M. J. M. Smulders · G. Bredemeijer · W. Rus-Kortekaas
P. Arens · B. Vosman

Use of short microsatellites from database sequences to generate polymorphisms among *Lycopersicon esculentum* cultivars and accessions of other *Lycopersicon* species

Received: 9 August 1996 / Accepted: 23 August 1996

Abstract A search of nearly 2000 sequences from *Solanaceae* species in the EMBL and Genbank databases yielded 220 microsatellites. Among these were 80 microsatellites from 675 *Lycopersicon* entries. Dinucleotide repeats, as well as $(CAA)_n$ and $(TAA)_n$ repeats, were over-represented in non-coding DNA. The other trinucleotide repeats were predominantly found in exonic DNA. PCR analysis of 44 of the microsatellite-containing *Lycopersicon* loci identified 36 primer pairs that yielded well-scorable fragments, or groups of fragments, in *L. esculentum* cultivars and accessions of *Lycopersicon* species. Twenty-nine of these amplified bands that were polymorphic among the four *Lycopersicon* species. Ten primer pairs generated polymorphic bands among seven tomato cultivars. Upon examining the number of microsatellites and the degree of polymorphisms in relation to the repeat type and motif, the type of DNA the microsatellite resided in, the length of the microsatellite, and the presence of imperfections in the microsatellite, only two significant correlations were found. (i) Imperfect repeats were less polymorphic among species than perfect repeats. (ii) The percentage of loci polymorphic among cultivars increased from 6% for the shortest loci (with eight or less repeat units) to 60% for the group with the longest repeats (12 repeat units or longer). Among the species, however, all length classes contained about 83% polymorphic loci. In general, 2–4 alleles were found for each locus among the samples of the test set. In a few cases, up to eight alleles were found. A combination of these microsatellite loci can therefore be useful in distinguishing cultivars of tomato, which are genetically very closely related to each other.

Communicated by H. F. Linskens

M. J. M. Smulders (✉) · G. Bredemeijer · W. Rus-Kortekaas
P. Arens · B. Vosman
DLO-Centre for Plant Breeding and Reproduction Research
(CPRO-DLO), P.O. Box 16,
NL-6700 AA Wageningen, The Netherlands

Key words Simple sequence repeat · Tandem repeat · Variety identification · STMS · Sequence tagged microsatellite site

Introduction

Short tandem repeats (STRs) are very common in the genomes of eukaryotes (Hamada et al. 1982, 1984). Such repeats display high levels of polymorphism due to variation in repeat length, which renders them very useful as genetic markers. Since the sequences flanking STRs are often conserved, PCR-based detection of repeat length variation is perfectly possible. To obtain microsatellite-containing DNA sequences from a given species, they can be cloned and sequenced. However, many microsatellite repeats have also been found in sequences already present in databases, both in mammals (Stallings 1995), such as humans, pig and chicken (Moran 1993), rat (Serikawa et al. 1992) and mouse (Love et al. 1990), and in plants (Lagerkrantz et al. 1993; Wang et al. 1994), including soybean (Morgante and Olivieri 1993; Maughan et al. 1995), maize (Lynn and Heun 1993) and *Arabidopsis* (Bell and Ecker 1994, Depeiges et al. 1995). Sequences containing $(AT)_n$ were found to be the most frequent dinucleotide repeat in plants.

In general, the degree of polymorphism increases with the total length of the microsatellite (i.e. the number of repetitions of the repeat unit). A limit of six repeat units was used by Weber and May (1989) and Lagerkrantz et al. (1993). Dinucleotide sequences with ten or fewer repeat units had a low information content in humans (Weber 1990). Love et al. (1990) found a weak correlation between the number of alleles and the number of repeats. Because of this, some database searches were done with a cut-off at eight (Stallings 1994 1995) or ten (Morgante and Olivieri 1993) repeat units, or 20 nucleotides (Weber 1990; Serikawa et al. 1992; Moran 1993; Bell and Ecker 1994; Wang et al. 1994).

To identify molecular markers for use in the identification of tomato varieties and accessions of *Lycopersicon* species (Vosman et al. 1992; Rus-Kortekaas et al. 1994; Arens et al. 1995), microsatellite-containing sequences from *Lycopersicon* species were extracted from the EMBL and Genbank databases. The cut-off was set as low as six repeat units for di- and tri-nucleotide repeats, and four repeats for the tetranucleotide repeats employed. In the present study the loci containing microsatellites are analyzed using the sequence tagged microsatellite (STMS) approach, and amplification results using primers designed for flanking sequences are shown and analyzed.

Materials and methods

Database search

In February 1995, sequences from the family *Solanaceae* were selected from the EMBL (Version 41.0) and Genbank (accessions additional to EMBL) databases using SRS (Sequence Retrieval System, version 3.1), at the CAOS-CAMM Centre in Nijmegen, The Netherlands. Of the 1854 entries, 1850 came from four genera: *Lycopersicon*, *Solanum*, *Nicotiana* and *Petunia*. On 25th October 1995, the EMBL (Version 44) and Genbank databases were searched again for *Lycopersicon* sequences, yielding 25% more sequences. These additional microsatellite motifs were included in the study described below.

The sequences retrieved were subsequently screened for homology to an artificial sequence containing 12 repeats of the two mononucleotide repeats, six repeats of all possible di-, and tri-nucleotide motifs (see Table 1), plus two tetranucleotide repeats (ATCT and ATGT), using FastA (Pearson and Lipman 1988). Included were those sequences that had at least 12 mononucleotide repeat units, six repeat units for di- and tri-nucleotide repeats, or four repeat units for the tetranucleotide repeats, allowing for one base pair mismatch internally, i.e. flanked on both sides by at least one perfect repeat unit. Obvious doublets in the database were reduced to one hit, and the poly(A) tails of cDNA clones were omitted. These microsatellites were used to determine the distribution of repeat motifs and length among the Solanaceae (see Table 1).

In the ensuing part of this study, microsatellites from the *Lycopersicon* sequences were used to determine the degree of polymorphism that can be detected using the STMS approach. Primer pairs were designed for DNA flanking microsatellite loci using PRIMER (Version 0.5). In this, some compound repeats, consisting of repeats shorter than the criteria listed above, were also included (see Table 2).

