Available online at www.sciencedirect.com
SCIENCE DIRECT®
Scientia Horticulturae 101 (2004) 255–267
SCIENTIA HORTICULTURAE
www.elsevier.com/locate/scihorti

Microsatellite markers in avocado (*Persea americana* Mill.): development of dinucleotide and trinucleotide markers

V.E.T.M. Ashworth [a,*], M.C. Kobayashi [b], M. De La Cruz [a], M.T. Clegg [a]

[a] *Department of Botany and Plant Sciences, University of California, Riverside, CA 92521, USA*
[b] *Department of Plant and Microbial Biology, University of California, Berkeley, CA 94720, USA*

Accepted 27 November 2003

Abstract

This paper compares the development of microsatellite markers from two genomic DNA libraries of avocado cultivar Hass enriched for dinucleotide and trinucleotide repeats. Sequencing showed 86 and 31% of clones from the respective libraries to contain microsatellites. However, banding profiles of trinucleotide loci were easier to interpret than those of dinucleotide loci. Of 376 dinucleotide-containing clones, 81% harbored AG repeats and 19% contained AC repeats. A subset of 104 trinucleotide-containing clones consisted of ATG repeats (44%), AGT repeats (30%), and AAG repeats (16%). Array lengths of up to 34 were attained in the dinucleotide repeat microsatellites, whereas trinucleotide arrays never exceeded 11 elements. Typing 37 genotypes at 25 loci (11 dinucleotide and 14 trinucleotide loci) revealed a total of 204 alleles, of which 60% originated from the dinucleotide loci. Average heterozygosity of the di- and trinucleotide loci was 73.4 and 52.6%, respectively. Many loci, especially dinucleotide loci, exhibited allele size differences that were not multiples of the repeat unit, suggesting nonstepwise mutational mechanisms. Several loci were additionally characterized by large gaps in the allele size distribution.

Keywords: Lauraceae; Microsatellite; Mutation; *Persea americana*; SSRs

1. Introduction

Microsatellites (simple sequence repeats, SSRs) are a form of repetitive DNA first discovered in the early 80s (Hamada et al., 1982). Their great potential as powerful genetic

* Corresponding author. Present address: 12 Glenview, Galway Road, Roscommon Town, Ireland. Tel.: +1-909-787-3543; fax: +1-909-787-4437.
E-mail address: vanessaa@citrus.ucr.edu (V.E.T.M. Ashworth).

0304-4238/$ – see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.scienta.2003.11.008

markers, combining the useful properties of high variability, co-dominant inheritance, and good reproducibility, was recognized simultaneously by several groups of researchers (Litt and Luty, 1989; Smeets et al., 1989; Tautz, 1989; Weber and May, 1989). Over the past decade microsatellites have assumed a central role in population-level studies as well as in breeding applications because their co-dominance makes them suitable for tracing paternity and tracking pollen movement (see for reviews Queller et al., 1993; Jarne and Lagoda, 1996; Goldstein and Pollock, 1997; Goldstein and Schlötterer, 1999; Sunnucks, 2000). Additionally, they are abundant and evenly distributed across the genome (Hamada et al., 1982; Stallings et al., 1991; Weissenbach et al., 1992), making them amenable to the study of genealogical relationships (Goldstein et al., 1995).

The repetitive unit of microsatellites is generally defined as being one to five bases long. Dinucleotide repeats are the most common category of repeat in a majority of organisms (Jurka and Pethiyagoda, 1995; Tóth et al., 2000; Katti et al., 2001) and are usually associated with non-coding regions of the genome (Young et al., 2000; Temnych et al., 2001). Trinucleotide repeats are often found within ORFs (Young et al., 2000) due to their triplet structure. Although they do not interfere with the reading frame they are nonetheless responsible for several human diseases, once repeated arrays exceed a threshold limit. In plants, trinucleotide microsatellites are relatively infrequent (Lagercrantz et al., 1993; Ma et al., 1996), compared with vertebrates and some other organisms.

Our goal is to develop microsatellite markers in order to expedite avocado breeding and selection that is currently hampered by a long generation time and the inability to perform controlled pollinations. A (diploid) member of the family Lauraceae, avocado belongs to one of the earliest lineages in the angiosperm phylogeny that predates the origin of the Eudicots (Zanis et al., 2002; http://www.mobot.org/MOBOT/Research/Apweb/html). Few DNA sequences of Lauraceae are deposited in the public sequence databases, precluding their use in the development of microsatellite loci. In this paper we compare the efficiency of marker development from a dinucleotide- and trinucleotide-enriched library. Respective marker types serve different—albeit overlapping—functions, viz. mapping using segregating progeny versus diversity studies over a range of germplasms. Allelic diversity at 25 marker loci is compared for a panel of 37 avocado genotypes, including 35 cultivars and two wild relatives (*P. schiedeana* and *P. steyermarkii*). Data on genealogical relationships are presented elsewhere (Ashworth and Clegg, 2003).

2. Materials and methods

2.1. Library construction

Due to the paucity of avocado DNA sequences in the DNA sequence databases we commissioned two microsatellite-enriched libraries from GIS (Genetic Identification Services Inc., Chatsworth, CA). The libraries were generated from total genomic DNA of avocado cultivar Hass digested with restriction enzymes *Hae*III, *Pvu*II, *Sca*I, *Stu*I, *Bsr*bI, *Rsa*I and *Eco*RV. A preliminary screen (data not shown) of the genomic DNA using probes to three dinucleotide and three trinucleotide repeat motifs showed GA, AAG and ATG to be the most abundant. The dinucleotide-enriched library was prepared against the motif GA. The

trinucleotide-enriched library was prepared against the motif ATG. Genomic fragments enriched for microsatellites were cloned into the *Hin*dIII cut site of plasmid pUC19 that was incorporated into *Escherichia coli* strain DH5α by electroporation.

