Journal of Heredity 2008:99(4):382–389
doi:10.1093/jhered/esn016
Advance Access publication March 14, 2008

Nucleotide Diversity and Linkage Disequilibrium in Wild Avocado (*Persea americana* Mill.)

Haofeng Chen, Peter L. Morrell, Marlene de la Cruz, and Michael T. Clegg

From the Department of Ecology and Evolutionary Biology, University of California, Irvine, CA 92697 (Chen, Morrell, and Clegg); and the School of Biological Sciences, University of California, Irvine, CA 92697 (de la Cruz).

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

Abstract

Resequencing studies provide the ultimate resolution of genetic diversity because they identify all mutations in a gene that are present within the sampled individuals. We report a resequencing study of *Persea americana*, a subtropical tree species native to Meso- and Central America and the progenitor of cultivated avocado. The sample includes 21 wild accessions from Mexico, Costa Rica, Ecuador, and the Dominican Republic. Estimated levels of nucleotide polymorphism and linkage disequilibrium (LD) are obtained from fully resolved haplotype data from 4 nuclear loci that span 5960 nucleotide sites. Results show that, although avocado is a subtropical tree crop and a predominantly outcrossing plant, the overall level of genetic variation is not exceptionally high (nucleotide diversity at silent sites, $\pi_{sil} = 0.0102$) compared with available estimates from temperate plant species. Intralocus LD decays rapidly to half the initial value within about 1 kb. Estimates of recombination rate (based on the sequence data) show that the rate is not exceptionally high when compared with annual plants such as wild barley or maize. Interlocus LD is significant owing to substantial population structure induced by mixing of the 3 botanical races of avocado.

Avocado (*Persea americana* Mill. *Lauraceae*) is an important subtropical crop in the Americas where it has been cultivated for several thousand years. The wild progenitor of cultivated avocado is a polymorphic tree species that is comprised of 3 botanical races: *Persea americana* var. *americana* Mill (the West Indian race), *Persea americana* var. *drymifolia* Blake (the Mexican race), and *Persea americana* var. *guatemalensis* Williams (the Guatemalan race) (Bergh and Ellstrand 1986). *Persea americana* has a synchronously dichogamous breeding system that tends to enforce outcrossing (Davenport 1986), although self-fertilization does occur in orchards composed of a single genotype (Davenport et al. 1994; Kobayashi et al. 2000). Avocado is a branched evergreen tree 40–80 ft tall with elliptic leaves 3–10 inches long. The avocado fruit has long been used as a rich source of human nutrition, has a high lipid content from 5% to 30% depending on the cultivar, and also is a good source of vitamins and antioxidants (Sharma and Martinez 1972). A recent comprehensive review of avocado biology, genetics, and future prospects in genomics can be found in Chanderbali et al. (2007).

Nucleotide sequence diversity has been extensively studied in several short-lived temperate plant and animal species, including *Arabidopsis thaliana* (Aguade 2001; Nordborg et al. 2005), *Drosophila melanogaster* (Haddrill et al. 2005; Thornton and Andolfatto 2006), wild barley (Lin et al. 2002; Morrell et al. 2003, 2005), and maize (Gaut et al. 2000; Tiffin and Gaut 2001a, 2001b). Resequencing studies of perennial woody plants with long generation times are still rare; exceptions include studies of aspen (Ingvarsson 2005), cypress (Kado et al. 2003), and *Pinus sylvestris* (Dvornyk et al. 2002; Garcia-Gil et al. 2003). The results of these somewhat limited studies indicate that outcrossing trees from temperate latitudes have low nucleotide diversities compared with selfing annual species like *Arabidopsis thaliana* and *Hordeum vulgare* ssp. *spontaneum* (wild barley). This stands in contrast to meta-analyses of allozyme data, which tend to suggest higher diversity in tree species (Hamrick and Godt 1997).

To date, nucleotide diversity has not been explored in tropical tree species. In this paper, we examine nucleotide sequence diversity at 4 loci in wild avocado accessions. We address the following questions: How much nucleotide sequence diversity is present in this subtropical tree species? What is the relative importance of recombination versus mutation in generating haplotypic diversity in avocado? And

Table 1. Wild avocado (*Persea americana*) accessions used in this study and their origins

Wild avocado name	Locations	Country	Botanical race[a]
46	Tepetl, State of Mexico	Mexico	M
63	Comonfort, State of Mexico	Mexico	M
65	Vargas, State of Mexico	Mexico	M
139	Ocampo, Michoacan	Mexico	M
184	Guanajuato	Mexico	M
229	Amatenango, Chiapas	Mexico	G
244	Tochimilco, Puebla	Mexico	M
Ch35	Amatenango, Chiapas	Mexico	G
QRO1	Villa Corregidora, Queretaro	Mexico	M
VER3	Calcahualco, Veracruz	Mexico	M
VER16	Coscomatepec, Veracruz	Mexico	M
VER22	Calcahualco, Veracruz	Mexico	M
Yu60	Yucatan	Mexico	G
Costa Rica		Costa Rica	W
Ecuador	Guayaquil	Ecuador	M
41	Guayaquil	Ecuador	M
JM1		DR	W
MB1		DR	W
MC1		DR	W
Ch3B		DR	W
MG1		DR	W

[a] M, Mexican; G, Guatemalan; W, West Indian; DR, the Dominican Republic.

how extensive is the intra- and interlocus linkage disequilibrium (LD) in this predominantly outcrossing tree?