Plant DNA

Seven *Lycopersicon esculentum* cultivars and accessions of three wild *Lycopersicon* species (see legend of Table 3) were obtained from the tomato collection of the Centre of Genetic Resources (CGN, part of CPRO-DLO, The Netherlands). The potato breeding line VSW 5337.3 was derived from the cross *S. phureja* 225696.1×*S. tuberosum* L. dihaploid VSW 42. For pepper, *Capsicum frutescens*, accession RU72-7 was obtained from CGN. Nuclear DNA was extracted from frozen leaves essentially as described by Bernatzky and Tanksley (1986) with some slight modifications (Vosman et al. 1992).

Amplification reactions

Each 25 ml amplification reaction contained: 10 ng genomic DNA, 0.32 mM of each primer (Isogen, Maarssen, the Netherlands), 100 mM deoxyribonucleotides, 50 mM KCl, 20 mM Tris-HCl (pH 8.4), 1.5 mM $MgCl_2$, 0.05% (v/v) polyoxyethylene ether (W-1), and 0.5 U *Taq* DNA polymerase (Life Technologies). Amplifications were performed in tubes or microtiter plates using a Hybaid Omni Gene thermal cycler. Basically, the amplification conditions were: 1 cycle of 94°C for 3 min; 30 cycles of 55°C for 45 s, 72°C for 1 min 45 s, and 94°C for 45 s. After the final cycle, one cycle of 55°C for 45 s and 72°C for 3 min was added (Arens et al. 1995). If necessary, the annealing temperature was lowered to 50 °C, and the number of cycles increased to 35 (see Table 3).

Table 1 Microsatellite-containing sequences from *Solaneceae* species in the EMBL/Genbank databases. The search was conducted in February 1995, except for the *Lycopersicon* sequences, for which the results of a later search in October 1995 are shown. In February, there were 65 microsatellites in *Lycopersicon*, in a total of 542 sequences. Number of repeat units used as cut-off: mononucleotides ≥10, di- and tri-nucleotides ≥6, tetranucleotides or higher ≥4, one internal mismatch allowed. In the case of compound repeats, each core is listed separately only if of sufficient length. Excluded from the list are *Lycopersicon* entries containing long microsatellite-like stretches of (combinations of) GATA, TATA, GACA and GGTA (Vosman and Arens (1997); EMBL Accession numbers X90770, X90937, X91107, and X91108)

Microsatellite repeat core motif	Number of entries				
	Lycopersicon	*Solanum*	*Nicotiana*	*Petunia*	Total
A/T	5	2		1	8
C/G	3	2	6		11
AT/TA	33	13	32	7	85
GA/TC	9	5	13	1	28
CA/TG	2		5	1	8
CG/GC					
CAA/TTG			9		9
GAA/TTC	7	7	7		21
TAA/TTA	9	2	2	2	15
CCA/TGG	6		11	1	18
CGA/TCG					
CTA/TAG					
CAG/CTG	2		2		4
GAG/CTC					
GAT/ATC	1			1	2
GCG/CGC			3		3
GATA/TATC	1				1
ATTT/AAAT	1	2	1		4
CCCCA/TGGGG	1				1
TAGATA/TATCTA			1		1
CCCCCCA/TGGGGGG		1			1
Total	80	34	92	14	220
# Sequences searched	675	404	735	169	1983
kbp searched	875.3	551.7	1145	229.9	2802
kbp/microsatellite	10.9	16.2	12.4	16.4	12.7

Table 2 *Lycopersicon* microsatellite loci from the database for which primers have been designed