2.2. Colony screening

Cells were spread on agar plates containing 100 μl of 2% BluoGal or X-Gal, 10 μl of 100 mM IPTG, and ampicillin (50 mg/ml) to eliminate vectors lacking an insert (blue). Additionally, inserts were screened for the correct size (ca. 350–1000 bp) by PCR using the M13 universal primers located in the plasmid flanks. Cells from each positive (white) colony were transferred to 0.2 ml microfuge tubes containing 10 μl aliquots of 1 μl 10× buffer (Qiagen, Valencia, CA), 3 mM dNTPs, 0.15 μM forward and reverse primer, and 0.05 units Taq DNA polymerase. PCR settings were 1 min at 94 °C, 35 cycles of 1 min at 94 °C, 1 min at 57 °C, 1 min at 72 °C, and a final extension of 2 min at 72 °C. PCR products were run out on a 2% agarose gel with a 100 bp DNA ladder (Promega, Madison, WI) as the size standard. Clones devoid of an insert, clones containing inserts that were too small or too long, or clones that harbored two or more inserts were excluded from further analysis.

2.3. Sequencing of inserts

For sequencing, suitable clones were cultured overnight in LB medium + amp and purified (Qiagen Spin Miniprep Kit; Qiagen, Valencia, CA). The sequencing reaction consisted of 1 μl fluorescently labeled IRDye primer (1 μM; LI-COR, Lincoln, NE), 7.2 μl 3.5× DNA sequencing buffer, 2 μl of each dNTP mix (G mix: 0.03 mM ddGTP, 135 μM each of dATP, dCTP, dTTP, and 7-deaza-dGTP; A mix: 0.2625 mM ddATP, 67.5 μM each of dATP, dCTP, dTTP, and 7-deaza-dGTP; T mix: 0.45 mM ddTTP, 67.5 μM each of dATP, dCTP, dTTP, and 7-deaza-dGTP; C mix: 0.15 mM ddCTP, 67.5 μM each of dATP, dCTP, dTTP, and 7-deaza-dGTP), and 1 μl (5 units/μl) Taq DNA Polymerase (buffer, dNTP mix and Polymerase supplied in the SequiTherm Excel™ II DNA Sequencing Kit-LC for 66 cm gels, Epicenter Technologies, Madison, WI), and 5–7 μl of purified DNA. Thermocycler conditions consisted of 5 min at 95 °C, and 30 cycles of 1 min at 95 °C, 1 min at 55 °C, and 1 min at 70 °C. All sequences were generated on a LI-COR DNA 4200 Long Read Sequencer (66 cm plates) on 6% acrylamide gel (Long Ranger, FMC Bioproducts, Rockland, MN; 1% ammonium persulfate, TEMED). In order to maximize efficiency, 12 reactions were generated using only the forward primer (labeled with IRD 800) and 12 using only the reverse primer (labeled with IRD 700), subjected to PCR and then pooled prior to denaturing with formamide + Basic Fuchsin. This allowed 24 distinct cloned fragments to be sequenced in each sequencing run.

2.4. Primer design

Sequences were edited in Sequencher version 3.1 (Gene Codes Inc., Ann Arbor, MI) and stored in a sequence library in BBEdit Lite 4.6 (Bare Bones Software Inc., Bedford, MA). Sequences were screened for multiple representation using FASTA version 3.2t06 (Pearson and Lipman, 1988). Primers were designed to unique sequences having sufficiently long

flanks either side of the microsatellite locus and at least four perfect trinucleotide repeats or at least seven perfect dinucleotide repeats. All primer pairs were selected using Oligo Primer Analysis Software version 4.01 (National Biosciences Inc., Plymouth, MN). Base composition, annealing temperature, and product size for the 25 microsatellite loci used in this study are presented in Table 1.

2.5. Fragment amplification

PCR reactions were visualized by either radioactivity or fluorescence. In the radioactive method reaction mixtures (10 μl total volume) contained 1 μl 10× buffer (containing 15 mM $MgCl_2$; Qiagen Taq DNA Polymerase kit, Valencia, CA), 0.19 μM dATP, dGTP and dTTP, 0.067 μM dCTP, and 0.017 μM α-^{32}P-dCTP (NEN EasyTides, Perkin Elmer Life Sciences), 0.8 μl $MgCl_2$ (25 mM), 1 μl each primer (1.5 pmol/μl), 0.5 μl Taq polymerase (5 units/μl), and 1–2 μl DNA (10–20 ng/μl). In the fluorescent method 10 μl reactions consisted of 1 μl 10× buffer (containing 15 mM $MgCl_2$; Qiagen Taq DNA Polymerase kit, Valencia, CA), 1.15 μl dNTPs (1.25 mM), 0.5 μl $MgCl_2$ (25 mM), 1.15 μl labeled fluorescent forward primer (1.5 pmol/μl; ABI fluorescent dyes 6-FAM, VIC, NED), 1.15 μl unlabelled reverse primer (1.5 pmol/μl), 0.08 μl Taq Polymerase (5 units/μl), and 1–2 μl DNA (10–20 ng/μl). All reactions were performed on a Stratagene Robocycler (Stratagene Inc., La Jolla, CA). The following settings were used: 2 min at 95 °C, 30 cycles of 95 °C for 1 min, 50–68 °C (depending on the primer) for 1 min, and 72 °C for 1 min, with a final extension of 45 min at 72 °C.

2.6. Fragment visualization

Radioactively labeled reaction products were denatured with 3 μl of a solution containing 0.03% Bromophenol Blue and 20 mM EDTA in formamide, and run out on an acrylamide gel for 1.5–2.5 h. After electrophoresis, the gel plates were separated and the gel allowed to dry. Bands were visualized by exposure of X-Omat AR film (Eastman Kodak, Rochester, NY) to the gel for 2–24 h and scored manually. Fluorescently labeled reaction products were combined with ROX 500 size standard and denatured with ABI Loading Buffer (50 mg/ml Blue Dextran, 25 mM EDTA, and de-ionized formamide) and electrophoresed on an ABI 377 Automated DNA Sequencer (Perkin Elmer Applied Biosystems). Fragment sizes were calculated using ABI GeneScan software, and adjusted using the known product size in cultivar Hass. Distances between called fragment sizes were compared with actual band distances appearing on the electrophoretogram, as recommended by Haberl and Tautz (1999).

