Online journal ISSN - 1676-5680
GMR Genetics and Molecular Research
EVOLUTION AND TECHNOLOGY

Construction of a molecular database for soybean cultivar identification in Brazil

M.B. Oliveira[1], E.S.N. Vieira[1,2] and I. Schuster[1,2]

[1]Departamento de Biologia, Universidade Paranaense, Umuarama, PR, Brasil
[2]Núcleo de Biotecnologia, Cooperativa Central de Pesquisa Agrícola, COODETEC, Cascavel, PR, Brasil

Corresponding author: I. Schuster
E-mail: ivan@coodetec.com.br

Genet. Mol. Res. 9 (2): 705-720 (2010)
Received October 20, 2009
Accepted January 28, 2010
Published April 20, 2010
DOI 10.4238/vol9-2gmr706

ABSTRACT. The narrow genetic base of soybean makes cultivar characterization based on morphological descriptors difficult; this characterization is mainly done for registration and protection. Correct characterization of cultivars could be achieved through molecular markers, since the frequencies of each allele in the population are known. Consequently, we developed a molecular characterization method and initiated the construction of a molecular database for soybean cultivar identification. Thirty-two soybean cultivars were analyzed with 48 fluorescent-labeled microsatellite markers. The reactions were carried out in singleplex, and genotyping in

 ©FUNPEC-RP www.funpecrp.com.br

quadriplex, using a capillary electrophoresis system in an automated sequencer. Probabilities of random identity and probabilities of random identity exclusion were calculated through estimated allele frequencies. A characterization profile was considered when the probability of random identity exclusion was equal or superior to 99.9999%. All steps of the experiment were doubled, using two independent sets of the same cultivar to evaluate the reproducibility of the method. A set of 13 microsatellite markers identified all 32 cultivars with 99.9999% certainty. The method was efficient and precise, with high reproducibility for cultivar characterization. These data are the beginning of a molecular database for soybean, and they can be used for cultivar characterization for registration and protection purposes and for cultivar identification in cases of intellectual property enforcement.

Key words: *Glycine max*; Molecular characterization; Fingerprinting; Genotyping method; Exclusion probability; Random identity probability

INTRODUCTION

Soybean is one of the major agriculture commodities worldwide, and Brazil is the second largest producer, with 57 million tons produced on 21.7 million ha in 2009 (Conab, 2009). Adaptation of soybean to the wide variety of climates in Brazil, from latitude 32° South to latitude 4° North, is mainly due to breeding programs. Breeding programs for any species require large investments in research, which are recovered with the release of new cultivars and seed commercialization. In order to guarantee recovery of the investment, it is necessary to protect the cultivars. Consequently, various countries have been creating cultivar protection systems. In order to be protected, a cultivar is normally described by morphological describers; it needs to be homogeneous and stable, and distinguishable from any other cultivar. Because of the great number of available soybean cultivars and the low variability of morphological descriptors, their distinction becomes difficult. Molecular characterization of cultivars has the potential to guarantee precise discrimination and genetic identification (Garcia et al., 2007; Schuster et al., 2009b).

Molecular markers detect variation directly in the DNA sequences; they are not affected by genotype and environment interaction, and methods for their detection can be automatized (Ferreira and Grattaplagia, 1998; Alcântara Neto, 2001; Caixeta et al., 2009). Microsatellite markers or SSRs (single sequence repeats) are the most recommended markers for cultivar characterization because they are co-dominant and multiallelic.

Several studies have focused on soybean cultivar characterization using SSR markers (Song et al., 1999; Narvel et al., 2000; Garcia et al., 2007).

Capillary electrophoresis in an automatic DNA sequencer has been used for fragment analysis, allowing high precision and reliable results, which would be useful for cultivar characterization and for the protection of intellectual property (Diwan and Cregan, 1997). For precise cultivar characterization, it is necessary to identify a

set of informative markers and to know the frequencies of alleles of these markers (Schuster et al., 2009a).

We characterized a set of 32 soybean cultivars using microsatellite markers detected with an automatic sequencer, calculating the allelic frequencies of 48 microsatellite markers, in order to estimate the minimum number of loci for individual characterization of these 32 cultivars.

MATERIAL AND METHODS

Genetic material

A set of 32 soybean cultivars from the Cooperativa Central de Pesquisa Agrícola, COODETEC, were used. Two samples of 50 seeds from each cultivar were ground and the DNA extracted according to the protocol described by McDonald et al. (1994), with some modifications (Schuster et al., 2004). The two samples of each genotype were used as proof and counterproof samples. Proof and counterproof samples were independently processed, on different days, for DNA extraction, amplification, electrophoresis, and genotyping.

This procedure was carried out to evaluate reproducibility and to estimate the confidence interval for allele sizing.

Amplification of SSR loci and capillary electrophoresis

Forty-eight microsatellite markers, distributed on 18 of the 20 soybean chromosomes, were selected according to their informativeness, previously detected using agarose gels (Vieira et al., 2009; Table 1). Sense primers were labeled with 6-FAM, PET, VIC, and NED dyes. The sequences of the primers are available in the Soybase databank (http://soybase.org/index.php).

Polymerase chain reactions (PCR) were prepared for a total volume of 20 µL. The reaction mixture consisted of 30 ng DNA, 3 mM $MgCl_2$, 1X buffer (2 mM Tris and 5 mM KCl), 250 µM dNTP, 0.4 µM of each primer (sense and antisense) and one unit of Taq DNA polymerase. The amplifications were run in Thermo Hybaid thermocyclers (Ashford, Middlesex, UK) programmed for a cycle at 94°C for 3 min; 35 cycles at 94°C for 30 s, 55°C for 30 s and 72°C for 45 s, and one final extension step at 72°C for 20 min.

PCR was run in singleplex and capillary electrophoresis in multiplex. Multiplex consisted of a PCR fragment combination, obtained with different dyes, after amplification. Capillary electrophoresis was performed in an ABI3130xl automatic sequencer, according to manufacturer instructions. The samples were genotyped using the Gene Mapper version 4.0 software (Applied Biosystems).

Repeatability

Proof and counterproof genotyping results were compared. The difference between the same allele, in base pairs, in the two independent genotypings, their standard deviations and the confidence interval for the estimated allele sizes, were used as repeatability parameters for the genotyping system.

