Research Article *Journal of Plantation Crops, 2013, 41(3): 277-286*

Development of SSR markers for Robusta coffee (*Coffea canephora*)

Jeena Devasia*, J.N. Madhura[1], R. Raj Kumar[2], M.S. Sheshasayee[1], R.S. Kulkarni[3], M. Udayakumar[1] and Jayarama

Central Coffee Research Institute, CRS Post, Chikmagalur Dist., Karnataka - 577 117
[1]Department of Crop Physiology, UAS, GKVK, Bangalore
[2]ITC, R&D Center, Peenya, Bangalore
[3]Department of Genetics and Plant Breeding, UAS, GKVK, Bangalore

(Manuscript Received: 02-02-13, Revised: 20-07-13, Accepted: 05-08-13)

Abstract

Coffee has long been bred with the view of improving important agronomic characteristics such as yield, bean size, cup quality, caffeine content, disease, drought resistance *etc*. However, the progress in coffee breeding using conventional approaches has been slow due to the narrow genetic base of cultivated coffee and the long generation time. The use of modern tools of molecular biology holds great promise for the faster development of improved varieties. A primary prerequisite is the availability of suitable marker systems. Co-dominant maker systems like SSRs provides comprehensive genome coverage, are locus specific and multi allelic. However, the number of SSR markers available for coffee is limited and there is an urgent need for generating large number of microsatellite markers. Aim of the study was to develop and characterize a comprehensive set of genomic and genic SSR markers for Robusta coffee by pre-cloning enrichment strategy and also by annotating Robusta specific unigene sequences. The pre-cloning enrichment (selective hybridization) strategy followed in the study resulted in identification of 405 SSRs in 267 sequences. The 405 SSRs isolated consisted of more of mono-nucleotide repeats (40.2%) followed by penta (33.3%), di (12.1%), tri (10.6%) and tetra (3.7%) nucleotide repeats. Among the genic SSRs identified, 43.7 per cent contained penta-repeat motifs followed by 22.5 per cent and 22.5 per cent sequences with hexa and mono repeat motifs respectively. The remaining identified motifs consisted of 5.5 per cent tri nucleotide repeat motifs, 3.5 per cent di repeat motifs and 2.2 per cent tetra repeat motifs. The study resulted in development of 31 genomic SSRs and 86 genic SSRs which were validated for locus specific amplification.

Keywords: Bioinformatics, *Coffea canephora*, ESTs, Robusta coffee, SSR marker

Introduction

Microsatellites or simple sequence repeats (SSRs) also known as variable number of tandem repeats (VNTRs) are elements on DNA sequences organized into clusters of tandem repeats with motif sizes ranging from one to many base pairs (bp). They are the most popular of marker systems and are widely used in genetic studies because of their abundant, distribution in the genome, multi-allelic nature, locus specificity, high polymorphism, reproducibility, inter-lab transferability and ease for automation (Powell *et al.*, 1996; Hendre *et al.*, 2008). Microsatellite markers have been developed for a large number of plant species including coffee and are increasingly being used for germplasm diversity, linkage analysis and molecular breeding (Gupta and Varshney, 2000; Combes *et al.*, 2000; Baruah *et al.*, 2003; Aggarwal *et al.*, 2007). The use of these marker system in *C. arabica* and related *Coffea* species indicated higher levels of diversity (Combes *et al.*, 2000; Moncada and McCouch, 2004) when compared to studies using RFLP (Paillard *et al.*, 1993) and RAPD (Lashermes *et al.*, 1993, 1996). However, in spite of the apparent advantages in using

*Corresponding Author: jeena.devasia@gmail.com

SSR markers, only about 224 SSRs were reported for coffee (Poncet *et al*., 2006) at the time of initiation of this study, signifying the need for increasing the availability of these genetically informative markers. The present study aimed at generating more number of informative SSR markers, both genomic and genic, for Robusta coffee.

The genomic SSRs are developed from the genomic DNA sequence of the species. The pre-cloning enrichment strategy was followed to make an enriched library to find out microsatellites from the coffee genome. This strategy employs the whole pool of digested DNA by subjecting them to repeat oligomer hybridization, thereby leading to selective enrichment. Among the various strategies followed for the development of genomic SSRs, the pre-cloning enrichment strategy was reported to be more successful and widely used approach (Karagyozov *et al*., 1993). The present study used this procedure to develop genomic SSRs.

The genic-SSRs were developed from sequencing data or EST database. Hence, this provides a cost-effective alternative when compared to genomic SSRs. Identification of SSRs in gene sequences of plant species was reported quite early (Morgante and Olivieri, 1993). The amount of public sequence data available for SSR marker development initially was limited (15,000 kb) and therefore, only a few genic SSRs were reported (Wang *et al.,* 1994). The rapid increase in sequence data generated from EST projects in several plant species, have allowed the development of genic SSRs in large numbers (Varshney *et al*., 2005). The EST database generated for coffee, from approximately 47,000 cDNA clones of *C. canephora* varieties includes 13,175 unigenes (Lin *et al.*, 2005). In the present paper, an effort was made to develop genic SSRs by annotating these ESTs by Bioinformatics approach.

Materials and methods

Genomic SSR development

Genomic SSR markers were developed following the pre cloning enrichment strategy. The basic protocol was proposed by Karagyozov *et al*. (1993), Armour *et al*. (1994) and Kijas *et al*. (1994). DNA was extracted from leaves of *Coffea canephora* variety S.3334 following the modified CTAB method as described by Bhat (2002). The selective hybridization protocol involved hybridization of size-fractionated adaptor ligated genomic DNA with complementary repeat oligomers followed by elution of the captured DNA fragments and subsequent amplification using adaptor primers and cloning in a competent vector.

The size fractionation of the DNA was achieved by restriction digestion with *Rsa*I (1U) to get approximately 500-1000 bp fragments. The size fractionated genomic DNA was ligated to appropriate adaptors and then hybridized with biotinylated repeat oligomers to enrich the microsatellite containing regions, which were further captured using streptavidin coated magnetic beads. After appropriate washes, the captured DNA was eluted by boiling and was subjected to amplification using adaptor primers to recover the enriched DNA fragments and the PCR product was ligated into T/A cloning vector (pTZ57R/T) and incorporated into a bacterial host. The transformed bacteria were plated on ampicillin selection media. The colonies were screened by colony PCR in a 15 µl reaction volume using M13 forward and reverse primers (3 µM), 100 µM dNTPs, 2 mM $MgCl_2$, 1 U *Taq*DNA polymerase and 1X PCR buffer, at annealing temperature of 58 °C. The PCR products were resolved on 0.8 per cent agarose gel using 1 Kb ladder and the inserts ranging from 500-1000 bp were again amplified in a 50 µl reaction and purified for sequencing. The purified PCR products were further sequenced using the nested M13 forward and reverse primers.

