Plant Varieties and Seeds (1999) **12**, 207–220

A selected set of trinucleotide simple sequence repeat markers for soybean cultivar identification

Q.J. SONG[1], C.V. QUIGLEY[1], R.L. NELSON[2], T.E. CARTER[3], H.R. BOERMA[4], J.L. STRACHAN[5] *and* P.B. CREGAN[1]

[1]*Soybean and Alfalfa Research Laboratory, Bldg. 006, Rm. 100, USDA, ARS, BARC-West, Beltsville, MD 20705 USA;* [2]*USDA, ARS, Department of Crop Sciences, University of Illinois, Urbana, IL 61801 USA;* [3]*USDA, ARS, North Carolina State University, Raleigh, NC 27695 USA;* [4]*Department of Crop Sciences, University of Georgia, Athens, GA 30602 USA, and* [5]*Plant Variety Protection Office, National Agricultural Library Building, USDA, Agricultural Marketing Service, Beltsville, MD 20705 USA*

Soybean [Glycine max *(L.) Merr.] cultivars are described for purposes of Plant Variety Protection (PVP) by standard pigmentation and morphological traits. However, many commercial soybeans arise from a limited number of elite lines and are often indistinguishable based on these traits. A system based on DNA markers could provide unique DNA profiles or fingerprints of cultivars. Simple Sequence Repeat (SSR) or microsatellite allele size profiling is used in human forensics to provide unique DNA fingerprints of an individual. Allele sizing technologies are well established and can be readily used to size SSR alleles from any organism. The purpose of the work presented here was to select and evaluate a small set of trinucleotide SSR markers with maximum reliability and repeatability that would provide a high level of discriminatory power to distinguish soybean genotypes. A total of 48 fluorescently labelled SSR primer sets was used to amplify genomic DNA of the 35 ancestors of N. American soybeans as well as a diverse group of elite N. American soybean cultivars. Only loci with allele size ranges that showed no overlap in size over a series of analyses and in which adjacent alleles differed by at least three basepairs were maintained for further statistical analysis via a clustering procedure. Cluster analysis was performed on the remaining loci and resulted in the identification of a subset of 13 loci, from 12 different linkage groups, that easily produced unique SSR allele size profiles for each of the 66 elite N. American soybean cultivars. This set of 13 loci was used to characterise four independent sets of elite cultivars that were selected based upon identical maturity, morphological, and pigmentation traits. Based upon these analyses, all cultivars could be distinguished using the set of 13 selected SSR loci. This set of loci is proposed as a standard set for use in DNA profiling of soybean cultivars for purposes of obtaining PVP.*

INTRODUCTION

The U. S. Plant Variety Protection Act (PVP Act) of 1970, amended 1994, was established to enable plant breeders to obtain intellectual property rights similar to patents for their crop cultivars, and thus protect them from commercial exploitation

by others. Unlike patent protection, the PVP Act includes an explicit research exemption which allows specific use of protected varieties for research and further plant breeding. Nonetheless, the PVP Act offers the owner of a plant cultivar 20 years of legal protection for the exclusive sale and reproduction of a protected cultivar. The Plant Variety Protection Office (PVPO) which is part of the Agricultural Marketing Service, U.S. Department of Agriculture, is charged with determining whether new crop cultivars qualify for protection, and issuing the PVP certificates when applicable. To qualify for PVP, a new cultivar must be demonstrated to be distinct from all other varieties within that species, and to be genetically uniform and stable within commercially acceptable limits.

When PVP is applied for, based upon the information supplied by the applicant, a computer driven comparative search is made to determine whether the cultivar is distinct, uniform, and stable. Over the past several decades, combinations of morphological data, resistance to diseases, resistance to insect pests and nematodes, physiological data, herbicide reaction, transgenic characteristics, and genetic markers have all been used to demonstrate distinctness. These various descriptors are listed in Table 1. The comparative search is undertaken using STAR, a database management program (Cuadra Associates, Inc., Los Angeles, CA). As the number of soybean cultivars in the PVP database has increased, the number of combinations of characteristics available to uniquely describe a new cultivar has decreased. A new cultivar must be distinguishable from the more than 1200 cultivars for which PVP certificates have already been issued, as well as the 1900 additional cultivars described in the PVPO database. The usual outcome of the comparative searches range from the candidate cultivar being uniquely described, to instances in which a candidate is indistinguishable from as many as 15 entries in the database. If a candidate is indistinguishable from others in the database, the PVP applicant is required to provide additional information that may be used to distinguish his/her new variety. This additional information may take the form of grow-out trials in which the candidate cultivar is grown in the field adjacent to cultivars from which it cannot be distinguished. The circumstances whereby an applicant has had to perform additional grow-out trials to demonstrate distinctness have increased in recent years. Molecular genetic markers provide an alternative to the standard morphological, pigmentation, disease resistance, and other characteristics used for defining the uniqueness of a new soybean cultivar, thereby relieving the soybean breeder of at least some of the burdens associated with expensive and time-consuming grow-out trials. Thus, the candidate cultivar and those cultivars from which it is indistinguishable can first be assayed with molecular genetic markers to determine if the candidate is unique. Using this approach, the candidate can be distinguished from some or all similar cultivars thereby reducing the size of grow-out trials or eliminating such testing altogether.