Materials and Methods

Plant Materials

A total of 21 wild accessions of *P. americana* were sampled in this study. The accessions were obtained from avocado germplasm collections located at the Centro de Investigaciónes Cientificas y Tecnológicas del Aguacate, in Coatepec de Harinas and Ixtapan de la Sal in the State of Mexico, Mexico, and at the National Institute for Forestry, Agriculture and Livestock Research Germplasm collection in the state of Guanajuato, Mexico. The wild material originated from multiple states within Mexico, from the mountainous central states of Guanajuato and Querétaro to the eastern coastal state of Veracruz, and from southern states of Chiapas and Yucátan. Accessions were also obtained from Central America (Costa Rica), South America (Ecuador), and the Dominican Republic (DR) on the Caribbean island of Hispaniola (Samples from DR were kindly provided by J. Nunez). The sample includes all 3 botanical races. The locations and putative racial designations of the wild avocado accessions are listed in Table 1.

Loci Sequenced

Four nuclear genes were sequenced in this study. Flavanone-3-hydroxylase (*F3H*) and chalcone synthase (*CHS*) are 2 of the core enzymes in the flavonoid biosynthesis pathway. The function of *F3H* is to convert (2S)-flavanones to (2R, 3R)-dihydroflavonols, which is a direct intermediate in the biosynthesis of anthocyanidins and flavonols (Forkmann and Dangelmayr 1980; Britsch et al. 1992). *CHS* synthesizes a precursor for the biosynthesis of flavonoids and isoflavonoids (Durbin et al. 2000). *Cellulase* (endo-1,4-D-Glucanase) is a gene related to ripening, whose activity increases throughout the avocado fruit ripening process (Tucker et al. 1987). Serine–threonine–kinase (*STK*) is one of the known classes of disease-resistance genes in plants (Di Gaspero and Cipriani 2003).

It was not possible to obtain complete sequences from all accessions for all 4 loci. Sample sizes for *F3H*, *CHS*, *Cellulase*, and *STK* were 16, 17, 20, and 20, respectively. For the *F3H* locus, variable length insertions/deletions (indels) within a large intron (~1 kb) precluded sequencing all accessions. For the *CHS* locus, a poly "A" polymorphism located just before the second exon of the locus caused unreliable reads, so the alignment was truncated after the first indel and resulted in a final sequence of 1.2 kb.

DNA Extraction, Polymerase Chain Reaction Amplification, and Sequencing

Avocado genomic DNA was extracted from silica gel–dried wild avocado leaves using the Cetyltrimethylammonium Bromide method and purified by cesium chloride equilibrium density gradient ultra centrifugation (Ausubel et al. 1994). All loci were sequenced in both directions, with a minimum 2-fold coverage and 4- to 6-fold coverage for most sequenced regions. Amplification primers were designed according to the published avocado mRNA sequences from GenBank. All primers used in these reactions are listed in Table 2. Polymerase chain reaction (PCR) amplication conditions were as follows: 95 °C preheating for 3 min, then 95 °C for 45 s, annealing temperatures ranged from 53 to 57 °C for 1 min according to the melting temperature (T_m) of individual primers, extension temperature was 72 °C, and the time ranged from 1 to 2 min 30 s depending on the length of the amplified products. Amplification included 35 cycles with a final extension of 72 °C for 30 min for poly-A tail elongation. PCR products were purified using the QIAquick PCR purification kit (Qiagen, Valencia, CA). Cleaned PCR products were direct sequenced using BigDye version 3.1 chemistry mix (Applied Biosystems, Foster City, CA) and Better Buffer sequencing buffer (The Gel Company, San Francisco, CA) following the manufacturer's protocols. The initial amplification primers and internal sequencing primers were used for sequencing, and each primer provided 500- to 600-bp high-quality DNA sequence. The sequence products were run on both ABI 377 and ABI 3100 sequencers (Applied Biosystems).

Sequence reads were assembled using PHRED/PHRAP/CONSED (Ewing and Green 1998; Gordon et al. 1998) to generate complete contigs, and POLYPHRED was used to detect the single-nucleotide polymorphism (SNP) sites and indels (Nickerson et al. 1997). The sequences were aligned using the program CLUSTALW