Genic SSR Development

The robusta coffee specific unigene sequences from the Cornell website were analyzed for the development of genic or EST-SSRs.

SSR mining

The 13,175 *C. canephora* specific Unigene sequences were screened for the presence of short repeat motifs using MISA perl scripts (Micro Satellite identification tool) with parameters configured to have minimum threshold of 1-10, 2-5, 3-4, 4-3, 5-2, 6-2 (*i.e*., mono repeat-10 times). The sequences containing repeat motifs of more than

15 bp in length were selected and primers were designed on the flanking region of the repeat motifs.

Primer designing and standardization

The primers were designed using Primer3 (Rozen and Skaletsky, 2000) software, considering the parameters of 50-55 per cent GC content, annealing temperature of 59±5 °C. The entire pipeline was automated using customized perl scripts using the Bioperl modules.

The synthesized oligos were standardized using the genomic DNA of *C. canephora* line, S.3334 as template. PCR amplifications were carried out in a 15 ml reaction volume consisting of genomic DNA (5 ng/µL), PCR Buffer (1X), $MgCl_2$ (2.5 mM), dNTPs (0.2 mM), 2 µM forward and reverse primers and 1U *Taq*DNA Polymerase. The PCR was programmed for initial denaturation at 94 °C for 5 min, followed by 35 cycles and each cycle consisting of denaturation at 94 °C for 1 min, gradient annealing temperature (± 5 °C of oligos Tm) for 30 sec, primer extension at 72 °C for 45 sec and a final extension of 8 min at 72 °C. The primers standardized for locus specific amplification were tested for polymorphism in the two parental lines of L1 Valley and S.3334. The products were resolved on 3 per cent agarose gel, to detect the polymorphism. The primers, monomorphic on the agarose gel, were resolved using 4.5 per cent poly acrylamide gel electophoresis (PAGE) and checked for polymorphism.

Results and discussion

From the enrichment experiment, 868 colonies were screened, 409 recombinant clones with varied fragment sizes ranging from 250 to 500 bp were selected among which 384 clones were sequenced with M/s Macrogen, Korea. Figure 1 depicts the agarose gel picture of the recombinant clones which were purified after PCR for sequencing reaction.

Of the 384 sequences, 267 sequences were found to contain a total of 405 SSRs. An account of 69.53 per cent of clones sent for sequencing were positive for SSRs and 38 per cent of the sequences were found to contain more than one SSR and 38.5 per cent of SSRs were found to be in compound formation (Table 1).

The 405 SSRs identified consisted of more of mono-nucleotide repeats (40.2%) followed by penta (33.3%), di (12.1%), tri (10.6%) and tetra (3.7%) nucleotide repeats (Table 2, Fig. 2).

In mono repeats, 'A' motif was most abundant (51%) followed by 'T' (33%), 'C' (22%) and 'G'(4%) repeats. In di repeats, TC motif formed the majority (24%) followed by CT (14%), GA (14%) and AG (12%) repeats. CG motif was only 2 per cent of di repeats. Amongst the trinucleotide repeats, TTG motif was found to be most abundant (7.16%). In the study, about 15 tetranucleotide repeats were

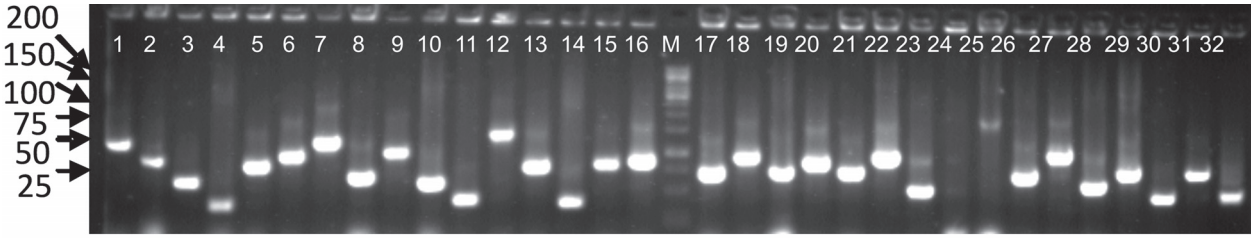

Fig. 1. Agarose gel (0.8%) depicting the colony PCR of the selected recombinant and transformed clones of 500- 1000 bp size

Table 1. Genomic sequences generated from the enrichment library

Library	No. of colonies screened	Colonies selected (250- 500bp)	No. of clones sequenced	Clones containing SSR repeats	Sequences containing more than one SSR	Total no. of repeats	SSRs in compound formation
Enrichment	868	409	384 (44.8% of recombinant clones)	267 (69.5% of sequenced clones)	102 (38.2% of total sequences with SSR)	405	156 (38.5% of total SSRs)

Table 2. Classification of genomic SSR repeats generated from the enrichment library

Repeat type	Total no.	% of total no. of repeats	No. >15 bp size	% of total no. of bp repeats
Mono	163	40.3	29	17.79
Di	49	12.1	19	38.78
Tri	43	10.6	18	41.86
Tetra	15	3.7	6	40.00
Penta	135	33.3	5	3.70
Total	405	100.0	77	

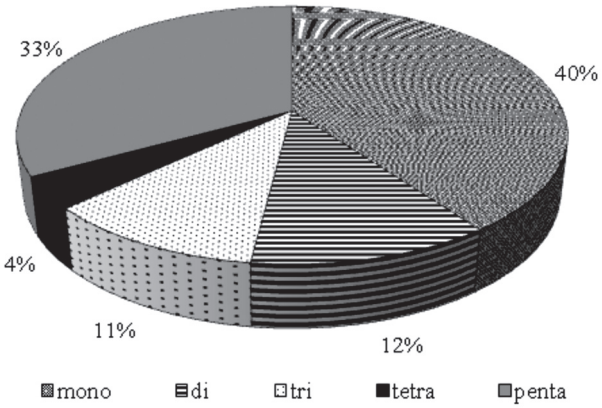

Fig. 2. Distribution of different repeat motifs of genomic SSRs developed by pre cloning enrichment strategy

isolated. Six repeats were more than 15 bp in size and TATG repeat motif was the most abundant repeat among tetranucleotides. About 135 penta repeat motifs were also isolated in the present enrichment experiment. Among the penta repeats, AAAAT motif was found to be most abundant.