Accession number	Position of first base of micro-satellite	Locus	Repeat[a]	Location with respect to gene[b]	Frag-ment size (bp)	Primer sequences (forward, reverse)	Calcu-lated primer Tm
A15983	1521	A15983	(AT)$_{6-1}$	o	70 bp	5′-TTCAGTTAAGGGGTTCATAAG-3′ 5′-TCATCAGTTTCAGCTTTATCG-3′	53.7 55.7
U20592	0698	LE20592	(TAT)$_{15-1}$(TGT)$_4$	o	166 bp	5′-CTGTTTACTTCAAGAAGGCTG-3′ 5′-ACTTTAACTTTATTATTGCCACG-3′	54.5 54.8
U21085	0058	LE21085	(TA)$_2$(TAT)$_{9-1}$	m	104 bp	5′-CATTTTATCATTTATTTGTGTCTTG-3′ 5′-ACAAAAAAAGGTGACGATACA-3′	55.0 54.9
M21775	0373	LE2A11	(ATCT)$_{5-1}$	m	157 bp	5′-AATTTTGTAAGGAGAAGACGG-3′ 5′-TCATATTCTTCACACCAAAGG-3′	55.3 55.2
X59139	2013	LEACC2G	(AAAT)$_3$	o	147 bp	5′-TTCCCAGGAAAGTAATTATCC-3′ 5′-GTTCAAGCTAGAAGCTACACG-3′	55.0 54.7
M88487	0232	LEACS4A	(TA)$_7$	o	128 bp	5′-TACAGAATAGGGTTTGCCATA-3′ 5′-GTTTTAGTGGGTTGTGTTGAA-3′	55.0 55.2
M96324	1352	LEATPACAa	(TA)$_7$	o	189 bp	5′-TTACTTACTCCCCTCCAACTC-3′ 5′-CGTTTGGTTACAAGAGAATTG-3′	55.1 55.1
M96324	0945	LEATPACAb	(GA)$_7$	o	184 bp	5′-GTATGTCAAATCTCTCTTGCG-3′ 5′-ACTCTCCATCGTCTCTTTCAC-3′	55.1 56.0
X61287	1302	LECAB9	(TA)$_6$(CA)$_3$	i	118 bp	5′-TTTATTATCCCAGAAGCCTTC-3′ 5′-CCTCACATTTAAACAAATTGC-3′	55.3 55.0
M84744	1567	LECBPE3	(TA)$_{10}$	m	122 bp	5′-CCTACAAAAACTGCCTCT-3′ 5′-TTATATCAATACAACAACATT-3′	
Z15141	0776	LECHI3	(TA)$_{6-1}$(GA)$_4$	m	128 bp	5′-TAACAATCAAAAGAACTTCGC-3′ 5′-ATCCCCTTATTGATTACATCC-3′	54.9 54.9
X14041	0073	LECHSOD	(CTT)$_6$	i	195 bp	5′-TTATCAATTCATCATTGTGGC-3′ 5′-AGGGGTAGTGACAGCATAAAG-3′	56.0 56.0
Z18277	1605	LEDIH4RE	(AAT)$_5$(AAG)$_2$	o	90 bp	5′-TTTTGTAATCATCTTGGAAAC-3′ 5′-ATTGTGTTATGATGATATTTG-3′	52.1 47.6
X13437	1738	LEE8	(TA)$_6$	i	152 bp	5′-TCTTTAGTAGCTCAGTGGCAG-3′ 5′-GGCCAACTAAATCGTTTATTC-3′	55.1 55.6
X53043	0492	LEEF1Aa	(TA)$_8$(ATA)$_9$	o	131 bp	5′-AAATAATTAGCTTGCCAATTG-3′ 5′-CTGAAAGCAGCAACAGTATTT-3′	54.1 55.0
X53043	0711	LEEF1Ab	(TTA)$_{4-1}$A(TTA) A(TTA)$_4$	o	245 bp	5′-AATTTAACAATTGCCAAGTGA-3′ 5′-TGGCTGAAGAATTTTAAATGA3′	55.0 55.2
X53043	1104	LEEF1Ac	(TAT)$_{7-1}$	o	184 bp	5′-CTTGCTGGCTAATCACAATAC-3′ 5′-CCTGCAAAAGAATTCTAAACG-3′	55.2 56.8
X63093	0392	LEGAST1	(TA)$_{12}$, (TG)$_{8-2}$	i	143 bp	5′-ATCTCTATTGTTTTCGACTCG-3′ 5′-TCTGTTGTTGCTGCTGCTC-3′	
X60441	3525	LEGTOM5	(TA)$_{10}$	o	181 bp	5′-AAAGATAAAGCATGAAATGAA-3′ 5′-GGAGTTGAGATAAAGTGAAGA-3′	
M96549	0390	LEHMG2A	(AAC)$_5$	c	254 bp	5′-ATCTGAAGAGCCTGTTTATCC-3′ 5′-AAAGCGTAACGACATGTAAAG-3′	55.1 54.9
L41253	0149	LEHSC70A	(TA)$_{15}$	o	244 bp	5′-CTCTTTGCTCCAATTCAGTTAACA-3′ 5′-TACTCTTCCCCGTAGATTTAGGTG-3′	59.8 59.9
M96549	1911	LEHSC80P	(GAA/G)$_5$, (GAA)$_{5-1}$	c	195 bp	5′-CCTGATTAAGAAGCACTCTGA-3′ 5′-CACTCATTGGAAACTTCTTTG-3′	54.9 55.0
M61915	0663	LEILV1B	(T)$_8$(TA)$_{10}$(T)$_5$	i	143 bp	5′-GATCGACACATTTGAATTTGT-3′ 5′-GGTCACTAATTAATTGATTCC-3′	55.1 50.6
X15855	1192	LELAT52	(TA)$_{8-1}$	i	157 bp	5′-CATTCACTTCGTTCTATTCAG-3′ 5′-TGCTGATGTTCCTGCATTG-3′	52.7 59.3
X56488	1305	LELAT59	(A)$_{26}$	o	75 bp	5′-AACAACATTTCACAAAGTGCT-3′ 5′-CGTCTCAATGAGACAACAAGT-3′	55.0 55.4
X15499	0618	LELAT59G	(TA)$_{9-1}$, (TA)$_{11}$	i	168 bp	5′-AAAAGGGGTATGAACATTAGG-3′ 5′-GCATCTATCGTCTTGTCACTC-3′	54.9 54.9
M76552	0519	LELE25	(TA)$_{13-1}$	o	225 bp	5′-TTCTTCCGTATGAGTGAGT-3′ 5′-CTCTATTACTTATTATTATCG-3′	
Z12127	0488	LELEUZIP	(AAG)$_{6-1}$TT (GAT)$_7$	c	105 bp	5′-GGTGATAATTTGGGAGGTTAC-3′ 5′-CGTAACAGGATGTGCTATAGG-3′	55.1 55.1
L35306	0061	LEMDDNa	(TA)$_9$	o	211 bp	5′-ATTCAAGGAACTTTTAGCTCC-3′ 5′-TGCATTAAGGTTCATAAATGA-3′	54.5 53.5
L35306	1157	LEMDDNb	(TG)$_4$(TA)$_4$	o	280 bp	5′-TAAATACAAAAGCAGGAGTCG-3′ 5′-GAGTTGACAGATCCTTCAATG-3′	54.9 54.5
X14060	4930	LENIA	(TA)$_6$, (TG)$_5$	o	210 bp	5′-TTAAGATTGTATTCATCATGG-3′ 5′-CTTTAGGCTTGTAATGGAGTG-3′	50.3 54.4
M69247	0580	LEPRP4	(TAT)$_3$(TGT)$_5$	o	200 bp	5′-TTCATTTCTTGCAACTACGAT-3′ 5′-CATACTAGCAACATCAAAGGG-3′	55.1 55.0

Table 2 (Continued)

Accession number	Position of first base of micro-satellite	Locus	Repeat[a]	Location with respect to gene[b]	Frag-ment size (bp)	Primer sequences (forward, reverse)	Calcu-lated primer Tm
X05985	0458	LERBCS3B	$(TG)_{6\text{-}1}(TA)_{8\text{-}1}$	i	198 bp	5′-AAACCTTGACATTACCTCCAT-3′ 5′-AGGAAGGTACGACAGAGTCTC-3′	55.2 55.1
Y06521	1455	LERIPE	$(TA)_{10}$	m	184 bp	5′-AAAATGCCTCTCTTCAAAGAT-3′ 5′-ACTACGGAAGTCTCTCAAGAT-3′	54.9 52.3
X79338	0047	LERNALX	$(ATT)_{6\text{-}1}$	m	126 bp	5′-CTCACCCACAAAGAAAATTC-3′ 5′-CTAACAAACATTGTACAACAATAATC-3′	54.8 54.2
M37151	0091	LESODB	$(TTC)_6$	m	207 bp	5′-TTATCAATTCATCATTGTGGC-3′ 5′-AGTAAGGGGTTTAGGGGTAGT-3′	56.0 55.1
L19762	2575	LESSF	$(CCCCA)_4$	o	216 bp	5′-TACGCTCTCAAGTACCGTAAG-3′ 5′-CCTACATTGACATGACCAAAT-3′	55.0 54.9
Z34518	1107	LESSRPSPGa	$(TATT)_5$	o	219 bp	5′-GAATATATCGGGGACAATCTC-3′ 5′-AACGAAATCTTTGTTCAGTGA-3′	55.1 55.0
Z34518	7761	LESSRPSPGb	$(C)_{16}$	i	332 bp	5′-AACATTAGTTTGATTGGATGG-3′ 5′-TTAAACTTTGCTTGACTTTCC-3′	54.2 54.1
X58253	1619	LEUBI3	$(AAG)_{6\text{-}1}$	m	149 bp	5′-GACTACAACATCCAGAAGGAG-3′ 5′-TTATAGAACTGCAACACAGCG-3′	53.9 56.8
M13938	0774	LEWIPIG	$(CT)_{8\text{-}1}(AT)_4$	i	254 bp	5′-GAGTCAAAGTTTGCTCACATC-3′ 5′-CTCTTCTGAACTTGCTTTGAG-3′	55.0 54.6
X51347	1419	LPHFS24	$(TA)_6$	m	149 bp	5′-TTGGATTTACAAGTTCGATGT-3′ 5′-GCATTTGACTTGATAGCAGTC-3′	54.8 55.1
M59427	0982	LPTRYINH	$(CT)_{9\text{-}1}(AT)_{3\text{-}1}$ $(TA)_3(CA)_{5\text{-}1}$	i	171 bp	5′-AAGTTTGCTCACATCATTCTG-3′ 5′-TAAAAGTTCTTCTCCCTCACC-3′	55.5 55.2
S65047	0499	S65047	$(TCTT)_3(CT)_4$	m	140 bp	5′-GATCAACCTAAAACATGCGAC-3′ 5′-TAAGCCTGATGGACTTGATTC-3′	57.2 56.9