Genetic interpretation

The alleles were described in base pairs, in whole number approach/proximity. Proof and counterproof results were compared, and allele sizes for unity proximity were determined, considering the molecular nature of the microsatellite loci. For di-nucleotides, the minimum difference between the sizes of the alleles was two nucleotides, whereas for the tri-nucleotides this difference was three nucleotides. Based on these results, we constructed a database in which each cultivar was characterized by its allele for each locus.

Microsatellite marker informativity

Genetic informativity of each microsatellite locus was evaluated by determining the allele frequency, using the expression of polymorphism information content (PIC):

$$PIC = 1 - \sum_{j=1}^{n} p_{ij}^{2} \qquad \text{(Equation 1)}$$

where p_{ij} is the frequency of the jth allele of the ith primer (Anderson et al., 1993).

Marker selection for cultivar identification

A minimum marker set was selected to characterize each cultivar individually, and another marker set was used to characterize all cultivars simultaneously. In order to characterize each cultivar with the smallest number of markers, the selected markers were those that presented alleles with the lowest frequency in the cultivar. The probability of random identity (PRI) was calculated as described by Schuster et al. (2009a):

$$PRI = \left(\prod_{j=1}^{n} P_{ij} \right) x100 \qquad \text{(Equation 2)}$$

where P_{ij} is the frequency of the ith allele in the jth locus and n the number of evaluated loci. The product of the allele frequencies is multiplied by 100 so that it can be expressed as a percentage. The minimum number of markers for cultivar characterization was the number needed to obtain a random identity probability of at least 0.0001%, i.e., another cultivar can randomly present the same allele profile as the cultivar-specific markers set in less than 0.0001% of the cases.

Probability of exclusion (PE) was estimated as a complement of the PRI: PE = 100% - PRI.

Thus, if the molecular profile of a specific marker set in a cultivar has a probability of random identity of 0.0001%, the probability of exclusion will be 99.9999%. When this molecular profile is obtained in any pair of samples, it indicates the probability that this identity is not random and that the samples are the same cultivar.

RESULTS AND DISCUSSION

All 48 loci were polymorphic, as they were chosen from a preliminary study (Vieira et al., 2009). All steps of this study were doubled, with samples from two independent DNA extractions. The genotyping results of proof and counterproof were similar, demonstrating the accuracy and reproducibility of the genotyping method used in this study. The standard deviation values ranged from 0 to 0.93, and the confidence intervals for the allele size estimates ranged from 0.0003 to 0.04.

The differences observed between the allele sizes ranged from 0 to 1.32 nucleotides, with an average value of 0.22. These values are smaller than the minimum repetitive unity, which is two nucleotides for dinucleotide loci and three for trinucleotide loci (Table 1). Altogether, 1605 genotype data points were obtained from the evaluations (proof and counterproof) of 32 cultivars with 48 microsatellite loci. In this data set, only 15 genotyping data points presented a difference larger than 1 bp between two independent evaluations (0.93%).

Table 1. Microsatellite markers used to characterize 32 soybean cultivars, nature of microsatellite replication, primer marked fluorescence, and linkage group.

Marker[1]	Nature	Fluorescence	L.G.	Marker[1]	Nature	Fluorescence	L.G.
Sat_085	Di	6FAM	C1	Satt302	Tri	VIC	H
Sat_141	Di	6FAM	G	Satt303	Tri	NED	G
Sat_168	Di	VIC	G	Satt307	Tri	6FAM	C2
Sat_294	Di	NED	A2	Satt309	Tri	6FAM	G
Satt020	Tri	6FAM	B2	Satt311	Tri	NED	D2
Satt030	Tri	6FAM	F	Satt335	Tri	NED	F
Satt070	Tri	NED	B2	Satt352	Tri	NED	G
Satt079	Tri	VIC	C2	Satt358	Tri	PET	O
Satt080	Tri	PET	N	Satt371	Tri	PET	C2
Satt114	Tri	NED	F	Satt386	Tri	VIC	D2
Satt173	Tri	6FAM	O	Satt406	Tri	6FAM	J
Satt175	Tri	PET	M	Satt417	Tri	VIC	K
Satt177	Tri	PET	A2	Satt426	Tri	VIC	B1
Satt181	Tri	NED	H	Satt431	Tri	VIC	J
Satt184	Tri	PET	D1a	Satt464	Tri	PET	D2
Satt191	Tri	6FAM	G	Satt485	Tri	NED	N
Satt197	Tri	VIC	B1	Satt540	Tri	NED	M
Satt200	Tri	PET	A1	Satt545	Tri	6FAM	A1
Satt216	Tri	NED	D1b	Satt579	Tri	PET	D1b
Satt231	Tri	VIC	E	Satt600	Tri	VIC	D1b
Satt233	Tri	NED	A2	Satt663	Tri	VIC	F
Satt253	Tri	PET	H	Satt685	Tri	VIC	E
Satt285	Tri	NED	J	Satt703	Tri	VIC	D1b
Satt301	Tri	NED	D2	Satt728	Tri	NED	M

[1]Primer sequences are available at Soybase (http://soybase.org/index.php); Di = dinucleotide; Tri = trinucleotide; L.G. = linkage group: Source: Soybase (http://soybase.org/index.php).

Most of the variations between the genotyping repeats ranged from 0 and 0.2 bp, and 90% of the genotyping data had a variation smaller than 0.5 bp between two genotyping repeats (Figure 1). These results indicate high genotyping accuracy in the independent assays.