Analysis of the 13,175 unigene sequences with MISA perlscripts (http://pgrc.ipk-gatersleben.de/misa/download/misa.pl), revealed 23,274 simple sequence repeat (SSR) motifs in 10,102 sequences. Of the 10,102 unigene sequences, 6270 sequences (47.59%) contained more than one SSR repeats and among the SSRs, 6380 (27.41%) were found to be in compound formation. Summary of the results obtained in the SSR search is given in Table 3. The success rate of discovering SSR containing sequences in unigenes or non-redundant ESTs in the present studyof coffee ESTs was about 77 per cent, which was relatively higher abundance of SSRs compared to the earlier reports in other crops, *viz*., grapes (Scott *et al.*, 2000), sugarcane (Cordeiro *et al.*, 2001) and cereals (Varshney *et al.*, 2002; Thiel *et al.*, 2003).

Amongst the genic SSRs identified, 10172 (44%) contained pentarepeat motifs followed by 5248 (23%) and 5242 (22%) sequences with hexa and monorepeat motifs respectively. Only 1285 (5%) sequences contained tri-nucleotide repeat motifs, while 824 (4%) sequences had di repeat motifs and 503 sequences (2%) had tetra repeat motifs. A pie chart depicting the distribution of the repeat motifs among the total SSRs identified is given in Figure 3.

Among the total mono repeats, 'A' motif was the most abundant (68%) followed by 'T' (24%), 'C' (5%) and 'G' (4%) repeat motifs. Among the di repeats, CT/GA motif formed the majority (19.5%),

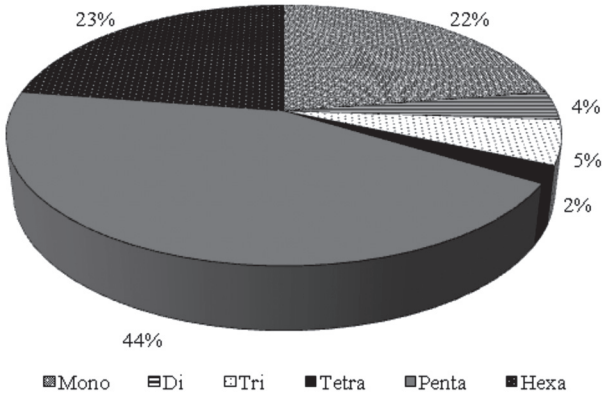

Fig. 3. Distribution of different repeat motifs in genic SSR identified from unigene sequences

Table 3. Genic sequences generated from the enrichment library

Sequences used	Total no. of sequences examined	SSR containing sequences	Sequences containing > 1 SSR	Total no. of repeats	No. of SSRs in compound formation
Unigene sequences	13175	10102 (76.68 % of total no. of sequences examined)	6270 (47.59 % of SSR containing sequences)	23274	6380 (27.41 % of total SSRs)

followed by TC/AG (15.4%), AG/TC (14.7%) and AT/TA (12.5%) repeats. CG/GC and GC/CG repeat motifs accounted to only 5 per cent each. The trinucleotide repeats were abundant in the following order: TTC/AAG>CAC/GTG>AAG/TTC>GAA/CTT>CCA/GGT>AGA/TCT repeats while ACA/TCT, ACG/TCG, CGA/GCT, TCG/AGC, GTT/CAA was found to be least in number. Poncet *et al.* (2006) reported similar results with GA/CT being the most abundant and CG motif as the least abundant di-repeat motif, while AGG/TCC (23%) and AAG/TTC (20.3%) were most abundant and AAC/TTG and AAT/TTA were the least abundant tri repeat motifs. Gao *et al.* (2003) and Aggarwal *et al.* (2007) reported more of GA repeats followed by AT and AC di-repeat motifs and AAG tri-repeat motif as the most abundant followed by ACT, ACC and AAT.

Further, about 503 tetra-nucleotide repeats were identified, of which, 80 were >15 bp in size. TTTC repeat motif was the most abundant tetra repeat followed by AAAT and AGAA. Amongst the 10,172 penta repeats documented, TTTTC>AAAAG>GAAAA were most abundant followed by CTTTT>TTTTG>TTCTT. Of the 5248 hexa repeat motifs, 153 were >15 bp in size. The order of abundance of hexa repeat motifs was AAAAAG>TTTTCT>TTTTTC>GAAAAA>CTTTTT. The present analysis of Unigene sequences revealed maximum number of penta repeat motifs, followed by hexa and mono repeat motifs (Fig. 3).

The SSR search criteria used for EST database mining can significantly alter the relative estimates of frequency/distribution of EST-SSRs, online with the earlier reports of Aggarwal *et al.* (2007) and Varshney *et al.* (2005) and hence there is a need for fixing a universally acceptable search criterion for identification of SSRs, in order to avoid wide variations in the per cent values obtained under each type of SSRs in different studies.

Sequences with repeat motifs of more than 15 base pairs (bp) were chosen for designing the primers. Of the total genomic SSRs detected, only about 20 per cent were more than 15 bp size. This could be because of the search criterion used to identify the SSR regions (minimum threshold of mono-10 repeat units, di-5, tri-4, tetra-3, penta-2, hexa-2 repeat units) as mentioned elsewhere in this paper. Further, among the 23,274 genic SSRs identified, around 3,991 SSRs (17%) were appropriate for primer designing. Most of the sequences unsuitable for marker development was due to the presence of SSR motif towards the end of the sequences and hence absence of enough nucleotides on the flanking regions of SSR motifs to design the primers. A comparison of the various types of genomic and genic SSRs detected and those with more than 15 bp size is depicted (Fig. 4 and 5).