[a] A minus-sign indicates deviation from perfect repeat, e.g. $(TA)_{6–1}$ is a $(TA)_6$ repeat with one internal base pair different from the repeat

[b] Based on the description of the sequence in the database; c=coding region (not further specified), i=intron, e=exon, o=outside coding region (3′ or 5′ of the gene), m=messenger RNA/cDNA (not further specified)

Detection of STMS polymorphisms

Samples were prepared for PAGE electrophoresis by adding an equal volume of formamide, containing 10 mM NaOH and 0.05% bromphenol blue, to the reaction mixtures. After denaturation at 80°C for 5 min followed by quenching on ice, samples were analyzed on vertical gels (6% polyacrylamide, 8 M urea, Tris-borate buffer) using a Model S2 sequencing gel electrophoresis apparatus (Life Technologies). The DNA bands were visualized by silver staining according to the Silver-sequence DNA sequencing system (Promega). The sizes of the PCR products were determined by comparison to an accompanying sequence reaction using pGEM-3Zf(+) control DNA (Promega).

Results

Database search

The *Solanaceae* contains four genera from whose members a significant number of sequences are present in the EMBL and Genbank databases: *Lycopersicon* (mainly *L. esculentum* Mill.), *Solanum* (mainly *S. tuberosum*), *Nicotiana* (*N. tabacum* L. and others) and *Petunia* [*P. hybrida* (Hook) Vilm] (Table 1). More than 200 microsatellites were identified among almost 2000 sequences. The absolute frequency varied among the genera, which may be due to the different categories of DNA sequenced. In all, about half of the microsatellites were dinucleotides, mainly $(TA)_n$, and over 30% were trinucleotides, mainly $(AAG)_n$, $(AAT)_n$ and $(ACC)_n$. $(AAC)_n$ repeats were found only in *Nicotiana*.

Based on the descriptions in the database, four categories of genomic sequences were distinguished: exons (coding sequences), introns, DNA upstream of or downstream from a gene, and messenger RNA/cDNA. The latter category contains sequences for which the position of the mature mRNA was not explicitly indicated. Therefore, this category may contain a mixture of microsatellites from transcribed coding and non-coding sequences. Forty two per cent of the microsatellites were found upstream of or downstream from a gene, 26% in introns, 22% in cDNA, and only 10% in coding DNA.

The types of microsatellite repeat unit were not evenly distributed over the different categories of genomic DNA. Upstream of or downstream from the gene and in intronic DNA, 61% of the repeats were dinucleotide repeats. In cDNAs, only 37% of the repeats were dinucleotides, and in exons only 13%. The trinucleotide repeats in general had the opposite distribution, but there were significant differences among repeat motifs. $(CAA)_n$ and $(TAA)_n$ repeats made up 69% of the trinucleotide repeats 3′ and 5′ of genes and in introns, but only 5% of the trinucleotide repeats in exons, whereas $(GAA)_n$ and $(CCA)_n$ repeats made up only

Table 3 Characterisation of PCR products of microsatellites in tomato and potato by the PAGE/silver-staining system

Locus	Amplification in				Quality of electrophoretic patterns for cultivars[d]	Number of different alleles[e] s	PCR conditions	
	Seven[a] *L. esculentum* cultivars	Four[b] *Lycopersicon* species	One *Solanum* accession	One *Capsicum* accession			annealing T (°C)	cycles
A15983	+[c]	+p	+	–	5	n.s.	50	30
LE20592	+p	(+)p	+	–	3	7	55	30
LE21085	+p	+p	–	–	2/3	5	50	30
LE2A11	(+)	+p	–	–	1	4	55	30
LEACC2G	+	+p	+	–	2	3	50	30
LEACS4A	+p	(+)	–	–	4	n.s.	55	30
LEATPACAa	(+)	+p	+	–	3	3	50	30
LEATPACAb	+	+p	–	–	1	3	55	30
LECAB9	+	+p	+	–	2	3	50	30
LECBPE3	+	(+)	–	–	3	1	50	30
LECHI3	+p	+p	+	–	2	2	55	30
LECHSOD	(+)	+p	–	–	1	4	55	30
LEDIH4RE	+	(+)p	+	–	2	3	55	30
LEE8	+	+p	+	–	1	4	55	30
LEEF1Aa	+p	+p	–	–	3	8	55	30
LEEF1Ab	–	+p	–	–	2	2	55	30
LEEF1Ac	–	(+)	–	–	5	n.s.	50	30
LEGAST1	+	+p	+	–	2	4	55	25
LEGTOM5	(+)	(+)p	–	–	2	2	50	30
LEHMG2A	+	+p	+	–	2/3	4	55	30
LEHSC70A	–	–	–	–	5	n.s.	50	35
LEHSC80P	+	+	–	+	1	1	55	30
LEILV1B	(+)	(+)p	–	–	3	3	50	30
LELAT52	+	(+)p	+	–	4	n.s.	55	30
LELAT59	+	–	–	–	3	1	55	30
LELAT59G	+	+p	+	–	2	4	55	30
LELE25	+p	–	–	–	1	3	50	30
LELEUZIP	+p	(+)p	+	–	2	4	55	30
LEMDDNa	+p	+p	–	–	1	4	55	30
LEMDDNb	+	+p	+	–	2	4	55	30
LENIA	+	+p	–	–	3	2	55	30
LEPRP4	(+)	+p	–	–	4	n.s.	50	30
LERBCS3B	+	+p	+	–	2	4	55	30
LERIPE	–	–	–	–	5	n.s.	50	35
LERNALX	+	+	+	–	2	1	55	30
LESODB	+	+p	+	+	3	4	55	30
LESSF	+p	+p	–	+	1	4	55	30
LESSRPSPGa	+	(+)p	–	–	1	2	50	30
LESSRPSPGb	+p	+p	+	+	3	5	50	35
LEUBI3	+	+	+	+	3	1	55	30
LEWIPIG	+p	+p	–	–	1	4	55	30
LPHFS24	+	+p	–	–	1	3	55	30
LPTRYINH	+	(+)	–	–	2	1	55	30
S65047	–	–	–	+	5	n.s.	50	35