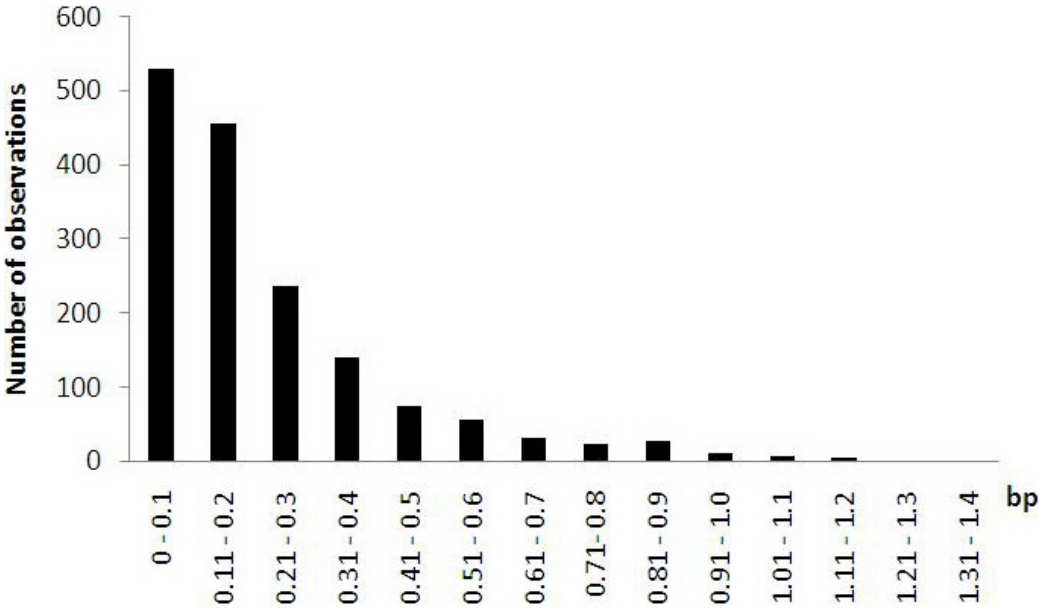

Figure 1. Frequency distribution of allele size differences, in base pairs, obtained from two independent genotyping of 32 soybean cultivars in 48 simple sequence repeat loci.

Soybean genotyping by fluorescent-labeled SSR with automated sizing of alleles was used for the first time by Diwan and Cregan (1997). Since then, there have been no published studies using fluorescent-labeled SSR and automated sizing to characterize soybean germplasm. Also, genotyping is not normally done in duplicate to check the precision of allele sizing. The results we obtained demonstrate high repeatability in the estimates of allele size at each locus. It is essential that a highly precise and reproducible genotyping system be used to build a molecular database for cultivar characterization. This precision in allele sizing cannot be obtained with genotyping based on agarose or acrylamide gel systems. This is the first time that a genotyping system using fluorescent-labeled molecular markers in a capillary gel system and automated sizing of alleles has been used to characterize soybean cultivars. Also, it is the first time that a genotyping system is evaluated for precision of sizing estimates of alleles.

Considering the proof and counterproof sample data and the nature of the microsatellite locus (di- or tri-nucleotide), a genotype for each cultivar was attributed, in base pairs, for each locus (Table 2). The data of this table constitute a reference database for comparison studies for genetic identity analyses. Furthermore, they are also a reference for the comparison of new cultivars and for genetic certification of seed lot origin. Also, the data in Table 2 represent the initial step for molecular database construction for soybean cultivars in Brazil.

Table 2. Allele size in base pairs observed in 32 soybean cultivars evaluated at 48 microsatellite loci, considering the di- or trinucleotide nature of simple sequence repeat loci.

Cultivar	Sat_085		Sat_141		Sat_168		Sat_294		Satt020		Satt030		Satt070		Satt079		Satt080		Satt114		Satt173		Satt175	
	Al. 1	Al. 2	Al. 1	Al. 2	Al. 1	Al. 2	Al. 1	Al. 2	Al. 1	Al. 2	Al. 1	Al. 2	Al. 1	Al. 2	Al. 1	Al. 2	Al. 1	Al. 2	Al. 1	Al. 2	Al. 1	Al. 2	Al. 1	Al. 2
CD 201	174		183		155		256		101		152		172		149		160		93		251		17	
CD 202	200		183		177		206		101		152		148		125		184		102		251		191	
CD 203	200		183		155		256		101		152		148		143		184		78		251		191	
CD 204	174		183		155		206		101		152		172		125		157		105		206		176	
CD 205	174		183		155		206		119		167		148		125		157		78		251		176	
CD 206	174	200	183		155		206		119		158		148		143		154		78		206		185	
CD 207	174		183	235	169		206		101		167		163		125		157		78		206		191	236
CD 208	200		183		155		206		101		152		172		149		160		93		251		176	
CD 209	174		183		155		206		119		167		148		149		157		93		206	251	176	
CD 210	174		203	235	169		206		101		158		163		149		184		78		251		185	
CD 211	174		183		155		206		101		152		172		125		157		78		263		176	
CD 212RR	174		183	205	155	169	256		101	119	158		148		149		154		105		206		176	
CD 213RR	174		205	235	169		186	206	119		158		148		149		181		93		206		161	
CD 214RR	174		183		155		256		101		152		148		149		160		105		251		176	
CD 215	200		183		177		206		101		152		148		125		184		93		251		191	
CD 216	174		183		155		256		119		167		148		125		157		78		197		167	
CD 217	174		211	235	155		190		101		149		172		125		184		78		206		176	
CD 218	200		183		177		190		125		158		148		125		184		102		206		191	
CD 219RR	174		183		155		206		101		152		172		149		157		93		251		176	
CDFAPA 220	174		183		155		206		101		167		163		143	149	154		78		251		185	
CD 221	200		183		169		222		119		158		148		149		154		78		206		191	
CD 222	174		183		155	169	206		101		167		163	172	125		157		105		206	263	176	191
CD 223AP	174		183		155		256		101		149		163		143		154		78		206		161	
CD 224	174		183		155		206		101		152		163		149		160		93		251		176	
CD 225RR	174		183		155		256		119		161		148		149		181		105		251		176	
CD 226RR	174		183		155		256		101		158		172		149		160		93		251		176	
CD 227	174		183		155		206		101		167		172		125		184		105		206	251	176	
CD 228	174		183		155		206		113		167		175		146		181		105		251		167	
CD 229RR	174		181	183	155		206		119		161		148		125		181		78		251		176	
CD 230RR	174		183		155		206		101		161		148		125		157		78		251		176	
CD 231RR	174		181		155		256		101		167		148		125		181		78		251		176	
CD 232	174		183		155		256		119		167		148		125		157		78		206		185	

Continued on next page

©FUNPEC-RP www.funpecrp.com.br

Table 2. Continued.