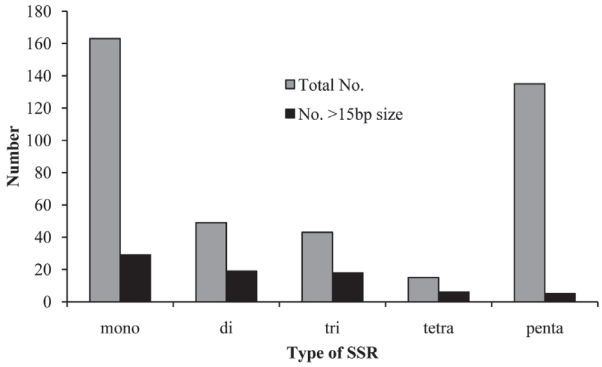

Fig. 4. Graph indicating the total number of genomic SSRs with >15bp detected under each type from the sequences screened

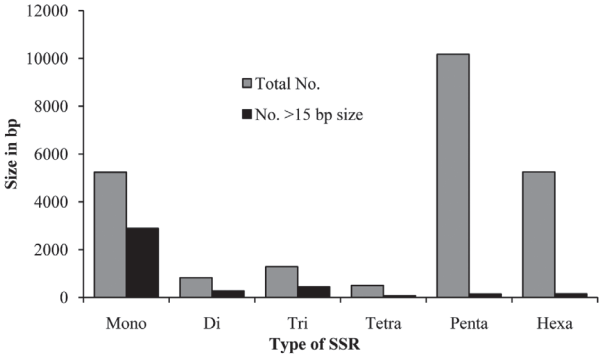

Fig. 5. Graph indicating the total number of genic SSRs with >15bp detected under each type from the Unigene sequences screened

Among the genomic SSRs screened, 77 sequences were selected for primer designing. For the genic SSRs, 1,097 sequences constituting di, tri, tetra, penta and hexa repeat motifs were considered for primer designing. Primers were successfully designed for 61 genomic and 783 genic repeat motifs. The 61 genomic and 100 genic primers were synthesized as desalted oligos from Sigma Aldrich, India.

About 31 genomic and 86 genic primers were standardized for locus specific amplification. Figure 6 depicts a representative gel of genic SSR markers standardized. A total of 117 SSR markers were submitted to the GenBank database and are available for public use. The details of the SSRs developed along with the sequence information and accession numbers are detailed in Table 4.

Table 4. Details of the genic (prefixed CCES) and genomic (prefixed CCGS) SSR markers developed in the study

Sl. No.	Primer Name	SSR Repeat	Primer Sequence		AT (°C)	Prod size	Gen Bank Accession No.
			Forward	Reverse			
1	CCES1	$(CTC)_8$	AGTCAGGTATGCTGCCATTG	AGGCAGCTAAATCAGCCAAG	51.7	254	GF1110660
2	CCES3	$(TC)_8$	GGATAGAATGCTTTGCAGCAG	TCTTTAACGGCGGAAAAC	51.7	129	GF1110744
3	CCES4	$(CCA)_{7a}(AAC)_6$	CAAAGCAGCTTCTTGGCTG	ACAACCTCGCCACCAAATAG	64.9	120	GF1110661
4	CCES5	$(GA)_{10}$	GCAGGCTTCACCGGTTTG	ACTGGGATGATGTCTGGCTC	61.7	270	GF1110662
5	CCES6	$(TCCAC)_5$	AGCGAAGCTGCGTTTATC	TTTTTATGCAAAATTGCTGAC	54.8	276	GF1110663
6	CCES9	$(CT)_9$	AGGATTTGTGCTTCCCTGTG	ACCAGCAAAAAGACGTAGGG	56.7	237	GF1110664
7	CCES10	$(TCT)_6$	GCTCTTCAGCAGCCAGAGA	ACCCATAAGACAAAAGGGGG	56.7	216	GF1110665
8	CCES12	$(CT)_{12}$	TCGGCTCCCAAATATTCATC	TTCTCAATGGTTTCGCATG	51.7	187	GF1110666
9	CCES13	$(AT)_{13}$	TTGCTTGAAAAATCAAGAGGC	TCTCTTCAAGTAAAATGTTTGA	51.7	251	GF1110667
10	CCES14	$(CCA)_6$	CACCACCACCTCCAGAG	GAGGGTAAGGAGGAGGAGC	64.9	141	GF1110668
11	CCES15	$(CT)_8$	CCTCCTAATAGTCCTTGCCTTG	ACGGAATTTTCCCGGACTAC	56.7	274	GF1110669
12	CCES16	$(AGC)_6$	ATCCCCATCAGAAGACCTC	GTTTACAATCGTCGAGCCG	59.9	226	GF1110670
13	CCES17	$(GCA)_6$	TTTCACTACTACCGGAGGCG	GCCGAGAAATCTGTCAGAGG	54.8	199	GF1110738
14	CCES18	$(CA)_8$	ACCCAAAACTGCCTTCC	TGTAGAAGCACCCGGAGAG	61.7	106	GF1110671
15	CCES19	$(AGA)_{10}$	AAACAGCAGGGAAAGCTG	AACCAAGGTCCATAAACCCC	61.7	189	GF1110672
16	CCES20	$(CT)_8$	GGAACGAACGTATCTCCTCG	GCTTGTAGCAGGCAGGAAAC	59.9	244	GF1110673
17	CCES21	$(AGGA)_8$	CTTCTGGAGTGAAAGCTTC	CCAATCGGGACACTCAC	51.7	202	GF1110674
18	CCES22	$(CT)_9$	AGGCTCTATCCTCCTCTCGG	GGTCCGACCTTGTTAGCGTA	64.9	235	GF1110675
19	CCES23	$(GAA)_6$	ATCATTGAAAGTTCAGCAGAC	TCCCAGTGATCTTCAGGGAC	64.9	237	GF1110676
20	CCES24	$(CCA)_6$	CTTCTCTCCAGACTGCCCAC	CAATCTCAATAACGGCAGC	64.9	267	GF1110677
21	CCES25	$(CT)_8$	TCACACCAAGCCTAGCACAC	ACCTCCCCGAGGACTTCTT	54.8	275	GF1110678
22	CCES26	$(GAG)_7$	AGGTTGTTCTTGTACTTGTTTG	CGGAGAGCATGATGAGTG	59.9	211	GF1110739
23	CCES27	$(GCA)_6$	ACAAAGCCTGTCGCTTG	TGCTGTAGTTGTGCCAGAGG	61.7	256	GF1110679
24	CCES28	$(AG)_9$	TTTCATGCCTCCATCTTTCC	GCAAAACTAATCTTTTCCAAG	54.8	164	GF1110680
25	CCES29	$(AC)_8$	TTATTGACGCAATCCCATG	TCCAGGCGTCTGTCTG	55	236	GF1110681
26	CCES31	$(TTA)_{6tt}(TTC)_6$	GTTTCAGGGTGTCGTTCG	TGGGAAGTCCAAAGAGGC	61.7	245	GF1110682
27	CCES32	$(CTCC)_5$	GTCGTCTGTTCCTCCTCGAC	CTAGGGTTTCGCAAGTCAGC	56.7	227	GF1110683
28	CCES33	$(AAATCA)_5$	GATGGATTCACCACCAAAGG	AAAAACATCAAGGGAGTGCG	64.9	247	GF1110684
29	CCES34	$(TG)_8$	CACGGTCGCACACTAACAG	CCACCAACAACCTTGC	61.7	265	GF1110685
30	CCES35	$(ATC)_8$	CAAGAGGTGCCTATCACCG	TTCTCGAGGACAATGGGAAC	57.9	127	GF1110686
31	CCES36	$(AAG)_8$	GCTGCTAAAGCATCCGAAAC	CTTGGTAAACCTCAACCCG	61.7	151	GF1110740
32	CCES37	$(TTA)_6$	CCACAAATCCTTCTTGCC	TGTCCCCAAGTAGGAAGC	59.9	271	GF1110687
33	CCES38	$(CCT)_6$	GCTGCCGTCAATAAATTGG	CTTGTCCTCAAAGGTAGCGG	46.8	192	GF1110688
34	CCES39	$(AAG)_6$	CAACACTCCTGACCAAGACG	CAGCCACAGGGTCTTC	46.8	236	GF1110689
35	CCES41	$(AAT)_8$	CAACAGAAAAGTTGGGGACG	TTATTCCCGCGTGGTAG	59.9	150	GF1110690
36	CCES42	$(AG)_{10}$	CCGTTACAGAATTTGCGG	AGCTCTGGTCGTTTCCAC	54.8	243	GF1110741
37	CCES43	$(TG)_9$	CAGCAACCAAACCTGCTG	GGGCATACATGAAAAAGCC	61.7	262	GF1110691
38	CCES44	$(GCA)_6$	GATAAAGAAAGAGGGGCTGG	TTGAAGTTGAGACGGCTGTG	64.9	100	GF1110692
39	CCES45	$(TGC)_6$	TCCCAAGATCCCTTTTGATG	TACCGCCATAACCAGACTCC	59.9	187	GF1110693
40	CCES46	$(ATG)_6$	TCGAGGAGTCTGGTTATGG	TCGTCAATAATTACATGGCAC	59.9	221	GF1110694
41	CCES48	$(GA)_{11}$	GCCGCGGTCAGTCTTACTAC	CACAAATCAACACCCATCCC	56.7	266	GF1110695
42	CCES49	$(TA)_8$	ACACCCCTTTGCTTGATGAC	TGGAGTTGAATTACATGAAGGTG	57.9	254	GF1110696
43	CCES50	$(TCA)_6$	GGCTCTTGGAGAGCTCAGG	CCCATCTGCTGACTCTGG	55	253	GF1110697
44	CCES51	$(AC)_{23}$	CGTTACCTAACCCTCCCTCC	AAGAGGGTTTTGCAGGGTC	59.9	210	GF1110698
45	CCES52	$(GGA)_6$	ATCTGGAGGAGGGGTTGTTC	GCTTCTCCAAGAATTGCTGC	61.7	153	GF1110699
46	CCES53	$(TTC)_6$	ATCCCGAATATGTTCTGCC	AAGAAAAACGGTGTTGCTGC	59.9	223	GF1110700
47	CCES54	$(CT)_{10}$	GGCACTGCTGCTTCTAGGAC	GGCTCCTTGTGTTTGGG	54.8	176	GF1110701
48	CCES55	$(CTC)_7$Gtctcctctg $(TTC)_7$	TCTCTTCCAATTCCAATGGC	TCCATAGTCCCCAAAAGCTG	61.7	126	GF1110702