3. Results

3.1. Microsatellite marker yields

Table 2 contrasts marker yields from the dinucleotide- and trinucleotide-enriched DNA libraries at successive stages of marker development. The main difference between the two marker development strategies is the number of clones containing microsatellites. We

Table 1
Primer and PCR product specifications for 25 microsatellite loci of avocado

Primer name	Base composition	T_m (°C)	Repeat type	Fragment size
AVT.005b.F	TTAGCAGCAGATAGAGGGAGAG	62	CAT5	186
AVT.005b.R	GGACCTGCCTTGTGGATTAG			
AVT.020GAT.F	CTACATAGATCGAAATAAGG	54	GAT9	164
AVT.020GAT.R	ATCTGGCTATGAAATGTTGG			
AVT.021F	ACTCTCGCCTCTGCGTTGAT	65	ATC8	136
AVT.021R	GACTCAACATGGTTAGAACAAGGC			
AVT.038F	GATTAAAGATGACCCTGAAG	56	TCA8	190
AVT.038R	GATTTGGCTCAAGATAGATC			
AVT.106F	CCAATCAAAAGGCAAACGAAGAAC	66	TCA6	309
AVT.106R	GCAAAGGAGGCGGTTTCGAGAT			
AVT.143F	CCCAACATCTACTTAGCGCAATAG	66	GAA8GAT6	211
AVT.143R	ATCATCATCGTCTTCACCGTCGTT			
AVT.158F	ACGAAGTTACGGGCTTATTTCACA	62	GAT7	267
AVT.158R	TTCTTCCCCTTCTCTCACATAATC			
AVT.191F	TCCACAACTTCTACAGGGTCGT	68	ATG7TGG4	170
AVT.191R	GGAAGATAACGCACCTTGAGTTC			
AVT.226F	GGCTGACTTTTATAGTCGATGT	62	TCA6.CTT4	298
AVT.226R	TCCGATTGACAGTGGATTGTT			
AVT.372F	GCGAACACTACTCACATAGG	58	TGA10	182
AVT.372R	ATTTTAACTAATGGATTGGATG			
AVT.386F	ACAACCCAAACATAAATGCT	60	TGA8	229
AVT.386R	AATAGAAGTGACATCCGACC			
AVT.436F	ACTAAAATGAGGGGAGACTAG	56	ATC9	152
AVT.436R	GAGTGTAGTGAGGAGTTTGG			
AVT.448F	ACGGTGTTTGGAAGAAGATG	60	GAT8	192
AVT.448R	GCACTTCAATCAATGCTTAC			
AVT.517F	AATCCTTCCACTCAGAAACT	59	GAT6	229
AVT.517R	TACACAAACGACAAGAATGG			
AVO.102F	TTCGCCTTATCAGCGTTAG	58	GA12	153
AVO.102R	TCTTGGAAAGCCCTACTCC			
AVD.001F	GTTTCCAAGCGACTCACGAG	66	CT12	226
AVD.001R	GATTCCATGCTGAATTGCCG			
AVD.002F	TGCATACTCTTAGCCCCATATGT	66	CT15CA13	327
AVD.002R	GGATCATTGGTTTTGATAACAGG			
AVD.003II.F	TCCCTTCAGTCTAAGATTAGCC	63	TC19	192
AVD.003R	GACCAACACTATTTGCCCCAC			
AVD.006F	GGGAGAGATGTATTGAGCA	58	TC9AC19	314
AVD.006R	ACTTGGTCGTAGATTGTAAAT			
AVD.013F	TTGCCAGTGGAACTTCAAAA	63	AG7.GA3.TCT4	216
AVD.013R	ACCCAACCAAAGATTTCAAT			
AVD.015F	GACCCCTACCCTAACTCTCA	61	GT26	258
AVD.015R	CTTCTAAACATTCCCTACAAAG			
AVD.017F	GCTCACAAGCGAACTTTCAT	64	TC18.AC8	211
AVD.017R	TAAATCCCCTTTCCCACCTT			
AVD.018F	TGCTGGCATAATGGCTGCTA	67	GA20	224
AVD.018R	CAAACATCTTCAGAACCGCC			
AVD.022F	CCACTTGGATTCTTGTTGGA	65	TC13	228
AVD.022R	ATTTGGGTTCGGCTTAGGAA			
AUCR.418F	AGATGGCTTTCTCCTTCTGA	56	GT12GA13	379
AUCR.418R	TTTGACACACAATCCAACTATG			

The repeat type and size of the amplification product are based on sequenced clones from a DNA library of cultivar Hass. In column 4 (repeat type) a period between motifs denotes an interrupted microsatellite locus.

Table 2
Comparison of microsatellite development from a dinucleotide (2*n*) and trinucleotide (3*n*) enriched DNA library of avocado

Successive stages in marker development	2*n*	3*n*
Clones sequenced	233	493
Duplicated clones[a] (%)	7	9
No suitable repeats present (%)	14	69
Primers designed (%)	64	12
Promising loci[b]	20 (57% of loci screened)	26 (45% of loci screened)

Percentages for the number of clones sequenced, clones duplicated (redundant), clones lacking repeats, and clones for which primers could be designed are expressed relative to the total number of clones sequenced.

[a] This comparison is inaccurate as it becomes progressively more likely for a new sequence to be identical/similar to a previous one (the dinucleotide-enriched library screened after completion of the 3*n* library screen).

[b] Data for the two repeat categories are based on a different number of loci screened.

sequenced 493 clones from the trinucleotide-enriched library. Of these, 31% contained a trinucleotide repeat and primers could be designed to 12% (58) of the original trinucleotide clones sequenced. By contrast, 86% of the 233 clones sequenced from the dinucleotide-enriched library contained repeats and primers were designed to 64% (150) of the original dinucleotide clones sequenced. Clone redundancy was noted for both enriched libraries.

