
65	Vargas, México	2660
139	Ocampo, Michoacán	573
184	Guanajuato	2000
244	Tochimilco, Puebla	2070
QRO1	Villa Corregidora, Querétaro	2500
VER3	Calcahualco, Veracruz	1500
VER16	Coscomatepec, Veracruz	1521
VER22	Calcahualco, Veracruz	1600

Discussion

The potential impact of a domestication bottleneck on extant diversity in cultivated avocado is of major concern. Human imposed selection, both conscious and unconscious, restricts genetic diversity by limiting the number of lineages that are maintained for propagation. Ultimately, this

Figure 1. (**A–D**). Haplotype alignment for the 4 loci used in this study: (**A**) *Cellulase*, (**B**) *CHS*, (**C**) *F3H*, and (**D**) *STK*. Nucleotide positions are given for each informative site, and nucleotide substitutions characterizing the haplotypes of wild and cultivated accessions are indicated relative to the consensus sequence. Abbreviations: Under "Wild," C: Central Mexican; S: Southern Mexican/Guatemalan; W: West Indian. Under "Cultivar," M: Mexican; G: Guatemalan; G × M: Guatemalan and Mexican hybrid; W: West Indian. Private alleles of Southern Mexican/Guatemalan accessions are marked in gray.

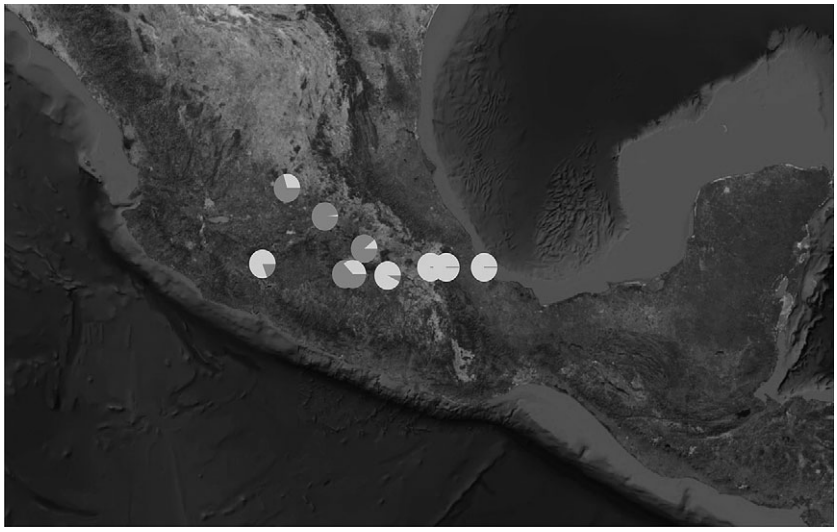


Figure 2. Population structure ($K = 2$) within Central Mexican wild avocado samples. Each of the 11 individuals is represented as a pie chart where gray and white represent the contributions from high elevation and low elevation haplotypes, respectively. Due to geographic proximity of some of the collection localities, 2 pairs of samples are partially or completely superimposed. The haplotype of the partially concealed individual that is fourth from the left (West) has a high/low elevation composition of 79.5% and 20.5%, respectively. The completely superimposed individual of the pair located third from the left and second from the top (North) (visible as a single, primarily gray, pie chart) has a high/low elevation composition of 1.7% and 98.3%, respectively.

translates into a reduced effective population size that affects all genes in the genome. Moreover, genes that carry useful mutations and are the direct targets of selection are expected to experience a selective sweep removing virtually all mutations not carried on the favored haplotype (assuming a low initial frequency for the useful mutation).