49	CCES57	$(TA)_8$	ATTTGGTCAAACCTTCACCG	CCTTTTCCCTTTTCCCACTC	59.9	227	GF1110703
50	CCES58	$(AAC)_6$	GAGAAGCGGTAGGAGGG	TCTTCTCGAGGTCATCCTGG	59.9	207	GF1110704
51	CCES59	$(GAA)_6$	CGATAGATTCTACTCTTTGC	AGTGCAGGATTCTATGGCG	61.7	127	GF1110742
52	CCES60	$(TA)_{11}$	TTTTGCCCTTGAGGTAATGC	ACTTGAGCTGGCATTTTCAG	59.9	166	GF1110705
53	CCES61	$(CT)_8$	AGCTTTCACAACACACGCAC	CAGTTGGCAAGACATCAACG	57.9	107	GF1110706
54	CCES62	$(TG)_8$	TTTTCAAGTGTGGGCAATG	TTGGAGAAAACCCGGAG	59.9	236	GF1110707
55	CCES64	$(ACA)_6$	ACGAACGAACCAAAATCGAG	CCCGCTTGCAAAGTAATC	61.7	258	GF1110708
56	CCES65	$(CT)_{11}$	GAAGATACGAAAACGCGCTC	CTGGCCTCTCGAGTCTCAAC	57.9	116	GF1110709
57	CCES66	$(AAC)_6$	TCTTCCACTTTCATCGGTCC	AAGGGCCAAAGGACTCTCTC	54.8	257	GF1110710
58	CCES67	$(AT)_{12}$	TTGGCCAAAAGATGGACCT	TTCAAACAAGCTGCAAACC	59.9	245	GF1110711
59	CCES69	$(TCA)_9$	AAGTTGCAGGTTACGTTGGG	AGGAGGAACAGTTGGAGAATG	61.7	279	GF1110712
60	CCES71	$(GT)_8$	GCCTGCCACTAAACGATTTC	TGAGTGCCATTCCACTGTG	54.8	237	GF1110743
61	CCES72	$(TC)_{11}$	TGTGATCGACTTGGGATC	CATCGTCAACCAGGTAACC	56.7	241	Not submitted
62	CCES73	$(TTTG)_5$	TGGATCCTTTTAATGTTTAGTAG	CCTGCATTAATGGCAAACAC	54.8	138	GF1110713
63	CCES74	$(TC)_{11}$	GTTACCCATCGGGAGG	TCGACATCGATGAAAAGCAG	54.8	247	GF1110714
64	CCES75	$(TA)_8$	CTCCTGGCCTTTTTATTCCC	GCCGTTCTTGTCGATG	59.9	203	GF1110715
65	CCES76	$(TC)_{14}$	CCTGGTTTCTCTTGCCTTG	AGCGGTTTCAAGCTAACG	59.9	266	GF1110716
66	CCES77	$(CAT)_6$	CATCATCAGCACCTCCATTG	CCTTTTGAGGTGCCCAAC	54.8	246	GF1110717
67	CCES78	$(AG)_8$	CTTGGTTTCCGTCGAAAGAG	TTTGGGAAGGCAGTCAAAAC	56.7	246	GF1110718
68	CCES79	$(TTC)_9$	GCAATTCCTAGTTTTACACA	ACATAGTCACCGTTCCCTG	55	232	GF1110719
69	CCES80	$(GCG)_8$	TGCTTCCCCATTTATTGAGC	GGAGGAGCATAGCGTTTGAC	59.9	233	GF1110720
70	CCES81	$(CAG)_6$	CTCCACACCAACAAAATCCC	CCTGAGTCTGCTGCTAAGCC	64.9	162	GF1110721
71	CCES82	$(CAC)_6$	GTACACGGAGACATTGGGC	CACCTGCTTTTCCTTCAACC	59.9	203	GF1110722
72	CCES83	$(TCT)_7$	CGCTCTATCTCTGGTCGGTC	TCAGAGCCTGAAGACGAGG	61.7	151	GF1110723
73	CCES84	$(CGG)_8$	ATATCATGGATGGTGCTGCC	AGCAGCTGGAACTAGAACCG	46.8	263	GF1110724
74	CCES85	$(TTC)_6$	GGCACGAGGCTTTCTCTC	TGCCTATGGCACAAGTTTC	55	229	GF1110725
75	CCES86	$(GT)_{10}$	CCTTGATTGTCACGTGTATGC	GCACATTCATTAGGGAGG	54.8	156	GF1110726
76	CCES87	$(TCGT)_7$	AGGAGCAATTCGTTCATTCG	GTAGGTTGGTGGAGACTGGC	55	176	GF1110727
77	CCES89	$(TTTA)_5$	TTTTGTGCCAAAGGGAAGAG	GACCGGAGAAGTCATTGG	51.7	108	GF1110728
78	CCES90	$(AG)_{12}$	TATATTTTCCCGCATCCCAG	CTCGACATTGACCTCACACC	59.9	201	GF1110729
79	CCES91	$(TA)_9$	CGAACAATCGGCTCAAGTC	TAGAAATTTGCGAACGCTG	59.9	245	GF1110730
80	CCES92	$(CTT)_6$	AAGGTTCGAGCCAGGAAG	ATCCTTTGCCTTGGTTGATG	61.7	147	GF1110731
81	CCES93	$(TG)_8$	ACACTTTGCGGGAATCAATC	CAGGGTTAAACAACATTGG	54.8	135	GF1110732
82	CCES94	$(TTTTTA)_5$	ATTACCGGCCTGAGACACAC	CAATTCCGACGTTTCATCAG	64.9	272	GF1110733
83	CCES96	$(GAG)_6$	TGACGAGTACCATTGGGATG	CGCCGTAAACCTAACTGGAG	59.9	192	GF1110734
84	CCES97	$(ACC)_6$	CATCTTCCCGCTTTACCAG	GAGGATAGCCGTGGTTG	64.9	252	GF1110735
85	CCES99	$(AT)_{13}$	TGAGTGGAGAATCGAAGGG	TGGATAATTCTCCGGTGAC	51.7	193	GF1110736
86	CCES100	$(CT)_{10}$	CGGGCTGCAGAAACAAG	TCCCTCATCCTTTTCTCCC	51.7	273	GF1110737
87	CCGS6	$(A)_{16}$	ACCATAGATTTTAAACATGCATC	TGAATTTTCTTTGTCTTTTTACC	56.7	298	GF110629
88	CCGS7	$(A)_{16}$	TGGACACCTACCACAACC	CCAGGGCCTTCTTC	61.7	283	GF110630
89	CCGS9	$(T)_{17}$	GTTTACGCTCAAGGGGTTC	GGGGTGTTATTAAGGGGTTATG	61.7	298	GF110631
90	CCGS12	$(A)_{18}$	CACGCTAAGTGGTCACGC	ACACCAAGGAGTGGCCTTC	51.8	243	GF110632
91	CCGS13	$(A)_{19}$	AGCCAAGGATCTTCCC	GGCCTAGATAGCAGAATCGC	55	207	GF110633
92	CCGS15	$(A)_{20}$	GGTCATTCCAACTCAGCAAC	TTACAGGTTTGGGAATGGG	64.9	244	GF110634
93	CCGS16	$(T)_{20}$	CAAGAATCGCGTTCAAAAC	GCAAACGTTCCCAACCTAAG	56.7	320	GF110635
94	CCGS22	$(CT)_{10}$	AATCACCAAGTTTCCATGCC	GAAGATCACGAAATCCCAGC	64.9	267	GF110636
95	CCGS24	$(TC)_{10}$	CTAGCCTTGTCGTGCCAC	AATCGCCCGACATTTATCAG	64.9	227	GF110637
96	CCGS26	$(AT)_9$	TTCTGTATCCGGTGATGGG	TGGGTTGGAGTCGATTCTG	59.9	280	GF110638
97	CCGS27	$(AC)_{8.5}$	TCACGTTACAACCAACC	ATTTTCGCTAATGCTGGCAC	55	279	GF110639
98	CCGS28	$(CT)_{8.5}$	ACCATTCTGACACCTCTCGC	GGAGAAGAAGGCCAAAATCC	59.9	198	GF110658
99	CCGS30	$(CT)_8$	CAAGTGGTGATAGAAGCC	CATGGGTCGGTCTACTTCGT	54.8	223	GF110640
100	CCGS31	$(AG)_8$	CGTTCAACTTATTCATCCTCTGG	TCCTTGTTTTGCACTTTTGC	59.9	200	GF110641
101	CCGS32	$(CT)_8$	ACCGAATCAATCCACC	TGGACTTGGCCATTTTC	56.7	210	GF110642
102	CCGS33	$(TG)_8$	GTTTGAAGGGAGGGGGTTC	TCTCGATCGTTCAATTTTGC	59.9	245	GF110643
103	CCGS35	$(GA)_8$	GAGAAAATGAAACACATTTAGG	GCCTCGCCTAGCAGTCATAG	45.1	280	GF110644
104	CCGS36	$(TC)_8$	GGCTGAAGAGAGCAGGC	TTCAGCCCTTTGTGGAAG	54.7	242	GF110645
105	CCGS37	$(GAG)_{7.67}$	CCCATTGGGATGATGAG	ACATTTGACTTCCCGGG	51.8	280	GF110646
106	CCGS41	$(GAA)_{6.67}$	GCGACTCACTTGGAAGATGC	GCGAATCTTCTGCCCAGATAC	56.7	300	GF110647
107	CCGS42	$(GCT)_6$	TCCCATTCCTCAGTCAATCC	CCTCCAGTAAGCTCCTTGG	50	300	GF110648