Simple Sequence Repeat (SSR) DNA markers, also known as microsatellites or Short Tandem Repeats (STR), have been used successfully by human geneticists for parentage testing, forensic identification and medical diagnostics (Edwards *et al.* 1992; Alford *et al.* 1994; and Hammond *et al.* 1994). In these studies, SSRs were found to provide accurate, reliable and rapid testing for these purposes. Likewise, in plants, SSR markers have been used for the purpose of genotype identification in species such as soybean [*Glycine max* (L.) Merr.] (Cregan *et al.* 1994; Diwan & Cregan 1997; Rongwen *et al.* 1995; Maughan *et al.* 1995), grape (*Vitis vinifera* L.) (Thomas & Scott 1993), rapeseed (*Brassica napus* L.) (Kresovich *et al.* 1995), apple (Hokanson *et al.* 1998), and many other species. In soybean, high levels of SSR allele length polymorphism have been

Table 1. Morphological, disease and nematode reaction, physiological responses, insect reactions, herbicide reactions, transgenic characteristics, and genetic markers that may be used in the description of soybean cultivars for purposes of distinguishing a candidate cultivar from those in the database of the Plant Variety Protection Office, Agricultural Marketing Service, U.S. Department of Agriculture

Morphological features	*Disease reactions*			*Reaction to nematodes*	*Physiological responses*	*Insect reactions*	*Herbicide reactions*	*Transgenic composition*	*Genetic markers*
	Bacterial	*Fungal*	*Viral*						
Seed shape	Bacterial pustule	Brown spot	Yellow mosaic	Soybean cyst nematode	Iron chlorosis on calcareous soil	Mexican bean beetle	Metribuzin	Identity of vector	Isozymes
Seed coat color	Bacterial blight	Frogeye leaf spot	Cowpea mosaic	Lance nematode	Phosphorus	Potato leaf hopper	Bentazone	Vector components	Restriction fragment length polymorphisms
Seed coat luster	Wildfire blight	Target spot	Bean pod mottle	Root knot nematode	Boron		Sulfonylurea	Transferred genetic material	Simple sequence repeats
Seed size in grams/100 seeds		Downey mildew	Soybean mosaic	Peanut root knot nematode	Aluminum		Glyphosate		
Hilum color		Powdery mildew		Reniform nematode			Glufosinate		
Cotyledon color		Brown stem rot					Pendimethalin		
Seed protein peroxidase activity		Stem canker							
Hypocotyl color		Pod and stem blight							
Leaf shape		Purple seed stain							
Flower color		Rhizoctonia root rot							
Pod color		Phytophthora root rot							
Pubescence color									
Plant growth habit									

reported. Akkaya *et al.* (1992) reported as many as eight alleles per SSR locus among a group of 43 soybean and wild soybean (*G. soja* Sieb. & Zucc.) genotypes. Morgante & Olivieri (1993) found similar levels of allelic diversity in soybean. Rongwen *et al.* (1995) defined allelic diversity at seven SSR loci in a group of 96 soybean genotypes. Maughan *et al.* (1995) detected a total of 79 alleles at five SSR loci in a sample of 94 accessions of cultivated and wild soybean. Diwan & Cregan (1997) determined allele numbers and gene diversity of di- and trinucleotide loci among the 35 ancestors of North American cultivars using fluorescently labelled polymerase chain reaction products whose sizes were determined on an ABI Prism 373A DNA sequencer. A total of 20 SSR loci was used in this study including 13 trinucleotide and 7 dinucleotide loci. Diwan & Cregan (1997) concluded that the 'stuttering' associated with the dinucleotide loci made it difficult to determine the main peak of the allele upon which the size was established. This was much less of a problem with trinucleotide SSRs and therefore, the authors recommended the use of trinucleotide loci for genotype identification. Thus, it was the objective of the work reported here to evaluate a group of trinucleotide loci and to select a subset that is maximally effective at providing unique SSR allele size profiles among a group of elite North American soybean cultivars.

MATERIAL AND METHODS

Soybean plant material and DNA isolation

Ancestors of North American Soybean Cultivars: Based upon pedigree analysis, Gizlice *et al.* (1994) identified a group of 35 genotypes to represent more than 95% of the allelic variation present in North American cultivated soybean germplasm (Table 2). Seeds of each of the 35 ancestors were obtained from the USDA Soybean Germplasm Collection courtesy of Dr. Randall Nelson (USDA-ARS, Univ. of Illinois, Urbana, IL).