Upon amplification, not all of the loci turned out to be useful (interpretable). Only 26 (45%) of the 58 amplified loci from the trinucleotide-enriched library contained highly informative bands, corresponding to a 5% yield of trinucleotide loci of 493 inserts sequenced. Comparable figures for the dinucleotide library are preliminary. Based on a screen of 35 dinucleotide loci we found that 20 (57%) yielded informative bands. Extrapolating from this sub-sample we would expect about 133 informative loci among our 233 inserts. Overall, the dinucleotide-enriched library yielded 11 times as many informative microsatellite loci per insert sequenced than the trinucleotide-enriched library.

Although less frequent the trinucleotide repeat loci were generally easier to interpret than the dinucleotide repeat loci (Fig. 1). Common problems in the interpretation of amplification products are the appearance of excess numbers of bands, smears, and amplification failure (null alleles).

3.2. Microsatellite motifs

Among 235 repeats (in 147 inserts) from the dinucleotide-enriched library, we found 190 AG repeats (81%), 43 AC repeats (18%), two AT repeats (1%), and GC repeats were absent. An earlier screen of the same dinucleotide-enriched library (141 repeats in 116 inserts; unpublished data) revealed similar distributions, albeit with both AT and GC repeats absent. When data from both dinucleotide screens were pooled (376 repeats), 81% of the dinucleotide clones contained AG repeats and 19% contained AC repeats, with the two AT repeats contributing 0.5%.

Fig. 2(1) shows the frequency distribution of the two main dinucleotide motif categories. The AG repeats were both more frequent and had longer repeat arrays compared with the AC repeats. Repeat numbers of 15 were the most common among AG arrays while 12 repeats were most common among AC arrays. The longest repeat array found was an AC repeated

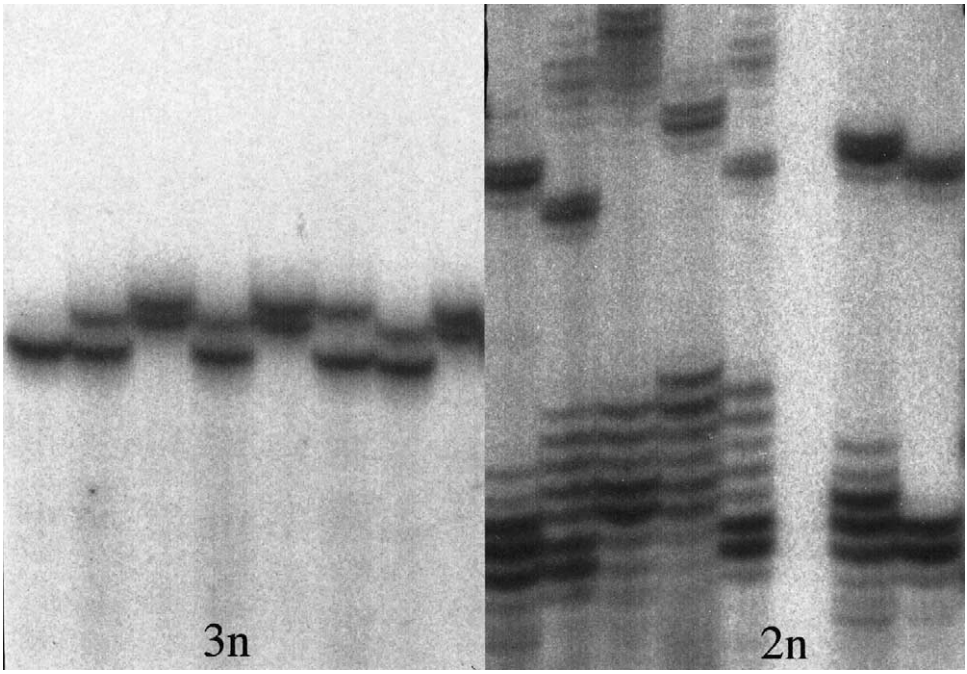

Fig. 1. Autoradiogram illustrating banding patterns of a trinucleotide (3*n*) microsatellite marker and a dinucleotide (2*n*) microsatellite marker. The banding at the 3*n* locus is unambiguous, whereas that of the 2*n* locus was deemed uninterpretable, excluding it from use as a marker.

34 times. Interestingly, the distributions of AC repeats, but especially of AG repeats, did not describe a bell-shaped curve. Instead, frequencies of 9–10 units (AC) and 8–14 units (AG) were relatively scarce, corresponding to a dip in the curve. A single, perfect, repeat array was found in 64% of the dinucleotide loci, 15% of loci were compound (repeats of two or more different motifs arranged in tandem or within the same insert), 13% were interrupted by short regions of non-repetitive imperfections, and 8% were both compound and interrupted.

In a subset of 104 trinucleotide repeats (present at 98 distinct loci), 46 (44%) contained ATG repeats, 32 (30%) contained AGT repeats, and 17 (16%) contained AAG repeats (counts for each repeat motif include corresponding permutations and reverse complements thereof; Fig. 2(2)). Motifs based on AGG and ACC each occurred five times; AGC- and ACG-type motifs occurred in two clones, and a single occurrence was noted for the AAT-type motif. AAC-type and GGC-type motifs were not detected. The longest trinucleotide array contained 11 elements and the most frequently represented array length was six elements. Arrays of nine and above were rare.