[a] *L. esculentum* cultivars used: Moneymaker, San Marzano Lampadone, Vision, Roma, Evita, Calypso, and UC82B. For some loci, only five cultivars were tested

[b] *L. esculentum* cv. Moneymaker, *L. pennellii* LA 333, *L. peruvianum* LA 462 and *L. hirsutum* LA 1363

[c] +: fragments amplified in all samples; (+): fragments amplified in some samples; p: polymorphisms among samples

[d] Quality rating: 1: weak stutter bands; 2; stutter bands relatively strong; 3: ladders of bands of equal intensities; 4: bands of unexpected sizes; 5: very weak bands or no amplification at all

[e] n.s.=not scorable; only alleles in *L. esculentum* cultivars and *Lycopersicon* species are counted

21% of the trinucleotide repeats outside coding sequences, and 72% of the exonic trinucleotide repeats. With respect to stop codons, $(TAA)_n$ repeats were absent from exons. $(TAG)_n$ repeats have not been found at all, but two (TGA)-type repeats have been identified in exons: a $(TGA)_n$ repeat in tomato and a $(GAY)_n$ repeat in petunia.

In all, there appears to be a promising number of microsatellites in these sequences. The questions of whether these can be amplified and what degree of polymorphism they detect were studied further for the genus *Lycopersicon*.

Microsatellite amplification

For the 80 *Lycopersicon* sequences with a microsatellite, primer pairs were designed around the microsatellite site

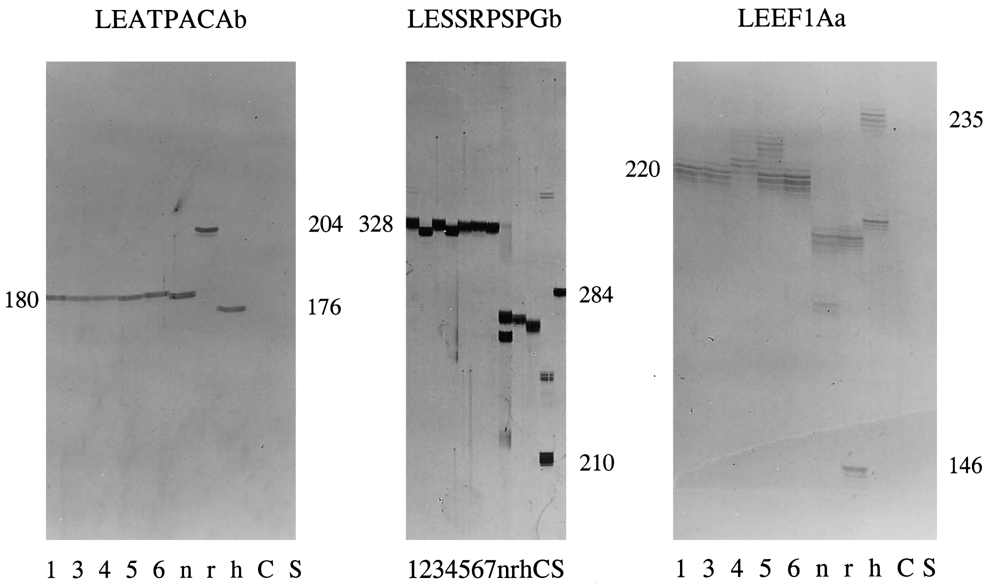

Fig. 1 Silver-stained denaturing PAGE electrophoresis of PCR amplification products of three tomato microsatellite loci in a test set consisting of *L. esculentum* cultivars (*1*=Moneymaker, *2*=San Marzano Lampadone, *3*=Vision, *4*=Roma, *5*=Evita, *6*=Calypso, *7*=UC82B), accessions of *Lycopersicon* species (*n*=*L. pennellii*, *r*=*L. peruvianum*, *h*=*L. hirsutum*), pepper (*C*) and potato (*S*). Primers and PCR conditions as in Table 3. For some fragments, the size is indicated in bp

to amplify the repeat. In a number of cases this was not possible, since the microsatellite was at the border of the sequenced DNA fragment, or was surrounded by repetitive DNA. Primer pairs have been designed and tested for 44 different microsatellite-containing sequences (Table 2).

PCR amplification of the 44 different microsatellites on genomic DNA of a small test-set of tomato cultivars and accessions (see legend of Table 3) was initially carried out under basic PCR conditions. In cases with no or very low reaction yield, the annealing temperature was decreased to 50°C and additionally, if necessary, the number of PCR cycles was increased to 35. In this way, 40 of the 44 primer sets amplified DNA in at least one cultivar or accession, and only four did not give any amplification at all.

Denaturing-PAGE electrophoresis in combination with silver staining was used for the detection of microsatellite polymorphisms. The quality of the PCR products was rated by a number from 1 to 5 (see legend of Table 3; slightly modified from Pepin et al. 1995). Of the 44 loci, 26 were well scorable as fragments of a specific molecular weight (quality 1 and 2; example: LEATPACAb in Fig. 1) while 10 gave ladders of bands of equal intensities, which are still scorable as specific patterns (quality 3; example: other loci in Fig. 1). Two primer pairs amplified not only the expected fragments but also additional fragments of many other sizes, while one primer pair amplified only fragments of unexpected sizes (quality 4). Finally, one locus gave fuzzy bands and four did not give any amplification at all (quality 5).