Elite North American Cultivars: A group of 66 elite cultivars, developed and released by public institutions in the U.S. and Canada, was selected to represent the complete range of cultivars grown in the U.S. and Canada. Seeds of each of the 66 cultivars were also obtained from the USDA Soybean Germplasm Collection. Four additional sets of elite cultivars were selected from the PVPO, USDA, soybean database in an effort to assemble groups of cultivars with similar morphological, pigmentation, and growth habit characteristics. Ten Maturity Group (MG) I, seven MG II, 10 MG IV, and nine MG VI cultivars were identified and seeds were obtained from the public or private developer of each as described by Diwan & Cregan (1997).

DNA Isolation: DNA was extracted from a bulked leaf tissue of 30 to 50 plants of each of the 35 ancestral cultivars as well as each of the 66 elite cultivars and the various sets of morphologically similar cultivars by the method described by Keim *et al.* (1988).

Simple Sequence Repeat allele sizing

Forward primers of 48 trinucleotide SSR loci in four sets were labelled with either blue (FAM), green (HEX), or yellow (NED) fluorescent tags (AB-PEC, Foster City, CA) (Table 3). In most cases, the PCR primer sequences were the same as those available on the World Wide Web in Soybase, the USDA, ARS Soybean Genome Database (http://129.186.26.94/SSR.html). However, in a number of instances primers were re-selected to provide allele size ranges that were more amenable to multiplex analysis. The reselected primer sequences are listed in Table 4. PCR reaction mixes contained 30 ng of soybean genomic DNA, 0.15 μM of 3′ and 5′ end primers, 200 μM of each nucleotide, 1X PCR Buffer containing 50 mM KCl, 10 mM Tris-HCl pH 9.0, 0.1% Triton X-100, 1 unit Taq DNA polymerase

Table 2. Thirty-five ancestral cultivars of North American soybeans and 66 elite North American cultivars (Maturity Group in parentheses) assayed to define allele size, number of alleles, and gene diversity using soybean simple sequence repeat markers

Ancestral cultivars				
FC 31745 (VI)	Bilomi 3 (X)	Flambeau (00)	Korean (II)	Peking (IV)
PI 71506 (IV)	Capital (0)	Haberlandt (VI)	Lincoln (III)	Perry (IV)
PI 88788 (III)	CNS (VII)	Illini (III)	Mandarin (Ottawa) (0)	Ralsoy (VI)
AK Harrow (III)	Dunfield (III)	Improved Pelican (VIII)	Manitoba Brown (00)	Richland (II)
Anderson (IV)	Fisk840 (00)	Jackson (VII)	Mejiro (IV)	Roanoke (VII)
Arksoy (VI)	Fiskeby III (00)	Jogun (III)	Mukden (II)	S-100 (V)
Bansei (II)	Fiskeby V (000)	Kanro (II)	Ogden (VI)	Strain # 18 (0)
Elite Cultivars				
Agassiz (0)	Dillon (VI)	Iroquois (III)	Maple Ridge (00)	Savoy (II)
Bay (V)	Evans (0)	Johnston (VIII)	McCall (00)	Sibley (I)
Benning (VII)	Gail (VI)	Kershaw (VI)	Narow (V)	Sprite (III)
Braxton (VII)	Gasoy 17 (VII)	KS4694 (IV)	OAC Aries (0)	Sturdy (I)
Brim (VI)	Glacier (00)	Lambert (0)	OAC Libra (0)	Thomas (VII)
BSR201 (II)	Glenwood (0)	Lawrence (IV)	OAC Musca (0)	TN4–86 (IV)
Burlison (II)	Gordon (VII)	Lloyd (VI)	OAC Pisces (0)	Toano (V)
Century (II)	Graham (V)	Logan (III)	Ozzie (0)	Weber (I)
Cisne (IV)	Hack (II)	Macon (III)	Parker (I)	Young (VI)
CN290 (II)	Harlon (I)	Manokin (IVS)	Pennyrile (IV)	Zane (III)
Conrad (II)	Haskell (VII)	Maple Donovan (0)	Perrin (VIII)	
Cook (VIII)	Hoyt (II)	Maple Glen (00)	Pershing (IV)	
Dassel (0)	Hutchson (V)	Maple Isle (00)	Preston (II)	
Dawson (0)	IA2021 (II)	Maple Presto (000)	Ripley (IV)	

in a total volume of 10 μL. The Mg^{++} concentration (Table 3) in the PCR cocktail for each locus was optimized to give clearly defined amplification products with minimal production of spurious products. Cycling consisted of 25 sec denaturation at 94°C, 25 sec annealing at 46°C, and 25 sec extension at 68°C for 32 cycles followed by an additional 7 min extension at 68°C on a MJ Research model PTC-100 thermocycler (MJ Research, Inc., Watertown, MA).