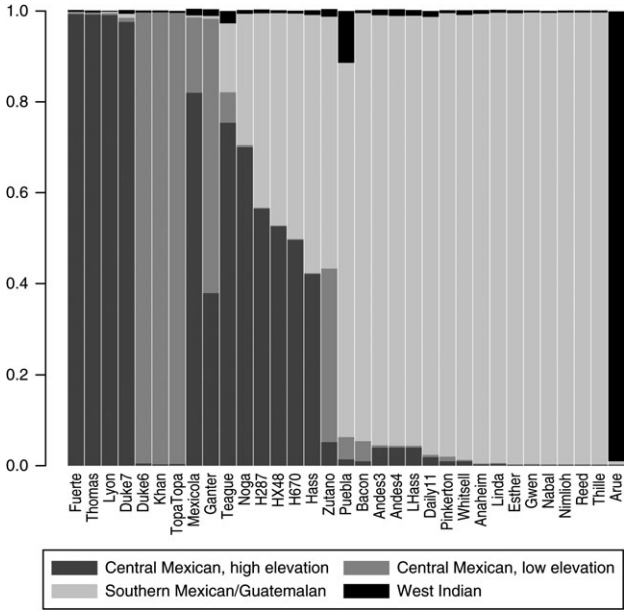


Figure 3. The assignment of 33 avocado cultivars to clusters inferred from DNA sequence haplotype data of 4 nuclear genes in 21 wild accessions. *x* axis: cultivar names; *y* axis: probabilities of assignment to the Mexican population (LHass = Leaven's Hass).

In the case of avocado, the picture is more complex, owing to separate domestication events and subsequent hybridization. The separate domestication events probably targeted independent samples from the wild gene pool and therefore would be expected to mitigate a bottleneck effect, a prediction supported by the diversity statistics (Table 3). The mean estimate of θ_W (Watterson 1975) is 5.65×10^{-3} for the pooled cultivated sample versus $\theta_W = 7.09 \times 10^{-3}$ for the wild avocado sample. When viewed in aggregate, the ratio of θ_W statistics averaged over all sites in the wild and cultivated samples is 0.797 (≈80%), indicating that only about 20% of average sequence diversity has been lost in the cultivars, consistent with the above expectation. If the frequency-based diversity measure, π (Tajima 1983), is used to estimate diversity loss through the domestication bottleneck, the ratio is 0.915. Calculations involving haplotype number or number of segregating sites fall between 93% and 87%. So despite undergoing several 1000 years of human selection and improvement by hybridization, the great majority of mutational genetic diversity found in wild avocado is retained in the cultivars, reflecting both the effect of multiple domestications and the role of subsequent hybridization.

When examined on a per-locus basis, the largest reduction in diversity is at the *Cellulase* locus where between about 50% (based on θ_W) and 35% (based on π) of diversity has been lost since domestication. Interestingly, the reduction is nearly equivalent in both Mexican and Guatemalan cultivars. Given the important role of the *Cellulase* locus in

Table 5. Summary of measures of nucleotide diversity at 4 loci in 3 avocado cultivar populations

Gene	Aligned length (bp)	Sample	Sample size	Segregating sites	Haplotype number	$\theta_W \times 10^3$	$\pi \times 10^3$	Tajima's *D*	Wall's *B*	R_m
Cellulase	1540	Guatemalan	30	11	6	1.81	1.42	−0.687	0.200	2
		Mexican	16	8	3	1.57	2.15	1.329	0.286	0
		Mixed	14	11	5	2.25	1.87	−0.661	0.300	0
CHS	1210	Guatemalan	24	20	12	4.65	4.49	−0.130	0.278	1
		Mexican	14	16	7	4.16	4.80	0.629	0.067	1
		Mixed	12	14	5	4.11	4.54	0.448	0.333	0
F3H	1812	Guatemalan	22	25	8	3.82	3.70	−0.120	0.125	1
		Mexican	14	42	8	7.58	7.76	0.101	0.400	2
		Mixed	12	41	8	7.80	9.17	0.801	0.410	3
STK	1398	Guatemalan	30	19	5	3.43	4.45	1.034	0.278	0
		Mexican	18	19	7	3.95	4.25	0.293	0.167	0
		Mixed	16	14	5	3.02	3.18	0.205	0.385	0

fruit maturation, it is possible there has been some degree of selection at this locus during domestication. But this would seem to require parallel selection in both Mexican and Guatemalan domesticates. The *STK* locus accounts for the remaining difference between wild and cultivated samples with a reduction in diversity ranging from 44% (based on θ_W) to 20% (based on π). Other statistical measures such as Tajima's *D* (Tajima 1989) (Table 3) do not, however, suggest extreme changes in the SNP frequency spectrum, so there is little compelling evidence for positive selection at the *Cellulase* or *STK* loci. Much larger samples of genes would be needed to test the selection hypothesis.