3.3. Allelic composition

The repeat types of the 25 loci used to compare a panel of 37 avocado genotypes are presented in Table 1. In this data set, 11 of 14 trinucleotide loci contained simple repeats of no more than 10 elements. Two loci contained compound repeats, and another locus contained

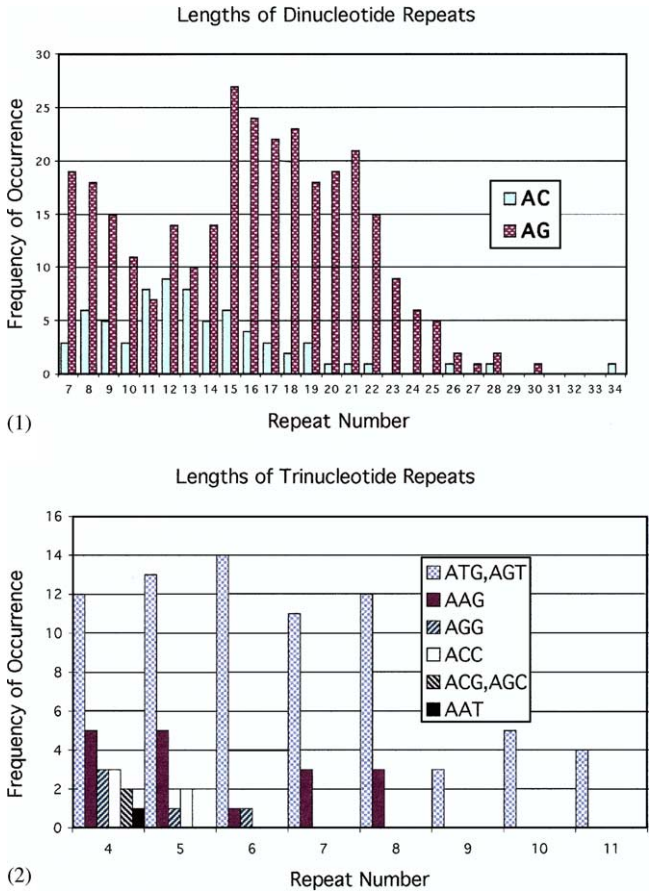

Fig. 2. Histograms depicting the frequency of different microsatellite repeat motifs in clones sequenced from a dinucleotide-enriched library (Fig. 2(1)) and trinucleotide-enriched library (Fig. 2(2)) of avocado. (1) Frequency of AG and AC dinucleotide repeats as a function of array length; (2) frequency of various trinucleotide repeat motifs as a function of array length.

two interrupted repeats. Among the 11 dinucleotide loci, six loci contained simple repeats, the longest consisting of 26 repeat elements. Two loci had compound repeats and two contained interrupted repeats (one included a short trinucleotide repeat). Average repeat lengths for the trinucleotide and dinucleotide categories are 8.5 and 19.4 repeat elements, respectively.

Consistent with their greater repeat lengths, dinucleotide loci were more variable than trinucleotide loci, with allele numbers of 12–21 and 3–10 per locus, respectively. The total number of alleles was 204, of which the 11 dinucleotide loci contributed 124 (60%). Many loci, especially dinucleotide loci, exhibited allele size differences that were not multiples of the repeat unit (e.g., a 1-base difference or $\geq$3 bases between consecutive dinucleotide alleles; Table 3). Several loci were additionally characterized by large gaps between successive alleles (e.g., AVT143, AVO102, AVD002, AVD013, and AVD015).

Table 3
Allele sizes and size intervals at 25 microsatellite loci for 37 avocado genotypes

Locus	Allele size range	Alleles of Hass	Heterozygosity (%)	Sequential allele size intervals (nucleotides)
AVT005b	186–201	186, 190	64.9	2 1 1 1 1 2 3 4
AVT020gat	152–179	161, 164	48.6	3 6 3 15
AVT021	130–139	130, 136	59.5	3 3 3
AVT038	175–202	187, 190	56.7	12 3 4 2 6
AVT106	303–315	309	35.1	3 3 6
AVT143	202–259	211	54.1	3 3 3 36 9 3
AVT158	261–272	267, 272	27.8	3 1 1 1 1 1 1 2
AVT191	167–173	170	40.5	3 3
AVT226	293–325	298, 302	62.2	3 2 1 1 2 2 4 17
AVT372	170–182	170, 182	64.9	3 6 3
AVT386	219–233	229	62.2	1 1 2 1 5 2 2
AVT436	137–155	152, 155	67.6	3 3 3 1 2 3 3
AVT448	163–202	192	43.2	15 3 1 2 6 2 1 6 3
AVT517	217–233	229	48.6	3 3 2 1 3 3 1
AVD001	218–264	224, 226	86.5	4 2 2 2 6 4 2 2 2 2 2 2 12
AVD002	301–341	327, 341	60.0	12 1 1 3 2 2 2 2 1 3 5 2 2 1 1
AVD003II	179–217	192, 211	73.0	2 11 3 1 1 1 1 1 1 2 2 2 2 2 6
AVD006	309–365	314, 351	75.7	3 2 1 3 9 1 2 1 5 2 2 5 2 1 1 1 1 3 9 2
AVD013	197–250	216, 242	78.4	18 1 1 1 2 1 1 8 7 4 1 1 5 2
AVD015	215–284	258, 260	34.4	9 1 4 6 3 14 2 4 2 22 2
AVD017	208–272	211, 254	69.4	1 1 1 5 2 10 9 3 4 10 11 7
AVD018	202–240	224	74.3	12 2 2 2 2 2 2 2 4 2 6
AVD022	216–256	222, 228	88.2	2 4 2 2 2 6 8 2 1 3 2 3 3
AVO102	141–202	153, 200	86.5	8 2 2 4 3 2 2 4 2 20 2 4 4 2
AUCR418	355–403	379, 381	80.6	2 4 2 2 8 6 1 1 2 2 4 2 10 2

Allele sizes for variety Hass are known, as Hass constitutes the focal species used for marker development. Allele size intervals were derived from fragment analysis using the GeneScan software. Loci AVT005b-AVT517 are trinucleotide arrays and the remainder is dinucleotide arrays. Heterozygosity is expressed as the percentage of heterozygous individuals of the total number of genotypes.

No significant correlation was found between allele number and repeat length within each class of repeat. Allele sizes in avocado cultivar Hass, the source of DNA used to generate the two enriched libraries, were not noticeably skewed towards larger allele size. Table 3 indicates that alleles in this cultivar are generally of intermediate size. Average heterozygosity of the trinucleotide and dinucleotide loci was 52.6 and 73.4%, respectively.