The presence of artifacts often interfered with allele designation. Especially, ladders of bands of nearly equal intensity, which occurred in almost one quarter of the loci, made allele sizing almost impossible. In an attempt to eliminate such ladder patterns, LEEF1Aa was used as a model locus for PCR optimization, including raising the annealing temperature, lowering the number of PCR cycles, minimizing the annealing and extension periods during the amplification cycles (Ellegren 1992; Lavi et al. 1994), touchdown PCR (Kresovich et al. 1995), and reducing the concentration of primers, *Taq* polymerase, $MgCl_2$, or template DNA (Saiki 1989). None of these changes improved the electrophoretic patterns. Additionally, the design of seven new primer combinations, consisting of new forward or reverse primers, or both, had no positive effect; with some new primer combinations, no amplification was achieved.

Primer pairs that amplify bands of quality 1 or 2 can be used for genotyping with exact fragment sizes. Primer pairs which produce bands of quality 3 can also be used unambiguously when different patterns are scored, e.g. using one symbol or letter for each separate pattern. The use of quality 4 and 5 primer pairs in genotyping is difficult or even impossible. Therefore, in the remainder of the study attention will be given only to the 36 primer pairs yielding banding patterns of quality 1, 2 or 3.

Degree of polymorphism

The majority of the primer pairs of quality 1, 2 or 3 amplified bands that were polymorphic among the *Lycopersicon* species (29 of 36; Table 3). In contrast, less than one third (10 of 36) of the primer pairs generated polymorphic bands among the cultivars. In all, six microsatellite loci were monomorphic. At almost all polymorphic loci, 2–4 different alleles were amplified among cultivars and species together. In Figure 1, some microsatellite fragment patterns are shown, varying from locus LEATPACAb (monomorphic among cultivars but polymorphic among species) to locus LEEF1Aa (which showed the maximum of eight different alleles).

Between *L. esculentum* cv Moneymaker and the accession of *L. pennellii*, 44% of the loci were polymorphic. Between *L. esculentum* and *L. hirsutum* 58% of the loci, and between *L. esculentum* and *L. peruvianum* 81% of the loci, were polymorphic. For a thorough comparison of the amount of variation among and within species, more accessions will have to be used from more species; this is currently being carried out.

Several primer pairs did not generate amplification products in some tomato cultivars or species (Table 3). For example, no amplification products were observed in *L. pennellii* for the LE20592 locus. The lack of amplification of an allele in certain cultivars or accessions can be the result of divergence in the sequences flanking the microsatellite, creating a null-allele. However, it can also result from the production of an undetectable amount of PCR product (Lavi et al. 1994). In the latter case, the optimum PCR conditions for the amplification of a fragment will differ between genotypes (Bell and Ecker 1994). Therefore, optimization of PCR conditions would be necessary for individual cultivars or species from which no fragments have yet been amplified. Until this is done, we do not accept the cases of no amplification to indicate a true negative result.

The primer sets have also been used to amplify genomic DNA from potato and pepper. In total, almost 50% of the primer sets generated amplification products in potato (Table 3), including LESSRPSPGb in Fig. 1. This percentage was also found in the reverse situation: of seven primers based on sequences from potato, four amplified fragments in tomato (data not shown). This indicates that, for a limited number of loci, the use of microsatellite loci-amplifying primer sets across genera is possible. Obviously, this does depend on how closely genera are related, since only 14% of the tomato primers amplified products in the accession of the more distant genus *Capsicum* (Table 3).

Discussion

A search of 675 database entries from *Lycopersicon* sequences yielded 80 microsatellites, or 1 in every 8.4 sequences, which was close to the figure found for microsatellites among all sequences from *Solanaceae* species, and in the same range as the 1 in every 13.8 sequences in *Arabidopsis thaliana* (L.) Heynh. found by Depeiges et al. (1995), who excluded some smaller repeats that were included here. PCR analysis of 44 microsatellite-containing loci from *Lycopersicon* yielded 36 primer pairs with banding patterns of sufficient quality to distinguish alleles unambiguously. Twenty-nine of these primer pairs (86%) amplified bands that were polymorphic among the accessions of four *Lycopersicon* species in the test set, while ten primer pairs (28%) generated polymorphic bands among seven cultivars of *L. esculentum*. This is a good result given the relatively low amount of genetic variation detected with RFLP and RAPD markers among *L. esculentum* cultivars (Miller and Tanksley 1990; Van der Beek et al 1992; Rus-Kortekaas et al. 1994), and the fact that many of the database microsatellites consisted of short stretches of repeats (Table 2).

The percentage of loci polymorphic among cultivars does not differ significantly (P>0.05) from the 2 out of 10 microsatellite loci found to be polymorphic by Broun and Tanksley (1996) among ten cultivars, even though those ten loci contained microsatellites that were much longer than the ones used in the present study. Also the percentage of loci polymorphic among species was comparable. To what extent the polymorphisms among tomato cultivars are coupled to introgression of DNA from wild species into some cultivars, is not known.

Although the bands amplified for the microsatellite loci have not been sequenced, there are several reasons to assume that the polymorphisms found are due to length differences in the microsatellite, and are not the result of insertions or deletions. First, the fragments had approximately the length expected based on the database sequence. Second, for three loci studied in 16 cultivars (LEMDDNa, LELEUZIP, and LELE25), the different alleles varied from each other with the expected multiples of two or three bases (Arens et al. 1995). Third, stutter bands and ladders of bands – a typical feature of microsatellites – were clearly visible in as many as 25 primer pairs (designated quality 2 and 3; Table 3).

A clear relationship between total repeat length and the degree of polymorphism was observed in this study among the cultivars: the degree of polymorphism increased with the total length of the repeat (Table 4). Such a correlation between length and degree of polymorphism has been reported in other species (Grist et al. 1993; Thomas and Scott 1993). In contrast, Bell and Ecker (1994) failed to detect a correlation between polymorphism information content and a repeat length of up to 50 nucleotides among *A. thaliana* strains. Between the *Lycopersicon* species there was no clear effect of the length of the microsatellite: most loci generated polymorphic fragments (Table 4). The fact that longer microsatellites generated polymorphisms both among cultivars and among species, while short microsatellites did so only among species, may indicate that longer microsatellites are by nature also able to produce polymorphisms at a higher frequency among a genetically very ho-

Table 4 Effect of repeat length on the degree of polymorphism among cultivars and species

Number of repeats	# Of loci[a]	# Of polymorphic loci	% Polymorphic loci
	Among seven *L. esculentum* cultivars		
4–8	18	1	6
9–11	8	3	38
≥12	10	6	60
	Among four *Lycopersicon* species		
4–8	18	15	83
9–11	8	7	88
≥12	10	7	70

[a] Only the 36 scorable microsatellite loci (quality 1–3) were included in the analysis

mogeneous group. At the same time, the lower mutation frequency of short microsatellites may render them useful for phylogenetic studies, for which the average mutation frequency of microsatellites is considered too high. Such short microsatellites can be found in sufficient amounts in, or close to, coding sequences, as shown here. Since primer sites located in such regions may be more conserved, primer pairs may generate products in a wider range of species. Whether short microsatellite loci are indeed useful for phylogenetic studies within the genus *Lycopersicon* is currently being studied in detail.