Amplification products from different SSR loci within a set that carry the same fluorescent label can be simultaneously analysed in the same gel lane because loci were selected to avoid overlapping allele size ranges. Therefore, after cycling 1.5 μL of the four FAM labelled, 2.0 μL of the four HEX labelled and 2.0 μL of the four NED labelled PCR products within a set were combined and brought to a total volume of 20 μL. A total of 0.5 μL of fluorescent ROX (red) labelled internal size standard, and 0.5 μL of a loading buffer (AB-PEC, Foster City, CA) were added to a 4.0 μL aliquot of the combined PCR products. The sample of the combined PCR products was loaded and separated on an ABI Prism 373A DNA sequencer (AB-PEC, Foster City, CA). GeneScan 672™ software (AB-PEC, Foster City, CA) was used for gel analysis. The Local Southern option in GeneScan 672™ was used for computation of allele size, and Genotyper™ software (AB-PEC, Foster City, CA) was applied for accurate visualization of the alleles, and for automated data output.

Selection of SSR loci that provided unambiguous allele size data

The electrophoregram and allele size data produced by the Genotyper software for each data point (48 loci × 101 genotypes) were reviewed to ensure that each allele size

Table 3. Four sets of 12 soybean simple sequence repeat loci selected to cover the 20 soybean linkage groups, the approximate allele size range of each locus in the ancestors of North American cultivars, the fluorescent dye conjugated to the forward primer of each primer pair, the optimal Mg^{++} concentration used in polymerase chain reaction amplification of each locus, and the size of aliquot from the PCR reaction used in the analysis cocktail. Each set of 12 loci contain four loci labeled with each fluorescent dye. Primers within a set were in some cases reselected to avoid overlapping of the allele size ranges of loci labeled with the same fluorophore

Locus	*Linkage group*	*Approximate allele size range in the 35 N. American ancestral cultivars (bp)*	*Fluorescent dye label*	*Mg^{++} Conc. (mM)*	*Size of aliquot from PCR used in analysis cocktail (µl)*
Set 1					
Satt002	D2	133–152	HEX	1.5	1.5
Satt141	D1b + W	149–202	NED	1.5	0.3
Satt143	L	262–306	HEX	2.25	3.0
Satt175	M	156–189	HEX	1.5	1.5
Satt177	A2	107–122	NED	1.5	0.3
Satt180	C1	215–269	NED	1.5	0.5
Satt196	K	178–202	FAM	1.5	1.5
Satt226	D2	290–342	NED	1.5	0.5
Satt231	E	216–243	HEX	1.5	3.0
Satt236	A1	215–227	FAM	1.5	0.5
Satt253	H	137–153	FAM	1.5	0.5
Satt294	C1	249–294	FAM	1.5	0.5
Set 2					
Satt009	N	163–244	NED	2.25	2.0
Satt038	G	151–187	FAM	2.25	1.0
Satt114	F	82–121	HEX	1.5	2.0
Satt173	O	199–262	FAM	1.5	1.0
Satt184	D1a + Q	141–184	HEX	1.5	2.0
Satt243	O	260–290	NED	3.0	5.0
Satt276	A1	263–395	HEX	3.0	5.0
Satt281	C2	270–344	FAM	1.5	3.0
Satt324	G	202–241	HEX	1.5	5.0
Satt353	H	96–135	FAM	1.5	1.0
Satt358	O	316–366	NED	3.0	1.5
Satt373	L	89–161	NED	3.0	1.0
Set 3					
Satt146	F	286–328	NED	1.5	0.5
Satt168	B2	201–236	HEX	1.5	1.0
Satt172	D1b+W	221–236	FAM	1.5	0.5
Satt197	B1	135–189	HEX	1.5	1.0
Satt249	J	217–259	NED	1.5	0.5
Satt307	C2	118–139	NED	1.5	1.0
Satt308	M	134–174	FAM	1.0	0.75
Satt409	A2	246–283	HEX	1.5	2.0
Satt431	J	185–227	NED	1.5	1.0
Satt434	H	308–350	HEX	1.5	2.0
Satt441	K	243–311	FAM	1.5	0.5
Satt577	B2	101–122	FAM	1.5	0.75