4. Discussion

Developing microsatellite markers from an enriched DNA library is time consuming. It involves a winnowing process that sifts out unsuitable loci at each of several successive steps (Table 2). Our comparison of marker development from a dinucleotide- and trinucleotide-enriched library shows that both development strategies have their merits but that the former yields appreciably more loci. The same has been reported elsewhere, including the problem of (trinucleotide) clone redundancy (30%—Butcher et al., 2000; Métais

et al., 2002). Trinucleotide loci are scarce and hence their development expensive in terms of sequencing costs, despite savings at the level of primer purchase. Conversely, dinucleotide microsatellites are present in most clones sequenced but many of these loci tend to be fussy or of narrow applicability to a wider taxon sample (see also Métais et al., 2002).

We found that band interpretation was rarely problematic when interpreting trinucleotide loci but often a concern with dinucleotide loci. Loci that reveal too many bands may reside in duplicated genomic regions or result from unfortunate primer placement, e.g., within a minisatellite. The moderately large genome size of avocado (883 Mb, Arumuganathan and Earle, 1991) may be indicative of sequence redundancy. It is possible that the preferential occurrence of trinucleotide repeats in expressed parts of the genome (Baker et al., 1995; Richard and Dujon, 1996; Young et al., 2000) confers superior amplification properties. Temnych et al. (2001) found GA repeats to be more readily amplified than AT and AC repeats and detected a correlation with enhanced GC richness in the flanking regions. By analogy, triplet repeats may also reside in GC-rich regions of the genome that facilitate PCR assays.

The high percentage of AG motifs (81%) among dinucleotide repeats is consistent with findings from other studies in plants. Many reports cite AG as being more abundant than AC, including general species surveys (Condit and Hubbell, 1992; Lagercrantz et al., 1993; Morgante and Oliveiri, 1993; Wang et al., 1994) and taxon-specific studies, including *Quercus* (Dow et al., 1995), *Eucalyptus* (Byrne et al., 1996; Brondani et al., 1998), *Triticum* (Ma et al., 1996), *Actinidia* (Weising et al., 1996), *Populus* (Dayanandan et al., 1998), *Prunus* (Cipriani et al., 1999), *Melaleuca* (Rossetto et al., 1999), and *Sorghum* (Kong et al., 2000). However, AC was more common than AG in *Pinus* (Echt and May-Marquardt, 1997), *Acacia* sp. (Butcher et al., 2000) and *Avicennia* (Maguire et al., 2000). By searching sequences deposited in GenBank, Tóth et al. (2000) found a preponderance of AG repeat types within exon sequences among a cross-section of embryophyta (mainly *Arabidopsis thaliana*). AG motifs also outnumbered AC motifs in introns and intergenic regions, but overall AT repeats were the most abundant. Temnych et al. (2001) similarly demonstrated a preponderance of AG over AC in rice, but again AT was the most common motif detected.

Differences in the frequencies of different trinucleotide motifs have variously been ascribed to conformational properties of the DNA (Gacy et al., 1995; Bidichandani et al., 1998; Usdin, 1998), selection for hydrophilic or hydrophobic amino acids (in the case of trinucleotide repeats residing in coding regions; Katti et al. (2001)), and unspecified cellular factors (Tóth et al., 2000). Ultimately, trinucleotide repeat arrays and their corresponding amino acid polymers have been suggested to provide the variation upon which selection may act (Moxon et al., 1994; Young et al., 2000).

Whether the predominant occurrence of GA and ATG/AGT accurately reflects the composition of microsatellite motifs in the avocado genome as a whole is uncertain. We expected a high representation of AG and ATG since our libraries were prepared against these two motifs. However, a preliminary screen had revealed these motifs to be the most abundant from the outset. Nonetheless, artifacts of the library construction process may also be responsible for the profile of repeat motifs. Hamilton and Fleischer (1999) showed that motif composition is affected by the restriction enzyme used to size-fractionate the genomic DNA. Their recommendation was to use a cocktail of restriction enzymes to size-fractionate the genomic library, as was used here. Technical difficulties with probing and amplifying TA

repeats similarly tend to artificially decrease the frequency of these repeats. Additionally, TA repeats have been shown to be associated with transposable elements in rice (Akagi et al., 2001; Temnych et al., 2001). Within the trinucleotide motifs we found a bias in favor of combinations of two bases charged with two hydrogen bonds (A and T) with one base charged with three hydrogen bonds (G or C), possibly due to conformational benefits of the DNA double helix. Tóth et al. (2000) report a preponderance of AAG in both intergenic regions and exons, whereas AGT and ATG are far less common. Katti et al. (2001) also found AAG to predominate, followed by ATG, then AAC, AAT, AGG, ACC, and small amounts of AGC, AGT, ACG and GGC.

Allele size differences that deviate from the repeat unit size appear to be widespread in our data set. They have also been noted elsewhere (Huang et al., 1998) and are due to insertion/deletion events in the repeat region and/or the flanking regions. Single-base differences arise in this way or, artifactually, by poly-A addition due to polymerase error. Although we cannot exclude the possibility of polymerase error we feel that our 45 min PCR final extension time should at least have minimized the problem of uneven poly-A addition. The fact that amplification sometimes failed, especially in the phylogenetically more distant accessions, lends further support to the assumption that flanking regions are undergoing mutation. Such mutations are equally likely to modify fragment length as to result in non-amplification (null alleles). Uncertain homology between bands—especially those from distant genotypes—has been lamented elsewhere (e.g., Grimaldi and Crouau-Roy, 1997; Ortí et al., 1997; Whitton et al., 1997; Noor et al., 2001). In general, these and other studies have shown that, upon sequencing, comigrating fragments turn out to be identical in size but not identical by descent.