Cultivar	Satt177		Satt181		Satt184		Satt191		Satt197		Satt200		Satt216		Satt231		Satt233		Satt253		Satt285		Satt301	
	Al. 1	Al. 2	Al. 1	Al. 2	Al. 1	Al. 2	Al. 1	Al. 2	Al. 1	Al. 2	Al. 1	Al. 2	Al. 1	Al. 2	Al. 1	Al. 2	Al. 1	Al. 2	Al. 1	Al. 2	Al. 1	Al. 2	Al. 1	Al. 2
CD 201	122		207		186		225		182		228		192		226		199		152		204		244	
CD 202	113		198		186		228		185		246		222		226		199		152		240		199	
CD 203	113		207		171		228		185		246		156		226		199		152		240		262	
CD 204	107		198		150		225		185		228		156		220		199		137		204		199	
CD 205	110		216		141		207		134		228		192		226		187		155		240		244	
CD 206	113		177		150		225		185		228		156		220		187		137		204		259	
CD 207	110		216		141		207		188		246		192		220		187		155		240		244	
CD 208	122		207		186		228		182		228		192		226		199		152		204		244	
CD 209	110		177		150		207		134		228		192		220	226	187		137		240		244	
CD 210	110		216		141		225		182		228		192		220		208		137		240		247	
CD 211	122		198		150		207		185		246		192		220		199		137		204		244	
CD 212RR	113		207		141		228		188		246		168		226		187		137	152	204		199	
CD 213RR	113		207		150		228		188		228	246	222		220		187	208	137		204		199	
CD 214RR	122		207		150		225		182		228		192		226		199		152		204		244	247
CD 215	110		177		186		204		182		246		192		238		199		137	152	240		262	
CD 216	110		216		141		207	228	134		228		192		226		187		152		240		244	
CD 217	122		207		150		189		173		228		222		220		199		137		204		199	
CD 218	110		198		150		228		185		246		192		220	226	187		155		240		199	
CD 219RR	122		207		150		225		185		228		192		220		199		152		204		244	
CDFAPA 220	110		216		141		225		134		228		156		223		187		137		240		244	
CD 221	110		177		141		225		182		246		192		220		187		155		204		262	
CD 222	107		198	216	141		207	225	185	188	228	249	138		220		199		137	155	204	240	199	
CD 223AP	110		216		186		228		182		228		171		226		199		152		240		244	
CD 224	110		207		141		207		182		246		192		226		199		137		204	240	244	
CD 225RR	122		216		150		228		173		246		192		226		187	199	152		204		244	
CD 226RR	122		207		186		228		182		228		192		226		199		152		204		244	
CD 227	107		198		150		207		185		228		156		226		199		137		204		244	
CD 228	110		216		171		207		134		246		189		226		187		155		240		244	
CD 229RR	110		216		141		207		173		246		192	222	226		187		137		204		244	
CD 230RR	113		216		150		204		134		246		192		226		187		137		204		244	
CD 231RR	113		207		141		204		173		246		192		226		187		137		240		199	
CD 232	110	113	207		150		225		134		228		192		220		187		155		240		259	

Continued on next page

Table 2. Continued.

Cultivar	Satt302		Satt303		Satt307		Satt309		Satt311		Satt335		Satt352		Satt358		Satt371		Satt386		Satt406		Satt417	
	Al. 1	Al. 2	Al. 1	Al. 2	Al. 1	Al. 2	Al. 1	Al. 2	Al. 1	Al. 2	Al. 1	Al. 2	Al. 1	Al. 2	Al. 1	Al. 2	Al. 1	Al. 2	Al. 1	Al. 2	Al. 1	Al. 2	Al. 1	Al. 2
CD 201	257		237		163		130		187		159		182		161		251		199		242		325	
CD 202	206		255		172		124		199		159		191		203		275		196		242		325	
CD 203	257		255		163		124		199		150		182		203		251		196		323		325	
CD 204	206		255		172		130		232		159		191		194		275		166		242	323	283	
CD 205	206		246		184		130		232		165		185		194		275		199		242		325	
CD 206	206		255		172		130		232		150		191		194		275		166		323		283	
CD 207	206		237		184		124		232		150		182		194		275		166		242		325	
CD 208	257		237		163		130		187		159		182		161		251		199		242		325	
CD 209	206		246		184		130		232		165		185		194		275		199		242		325	
CD 210	206		255		184		124		187		150		191		194		251		166		242		325	
CD 211	206		255		172		130		232		165		191		194		275		199		242		325	
CD 212RR	257		237		172		124	130	187		150	165	182		194		275		166	199	242		325	
CD 213RR	206	257	237		172		124		187		165		182		194		275		199		242	323	325	
CD 214RR	257		237		163		130		187	199	150		182		161		251		166		242		325	
CD 215	206		255		172		124		232		150		191		194		275		199		242		325	
CD 216	257		246		163	184	130		232		165		185		194		251		199		242		325	
CD 217	206		222		184		133		232		159		167		203		260		199		245		283	
CD 218	206		255		172		124		232		159		191		203		275		196		242		325	
CD 219RR	206		255		172		130		232		159		191		194		275		166		242		325	
CDFAPA 220	206		237		172		130		232		159		182		194		275		199		242		325	
CD 221	206		255		184		124		232		150		191		194		251		199		242		325	
CD 222	206		237		184		124	130	232		159		182		194		275		166	199	242	245	283	
CD 223AP	257		246		163		130		232		150		185		194		251		196		242		283	
CD 224	206		246		172		130		232		159		185		161		251		166	199	245		325	
CD 225RR	257		237		163		130		199		159		182		194		275		166		326		325	
CD 226RR	257		237		172		130		199		159		182		161		275		199		242		325	
CD 227	206		255		184		130		187		159		191		194		254		166		242		283	
CD 228	206		246		163		145		232		165		185		194		275		199		242		325	
CD 229RR	206		237	246	184		145		232		159		167	194	194		275		199		323		325	
CD 230RR	206		237		184		130		232		150		194		194		275		199		242		325	
CD 231RR	257		237		184		145		232		159		194		194		275		199		242		325	
CD 232	257		237		184		130		232		150		182		194		275		199		242		325	

Continued on next page

Table 2. Continued.