108	CCGS45	$(TTG)_5$	CTTGGAGAATTTTCGGTGG	CCGAGGTGGGAGTTCATAG	54.8	193	GF110649
109	CCGS47	$(TCA)_5$	AATGGGACAGAGGCATTAGC	GGCAATGTTTGGTTTGATCC	64.9	280	GF110650
110	CCGS48	$(TGG)_5$	CCATCCACTTTGAACTCTCTC	AAAGAACATGTTGGTGGGATATG	64.9	274	GF110659
111	CCGS49	$(AAG)_5$	TTGCGGAGACAGAATGTTC	TTTGACGCTGGCTTTTTC	51.7	212	GF110651
112	CCGS50	$(TTG)_5$	CCAATTAACGTTGGTAGGCTG	TGCATACCACATCATTGAACG	54.8	250	GF110652
113	CCGS51	$(GAG)_5$	TGAGGGATTCTGAAAGAGC	TTTTTCCAAGCAAGCCAAC	56.7	269	GF110653
114	CCGS55	$(AGAA)_4$	TAAGCCTGCAGCGTGAAAAG	GGGCTTAGTCAGCATCAAG	64.9	210	GF110654
115	CCGS56	$(TC)_{10}$	TCAAAGGGGGAAAGGAG	CACAACACAAATTCGATCTTCC	56.7	292	GF110655
116	CCGS58	$(ATTGA)_{3.4}$	GGACCCCTTTCAATTCC	GCCTTTGAACTGAAGCATCC	50	222	GF110656
117	CCGS60	$(ATTAA)_3$	TCCCAATGAATCTTGCC	TATGACCCATGAATACGCC	42.2	260	GF110657