The geographic races of avocado are each distinguished by a suite of morphological traits; the populations identified in the STRUCTURE and STRUCTURAMA analyses closely correspond to these racial groupings. Indeed, substantial differences in haplotype composition among major racial groups suggest that these populations were separated for appreciable periods of time prior to the advent of human utilization of avocado (see Figure 1). The overall picture is one of modestly sized geographic populations that have retained their genetic integrity. Once humans entered the picture these different populations appear to have been domesticated more or less in situ and then disseminated more widely. In more recent times, these separate domesticates were admixed and provided the basis for many of the modern cultivated forms.

The suggestion of 2 genetic populations based on elevational and latitudinal differences in the highlands of Central Mexico is new and should stimulate further sampling and genetic analysis of avocado in Mexico and Central America. Avocado is one of the first trees domesticated in the Neotropics and can be used as model of tree domestication in areas of high biological diversity (Galindo-Tovar et al. 2007). The ability to unravel the complex hybrid origins of various cultivars should provide useful guidance to genetic resource managers and to breeders. It is also noteworthy that haplotype data can supplement archeological investigations by providing a finer resolution view of the domestication of major tree crops.

Funding

California Avocado Commission; University of California Discovery Program.

Acknowledgments

S. Hegde and J. Ross-Ibarra provided valuable comments on earlier versions of the manuscript. Sequence data from this article have been deposited with the GenBank Data Libraries under accession nos. EU335454–EU335749.

References

Ashworth VE, Clegg MT. 2003. Microsatellite markers in avocado (*Persea americana* Mill.): genealogical relationships among cultivated avocado genotypes. J Hered. 94:407–415.

Ausubel FM, Brent R, Kingston RE, Moore DD. 1994. Preparation of genomic DNA from plant tissue. Curr Protoc Mol Biol. 1: (Suppl 27):2.3.1.

Bergh B, Ellstrand N. 1986. Taxonomy of the Avocado. Calif Avocado Soc Yearb. 70:135–145.

Bhangale TR, Stephens M, Nickerson DA. 2006. Automating resequencing-based detection of insertion-deletion polymorphisms. Nat Genet. 38:1457–1462.

Chen H, Morrell PL, de la Cruz M, Clegg MT. 2008. Nucleotide diversity and linkage disequilibrium in wild avocado (*Persea americana* Mill.). J Hered. 99:382–389.

Davenport TL. 1986. Avocado flowering. Hortic Rev. 8:257–289.

Davis J, Henderson D, Kobayashi M, Clegg MT, Clegg MT. 1998. Genealogical relationships among cultivated avocado as revealed through RFLP analysis. J Hered. 89:319–323.

Ewing B, Green P. 1998. Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res. 8:186–194.

Furnier G, Cummings M, Clegg M. 1990. Evolution of the avocados as revealed by DNA restriction fragment variation. J Hered. 81:183–188.

Galindo-Tovar ME, Ogata-Aguilar N, Arzate-Fernandez AM. 2007. Some aspects of avocado (*Persea americana* Mill.) diversity and domestication in Mesoamerica. Genet Resour Crop Evol. doi: 10.1007/s10722-007-9250-5.

Gama-Campillo L, Gomez-Pompa A. 1991. An ethnoecological approach for the study of Persea: A case study in the Maya area. In: Lovatt CJ, Holthe

PA, Arpaia ML, editors. Proceedings of Second World Avocado Congress; 1991 Apr 21–26; California. Orange (CA). pp. 11–17.

Gordon D, Abajian C, Green P. 1998. Consed: A graphical tool for sequence finishing. Genome Res. 8:195–202.

Hudson RR, Kaplan NL. 1985. Statistical properties of the number of recombination events in the history of a sample of DNA sequences. Genetics. 111:147–164.

Huelsenbeck JP, Andolfatto P. 2007. Inference of population structure under a Dirichlet process model. Genetics. 175:1787–1802.

Lewis PO, Zaykin D. 2001. Genetic data analysis: computer program for the analysis of allelic data. Version 1.1. Storrs, CT: University of Connecticut.