Continued

Table 3. Continued

Locus	*Linkage group*	*Approximate allele size range in the 35 N. American ancestral cultivars (bp)*	*Fluorescent dye label*	*Mg^{++} Conc. (mM)*	*Size of aliquot from PCR used in analysis cocktail (µl)*
Set 4					
Satt147	D1a+Q	173–216	NED	1.5	1.0
Satt191	G	191–233	FAM	1.5	0.3
Satt194	C1	334–369	FAM	2.25	1.0
Satt242	K	119–162	NED	1.5	2.0
Satt259	O	128–154	FAM	1.5	0.3
Satt354	I	325–359	NED	1.5	2.0
Satt357	C2	215–322	HEX	1.5	0.3
Satt414	J	259–319	FAM	1.5	0.3
Satt440	I	161–212	HEX	1.5	3.0
Satt453	B1	349–392	HEX	0.75	3.0
Satt534	B2	224–260	NED	1.5	1.0
Satt588	K	134–169	HEX	1.5	1.0

Table 4. Primers to SSR loci used in this study that differ from those reported on the World Wide Web in Soybase, the USDA, ARS Soybean Genome Database (http://129.186.26.94/SSR.html). Primers were reselected to provide allele size ranges that were more amenable to multiplex analysis

SSR locus	*Forward primer (5′–3′)*	*Reverse primer (5′–3′)*
Satt002	GCGTGTGGGTAAAATAGATAAAAAT	GCGTCATTTTGAATCGTTGAA
Satt141	GGCTGGTGGTGTGCATAATAA	CCGTCATAAAAAGTCCCTCAGAAT
Satt143	GCGAAATCTATTAAAATGCTGAAATGAA	GCGCCACCTTGTCATTCTCTGAA
Satt243	GCGCATCAAATTGAAAGTAAGGAAAT	GCGTTGTAAGATCACGCCATTATT
Satt281	GCGTACACCTCTTTTGATGAC	GCGAGTAACATGAAGTCTACGATAACA
Satt353	CAACCGATTCACTATTCATTACTTCAAT	CCTAAGTTAATGGGAATGCCTTCTT
Satt358	GCGAGCTGTGCGCTTTATGT	GCGTGGCAGTACTTTAAATAAGACTT
Satt373	GCGTGGAAATACTGAATGAAAGCATATT	GCGTTATGCGCTCAAATTTTAAGTC
Satt307	GGCCTTTAGAACGTCTGACTTA	TTCTAGCATCAGAGGTAACCTTTGT
Satt409	TTACAAATTTACTCCTTAGACCAT	GACTTGAGTTGCCTTCTTTTCT
Satt431	GGCACCCTTGATAAATAAGAGA	GCGCATCTATCATTTCCCTTTTTATTAT
Satt194	TTGTAAATCATAAATTGTCA	TTATTGGAGAAAAAGAAATG
Satt242	GCGACTTTATTGAAACAATTTTTGACA	GCGCTGTGAGTGCCAACTAACTACTTTTA
Satt259	GCGACTTCCGATATACTATAATGTCTTG	GCGGAGTTTGTCATTTTGAAAAGGAT
Satt354	CAAATAAAAATGGACACCAAAAGTA	AATTGCCAAAAATAGCCACAC
Satt453	GCAGACGAGGATTCATTA	GTAGTGGGGAAGGGAAGTTA
Satt534	TTCATGCATATACATCACGTATTATT	TGTAAAAACTAAAGAATGGACTGTGG
Satt588	GCTGCATATCCACTCTCATTGACT	GCGAGAACCATTATTAATATTTTGCATTT

was properly determined. A size range for each allele at each locus across gels was determined. Based upon each individual allele size range, an exact allele size was assigned to each locus. Using these allele sizes and allele size range data, loci among the 48 were eliminated from inclusion in the final set of fingerprinting loci if there was overlap in any two individual allele size ranges. In addition, loci were generally

eliminated from further consideration if there were not at least three basepair differences between adjacent allele sizes. We felt that the use of these criteria would assure the selection of a set of SSR loci that would provide accurate, repeatable, and reliable genotype discrimination. Gene diversity, a measure of the relative informativeness of a marker (Weir 1990), was calculated as: $1 - \Sigma P_{ij}^2$ where P_{ij} is the frequency of the jth allele for ith locus.

Selection of a subset of maximally informative SSR loci

In order to identify a subset of SSR loci that would be maximally efficient in distinguishing the 66 N. American cultivars, it was first necessary to calculate a distance measure between each pair of loci. Such a measure would provide an estimate of the degree of similarity of the information content of each pair of loci in relation to the 66 N. American elite cultivars. The general procedure outlined by Nei & Li (1997) was adapted for this purpose as follows. At two SSR loci, x and y, genotypes were placed in groups based upon the allele(s) carried at each locus. The estimation of similarity (S_{ij}) between a particular allelic group at locus x and another at locus y is: $S_{ij} = 2^*Nx_iy_j/(Nx_i + Ny_j)$, where Nx_iy_j is the number of common genotypes in the ith group of locus x and the jth group of locus y, Nx_i is the number of genotypes in the ith group of locus x, and Ny_j is the number of genotypes in the jth group of locus y. The mean similarity between two loci ($\bar{S}_{ij}$) is the mean of all possible comparisons between allele groups at two loci and is calculated as:

$$\bar{S}_{ij} = \Sigma\ (2^*Nx_iy_j/(Nx_i + Ny_j))/G,$$

where G is the number of comparisons between a group at locus x and a group at locus y that possess at least one genotype in common. The distance between loci x and y (d_{ij}) was calculated as $1 - \bar{S}_{ij}$. The value of d_{ij} varies from 0 to 1.0. The matrix of mean distance values was calculated for each pair of SSR loci. The matrix was used to determine clusters of genotypes using the average linkage method of PROC CLUSTER in SAS (SAS 1989). the normalized root mean square distances from PROC CLUSTER were used to create a dendrogram using NTSYS (Rohlf 1992).

Relative effectiveness of selected markers to distinguish cultivars

To determine the effectiveness of the subset of selected SSR markers to distinguish the 66 elite N. American cultivars, simple genetic similarity coefficients were calculated between each pair of cultivars. These calculations were identical to those employed by Diwan & Cregan (1997) to calculate dissimilarity, except that similarity was 1 – dissimilarity. As described by Diwan & Cregan (1997) these calculations took into account the possibility of two alleles at a locus which would be expected to occur, particularly in elite cultivars that are derived from a single plant in the F_4 or F_5 generation. In the F_4 or F_5 generation 12.5 or 6.25% of the loci, respectively, would be anticipated to be heterozygous and in later generations this heterozygosity would be manifested as a mixture of two different homozygous types. The mean and variance of the similarity values were determined. As a basis of comparison, a similar matrix was calculated using the complete set of 30 loci and the mean and variance of these values were determined. A 't' test was used to compare the means of the two sets of similarity coefficients and an F test was used to compare the variances.

Effectiveness of selected SSR markers to distinguish independent sets of elite cultivars

To determine the effectiveness of the subset of SSR markers to distinguish individuals

within an independent sample of currently grown elite N. American cultivars, the cultivars previously described by Diwan & Cregan (1997) were analyzed at each SSR locus. Simple genetic similarity coefficients were calculated between each pair of cultivars within each of the four groups as described above. The mean, minimum and maximum values of the similarity coefficients within each group were determined.

RESULTS AND DISCUSSION

Observation of the allele size data and the electrophoregrams obtained from the Genotyper software allowed an assessment of the consistency of allele size determination across a series of sequencing gels that was required to analyze the 35 ancestral cultivars and the 66 elite N. American cultivars. Using these data, each locus was analysed to determine the size range of each individual allele at each locus and also to determine an 'exact' allele size to assign to each allele. Using the assigned allele sizes and the range of sizes within each allele size class, loci were discarded if allele sizes within a locus did not differ by three or more basepairs or if the range of an individual allele varied by more than two basepairs. Using these criteria, 18 of the 48 loci were discarded. This was mainly the result of allele sizes that did not vary by three bases or more. The remaining 30 loci and their allele sizes and numbers are given in Table 5. Allele numbers at these loci varied widely from four alleles in the case of Satt253, to a high of 16 alleles in the case of Satt009. Gene diversity is a function of the allele number at a locus as well as allelic frequency and is the probability that any two genotypes drawn at random from the population of individuals being assayed will be polymorphic. Gene diversity scores ranged from 0.59 to 0.83 (Table 5).

Identification of Maximally Informative SSR Loci

Cluster analysis was undertaken to group loci based upon the dissimilarity of the genetic information each provided about the 66 N. American elite cultivars (Figure 1). The analysis revealed loci such as Satt196 and Satt243 that gave relatively similar information about the genotypes. Likewise, it revealed that Satt009 provided information that was the most unique in relation to the other 29 loci. Based upon the cluster analysis, it was possible to choose a subset of loci that provided maximal information content. The size of such a subset is arbitrary and was selected to provide a number of markers that 1) was large enough to produce unique allele size profiles for the 66 N. American cultivars and 2) small enough to be easily analyzed on one or two gels. Therefore, the number of clusters in the dendrogram was arbitrarily determined at 13 (Figure 1). One marker was then selected from each cluster beginning with Satt009, followed by Satt534, Satt373, Satt038, Satt191, etc. In those cases in which there was more than one marker in a cluster, a marker was selected to maximize the number of linkage groups represented among the selected loci. In this manner, the 13 selected markers were drawn from 12 of the 20 soybean linkage groups (Table 5). Only in the case of loci Satt038 and Satt191 were loci selected from the same linkage group. Both loci map to linkage group G. Based upon the map of linkage group G derived from the 240 recombinant inbred lines of the University of Utah Minsoy x Noir 1 mapping population, a distance of 91.7 centimorgans (cM) separates Satt038 and Satt191 (Cregan *et al.* 1999). Thus, these loci segregate independently and would be anticipated to provide independent genetic information.