Development of microsatellite markers in avocado has been fairly slow due to the apparent scarcity of loci that are readily interpretable across a broad sample of genotypes. It is possible that we are missing potential loci by applying too high a cutoff when screening DNA sequences for the presence of a microsatellite. Additional loci might be procured by lowering the threshold to below four repeats (trinucleotide loci) and seven repeats (dinucleotide loci). We found that modifications in PCR conditions and re-designing primers rarely lead to significant improvements in band interpretability, suggesting genome-specific impediments. Given the appreciable genetic diversity present in avocado (Ashworth and Clegg, 2003) it is possible that the use of a single cultivar as the source of genomic DNA for the microsatellite-enriched library is too narrow for broad spectrum marker development. This might explain why we have been unable to use markers developed by Lavi et al. (1994) and Sharon et al. (1997).

Acknowledgements

This work was supported by a grant from the California Avocado Commission.

References

Akagi, H., Yokozeki, Y., Inagaki, A., Mori, K., Fujimura, T., 2001. *Micron*, a microsatellite-targeting transposable element in the rice genome. Mol. Genet. Genom. 266, 471–480.

Arumuganathan, K., Earle, E.D., 1991. Nuclear DNA content of some important plant species. Plant Mol. Biol. Rep. 9, 208–218.

Ashworth, V.E.T.M., Clegg, M.T., 2003. Microsatellite markers in avocado (*Persea americana* Mill.). Genealogical relationships among cultivated avocado genotypes. J. Hered. 94, 407–415.

Baker, R.J., Longmire, J.L., Van Den Bussche, R.A., 1995. Organization of repetitive elements in the upland cotton genome (*Gossypium hirsutum*). J. Hered. 86, 178–185.

Bidichandani, S.I., Ashizawa, T., Patel, P.I., 1998. The GAA triplet-repeat expansion in Friedreich ataxia interferes with transcription and may be associated with an unusual DNA structure. Am. J. Hum. Genet. 62, 111–121.

Brondani, R.P.V., Brondani, C., Tarchini, R., Grattapaglia, D., 1998. Development, characterisation and mapping of microsatellite markers in *Eucalyptus grandis* and *E. urophylla*. Theor. Appl. Genet. 97, 816–827.

Butcher, P.A., Decroocq, S., Gray, Y., Moran, G.F., 2000. Development, inheritance and cross-species amplification of microsatellite markers from *Acacia mangium*. Theor. Appl. Genet. 101, 1282–1290.

Byrne, M., Marquez-Garcia, M.I., Uren, T., Smith, D.S., Moran, G.F., 1996. Conservation of genetic diversity of microsatellite loci in the genus *Eucalyptus*. Aust. J. Bot. 44, 331–341.

Cipriani, G., Lot, G., Huang, W.G., Marrazzo, M.T., Pertlunger, E., Testolin, R., 1999. AC/GT and AG/CT microsatellite repeats in peach [*Prunus persica* (L.) Batsch]: isolation, characterisation and cross-species amplification in *Prunus*. Theor. Appl. Genet. 99, 65–72.

Condit, R., Hubbell, S.P., 1992. Abundance and DNA sequence of two-base repeat regions in tropical tree genomes. Genome 34, 66–71.

Dayanandan, S., Rajora, O.P., Bawa, K.S., 1998. Isolation and characterisation of microsatellites in trembling aspen (*Populus tremuloides*). Theor. Appl. Genet. 96, 950–956.

Dow, B.D., Ashley, M.V., Howe, H.F., 1995. Characterisation of highly variable $(GA/CT)_n$ microsatellites in the bur oak, *Quercus macrocarpa*. Theor. Appl. Genet. 91, 137–141.

Echt, C.S., May-Marquardt, P., 1997. Survey of microsatellite DNA in pine. Genome 40, 9–17.

Gacy, A.M., Goellner, G., Juranic, N., Macura, S., MacMurray, C.T., 1995. Trinucleotide repeats that expand in human disease form hairpin structures in vitro. Cell 81, 533–540.

Goldstein, D.B., Pollock, P.D., 1997. A review of mutation processes and methods of phylogenetic inference. J. Hered. 88, 335–342.

Goldstein, D.B., Schlötterer, C. (Eds.), 1999. Microsatellites: Evolution and Applications. Oxford University Press, Oxford.

Goldstein, D.B., Linares, A.R., Cavalli-Sforza, L.L., Feldman, M., 1995. Genetic absolute dating based on microsatellites and the origin of modern humans. Proc. Natl. Acad. Sci. U.S.A. 92, 6723–6727.

Grimaldi, M.C., Crouau-Roy, B., 1997. Microsatellite allelic homoplasy due to variable flanking sequences. J. Mol. Evol. 44, 336–340.

Haberl, M., Tautz, D., 1999. Comparative allele sizing can produce inaccurate allele size differences for microsatellites. Mol. Ecol. 8, 1347–1350.

Hamada, H., Petrino, M.G., Kakunaga, T., 1982. A novel repeated element with Z-DNA-forming potential is widely found in evolutionarily diverse eukaryotic genomes. Proc. Natl. Acad. Sci. U.S.A. 79, 6465–6469.

Hamilton, M.B., Fleischer, R.C., 1999. Cloned microsatellite repeats differ between 4-base restriction endonucleases. J. Hered. 90, 561–563.

Huang, W.G., Cipriani, G., Morgante, M., Testolin, R., 1998. Microsatellite DNA in *Actinidia chinensis*: isolation, characterisation, and homology in related species. Theor. Appl. Genet. 97, 1269–1278.

Jarne, P., Lagoda, J.L., 1996. Microsatellites—from molecules to populations and back. Trends Ecol. Evol. 11, 424–429.

Jurka, J., Pethiyagoda, C., 1995. Simple repetitive DNA sequences from primates: compilation and analysis. J. Mol. Evol. 40, 120–126.

Katti, M.V., Rangekar, P.K., Gupta, V.S., 2001. Differential distribution of simple sequence repeats in eukaryotic genome sequences. Mol. Biol. Evol. 18, 1161–1167.

Kong, L., Dong, J., Hart, G.E., 2000. Characteristics, linkage-map positions, and allelic differentiation of *Sorghum bicolor* (L.) Moench DNA simple-sequence repeats (SSRs). Theor. Appl. Genet. 101, 438–448.