Cultivar	Satt426		Satt431		Satt464		Satt485		Satt540		Satt545		Satt579		Satt600		Satt663		Satt685		Satt703		Satt728	
	Al. 1	Al. 2	Al. 1	Al. 2	Al. 1	Al. 2	Al. 1	Al. 2	Al. 1	Al. 2	Al. 1	Al. 2	Al. 1	Al. 2	Al. 1	Al. 2	Al. 1	Al. 2	Al. 1	Al. 2	Al. 1	Al. 2	Al. 1	Al. 2
CD 201	198		187		219		264		148		191		198		203	215	252		185		229	235	149	
CD 202	198		199		189	219	252		154		203		198		215		252		185	215	229		194	
CD 203	198		199		219		252		154		203		174		155		249		185		229		194	
CD 204	198		232		189		240		169		191		174		155		252		218		199		191	
CD 205	219		232		219		264		154		206		174		155		249		185		229		149	
CD 206	198	201	232		219		240	264	169		191		174		155		252		215		199		191	
CD 207	198		232		219		240		148		203		198		215		249		215		229		194	
CD 208	198		187		219		264		148		191		198		215		252		185		229		149	
CD 209	219		232		219		264		154		203		174		155		249		185		229		149	
CD 210	198		187		219		264		148		188		174		155		249		215		199		194	
CD 211	198		232		219		240		169		203		174		155		249		185		199		191	
CD 212RR	198		187	232	189		264		154		203		198		215		213	249	215		229		149	191
CD 213RR	198		187		189		240		148		203		198		215		213		215		229		191	
CD 214RR	198		187	199	219		264		166		191		198		215		252		215		229		149	
CD 215	198		232		219		240		154		203		198		215		213		215		229		194	
CD 216	219		232		219		264		154		203		192		203		249		215		229		149	
CD 217	198		232		189		240		166		191		174		203		249		218		229		194	
CD 218	198		232		189		252		154		203		198		215		252		185		229		149	
CD 219RR	198		232		219		240		169		191		174		155		252		185		199		149	
CDFAPA 220	219		232		219		264		154		203		174		155		249		185		229		191	
CD 221	198		232		219		264		169		188		198		215		249		215		199		194	
CD 222	198		232		189		240		148	169	191	203	174	198	155	215	249	252	215		229		149	
CD 223AP	201		232		219		264		154		203		174		203		249		185		199	229	149	
CD 224	198		232		219		264		148		203		198		215		249		185		229		149	
CD 225RR	201		199		219		264		166		203		192		203		249		215		235		149	
CD 226RR	198		199		219		264		148		203		198		215		252		185		229		149	
CD 227	198		187		219		252		169		191		174		155		252		215		229		191	
CD 228	201		232		219		264		154		203		174		155		252		215		235		149	
CD 229RR	201		232		189	219	264		154		203		174	198	203		249		185		229		149	194
CD 230RR	219		232		219		264		154		203		198		203		249		185		229		194	
CD 231RR	201		232		189		264		154		203		174		155		249		185		229		149	
CD 232	219		232		219		264		169		203		174		155		249		215		199		191	

Al. 1 = allele 1; Al. 2 = allele 2.

Using this set of SSR loci, characterized by the methodology used in this study, other cultivars can be added, enriching the database.

Several studies have been published revealing genetic diversity and germplasm characterization of soybean by molecular markers, such as RFLP (Keim et al., 1989, 1992), RAPD (Abdelnoor et al., 1995), AFLP (Bonato et al., 2006a,b), and SSR (Priolli et al., 2002, Yamanaka et al., 2007). None of them gave individual characterization (fingerprinting) of the cultivars. Knowledge of a molecular profile from the cultivars that we evaluated will allow the use of these data in other studies; this database can be increased with new data from other studies that use the same methodology.

In the set of 48 loci evaluated in the 32 soybean cultivar samples, 178 alleles were observed, ranging from two to seven alleles per locus, with a mean of 3.71. PIC values varied from 0.30 (Satt417) to 0.78 (Satt080), for a mean of 0.57 (Table 3). Only 11 of the 48 loci presented PIC values lower than 0.5. These values are relatively high, considering the number of samples and the fact that the cultivars came from the same breeding program. If a greater number of samples with greater genetic diversity were to be analyzed, the probability of detecting other alleles would increase, increasing the genetic informativity of each locus.

Narvel et al. (2000), evaluating the genetic diversity of 39 elite soybean cultivars and 40 plant introductions (PI) with 74 microsatellite markers, obtained PIC estimates ranging from 0.02 to 0.84 for all genotypes (mean of 0.56), 0 to 0.84 for PI (mean of 0.56) and 0 to 0.79 for elite cultivars (mean of 0.50). The number of alleles per locus varied from 2 to 11 for all genotypes (mean of 5.4), from 1 to 10 for the PI (mean of 4.9) and from 1 to 8 for elite cultivars (mean of 3.5). Song et al. (1999) used 48 microsatellite markers to characterize 101 soybean cultivars. PIC values ranged from 0.59 to 0.83 with four alleles per locus.

Priolli et al. (2002), evaluating a set of 186 Brazilian soybean cultivars with 12 SSR markers, obtained values of gene diversity, which is equivalent to PIC in autogamous species, from 0.41 to 0.82. In this set of 12 SSR, they found 62 alleles, a mean of five alleles per locus. Yamanaka et al. (2007), evaluating 272 soybean cultivars from Brazil, China and Japan, with 12 SSR markers, obtained PIC values from 0.22 to 0.84, with a mean of seven alleles per locus. All these studies used a representative germplasm set with potentially high genetic diversity. In our study, we used only soybean cultivars obtained from one breeding program, yet the values for genetic diversity were only slightly lower than from those obtained from apparently diverse germplasms. This shows that cultivated varieties of soybean obtained by a single breeding program can have a similar diversity to that found in all cultivated cultivars.

Significant allele diversity was found among the cultivars, even though the frequency of some alleles was high at some loci. Information about the allele frequencies at each locus (Table 3) allows calculations of probabilities of random identity and probabilities of random identity exclusion, indicating if two samples have the same genotype or not (Schuster et al., 2009a). This information can be used in cases where there is no distinction based on morphological descriptors, in registration processes and cultivar protection. The PRI for a cultivar is the product of the frequency of the alleles present in this cultivar (Schuster et al., 2009a). For this reason, it is necessary to know the frequency of the alleles in a reference population in order to calculate the PRI.

Few studies present the allelic frequency of evaluated populations. Priolli et al. (2002) reported the allelic frequencies of 12 SSR loci for 186 Brazilian soybean cultivars. However, they did not identify the alleles, and consequently the information about allelic frequency cannot be used to estimate PRI. Schuster et al. (2009b) presented the allelic frequency for 23 SSR loci in 32 Brazilian wheat cultivars. For each allele, a cultivar that contains this allele

Table 3. Number of alleles, allele frequencies and polymorphism information content (PIC) estimated for 48 microsatellite loci, obtained from the genetic profiles of 32 samples of soybean cultivars.