Table 2. Primers used in the 4 loci amplification and sequencing

Primer name	Tm	%GC	Direction	Sequence (5′-3′)
F3HPaF1[a]	60.6	59.09	Forward	GGTTGCTTACAACCAGTTCAG
F3HPaF2	60.1	44.44	Forward	ACCCTTGTCAGTTGTTCT
F3HPaF3	50.1	43.75	Forward	GAAACAAGGGAAGATG
F3HPaF4	63.4	61.11	Forward	CTCCAGGCCACCAAAGAC
F3HPaR1[a]	62.45	59.09	Reverse	GCTTCTCACCCTCCCTAATC
F2HPaR3	59.7	50.00	Reverse	GGCTAGACCCATTAGCTTGT
F3HPaR4	61.0	44.44	Reverse	TTTCATTTCCCATCTGCC
F3HPaR5	52.7	38.89	Reverse	GCTCATTTGATTAGGATG
CellulaseF1[a]	69.6	54.55	Forward	GCTGCTGCCTCCATCGTATTTG
CellulaseF3	63.0	55.56	Forward	TACCTTCCATGGCCCACT
CellulaseF4	60.2	50.00	Forward	TCATCTCCCTGTATGCTCAC
CellulaseR1[a]	68.2	54.55	Reverse	TCTGTCACTCCGTCACTGGGTT
CellulaseR3	60.8	50.00	Reverse	CCTCCGGAAGAGTTCAGATA
CellulaseR4	61.4	57.89	Reverse	TCTGCCCCTAGTGTGTGAC
CHSPaF1[a]	73.7	59.09	Forward	TCGGATCACCAACAGCGAGCAC
CHSPaF2	52.8	31.58	Forward	CTATTGATGCATGATTTTC
CHSPaR1[a]	73.9	59.09	Reverse	GCAACCCAAGGCTATCAAAGGA
CHSPaR2	61.3	55.56	Reverse	GGTACATCATGAGGCGCT
STKPaF1[a]	61.2	41.67	Forward	CGGGAGGACATAATATATTACTGC
STKPaF2	55.2	42.11	Forward	GTTGCTCCTGAAGTGATAA
STKPaF3	60.4	50.00	Forward	CAGGAAAGAAGCACGGTT
STKPaR1[a]	59.8	39.13	Reverse	CAGGCAGATCATAACAGATAGAA
STKPaR2	62.4	52.63	Reverse	CGTGCAGCAACACCAGTAT
STKPaR3	60.1	47.37	Reverse	TGAAGTAAACCATCCTGCC

[a] Amplification primers, also used in sequence reactions.

(Thompson et al. 1994), and alignments were manually adjusted.

Haplotype Phasing

Almost two-thirds of the accessions were heterozygous at each locus. We used allele-specific PCR (AS-PCR) amplification to experimentally phase haplotypes of heterozygous individuals. Paired AS-PCR primers were designed to terminate in a common SNP or indel site, with each of the primers ending in one of the SNP or indel states. For SNP-specific primers, an additional nucleotide mismatch was included to the 3′ end of the primer, 2 nucleotides from the final (selective) nucleotide site (Newton et al. 1989). AS-PCR primers were paired with one of the initial amplification primers for second-round PCR (i.e., AS-PCR), using purified product from the first-round PCR as template. Purified AS-PCR products were then direct sequenced to get individual haplotypes. A detailed description of the method is presented in Chen (Chen HF, 2006, unpublished dissertation, University of California, Riverside, CA). The accuracy of inferred haplotypes was assessed using Error Detection Using Triplets (EDUT) (Toleno et al. 2007).

Sequence Analysis

Sequence Diversity Estimation and Tests of Neutrality

Nucleotide polymorphism was analyzed using tools from the LIBSEQUENCE library (Thornton 2003). COMPUTE and DESCPOLY were used to calculate nucleotide diversity: Watterson's θ (θ_W) (Watterson 1975) and Tajima's π (Tajima 1983). POLYDNDS was used to estimate the level of polymorphism at silent and replacement sites. COMPUTE was also used to calculate Tajima's D (Tajima 1989) in order to test departures from neutrality of the standard coalescent model. The significance of Tajima's D was tested by 10 000 replicate coalescent simulations.

Linkage Disequilibrium

Two approaches were used to estimate levels of LD: first, the population recombination parameter $\rho = 4N_e r$ was estimated, where N_e is the effective population size and r is the recombination rate per generation (Hudson 1987; Andolfatto and Przeworski 2000). We employed 2 estimators of the recombination parameter ρ: Hudson's estimator ($\hat{\rho}_{H01}$), which uses an approximate likelihood method, based on pairwise LD between sites (Hudson 2001) and Thornton's estimator ($\hat{\rho}_{T05}$), in which the posterior distributions of $\hat{\rho}$ and $\hat{\theta}$ are jointly estimated on the basis of summary statistics of the observed data using rejection sampling (Haddrill et al. 2005). We also report the ratio $\hat{\rho}/\hat{\theta}$, a measure of the relative contribution of recombination and mutation to haplotypic diversity. The LD estimator Wall's B (Wall 1999) and the minimum number of recombination events (Hudson and Kaplan 1985) are also reported.

LD between pairs of polymorphic sites within each locus was calculated based on squared correlation in allele frequency, r^2, (Weir 1990), using the program RSQ (Thornton 2003). The decay of LD with physical distance was estimated using nonlinear regression of LD between polymorphic sites versus distance in base pairs between sites

(Remington et al. 2001). The statistical package R (www.r-project.org) was used to summarize the output file of RSQ and to make plots of LD versus distance. Low-frequency sites, even when there is complete association, will result in low values of r^2. This means that r^2 plots will contain many values near zero, and these contribute substantial background noise to the data. Thus, we calculated r^2 using a frequency filter of 5 chromosomes or roughly 10% frequency; polymorphic sites with a frequency of <5 were removed from the calculation.