Morrell PL, Clegg MT. 2007. Genetic evidence for a second domestication of barley (*Hordeum vulgare*) east of the Fertile Crescent. Proc Natl Acad Sci USA. 104:3289–3294.

Morrell PL, Lundy KE, Clegg MT. 2003. Distinct geographic patterns of genetic diversity are maintained in wild barley (*Hordeum vulgare ssp. spontaneum*) despite migration. Proc Natl Acad Sci USA. 100: 10812–10817.

Nickerson DA, Tobe VO, Taylor SL. 1997. PolyPhred: Automating the detection and genotyping of single nucleotide substitutions using fluorescence-based resequencing. Nucleic Acids Res. 25:2745–2751.

Pritchard JK, Stephens M, Donnelly P. 2000. Inference of population structure using multilocus genotype data. Genetics. 155:945–959.

Robinson TR. 1926. Avocado for Florida. Proc Fla State Hortic Soc. 39:182–191.

Rosenberg NA, Li LM, Ward R, Pritchard JK. 2003. Informativeness of genetic markers for inference of ancestry. Am J Hum Genet. 73: 1402–1422.

Rosenberg NA. 2005. Algorithms for selecting informative marker panels for population assignment. J Comput Biol. 12:1183–1201.

Rozas J, Sanchez-DelBarrio JC, Messeguer X, Rozas R. 2003. DnaSP, DNA polymorphism analyses by the coalescent and other methods. Bioinformatics. 19:2496–2497.

Sanchez-Colín S, Mijares-Oviedo P, Lopez-Lopez L, Barrientos-Priego AF. 1998. Historia del Aguacate en Mexico. Memoria Fundación Salvador Sánchez Colin 1998-2001. Coatepec Harinas (México): CICTAMEX. pp. 171–187.

Schnell RJ, Brown JS, Olano CT, Power EJ, Krol CA. 2003. Evaluation of avocado germplasm using microsatellite markers. J Am Soc Hortic Sci. 128:881–889.

Scora RW, Wolstenholme BN, Lavi U. 2002. Taxonomy and botany. In: Whiley A, Schaffer B, Wolstenholme B, editors. The avocado: Botany, production and uses. New York: CAB International. pp. 15–37.

Smith BD. 1997. The initial domestication of *Cucurbita pepo* in the Americas 10,000 years ago. Science. 276:932–934.

Smith CE. 1966. Archaeological evidence for selection in avocado. Econ Bot. 20:169–175.

Smith CE. 1969. Additional notes on pre-conquest avocados in Mexico. Econ Bot. 23:135–140.

Stephens M, Sloan JS, Robertson PD, Scheet P, Nickerson DA. 2006. Automating sequence-based detection and genotyping of SNPs from diploid samples. Nat Genet. 38:375–381.

Storey WB, Bergh BO, Zentmyer GA. 1986. The origin, indigenous range and dissemination of the avocado. Calif Avocado Soc Yearb. 70:127–133.

Tajima F. 1983. Evolutionary relationship of DNA sequences in finite populations. Genetics. 105:437–460.

Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics. 123:585–595.

Thompson JD, Higgins DG, Gibson TJ. 1994. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22:4673–4680.

Thornton K. 2003. Libsequence: a C++ class library for evolutionary genetic analysis. Bioinformatics. 19:2325–2327.

Toleno DM, Morrell PL, Clegg MT. 2007. Error detection in SNP data by considering the likelihood of recombinational history implied by three-site combinations. Bioinformatics. 23:1807–1814.

Wall JD. 1999. Recombination and the power of statistical tests of neutrality. Genet Res. 74:65–79.

Watterson GA. 1975. On the number of segregating sites in genetical models without recombination. Theor Popul Biol. 7:256–276.

Williams LO. 1976. The botany of the avocado and its relatives, pp. 9–15 In: Sauls JW, Phillips RL, Jackson LK, editors. Proceedings of the First International Tropical Fruit Short Course, The Avocado; 1976 Nov 5–10; Gainsville (FL): University of Florida.

Received May 28, 2008
Accepted July 29, 2008

Corresponding Editor: John Burke