Marker	Nº of alleles	Allele	Frequency	PIC	Marker	Nº of alleles	Allele	Frequency	PIC
Satt216	7	138	0.03	0.59	Sat_141	6	181	0.05	0.31
		156	0.16				183	0.83	
		168	0.03				203	0.02	
		171	0.03				205	0.03	
		189	0.03				211	0.02	
		192	0.61				235	0.06	
		222	0.11						
Satt175	6	161	0.06	0.64	Sat_294	5	186	0.02	0.56
		167	0.06				190	0.06	
		176	0.55				206	0.58	
		185	0.13				222	0.03	
		191	0.19				256	0.31	
		236	0.02						
Satt030	5	149	0.06	0.74	Satt080	5	154	0.16	0.78
		152	0.31				157	0.31	
		158	0.22				160	0.16	
		161	0.09				181	0.16	
		167	0.31				184	0.22	
Satt191	5	189	0.03	0.73	Satt197	5	134	0.22	0.77
		204	0.09				173	0.13	
		207	0.28				182	0.28	
		225	0.30				185	0.27	
		228	0.30				188	0.11	
Satt301	5	199	0.25	0.62	Satt352	5	167	0.05	0.72
		244	0.55				182	0.38	
		247	0.05				185	0.19	
		259	0.06				191	0.31	
		262	0.09				194	0.08	
Satt020	4	101	0.64	0.50	Satt070	4	148	0.53	0.62
		113	0.03				163	0.17	
		119	0.30				172	0.27	
		125	0.03				175	0.03	
Satt079	4	125	0.47	0.61	Satt114	4	78	0.47	0.67
		143	0.11				93	0.25	
		146	0.03				102	0.06	
		149	0.39				105	0.22	
Satt173	4	197	0.03	0.55	Satt177	4	107	0.09	0.70
		206	0.36				110	0.42	
		251	0.56				113	0.23	
		263	0.05				122	0.25	
Satt181	4	177	0.13	0.71	Satt184	4	141	0.34	0.68
		198	0.17				150	0.41	
		207	0.38				171	0.06	
		216	0.33				186	0.19	
Satt231	4	220	0.38	0.54	Satt303	4	222	0.03	0.66
		223	0.03				237	0.42	
		226	0.56				246	0.20	
		238	0.03				255	0.34	
Satt309	4	124	0.28	0.56	Satt371	4	251	0.28	0.49
		130	0.59				254	0.03	
		133	0.03				260	0.03	
		145	0.09				275	0.66	
Satt406	4	242	0.77	0.39	Satt540	4	148	0.23	0.69
		245	0.08				154	0.44	
		323	0.13				166	0.09	
		326	0.03				169	0.23	
Satt545	4	188	0.06	0.51	Sat_168	3	155	0.75	0.40
		191	0.27				169	0.16	
		203	0.64				177	0.09	
		206	0.03						
Satt200	3	228	0.53	0.51	Satt233	3	187	0.47	0.54
		246	0.45				199	0.48	
		249	0.02				208	0.05	

Continued on next page

Table 3. Continued.

Marker	Nº of alleles	Allele	Frequency	PIC	Marker	Nº of alleles	Allele	Frequency	PIC
Satt253	3	137	0.45	0.64	Satt307	3	163	0.23	0.65
		152	0.34				172	0.38	
		155	0.20				184	0.39	
Satt311	3	187	0.20	0.51	Satt335	3	150	0.33	0.63
		199	0.14				159	0.47	
		232	0.66				165	0.20	
Satt358	3	161	0.16	0.44	Satt386	3	166	0.30	0.56
		194	0.72				196	0.13	
		203	0.13				199	0.58	
Satt426	3	198	0.64	0.52	Satt431	3	187	0.19	0.49
		201	0.17				199	0.14	
		219	0.19				232	0.67	
Satt485	3	240	0.27	0.54	Satt579	3	174	0.50	0.55
		252	0.13				192	0.06	
		264	0.61				198	0.44	
Satt600	3	155	0.42	0.64	Satt663	3	213	0.08	0.55
		203	0.20				249	0.56	
		215	0.38				252	0.36	
Satt685	3	185	0.48	0.56	Satt703	3	199	0.23	0.47
		215	0.45				229	0.69	
		218	0.06				235	0.08	
Satt728	3	149	0.50	0.62	Sat_085	2	174	0.80	0.32
		191	0.23				200	0.20	
		194	0.27						
Satt285	2	204	0.53	0.50	Satt302	2	206	0.64	0.46
		240	0.47				257	0.36	
Satt417	2	283	0.19	0.30	Satt464	2	189	0.25	0.38
		325	0.81				219	0.75	

was presented as a reference cultivar. In this case, using one reference cultivar for each allele, it is possible to test, in an independent study, which allele is present in a cultivar that was not evaluated in the original study, and use allele frequency to obtain PRI. In the above publications, the allele size was not identified, because genotyping was made on acrylamide gels. In this type of genotyping system, precise determination of allele size is not possible because it can change from one gel to another, or when samples are from different experiments or different labs. In our study, the genotyping system was highly reproducible, permitting characterization based on the length of the amplified fragment, in base pairs. Therefore, information about allelic frequency can be used in other assays of cultivar characterization. One can obtain the genetic profile of any cultivar based on those we examined with SSR markers; using the allelic frequencies shown in Table 3, an estimated PRI can be calculated for each cultivar.

As soybean is an autogamous species, it is expected that all plants of a cultivar will be homozygotes. However, some cultivars presented two alleles at some loci. The presence of two alleles in the same cultivar characterizes a mixture of pure lines. Although in these cases the frequency of each allele in each cultivar was not estimated, two alleles with the same proportion was considered for the calculation of allele frequencies. This procedure must be considered because, in a case of genetic identity investigation, the presence of any of the two alleles cannot discard the identity hypothesis, regardless of its frequency.

At several loci, rare alleles (low frequency) were observed. In these cases, this information should be used in a conservative manner, changing the frequencies of these rare alleles to 5/2n, where n is the total number of evaluated cultivars (National Research Council, 1996). Thus, all the frequencies with estimates lower than 0.08 were increased to 0.08 (N = 32).

Using the information on allele frequencies, it was possible to identify a minimum set of markers to characterize each of the cultivars and select markers to characterize all cultivars simultaneously (Table 4). In cases in which a specific cultivar presented more than one allele per locus, the frequencies of both alleles were added to calculate the probability of random identity. We obtained a value of less than 0.0001% probability of random identity for all cultivars. The minimum number of markers to obtain this probability ranged from 6 to 11 for each cultivar, and a set of 13 markers was selected for the simultaneous characterization of the 32 cultivars (Table 4).

Table 4. Minimum set of microsatellite markers selected to characterize the 32 evaluated soybean cultivars, allele frequencies and probability of random identity (PRI).