The extent of interlocus LD was calculated by using randomizations of haplotype combinations at 2-locus pairs to test for an excess of LD in the empirical data sets relative to random configurations of the haplotypes at each locus (Morrell et al. 2005). The method compares the extent of LD between all pairs of sites between 2 loci, generally using a minimum frequency filter such that the rare SNP variant must occur at least twice in the sample. This produces a distribution of values for the LD statistic (we use r^2) with a total of ($S_1 \times S_2$) values, where S_1 and S_2 are the number of SNPs considered at each locus. The observed distribution of r^2 values in empirical data was compared with those observed in 1000 randomizations of sample order. For the present diploid data, the phase of haplotypes in the empirical data is arbitrary. Preservation of haplotypes when sample order is randomized results in a more conservative test in cases where some degree of nonindependence is observed between linked sites (within a locus)(Morrell et al. 2005).

Results

Nucleotide Sequence Polymorphism and Diversity

Sequencing of the 4 loci resulted in 5960 bp of aligned sequence data, including 3737 bp of coding regions (62.7% of total sequence) and 2223 bp of noncoding regions. There are 13 indels within sequenced introns. The length of coding and noncoding portions of each locus is shown in Table 3. Observed heterozygosity is high at the 4 loci. Thus, for *Cellulase*, *CHS*, *F3H*, and *STK*, observed heterozygosity was 40%, 70.6%, 50%, and 50%, respectively.

A total of 176 SNPs were detected over all 4 loci, resulting in an average of 1 SNP every 33.7 bp (excluding indel polymorphisms). Nucleotide diversity at each locus is shown in Table 3. The *F3H* locus has the highest diversity, where $\theta_W = 0.0109$ and $\pi = 0.0123$, almost twice as high as observed at the other 3 loci. Nonsynonymous diversity (π_{nonsyn}) is generally much lower than silent variation, the ratio of nonsynonymous to synonymous diversities (d_N/d_S) range from 0.194 for *F3H* to 0.441 for *STK*, indicating that there is some degree of purifying selection at these loci. The high level of diversity in *F3H* results primarily from high diversity within the intron, where $\pi = 0.018$; purifying selection at the protein level is moderately strong as judged by d_N/d_S ratios. If we consider these 4 loci together, the overall genetic diversity of wild avocado is at the same level as *A. thaliana*, (Nordborg et al. 2005), wild rice (Zhu et al.

Table 3. Descriptive statistics and estimates of nucleotide diversity in 4 nuclear loci of wild avocado

Gene	Sample size	Aligned length (bp)	Coding region (bp)	Noncoding region (bp)	Segregating sites	Number of haplotypes	Haplotype diversity	$\theta_W \times 10^{-3}$	$\pi \times 10^{-3}$	π_{sil}[a] $\times 10^{-3}$	$\pi_{nonsyn} \times 10^{-3}$	π_{nonsyn}/π_{syn}[b]	Tajima's *D*	R_m[c]
F3H	16	1812	819	993	76	16	0.897	10.9	12.3	17.9	1.19	0.194	0.503	5
Cellulase	20	1540	788	752	30	12	0.829	4.60	3.46	4.95	1.23	0.277	-0.850	1
CHS	16	1210	768	442	27	15	0.907	5.66	4.36	7.61	0.719	0.112	-0.812	2
STK	20	1398	1362	36	43	16	0.901	7.23	6.18	10.3	4.97	0.441	-0.514	3
Average	18	1490	934.25	555.75	44	14.75	0.884	7.09	6.58	10.2	2.02	0.256	—	—

[a] π_{sil}: nucleotide diversity estimate at silent sites (noncoding and synonymous sites).

[b] π_{nonsyn} and π_{syn}: nucleotide diversity estimate at nonsynonymous sites and synonymous sites.

[c] R_m: minimum recombination number.

2007), and wild barley (Morrell et al. 2003), all of which are self-fertilizing annual species of temperate origin.

The proportion of singletons (polymorphic sites present only once in the sample) is high at all 4 loci. For example, there are 13 singletons at the *Cellulase* locus (13 of 30 SNPs). Table 3 summarizes the measurements of the site frequency spectrum and Tajima's *D*; none of the values are significantly different from zero, indicating that none of the loci are significantly different from neutral expectations. Tajima's *D* is negative in *Cellulase*, *CHS*, and *STK*, although not significantly different from zero; the negative values indicate a slight excess of rare mutations at these loci.

Estimation of Recombination and LD

Hudson and Thornton's estimates of the recombination parameter $\rho(\hat{\rho}_{H01}, \hat{\rho}_{T05})$, Thornton's mutation estimator, $\hat{\theta}_{T05}$, and the ratio of $\hat{\rho}$ over (both $\hat{\rho}_{H01}/\hat{\theta}_W$ and $\hat{\rho}_{T05}/\hat{\theta}_{T05}$) and Wall's *B* are shown in Table 4. Figure 1 depicts point estimates and the 95% confidence interval of Thornton's $\hat{\theta}_{T05}$, $\hat{\rho}_{T05}$, and $\hat{\rho}_{T05}/\hat{\theta}_{T05}$ for all 4 loci estimated individually.