Lagercrantz, U., Ellegren, H., Andersson, L., 1993. The abundance of various polymorphic microsatellite motifs differs between plants and vertebrates. Nucleic Acids Res. 21, 1111–1115.

Lavi, U., Akkaya, M., Bhagwat, A., Lahav, E., Cregan, P.B., 1994. Methodology of generation and characteristics of simple sequence repeat DNA markers in avocado (*Persea americana* M.). Euphytica 80, 171–177.

Litt, M., Luty, J.A., 1989. A hypervariable microsatellite revealed by in vitro amplification of a dinucleotide repeat within the cardiac muscle actin gene. Am. J. Hum. Genet. 44, 397–401.
Ma, Z.Q., Roder, M., Sorrells, M.E., 1996. Frequencies and sequence characteristics of di-, tri- and tetra-nucleotide microsatellites in wheat. Genome 39, 123–130.
Maguire, T.L., Edwards, K.J., Saenger, P., Henry, R., 2000. Characterisation and analysis of microsatellite loci in a mangrove species, *Avicennia marina* (Forsk.) Vierh. (Avicenniaceae). Theor. Appl. Genet. 101, 279–285.
Métais, I., Hamon, B., Jalouzot, R., Peltier, D., 2002. Structure and level of genetic diversity in various bean types evidenced with microsatellite markers isolated from a genomic enriched library. Theor. Appl. Genet. 104, 1346–1352.
Morgante, M., Oliveiri, A.M., 1993. PCR-amplified microsatellites as markers in plant genetics. Plant J. 3, 175–182.
Moxon, E.R., Rainey, P.B., Nowak, M.A., Lenski, R.E., 1994. Adaptive evolution of highly mutable loci in pathogenic bacteria. Curr. Biol. 4, 24–33.
Noor, M.A.F., Kliman, R.M., Machado, C.A., 2001. Evolutionary history of microsatellites in the obscura group of *Drosophila*. Mol. Biol. Evol. 18, 551–556.
Ortí, G., Pearse, D.E., Avise, J.C., 1997. Phylogenetic assessment of length variation at a microsatellite locus. Proc. Natl. Acad. Sci. U.S.A. 94, 10745–10749.
Pearson, W.R., Lipman, D.J., 1988. Improved tools for biological sequence comparison. Proc. Natl. Acad. Sci. U.S.A. 85, 2444–2448.
Queller, D.C., Strassmann, J.E., Hughes, C.R., 1993. Microsatellites and kinship. Trends Ecol. Evol. 8, 285–288.
Richard, G.F., Dujon, B., 1996. Distribution and variability of trinucleotide repeats in the genome of the yeast *Saccharomyces cerevisiae*. Gene 174, 165–174.
Rossetto, M., McLauchlan, A., Harriss, F.C., Henry, R.J., Baverstock, P.R., Lee, L.S., Maguire, T.L., Edwards, K.J., 1999. Abundance and polymorphism of microsatellite markers in the tea tree (*Melaleuca alternifolia*, Myrtaceae). Theor. Appl. Genet. 98, 1091–1098.
Sharon, D., Cregan, P.B., Mhameed, S., Kusharska, M., Hillel, J., Lahav, E., Lavi, U., 1997. An integrated genetic linkage map of avocado. Theor. Appl. Genet. 95, 911–921.
Smeets, A.J.M., Brunner, H.G., Ropers, H.H., Wieringa, B., 1989. Use of variable simple sequence motifs as genetic markers: application to the study of myotonic dystrophy. Hum. Genet. 83, 245–251.
Stallings, R.L., Ford, A.F., Nelson, D., Torney, D.C., Hildebrand, C.E., Moyzis, R.K., 1991. Evolution and distribution of $(GT)_n$ repetitive sequences in mammalian genomes. Genomics 10, 807–815.
Sunnucks, P., 2000. Efficient genetic markers for population biology. Trends Ecol. Evol. 15, 199–203.
Tautz, D., 1989. Hypervariability of simple sequences as a general source for polymorphic DNA markers. Nucleic Acids Res. 17, 6463–6471.
Temnych, S., DeClerck, G., Lukashova, A., Lipovich, L., Cartinhour, S., McCouch, S., 2001. Computational and experimental analysis of microsatellites in rice (*Oryza sativa* L.): frequency, length variation, transposon associations, and genetic marker potential. Genome Res. 11, 1441–1452.
Tóth, G., Gáspári, Z., Jurka, J., 2000. Microsatellites in different eukaryotic genomes: survey and analysis. Genome Res. 10, 967–981.
Usdin, K., 1998. NGG-triplet repeats form similar intrastrand structures: implications for the triplet expansion diseases. Nucleic Acids Res. 26, 4078–4085.
Wang, Z., Weber, J.L., Zhong, G., Tanksley, S.D., 1994. Survey of plant short tandem DNA repeats. Theor. Appl. Genet. 88, 1–6.
Weber, J.L., May, P.E., 1989. Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction. Am. J. Hum. Genet. 44, 388–396.
Weising, K., Fung, R.W.M., Keeling, D.J., Atkinson, R.G., Gardner, R., 1996. Characterisation of microsatellites from *Actinidia chinensis*. Mol. Breed. 2, 117–131.
Weissenbach, J., Gyapay, G., Dib, C., Vignal, A., Morisette, J., Millaseau, P., Vaysseix, G., Lathrop, M., 1992. A second generation linkage map of the human genome. Nature 359, 794–801.
Whitton, J., Rieseberg, L.H., Ungerer, M.C., 1997. Microsatellite loci are not conserved across the Asteraceae. Mol. Biol. Evol. 14, 204–209.
Young, E.T., Sloan, J.S., Van Riper, K., 2000. Trinucleotide repeats are clustered in regulatory genes in *Saccharomyces cerevisiae*. Genetics 154, 1053–1068.
Zanis, M.J., Soltis, D.E., Soltis, P.S., Mathews, S., Donoghue, M.J., 2002. The root of the angiosperms revisited. Proc. Natl. Acad. Sci. U.S.A. 99, 6848–6853.