It was not possible to draw general conclusions about differences in the polymorphic nature of each of the repeat motifs, since most motifs were present in small numbers (one or two per class), while $(AT)_n$ and $(ATT)_n$ accounted for the majority of di- and tri-nucleotide type repeats (in accordance with Morgante and Olivieri 1993). With all motifs pooled, di- and tri-nucleotide motifs are comparable in their degree of polymorphism: 24% of the dinucleotide repeats and 31% of the trinucleotide repeats were polymorphic among cultivars, and 71% and 62%, respectively, were polymorphic among species. Given that trinucleotide repeats were found more often in coding DNA, while dinucleotide repeats were more often located outside genes or in introns, it is not surprising that no effect could be discerned between the location of the microsatellite site in coding or non-coding DNA and the degree of polymorphism of the repeat.

Interruptions in the stretch of the microsatellite repeat may prevent slippage of the polymerase, and a decrease in the degree of polymorphism, as was observed for cattle microsatellites (Pépin et al. 1995). We found a significant difference (P<0.05) in the percentage of polymorphic microsatellites between imperfect repeats (7 of 11 or 64% polymorphic) and perfect repeats (22 of 25 or 88% polymorphic).

Based on our data, the production of more stutter bands during the PCR reaction *per se* does not indicate a higher degree of polymorphism in vivo, since the average of three alleles per primer pair produced in our test set is the same both for primer pairs that produce ladders of bands (quality 3) and for primer pairs with hardly any (quality 1) stutter bands. An interesting observation, though, is that the two loci with seven and eight alleles are both from quality 3 patterns (ladders of bands). Also, a breakdown of the results in Table 3 in terms of quality and repeat length shows that patterns of quality 1 occur significantly more frequently (P<0.05) among microsatellites having a short repeat (8 of 18 microsatellites ≤8 repeats; 3 of 18 microsatellites ≥9 repeats). The fact that the short repeats are less prone to stuttering, in combination with their relatively low degree of polymorphism (see above), supports the view that strand slippage during replication is the basis for the generation of new length variants in vivo (Schlötterer and Tautz 1992) as well as for the generation of stutter bands during the PCR reaction (see Ellegren 1992).

In contrast with several other studies (Smith and Devey 1994; Rongwen et al. 1995), our results do not support the hypothesis that tri- and tetra-nucleotide repeats amplify with fewer stutter bands than dinucleotide repeats: 6 out of 18 dinucleotide repeats produced patterns of quality 1, which is about the same as 4 out of 14 tri/tetra-nucleotide repeats.

For 12 loci the ladders of bands amplified were of nearly equal intensity, making exact allele sizing impossible. Attempts to eliminate the ladder patterns by extensive optimization of conditions for the LEEF1Aa locus (Fig. 1) failed. In view of the intrinsic relation between stuttering and slippage in vivo, this is not surprising. The polymorphism of these loci was scored by the designation of specific patterns of stutter bands by a letter, instead of using exact allele length. Preliminary results indicate that the use of an automated system of fluorescence detection (see examples in Frégeau and Fourney 1993; Schwengel et al. 1994; Gill et al. 1995) simplifies the patterns, as only one strand of the template DNA is labelled and the amount of material per fragment can be quantified, thus allowing an exact allele length designation.

DNA profiling of cultivars and accessions

When choosing new microsatellite loci for identification purposes or for studies on genetic variation, both the level of polymorphism and the scorability of the banding patterns are important. On average, the number of alleles for the 36 well-scorable tomato microsatellite loci was 3.1. Among the 30 polymorphic loci, the average number of alleles was 3.8; 89% of these loci had 2–4 alleles (Table 3), only four loci had more, with a maximum of eight different alleles (LEEF1Aa; Fig. 1). In a comparable study of 30 polymorphic microsatellite loci, including some from database sequences, in a panel of six *A. thaliana* ecotypes, the average number of alleles was 4.1, and 60% of the primers had 2–4 alleles (Bell and Ecker 1994). Within the same species, Depeiges et al. (1995) found 12 of 26 primer pairs to be polymorphic among four ecotypes, of which 83% had two alleles.

In comparison to the *Arabidopsis* ecotypes, and in view of the low amount of genetic variation within cultivated *L. esculentum*, the results obtained with short database sequences are good, and open up the possibility to identify cultivars. However, data from a number of microsatellite loci will have to be combined to provide a unique DNA profile for individual cultivars or accessions. This is currently being tested.

Conclusion

Isolating microsatellite loci from database sequences appears useful, since most sequences lend themselves to the design of primers, a fair amount of primer pairs produce polymorphic bands among cultivars, and most primer pairs do so among species. As pointed out by Veillieux et al. (1995), this is the simplest way to perform a genetic analysis. Possibly, microsatellites that are cloned and sequenced after enrichment procedures – which generally are longer – may be polymorphic to a higher degree. Even then, the microsatellites obtained in the present study are very useful, especially since many are $(AT)_n$ repeats, which cannot be cloned using selective enrichment by hybridization.

Acknowledgements The database searches and the design of primers were done using the bioinformatics facilities of the CAOS-CAMM Center in Nijmegen, The Netherlands. Dr. A.W. van Heusden and Dr. J. M. Sandbrink are thanked for critical reading of the manuscript.