One method to examine the effectiveness of the subset of 13 markers versus the complete set of 30 loci is to compare the

Table 5. Linkage group, alleles sizes and number, and gene diversity of 30 simple sequence repeat loci based upon 35 North American ancestors and 66 North American elite soybean cultivars. The subset of 13 loci selected via the cluster analysis is shown in bold typeface

Locus	*Linkage group*	*Size of alleles (bp)*	*Number of alleles*	*Gene diversity*
Satt002	D2	133, 136, 139, 143, 146, 149, 152	7	0.59
Satt009	**N**	**163, 166, 184, 187, 190, 193, 199, 202, 214, 217, 220, 223, 229, 232, 241, 244**	**16**	**0.82**
Satt038	**G**	**151, 163, 166, 169, 172, 184, 187**	**7**	**0.71**
Satt114	**F**	**82, 88, 97, 106, 109, 118, 121**	**7**	**0.77**
Satt141	D1b	149, 164, 184, 187, 190, 193, 196, 199, 202	9	0.82
Satt147	**D1a+Q**	**174, 177, 191, 194, 204, 207, 210, 216**	**8**	**0.74**
Satt168	B2	201, 212, 227, 230, 233, 236	6	0.72
Satt173	O	199, 205, 208, 211, 214, 217, 223, 232, 238, 244, 247, 250, 262	13	0.82
Satt175	M	156, 159, 165, 168, 171, 174, 183, 189	8	0.78
Satt177	**A2**	**107, 110, 113, 116, 122**	**5**	**0.65**
Satt180	C1	215, 218, 242, 245, 260, 266, 269	7	0.74
Satt184	D1a+Q	141, 148, 166, 169, 181, 184	6	0.69
Satt191	**G**	**191, 206, 209, 221, 224, 227, 233**	**7**	**0.72**
Satt196	K	178, 181, 184, 187, 190, 202	6	0.60
Satt197	B1	135, 144, 174, 180, 183, 186, 189	7	0.72
Satt231	E	217, 220, 223, 226, 232, 235, 238, 241, 244	9	0.68
Satt236	A1	212, 215, 221, 224, 227	5	0.73
Satt242	**K**	**119, 134, 140, 143, 147, 150, 153, 156, 162**	**9**	**0.78**
Satt243	**O**	**260, 263, 269, 275, 290**	**5**	**0.67**
Satt253	H	137, 147, 150, 153	4	0.72
Satt259	O	129, 135, 145, 148, 151, 154	6	0.71
Satt294	**C1**	**249, 258, 264, 267, 282, 285, 288, 291, 294**	**9**	**0.70**
Satt307	C2	118, 127, 133, 136, 139	5	0.73
Satt308	**M**	**134, 137, 150, 153, 156, 159, 171, 174**	**8**	**0.73**
Satt324	G	202, 226, 229, 232, 238, 241	6	0.67
Satt353	H	96, 111, 117, 126, 129, 135	6	0.61
Satt373	**L**	**89, 92, 98, 101, 116, 122, 125, 128, 140, 155, 158, 161**	**12**	**0.80**
Satt414	**J**	**260, 284, 298, 301, 304, 307, 310, 316, 319**	**9**	**0.83**
Satt534	**B2**	**224, 227, 233, 236, 239, 242, 251, 254, 257, 260**	**10**	**0.76**
Satt577	B2	101, 110, 113, 116, 119, 122	6	0.74

two sets in terms of their ability to distinguish the 66 elite N. American cultivars. The mean dissimilarity coefficient calculated based upon the 13 loci was numerically higher (0.28 versus 0.30) than that for the 30 loci. The difference between the two means was statistically significant ($P<0.05$). The variances of the two matrices of values were 0.0185 (based upon 13 loci) and 0.0132 (based upon 30 loci) and did not differ significantly in magnitude. These results clearly indicate that the procedure used in the selection of loci was effective in identifying a subset of loci that was similar to the complete set in terms of ability to distinguish individual elite soybean cultivars.