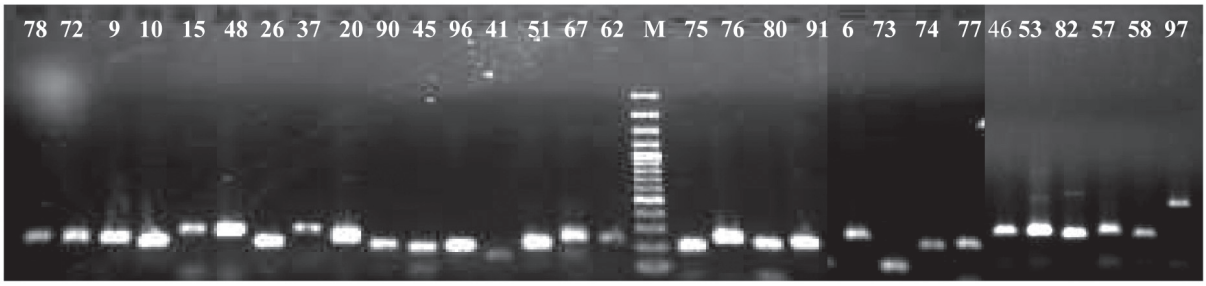

Fig. 6. Representative gel of standardised genic SSR primers (Numbers indicate the suffixed CCES primers) validated for locus specific amplification in S. 3334

The 31 genomic SSR primers and 86 genic SSR primers standardized for locus specific amplification were tested for polymorphism in two parental lines *viz*., L1 Valley and S.3334. Figure 7 depicts a representative gel of the primer products amplified in the two lines and compared on 3 per cent agarose gel.

genomic SSRs to be more polymorphic compared to genic or EST–SSRs in coffee. It was reported less polymorphism in genic SSR's because of the sequence conservation in the transcribed regions (Scott *et al*., 2000). However, the present study indicates more polymorphism in genic SSRs compared to genomic SSRs which might be due to

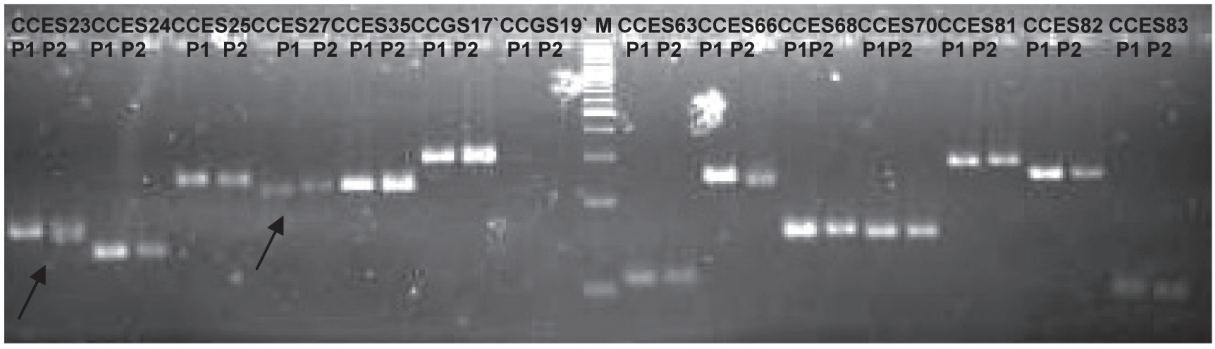

Fig. 7. Representative gel of standardized genic (CCES) and genomic (CCGS) showing polymorphismin the parental lines L1 Valley (P1) and S3334 (P2)

In the primer screening, of the 31 genomic SSR screened, three genomics SSRs, (CCGS 22, CCGS 35 and CCGS 41) were found to be polymorphic between the parental lines. Among the 86 genic SSRs screened, 22 were found to be polymorphic. Aggarwal *et al.* (2007) reported

the lesser number of genomic SSRs standardised and analysed for polymorphism (61) compared to that that of the genic SSRs analysed (86). Further, the polymorphism is with regard to only two lines and hence cannot be generalised across species. The information generated in the study offers further

scope of developing atleast around 170 genomic and another 400 genic SSRs. It is hoped the markers developed with prove beneficial to the community of molecular biologists in their efforts in identifying genes and trait specific markers.