References

Arens P, Bredemeijer G, Smulders MJM, Vosman B (1995) Identification of tomato cultivars using microsatellites. Acta Horticult 412:49–57

Bell CJ, Ecker JR (1994) Assignment of 30 microsatellite loci to the linkage map of Arabidopsis. Genomics 19:137–144

Bernatzky R, Tanksley SD (1986) Genetics of actin-related sequences in tomato. Theor Appl Genet 72:314–321

Broun P, Tanksley SD (1996) Characterization and genetic mapping of simple repeat sequences in the tomato genome. Mol Gen Genet 250:39–49

Depeiges A, Goubely C, Lenoir A, Cocherel S, Picard G, Raynal M, Grellet F, Delseny M (1995) Identification of the most represented repeated motifs in *Arabidopsis thaliana* microsatellite loci. Theor Appl Genet 91:160–168

Ellegren H (1992) Polymerase-chain-reaction (PCR) analysis of microsatellites: a new approach to studies of genetic relationships in birds. The Auk 109:886–895

Frégeau CJ, Fourney RM (1993) DNA typing with fluorescently tagged short tandem repeats: a sensitive and accurate approach to human identification. Biotechniques 15:100–119

Gill P, Kimpton CP, Urquhart A, Oldroyd N, Millican ES, Watson SK, Downes TJ (1995) Automated short tandem repeat (STR) analysis in forensic casework – a strategy for the future. Electrophoresis 16:1543–1552

Grist SA, Firgaira FA, Morley AA (1993) Dinucleotide repeat polymorphisms isolated by the polymerase chain reaction. BioTechniques 15:304–309

Hamada H, Petrino MG, Kakunaga T (1982) A novel repeated element with Z-DNA-forming potential is widely found in evolutionarily diverse eukaryotic genomes. Proc Natl Acad Sci USA 79:6465–6469

Hamada H, Petrino MG, Kakunaga T, Seidman M, Stollar BD (1984) Characterization of genomic poly(dT-dG) poly(dC-dA) sequences: structure, organization and conformation. Mol Cell Biol 4:2610–2621

Kresovich S, Szewc-McFadden AK, Bliek SM, McFerson JR (1995) Abundance and characterization of simple-sequence repeats (SSRs) isolated from a size-fractionated genomic library of *Brassica napus* L. (rapeseed). Theor Appl Genet 91:206–211

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 PB (1994) Methodology of generation and characteristics of simple sequence repeat DNA markers in avocado (*Persea americana* M.). Euphytica 80:171–177

Love JM, Knight AM, McAleer MA, Todd JA (1990) Towards construction of a high-resolution map of the mouse genome using PCR-analysed microsatellites. Nucleic Acids Res 18:4123–4130

Lynn M, Heun M (1993) Mapping maize microsatellites and polymerase chain reaction confirmation of the targeted repeats using a CT primer. Genome 36:884–889

Maughan PJ, Saghai Maroof MA, Buss GR (1995) Microsatellite and amplified sequence length polymorphisms in cultivated and wild soybean. Genome 38: 715–723

Miller JC, Tanksley SD (1990) RFLP analysis of phylogenetic relationships and genetic variation in the genus *Lycopersicon*. Theor Appl Genet 80:437–448

Moran C (1993) Microsatellite repeats in pig (*Sus domestica*) and chicken (*Gallus domesticus*) genomes. J Hered 84:274–280

Morgante M, Olivieri AM (1993) PCR-amplified microsatellites as markers in plant genetics. Plant J 3:175–182

Pearson WR, Lipman DJ (1988) Improved tools for biological sequence comparison. Proc Natl Acad Sci USA 85:2444–2448

Pépin L, Amigues Y, Lepingle A, Berthier JL, Bensaid A, Vaiman D (1995) Sequence conservation of microsatellites between *Bos taurus* (cattle), *Capra hircus* (goat) and related species. Examples of use in parentage testing and phylogeny analysis. Heredity 74:53–61

Rongwen J, Akkaya MS, Bhagwat AA, Lavi U, Cregan PB (1995) The use of microsatellite DNA markers for soybean genotype identification. Theor Appl Genet 90:43–48

Rus-Kortekaas W, Smulders MJM, Arens P, Vosman B (1994) Direct comparison of levels of genetic variation in tomato detected by a GACA-containing microsatellite probe and by random amplified polymorphic DNA. Genome 37: 375–381

Saiki K (1989) The design and optimization of the PCR. In: Erlich HA (ed) PCR Technology. Stockton Press, New York, pp 7–16

Schlötterer C, Tautz D (1992) Slippage synthesis of simple sequence DNA. Nucleic Acids Res 20:211–215

Schwengel DA, Jedlicka AE, Nanthakumar EJ, Weber JL, Levitt RC (1994) Comparison of fluorescence-based semi-automated genotyping of multiple microsatellite loci with autoradiographic techniques. Genomics 22:46–54

Serikawa T, Kuramoto T, Hilbert P, Mori M, Yamada J, Dubay CJ, Lindpainter K, Ganten D, Guénet J-L, Lathrop GM, Beckmann JS (1992) Rat gene mapping using PCR-analyzed microsatellites. Genetics 131:701–721

Smith DN, Devey ME (1994) Occurrence and inheritance of microsatellites in *Pinus radiata*. Genome 37:977–983

Stallings RL (1994) Distribution of trinucleotide microsatellites in different categories of mammalian genomic sequences: implications for human genetic diseases. Genomics 21:116–121

Stallings RL (1995) Conservation and evolution of $(CT)_n/(GA)_n$ microsatellite sequences at orthologous positions in diverse mammalian genomes. Genomics 25:107–113

Thomas MR, Scott NS (1993) Microsatellite repeats in grapevine reveal DNA polymorphisms when analyzed as sequence-tagged sites (STSs). Theor Appl Genet 86:985–990

Van der Beek H, Verkerk R, Zabel P, Lindhout P (1992) Mapping strategy for resistance genes in tomato based on RFLPs between cultivars: CF9 (resistance to *Cladosporium fulvum*) on chromosome 1. Theor Appl Genet 84:106–112

Veilleux RE, Shen LY, Paz MM (1995) Analysis of the genetic composition of anther-derived potato by randomly amplified polymorphic DNA and simple sequence repeats. Genome 38:1153–1162

Vosman B, Arens P, Rus-Kortekaas W, Smulders MJM (1992) Identification of highly polymorphic DNA regions in tomato. Theor Appl Genet 85:239–244

Vosman B, Arens (1997) Molecular characterization of GATA /GACA microsatellite repeats in tomato. Genome (in press)

Wang Z, Weber JL, Zhong G, Tanksley SD (1994) Survey of plant short tandem DNA repeats. Theor Appl Genet 88:1–6

Weber JL, May PE (1989) Abundant class of human polymorphisms which can be typed using the polymerase chain reaction. Am J Hum Genet 44:388–396

Weber JL (1990) Informativeness of human $(dC\text{-}dA)_n \cdot (dG\text{-}dT)_n$ polymorphisms. Genomics 7:524–530