For both $\hat{\rho}_{H01}$ and $\hat{\rho}_{T05}$, recombination rate estimates are lower than estimates of mutation rate θ_W and $\hat{\theta}_{T05}$. Although estimates of recombination rate from $\hat{\rho}_{T05}$ are larger than $\hat{\rho}_{H01}$, 3 of the 4 loci show a greater relative contribution of mutation than recombination to haplotype diversity. Point estimates of $\hat{\rho}_{T05}/\hat{\theta}_{T05}$ average 0.787 for the 4 loci, with an upper 95% confidence interval ranging from 1.2 to 5.4 (Figure 1). Thus the upper bound of 95% confidence on the estimate of $\hat{\rho}_{T05}/\hat{\theta}_{T05}$ is much lower than point estimates for *Drosophila melanogaster* that are as high as 17.5 (Pool and Aquadro 2006).

A plot of intralocus LD for all loci is shown in Figure 2. The average LD, measured as r^2, is summarized by a second-order loess line (Cleveland 1979, 1981) and decays to half the initial value within a distance of 1.0 kb. Wall's *B* values (0.1–0.2) also show that LD within the 4 loci is relatively low.

Despite the rapid decay of LD within loci, there is a significant excess of interlocus LD for all six 2-locus comparisons in the full sample of 21 accessions (42 chromosomes) from all 3 botanical races. The proportion of data sets with randomized haplotype configurations with median r^2 as great as that observed in the empirical data ranges from 0 to 0.003 and thus would all be significant at $P < 0.05$ after Bonferroni correction for the 6 tests performed. However, when the analyses are repeated on data from within each botanical race, no significant interlocus LD was detected. For 12 samples from Central Mexico (the Mexican Botanical Race), *P* values based on median r^2 values ranged from 0.072 to 0.916, with a median of 0.518. The number of SNPs segregating in accessions of the Guatemalan (Southern Mexican) race (from 6 chromosomes) resulted in only 3 comparisons. For all 3 comparisons, there was no significant interlocus LD. Among West Indian accessions (12 chromosomes), none of the 6 comparisons were significant; *P* values = 0.173 – 1.

Table 4. Hudson and Thornton's estimates of population recombination rate ($\hat{\rho}$), the ratio of $\hat{\rho}/\hat{\theta}$ and Wall's *B* statistics

Loci	$\hat{\rho}_{H01} \times 10^{-3}$	$\hat{\rho}_{T05} \times 10^{-3}$	$\hat{\rho}_{H01}/\hat{\theta}_W$	$\hat{\rho}_{T05}/\hat{\theta}_{T05}$	Wall's *B*
F3H	1.54	4.76	0.18	0.493	0.151
Cellulase	0.865	1.87	0.19	0.494	0.103
CHS	4.61	6.27	0.81	1.314	0.200
STK	2.32	4.81	0.33	0.845	0.119

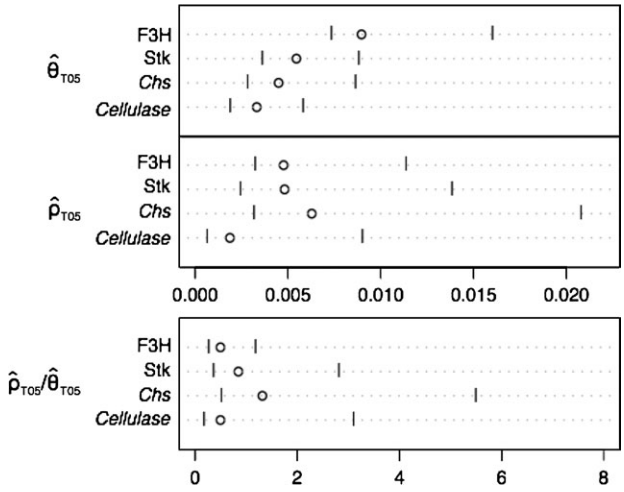

Figure 1. $\hat{\rho}_{T05}$ and $\hat{\theta}_{T05}$ estimates and $\hat{\rho}_{T05}/\hat{\theta}_{T05}$ ratio of 4 loci in wild avocados. The bars show the 95% confidential scope.

Discussion

Nucleotide Diversity

This is the first study of nucleotide diversity from a predominantly outcrossing tropical tree species. The 21 wild accessions used in this study represent all 3 botanical

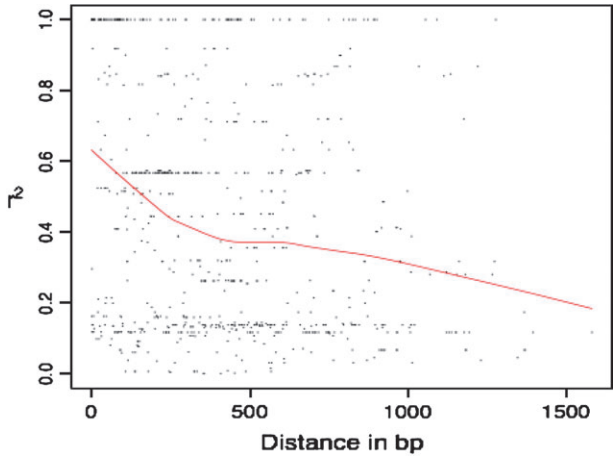

Figure 2. Plots of squared allele frequency correlations (r^2) by physical distance between sites in wild avocados. An SNP frequency filter of 10% was used. The red line in each plot depicts the lowess smoothed line that summarizes the observed data in LD.

races of avocado. A recent study by Chen et al. (Chen HF, Morrell PL, Ashworth VETM, de la Cruz M and Clegg MT, unpublished data) confirmed that 12 accessions are of Mexican origin, 3 are Southern Mexican/Guatemalan origin, and 6 that are of West Indian origin (Table 1).