Cultivar	SATT080	SATT197	SATT030	SATT191	SATT352	SATT181	SATT540	SATT184	SAT_294	SATT177	SATT114	SATT303	SATT307	PRI[1]
CD 201	0.16[1]	0.28	0.31	0.30	0.38	0.38	0.23	0.19	0.31	0.25	0.25	0.42	0.23	<0.0001%
CD 202	0.22	0.27	0.31	0.30	0.31	0.17	0.44	0.19	0.58	0.23	0.08	0.34	0.38	<0.0001%
CD 203	0.22	0.27	0.31	0.30	0.38	0.38	0.44	0.08	0.31	0.23	0.47	0.34	0.23	<0.0001%
CD 204	0.31	0.27	0.31	0.30	0.31	0.17	0.23	0.41	0.58	0.09	0.22	0.34	0.38	<0.0001%
CD 205	0.31	0.22	0.31	0.28	0.19	0.33	0.44	0.34	0.58	0.42	0.47	0.20	0.39	<0.0001%
CD 206	0.16	0.27	0.22	0.30	0.31	0.13	0.23	0.41	0.58	0.23	0.47	0.34	0.38	<0.0001%
CD 207	0.31	0.11	0.31	0.28	0.38	0.33	0.23	0.34	0.58	0.42	0.47	0.42	0.39	0.0001%
CD 208	0.16	0.28	0.31	0.30	0.38	0.38	0.23	0.19	0.58	0.25	0.25	0.42	0.23	<0.0001%
CD 209	0.31	0.22	0.31	0.28	0.19	0.13	0.44	0.41	0.58	0.42	0.25	0.20	0.39	<0.0001%
CD 210	0.22	0.28	0.22	0.30	0.31	0.33	0.23	0.34	0.58	0.42	0.47	0.34	0.39	<0.0001%
CD 211	0.31	0.27	0.31	0.28	0.31	0.17	0.23	0.41	0.58	0.25	0.47	0.34	0.38	<0.0001%
CD 212RR	0.16	0.11	0.22	0.30	0.38	0.38	0.44	0.34	0.31	0.23	0.22	0.42	0.38	<0.0001%
CD 213RR	0.16	0.11	0.22	0.30	0.38	0.38	0.23	0.41	0.66	0.23	0.25	0.42	0.38	<0.0001%
CD 214RR	0.16	0.28	0.31	0.30	0.38	0.38	0.09	0.41	0.31	0.25	0.22	0.42	0.23	<0.0001%
CD 215	0.22	0.28	0.31	0.09	0.31	0.13	0.44	0.19	0.58	0.42	0.25	0.34	0.38	<0.0001%
CD 216	0.31	0.22	0.31	0.58	0.19	0.33	0.44	0.34	0.31	0.42	0.47	0.20	0.62	0.0001%
CD 217	0.22	0.13	0.08	0.08	0.08	0.38	0.09	0.41	0.08	0.25	0.47	0.08	0.39	<0.0001%
CD 218	0.22	0.27	0.22	0.30	0.31	0.17	0.44	0.41	0.08	0.42	0.08	0.34	0.38	<0.0001%
CD 219RR	0.31	0.27	0.31	0.30	0.31	0.38	0.23	0.41	0.58	0.25	0.25	0.34	0.38	<0.0001%
CDFAPA 220	0.16	0.22	0.31	0.30	0.38	0.33	0.44	0.34	0.58	0.42	0.47	0.42	0.38	0.0001%
CD 221	0.16	0.28	0.22	0.30	0.31	0.13	0.23	0.34	0.08	0.42	0.47	0.34	0.39	<0.0001%
CD 222	0.31	0.38	0.31	0.58	0.38	0.50	0.46	0.34	0.58	0.09	0.22	0.42	0.39	0.0001%
CD 223AP	0.16	0.28	0.08	0.30	0.19	0.33	0.44	0.19	0.31	0.42	0.47	0.20	0.23	<0.0001%
CD 224	0.16	0.28	0.31	0.28	0.19	0.38	0.23	0.34	0.58	0.42	0.25	0.20	0.38	<0.0001%
CD 225RR	0.16	0.13	0.09	0.30	0.38	0.33	0.09	0.41	0.31	0.25	0.22	0.42	0.23	<0.0001%
CD 226RR	0.16	0.28	0.22	0.30	0.38	0.38	0.23	0.19	0.31	0.25	0.25	0.42	0.38	<0.0001%
CD 227	0.22	0.27	0.31	0.28	0.31	0.17	0.23	0.41	0.58	0.09	0.22	0.34	0.39	<0.0001%
CD 228	0.16	0.22	0.31	0.28	0.19	0.33	0.44	0.08	0.58	0.42	0.22	0.20	0.23	<0.0001%
CD 229RR	0.16	0.13	0.09	0.28	0.16	0.33	0.44	0.34	0.58	0.42	0.47	0.62	0.39	<0.0001%
CD 230RR	0.31	0.22	0.09	0.09	0.08	0.33	0.44	0.41	0.58	0.23	0.47	0.42	0.39	<0.0001%
CD 231RR	0.16	0.13	0.31	0.09	0.08	0.38	0.44	0.34	0.31	0.23	0.47	0.42	0.39	<0.0001%
CD 232	0.31	0.22	0.31	0.30	0.38	0.38	0.23	0.41	0.31	0.65	0.47	0.42	0.39	0.0001%

[1]Table data correspond to the allele frequencies of alleles shown in Table 2.

Garcia et al. (2007) selected a set of 10 loci with high PIC from 69 tested microsatellite loci and used them to identify 32 Brazilian soybean genotypes. Song et al. (1999) identified 66 lines of American elite soybeans, selecting a set of 13 microsatellite loci of 48 markers. These 13 loci were used to characterize four elite cultivars with the same maturity and morphological traits; they were able to distinguish all cultivars. In both cases, the researchers were only interested in differentiating the test cultivars, i.e., a single difference among the cultivars was enough for its differentiation from the others.

In our study, the objective was to identify a set of SSR loci that could identify cultivars with 99.9999% probability of random identity exclusion, based on allele frequencies. Besides not having any cultivar with the same molecular profile among the evaluated cultivars, the

probability of finding another variety with the same molecular profile among non-evaluated cultivars would be less than 0.0001%. The set of markers indicated in Table 4 guarantees that for each of the evaluated cultivars other cultivars with the same genetic profile will not be found, with a minimum probability of 99.9999%. This set of markers can be used in cases of intellectual property protection and for genetic purity certification of these cultivars.