Acknowledgements

The authors wish to acknowledge the Department of Biotechnology, New Delhi, India for the financial support in carrying out the study.

References

Aggarwal, R.K., Hendre, P.S., Varshney, R.K., Bhat, P.R., Krishnakumar, V. and Singh, L. 2007. Identification, characterization and utilization of EST-derived genic microsatellite markers for genome analyses of coffee and related species. *Theoretical and Applied Genetics* **114**(2): 359-72.

Armour, J.A., Neumann, R., Gobert, S. and Jeffreys A.J. 1994. Isolation of human simple repeat loci by hybridization selection. *Human Molecular Genetics* **3**: 599-565.

Bhat, P.R. 2002. Phenotyping and molecular analysis of *Coffea arabica* L. accessions and mapping population to identify RAPD markers associated with root characteristics and associated physiological traits. Ph.D thesis submitted to University of Agricultural Science, Bangalore.

Baruah, A., Naik, V., Hendre, P.S., Rajkumar, R., Rajendrakumar, P. and Aggarwal, K. 2003. Isolation and characterization of nine microsatellite markers from *Coffea arabica* L. showing wide cross-species amplification. *Molecular Ecology Notes* **3**: 647-650.

Combes, M.C., Andrzejewski, S., Anthony, F., Berthand, B., Rovellis, P., Graziosi, G. and Leshermes. P. 2000. Characterization of microsatellite loci in *Coffea arabica* and related coffee species. *Molecular Ecology* **9**: 1171-1193.

Cordeiro, G.M., Casu, R.E., McIntyre, C.L., Manners, J.M. and Henry, R.J. 2001. Microsatellite markers from sugarcane (*Saccharum spp.*) ESTs cross transferable to erianthus and sorghum. *Plant Science* **160**: 1115-1123.

Gao, L.F., Tang, J.F., Li, H.W. and Jia, J.Z. 2003. Analysis of microsatellites in major crops assessed by computational and experimental approaches. *Molecular Breeding* **12**: 245-261.

Gupta, P.K. and Varshney, R.K. 2000. The development and use of microsatellite markers for genetic analysis and plant breeding with emphasis on bread wheat. *Euphytica* **113**(3): 163-185.

Hendre, P.S., Phanindranath, R., Annapurna, V., Lalremruata, A. and Aggarwal, R.K. 2008. Development of new genomic microsatellite markers from robusta coffee (*Coffea canephora* Pierre ex A. Froehner) showing broad cross-species transferability and utility in genetic studies. *BMC Plant Biology* **8**: 51.

Karagyozov, L., Kalcheva, I.D. and Chapman, V.M. 1993. Construction of random small-insert genomic libraries highly enriched for simple sequence repeats. *Nucleic Acids Research* **21**: 3911-3912.

Kijas, J.M., Fowler, J.C., Garbett, C.A. and Thomas, M.R. 1994. Enrichment of microsatellites from the citrus genome using biotinylated oligonucleotide sequences bound to streptavidin-coated magnetic particles. *Biotechniques* **16**: 656-662.

Lashermers, P., Cross, P., Marmey and Charrier A. 1993. Use of random amplified DNA markers to analyse genetic variability and relationship of *Coffea* species. *Genetics Resources and Crop Evolution* **40**: 91-99.

Lashermes, P., Trouslot, P., Anthony, F., Combes, M.C. and Charrier, A. 1996. Genetic diversity for RAPD markers between cultivated and wild accessions of *Coffea arabica*. *Euphytica* **87**: 59-64.

Lin, C., Mueller, L.A., Carthy, J.M., Crouzillat, D., Petiard, V., Tanksley, S.D. 2005. Coffee and tomato share common gene repertoires as revealed by deep sequencing of seed and cherry transcripts. *Theoretical and Applied Genetics* **112**: 114-130.

Moncada, P. and McCouch, S. 2004. Simple sequence repeat diversity in diploid and tetraploid *Coffea* species. *Genome* **47**: 501-509.

Paillard, M., Duchateau, N. and Petiard, V. 1993. Diversite genetique de quelques groupes de cafeieres: utilization des outils moleculaires, RFLP et RAPD. 15 colloque. ASIC, Montpellier, France.

Poncet, V., Rondeau, M., Tranchant, C., Cayrel, A., Hamon, S., de Kochko, A. and Hamon, P. 2006. SSR mining in coffee tree EST databases: potential use of EST-SSRs as markers for the *Coffea* genus. *Molecular Genetics and Genomics* **276**(5): 436-449.

Powell, W., Morgante, M., Andre, C., Hanafey, M., Vogel, J., Tingey, S. and Rafalsky, A. 1996. The comparison of RFLP, RAPD, AFLP and SSR (microsatellite) markers for germplasm analysis. *Molecular Breeding* **2**: 225-238.

Rozen, S. and Skaletsky, H.J. 2000. Primer 3 on the WWW for general users and for biologist programmers. In: *Bioinformatics Methods and Protocols: Methods in Molecular Biology*. (Eds.) Krawetz, S. and Misener, S., Humana Press, Totowa, N.J. pp 365-386.

Scott, K.D., Eggler, P., Seaton, G., Rossetto, M., Ablett, E.M., Lee, L.S., Henry, R.J. 2000. Analysis of SSRs derived

from grape ESTs. *Theoretical and Applied Genetics* **100**: 723-726.

Thiel, T., Michalek, W., Varshney, R.K. and Graner, A. 2003. Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (*Hordeum vulgare* L.). *Theoretical and Applied Genetics* **106**: 411-422.

Varshney, R.K., Thiel, T., Stein, N., Langridge, P., Graner, A. 2002. *In silico* analysis on frequency and distribution of microsatellites in ESTs of some cereal species. *Cellular and Molecular Biology Letter* **7**(2A): 537-546.

Varshney, R.K., Balyan, H.S. and Langridge, P. 2005. Wheat. In: *The Genome: Cereals and Millets* (Ed.) Kole, C., Science Publishers, Inc., Enfield (NH), USA, pp. 121-219.

Wang, Z., Weber, J.L., Zhong, G. and Tanksley, S.D. 1994. Survey of plant short tandem DNA repeats. *Theoretical and Applied Genetics* **88**: 1-6.