The observed heterozygosity in the sample is 40–70.6%, consistent with the 63.5% heterozygosity reported in a study of microsatellite diversity in avocado (Ashworth and Clegg 2003), thus providing direct evidence of the outcrossing mating system of wild avocado. Avocado is not especially diverse for an outcrossing species with diversity θ_W = 0.0071, π = 0.0066, and π_{sil} = 0.0102 at silent sites compared with the self-fertilizing annual plant species *A. thaliana* (π_{sil} = 0.005~0.0175) (Aguade 2001), and it is somewhat less than wild barley (*H. vulgare* ssp. *spontaneum*) (π_{sil} = 0.012) (Morrell et al. 2005). Diversity in avocado is substantially less than that found in the outcrossing species, *Arabidopsis lyrata* and *Arabidopsis halleri* (π_{sil} = 0.023 and 0.015) (Wright et al. 2003; Ramos-Onsins et al. 2004), and less than reported in the outcrossing tree *Populus tremula* (π_{sil} = 0.016) (Ingvarsson 2005) and in the wild relatives of maize: *Zea diploperennis*, *Zea perennis*, *and Zea parviglumis* (π_{sil} = 0.012, 0.013, and 0.023, respectively) (White and Doebley 1999; Tiffin and Gaut 2001a; Tiffin et al. 2004).

Linkage Disequilibrium

A relatively large number of diploid samples have been fully experimentally phased in the present data, permitting the estimation of recombination and LD as a function of distance within genes. Under coalescent theory and assuming a standard neutral model, the relative importance of recombination versus mutation in generating novel haplotypes can be quantified as $\rho/\theta(=4N_e r/4N_e\mu = r/\mu)$ where μ is the mutation rate and r is the recombination rate. Despite an outcrossing mating system, the ratio of $\hat{\rho}/\hat{\theta}$ for avocado is not high, ranging from 0.18 to 0.81 ($\hat{\rho}_{H01}/\hat{\theta}_{w}$) and 0.49 to 1.3 ($\hat{\rho}_{T05}/\hat{\theta}_{T05}$). In contrast, species-wide samples from *D. melanogaster* yield estimates of $\hat{\rho}/\hat{\theta}=0.4-6.8$ (Haddrill et al. 2005) or even up to 17.5 (Pool and Aquadro 2006), whereas maize yields $\hat{\rho}/\hat{\theta}=1.5$ (Wright et al. 2005). For the self-fertilizing species *A. thaliana*, $\hat{\rho}/\hat{\theta}=0.05$ (Nordborg et al. 2005), but $\hat{\rho}/\hat{\theta}$ is nearly 1.5 in wild barley (Morrell et al. 2006). Thus, whereas high rates of heterozygosity provide the potential for effective recombination in wild avocado, the estimated relative contribution of recombination to haplotype diversity is lower than in many other species.

The extent of LD is influenced by effective population size, population mating patterns, and admixture. Generally speaking, LD decays more rapidly in outcrossing than in selfing species (Nordborg 2000) because recombination is less effective in selfing species, where individuals are more likely to be homozygous. In the current avocado data, intralocus LD decays rapidly within a distance of 1 kb. Compared with other species, avocado intralocus LD decays slower than LD observed in maize, an outcrossing species ($r^2 < 0.1$ within 1500 bp) (Remington et al. 2001), and LD in self-fertilizing wild barley, where LD decays very quickly within 300 bp (Morrell et al. 2005). However, LD is much more extensive in self-fertilizing *A. thaliana*, where it extends about 10 kb (Kim et al. 2007).

Significant interlocus LD was detected in the total sample of wild avocado but was absent for all 2-locus comparisons when the data were partitioned by botanical race. Significant interlocus LD can result from physical linkage among loci, a very recent species-wide reduction in effective population size, geographic structure, or selection. The absence of significant interlocus LD between loci within each botanical race suggests that physical linkage is unlikely to explain the high degree of interlocus LD in the full wild avocado data set. The genetic map position for these loci is unknown; however, it is unlikely that all 4 loci are in close physical proximity on the 12 avocado chromosomes. In the absence of strong evidence for selection at any of the loci or a demographic crash in the species history, geographic structure best explains the excess interlocus LD observed in the pooled data. In a much larger number of 2-locus comparisons in wild barley, loci with significant geographic structure were found to have higher levels of interlocus LD (Morrell et al. 2005). That is, LD is quite sensitive to the presence of geographic structure at one or more of the loci in a 2-locus comparison.

An important caveat is that sample size and SNP number are small for the interlocus LD estimates calculated within avocado races. Limited sample size reduces the power to test for significant interlocus LD because sample size limits the number of randomizations of haplotypes that are possible. However, the number of possible orders into which n unique chromosomes (or haplotypes) can be sorted is n factorial. Thus, with just 6 haplotypes, there are 720 unique orders. With 12 chromosomes (the number in the 2 subsamples considered here), there are 4.8×10^8 possible orders. However, the actual number of unique combinations (or orders) is determined by the number of observed haplotypes. A limited number of SNPs at any locus also limits the number of comparisons of SNPs between loci for any comparison in which the locus is involved (i.e., a small number of SNPs at one locus impacts multiple interlocus tests). In both the Guatemalan and West Indian races, the *CHS* locus has only 3 nonsingleton SNPs. Despite these considerations, interlocus LD in the full sample of wild avocado almost certainly results from geographic structure and the admixture of samples from across botanical races.