In Brazil, the intellectual property of the cultivar's owners is established by the Plant Variety Protection (PVP), granted by SNPC (Serviço Nacional de Proteção de Cultivares). While not providing an official registration/patent, the PVP offers a plant cultivar's owner legal protection for exclusive sale of a protected cultivar. In the case of non-authorized use of a protected cultivar, it is necessary to provide evidence for the genetic identity of the improperly used cultivars. This evidence can be provided easily and precisely through PRI. If an unknown soybean cultivar is evaluated by some of the markers that we used in this study, and it has the same alleles as a known cultivar at all loci, with PRI 0.0001% or less, this assures that the two cultivars (known and unknown) are the same cultivar, with 99.9999% probability or more. This molecular information can be used in judicial enforcement of PVP rights.

The construction of a molecular database for soybean cultivar characterization has thus been initiated. The method that we used was efficient, accurate and showed high reproducibility for this purpose. Construction and expansion of this database can have great impact for combating illegal use of seeds and for intellectual property protection. To include new cultivars in the present database, it is recommended that one of the cultivars used in this study be used in each PCR plate, working as a reference for the precision allele sizing. This is the first study done constructing a molecular database for soybean characterization in Brazil. This molecular database needs to be completed with information for other cultivars; this could be shared with many sectors interested in using this information, including breeding programs, seed producers, SNPC, the justice system, etc.

In Brazil, the SNPC began a program to establish a trustworthy, precise and reproducible genotyping method to be used in soybean cultivar characterization for cultivar rights protection. The method that we used here can be recommended for this purpose, because it meets all the requirements.

ACKNOWLEDGMENTS

The authors thank CAPES for a fellowship granted to M.B. Oliveira and COODETEC for financial support to carry out the research.

REFERENCES

Abdelnoor RV, Barros EG and Moreira MA (1995). Determination of genetic diversity within Brazilian soybean germplasm using random amplified polymorphic DNA techniques and comparative analysis with pedigree data. *Rev. Bras. Genet.* 18: 265-273.

Alcântara Neto F (2001). Marcadores Microssatélites na Identificação de Cultivares de Soja. Master's thesis, Universidade Federal de Viçosa, Viçosa.

Anderson JA, Churchill GA, Autrique JE, Tanksley SD, et al. (1993). Optimizing parental selection for genetic linkage maps. *Genome* 36: 181-186.

Bonato ALV, Calvo ES, Geraldi IO and Arias CAA (2006a). Genetic similarity among soybean (*Glycine max* (L) Merrill) cultivars released in Brazil using AFLP markers. *Genet. Mol. Biol.* 29: 692-704.

Bonato ALV, Calvo ES, Arias CAA, de Ferraz Toledo JF, et al. (2006b). Prediction of genetic variability through AFLP-

based measure of genetic distance in soybean. *Crop Breed. Appl. Technol.* 6: 30-39.
Caixeta ET, Oliveira ACB, Brito GG and Sakiyama NS (2009). Tipos de Marcadores Moleculares. In: Marcadores Moleculares (Borém A and Caixeta ET, eds.). 2nd edn. Folha de Viçosa, Viçosa, 11-94.
Companhia Nacional de Abastecimento (Conab) (2009). Acompanhamento da Safra Brasileira: Grãos - Safra 2008/2009 - Quarto Levantamento - Janeiro 2009. In: Companhia Nacional de Abastecimento. Conab, Brasília. Available at [http://www.conab.gov.br/conabweb/download/safra/3graos_08.09.pdf]. Accessed January 23, 2009.
Diwan N and Cregan PB (1997). Automated sizing of fluorescent-labeled simple sequence repeat (SSR) markers to assay genetic variation in soybean. *Theor. Appl. Genet.* 95: 723-733.
Ferreira ME and Grattapaglia D (1998). Introdução ao Uso de Marcadores Moleculares em Análise Genética. 3rd edn. Embrapa and Cenargen, Brasília.
Garcia AF, Alberini JL, Zucchi MI and Souza AP (2007). Microsatellite molecular markers in the cultivar identification of Brazilian soybean for human consumption. *Crop Breed. Appl. Biotechnol.* 7: 155-164.
Keim P, Shoemaker RC and Palmer RG (1989). Restriction fragment length polymorphism diversity in soybean. *Theor. Appl. Genet.* 77: 786-792.
Keim P, Beavis W, Schupp J and Freestone R (1992). Evaluation of soybean RFLP marker diversity in adapted germplasm. *Theor. Appl. Genet.* 85: 202-212.
McDonald MB, Elliot LJ and Sweeney PM (1994). DNA extraction from dry seeds for RAPD analyses in varietal identification studies. *Seed Sci. Technol.* 22: 171-176.
Narvel JM, Fehr WR, Chu WC, Grant D, et al. (2000). Simple sequence repeat diversity among soybean plant introductions and elite genotypes. *Crop Sci.* 40: 1452-1458.
National Research Council (NRC) (1996). The Evaluation of Forensic DNA Evidence. Committee on DNA Forensic Science: an Update. National Academy Press, Washington, DC.
Priolli RHG, Mendes-Júnior CT, Arantes NU and Contel EPB (2002). Characterization of Brazilian soybean cultivars using microsatellite markers. *Genet. Mol. Biol.* 25: 185-193.
Schuster I, Queiroz VT, Teixeira AI, Barros EG, et al. (2004). Determination of genetic purity of soybean seeds with the aid of microsatellite molecular markers. *Pesq. Agropec. Bras.* 39: 247-253.
Schuster I, Vieira ESN and Padilha L (2009a). Marcadores Moleculares no Pós-Melhoramento. In: Marcadores Moleculares. 2nd edn. Folha de Viçosa, Viçosa, 177-208.
Schuster I, Vieira ESN, Silva GJ, Franco FA, et al. (2009b). Genetic variability in Brazilian wheat cultivars assessed by microsatellite markers. *Genet. Mol. Biol.* 32: 557-563.
Song QJ, Quigley CV, Nelson RL, Carter TE, et al. (1999). A selected set of trinucleotide simple sequence repeat markers for soybean cultivar identification. *Plant Var. Seeds* 12: 207-220.
Vieira ESN, Schuster I, Silva RB and Oliveira MAR (2009). Genetic variability of soybean cultivars obtained through microsatellite markers in agarose gel. *Pesq. Agropec. Bras.* 44: 1460-1466.
Yamanaka N, Sato H, Yang Z, Xu DH, et al. (2007). Genetic relationships between Chinese, Japanese, and Brazilian soybean gene pools revealed by simple sequence repeat (SSR) markers. *Genet. Mol. Biol.* 30: 85-88.