These data have important implications for association mapping or LD mapping (Long and Langley 1999), a particularly appealing technique for gene or Quantitative Trait Loci (QTL) localization in perennial crops where traditional QTL and other breeding approaches are difficult, owing to long generation time (Ross-Ibarra et al. 2007). To be effective, association mapping in avocado, whether using domesticates or wild samples, will have to control for the interlocus LD resulting from population structure. This will require the characterization of structure within samples before undertaking an association mapping approach.

This is the first resequencing study of a subtropical tree species, and it provides a useful picture of nucleotide

sequence diversity in an important element of the flora. The picture is one of moderate diversity where recombination is less important than mutation as a source of haplotypic diversity, in contrast to wild barley and maize, for example. Despite this, LD decays relatively rapidly within the span of a typical gene, and there is no evidence for selection at the 4 loci investigated. These data show a rapid loss of LD within genes, but admixture among races is a confounding factor that induces substantial interlocus LD. This will complicate the use of association mapping in avocado.

Funding

California Avocado Commission.

Acknowledgments

We thank M. Durbin, K. Lundy, P. Robinson, and S. Zarate for experimental and technical assistance. Sequence data from this article have been deposited with the GenBank data libraries under accession numbers EU335454–EU335564.

References

Aguade M. 2001. Nucleotide sequence variation at two genes of the phenylpropanoid pathway, the FAH1 and F3H genes, in *Arabidopsis thaliana*. Mol Biol Evol. 18:1–9.

Andolfatto P, Przeworski M. 2000. A genome-wide departure from the standard neutral model in natural populations of *Drosophila*. Genetics. 156:257–268.

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 Yearbook. 70:135–145.

Britsch L, Ruhnau-Brich B, Forkmann G. 1992. Molecular cloning, sequence analysis, and in vitro expression of flavanone 3 beta-hydroxylase from *Petunia hybrida*. J Biol Chem. 267:5380–5387.

Chanderbali AS, Albert VA, Ashworth VETM, Clegg MT, Litz RE, Soltis DE, Soltis PS. 2008. *Persea americana* (avocado): bringing ancient flowers to fruit in the genomic era. Bioessays. 29:1–11.

Cleveland WS. 1979. Robust locally weighted regression and smoothing scatterplots. J Am Stat Assoc. 74:829–836.

Cleveland WS. 1981. LOWESS: a program for smoothing scatterplots by robust locally weighted regression. Am Stat. 35:54.

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

Davenport TL, Parnitzki P, Fricke S, Hughes MS. 1994. Evidence and significance of self-pollination of avocados in Florida. J Am Soc Hortic Sci. 119:1200–1207.

Di Gaspero G, Cipriani G. 2003. Nucleotide binding site/leucine-rich repeats, Pto-like and receptor-like kinases related to disease resistance in grapevine. Mol Genet Genomics. 269:612–623.

Durbin ML, McCaig B, Clegg MT. 2000. Molecular evolution of the chalcone synthase multigene family in the morning glory genome. Plant Mol Biol. 42:79–92.

Dvornyk V, Sirvio A, Mikkonen M, Savolainen O. 2002. Low nucleotide diversity at the pal1 locus in the widely distributed *Pinus sylvestris*. Mol Biol Evol. 19:179–188.

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

Forkmann G, Dangelmayr B. 1980. Genetic control of chalcone isomerase activity in flowers of *Dianthus caryophyllus*. Biochem Genet. 18:519–527.

Garcia-Gil MR, Mikkonen M, Savolainen O. 2003. Nucleotide diversity at two phytochrome loci along a latitudinal cline in *Pinus sylvestris*. Mol Ecol. 12:1195–1206.

Gaut BS, Le Thierry d'Ennequin M, Peek AS, Sawkins MC. 2000. Maize as a model for the evolution of plant nuclear genomes. Proc Natl Acad Sci USA. 97:7008–7015.

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

Haddrill PR, Thornton KR, Charlesworth B, Andolfatto P. 2005. Multilocus patterns of nucleotide variability and the demographic and selection history of *Drosophila melanogaster* populations. Genome Res. 15:790–799.

Hamrick J, Godt M. 1997. Allozyme diversity in cultivated crops. (Germplasm regeneration: Developments in population genetics and their implications). Crop Sci. 37:26–31.

Hudson RR. 1987. Estimating the recombination parameter of a finite population model without selection. Genet Res. 50:245–250.

Hudson RR. 2001. Two-locus sampling distributions and their application. Genetics. 159:1805–1817.

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.

Ingvarsson PK. 2005. Nucleotide polymorphism and linkage disequilibrium within and among natural populations of European aspen (*Populus tremula* L., Salicaceae). Genetics. 169:945–953.

Kado T, Yoshimaru H, Tsumura Y, Tachida H. 2003. DNA variation in a conifer, *Cryptomeria japonica* (Cupressaceae sensu lato). Genetics. 164:1547–1559.

Kim S, Plagnol V, Hu TT, Toomajian C, Clark RM, Ossowski S, Ecker JR, Weigel D, Nordborg M. 2007. Recombination and linkage disequilibrium in *Arabidopsis thaliana*. Nat Genet. 39:1151–1155.

Kobayashi M, Lin JZ, Davis J, Francis L, Clegg MT. 2000. Quantitative analysis of avocado outcrossing and yield in California using RAPD markers. Sci Hortic. 86:135–149.

Lin JZ, Morrell PL, Clegg MT. 2002. The influence of linkage and inbreeding on patterns of nucleotide sequence diversity at duplicate alcohol dehydrogenase loci in wild barley (*Hordeum vulgare ssp. spontaneum*). Genetics. 162:2007–2015.

Long AD, Langley CH. 1999. The power of association studies to detect the contribution of candidate genetic loci to variation in complex traits. Genome Res. 9:720–731.

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.

Morrell PL, Toleno DM, Lundy KE, Clegg MT. 2005. Low levels of linkage disequilibrium in wild barley (*Hordeum vulgare ssp. spontaneum*) despite high rates of self-fertilization. Proc Natl Acad Sci USA. 102:2442–2447.

Morrell PL, Toleno DM, Lundy KE, Clegg MT. 2006. Estimating the contribution of mutation, recombination and gene conversion in the genetation of haplotypic diversity. Genetics. 173:1705–1723.

Newton C, Graham A, Heptinstall L. 1989. Analysis of any point mutation in DNA. The amplification refractory mutation system (ARMS), amplification refractory mutation system (ARMS). Nucleic Acids Res. 17:2503–2516.

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.

Nordborg M. 2000. Linkage disequilibrium, gene trees and selfing: an ancestral recombination graph with partial self-fertilization. Genetics. 154:923–929.

Nordborg M, Hu TT, Ishino Y, Jhaveri J, Toomajian C, Zheng H, Bakker E, Calabrese P, Gladstone J, Goyal R, et al. 2005. The pattern of polymorphism in *Arabidopsis thaliana*. PLoS Biol. 3:e196.

Pool JE, Aquadro CF. 2006. History and structure of sub-Saharan populations of *Drosophila melanogaster*. Genetics. 174:915–929.

Ramos-Onsins SE, Stranger BE, Mitchell-Olds T, Aguade M. 2004. Multilocus analysis of variation and speciation in the closely related species *Arabidopsis halleri* and *A. lyrata*. Genetics. 166:373–388.

Remington DL, Thornsberry JM, Matsuoka Y, Wilson LM, Whitt SR, Doebley J, Kresovich S, Goodman MM, Buckler ES. 2001. Structure of linkage disequilibrium and phenotypic associations in the maize genome. Proc Natl Acad Sci USA. 98:11479–11484.

Ross-Ibarra J, Morrell PL, Gaut BS. 2007. Plant domestication, a unique opportunity to identify the genetic basis of adaptation. Proc Natl Acad Sci USA. 104(Suppl 1):8641–8648.

Sharma CB, Martinez GC. 1972. The component triacylglycerols of avocado fruit-coat. J Am Oil Chem Soc. 49:229–232.

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. 9:2325–2327.

Thornton K, Andolfatto P. 2006. Approximate Bayesian inference reveals evidence for a recent, severe bottleneck in a Netherlands population of *Drosophila melanogaster*. Genetics. 172:1607–1619.

Tiffin P, Gaut BS. 2001a. Sequence diversity in the tetraploid *Zea perennis* and the closely related diploid *Z. diploperennis*: insights from four nuclear loci. Genetics. 158:401–412.

Tiffin P, Gaut BS. 2001b. Molecular evolution of the wound-induced serine protease inhibitor wip1 in *Zea* and related genera. Mol Biol Evol. 18:2092–2101.

Tiffin P, Hacker R, Gaut BS. 2004. Population genetic evidence for rapid changes in intraspecific diversity and allelic cycling of a specialist defense gene in *Zea*. Genetics. 168:425–434.

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.

Tucker ML, Durbin ML, Clegg MT, Lewis LN. 1987. Avocado cellulase: nucleotide sequence of a putative full-length cDNA clone and evidence for a small gene family. Plant Mol Biol. 9:197–203.

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.

Weir BS. 1990. Genetic Data Analysis. Sunderland (MA): Sinauer.

White SE, Doebley JF. 1999. The molecular evolution of terminal ear1, a regulatory gene in the genus *Zea*. Genetics. 153:1455–1462.

Wright SI, Lauga B, Charlesworth D. 2003. Subdivision and haplotype structure in natural populations of *Arabidopsis lyrata*. Mol Ecol. 12:1247–1263.

Wright SI, Bi IV, Schroeder SG, Yamasaki M, Doebley JF, McMullen MD, Gaut BS. 2005. The effects of artificial selection on the maize genome. Science. 308:1310–1314.

Zhu Q, Zheng X, Luo J, Gaut BS, Ge S. 2007. Multilocus analysis of nucleotide variation of *Oryza sativa* and its wild relatives: severe bottleneck during domestication of rice. Mol Biol Evol. 24:875–888.

Received June 28, 2007
Accepted January 25, 2008

Corresponding Editor: James Hamrick