Effectiveness of 13 selected SSR markers to distinguish independent sets of elite cultivars

None of the 10 MG I, seven MG II, 10 MG IV, and nine MG VI cultivars examined with the 13 selected trinucleotide SSR loci was

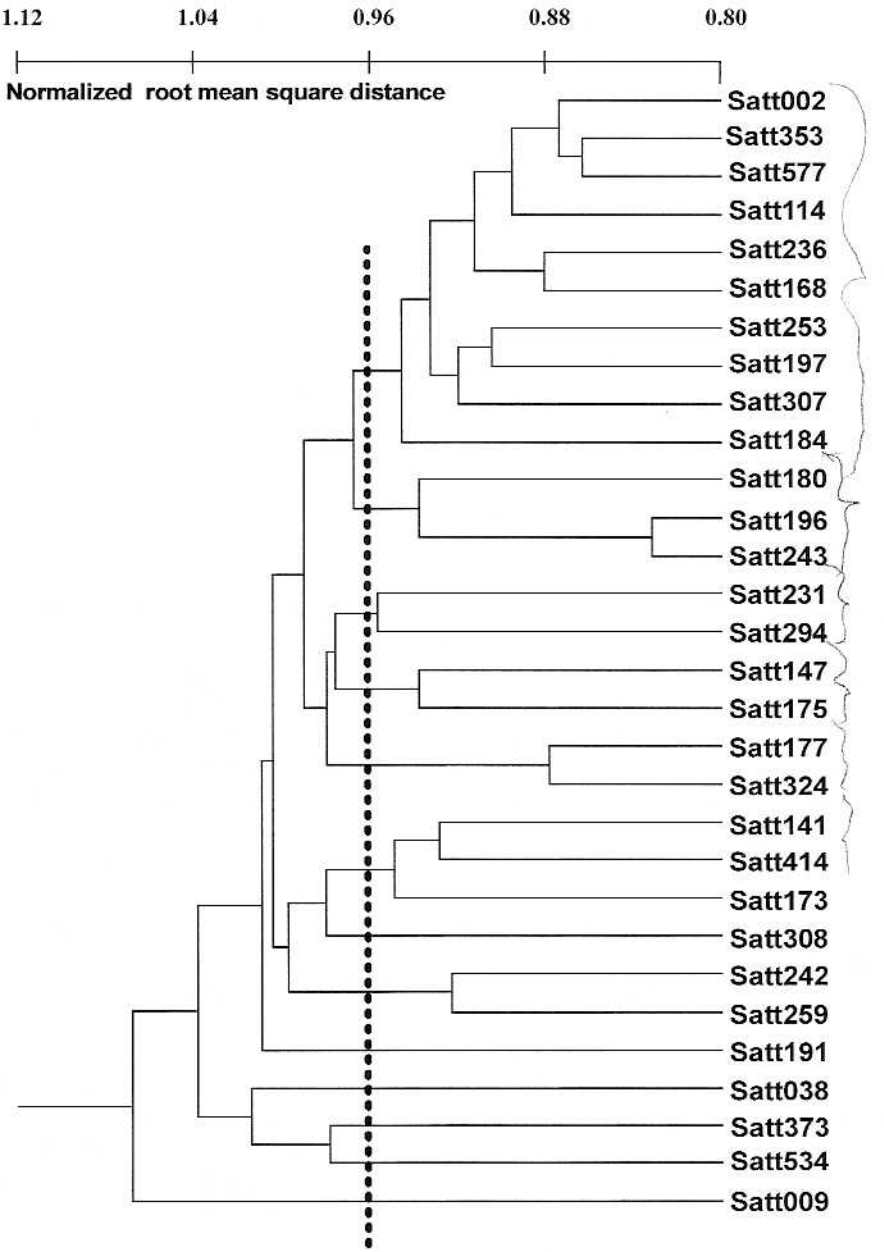

Figure 1. Dendrogram of cluster derived from 30 SSR markers used to determine allele size profiles in 66 elite N. American cultivars. The number of clusters from which markers were selected was arbitrarily determined and is indicated by the vertical hatched line.

Table 6. Linkage group and alleles sizes at 13 selected simple sequence repeat loci found in 10 Maturity Group (MG) I, seven MG II, 10 MG IV, and nine MG VI elite currently grown N. American soybean cultivars

Locus	*Linkage group*	*Size of alleles (bp)*	*Number of alleles*
Satt009	N	157*, 163, 220, 223, 229,	5
Satt038	G	172, 175*, 178*	3
Satt114	F	82, 97, 109, 121	4
Satt147	D1a+Q	174, 183*, 191, 194, 210,	5
Satt177	A2	110, 113, 122	3
Satt191	G	206, 209, 221, 224, 227	5
Satt242	K	134, 140, 143, 150, 153, 156, 162	7
Satt243	O	260, 269, 290	3
Satt294	C1	249, 258, 267, 282, 285, 288, 294	7
Satt308	M	137, 150, 153, 156, 171, 174	6
Satt373	L	77*, 89, 92, 98, 101, 104*, 125, 149*, 158, 161	10
Satt414	J	245*, 266*, 301, 304, 307, 310, 319	7
Satt534	B2	209*, 233, 251, 254, 257, 260	6

* Denotes an allele not present in the 35 N. American ancestors and 66 N. American elite soybean cultivars.

included among the 66 N. American elite cultivars used to select the set of 13 loci. Thus, these 36 cultivars represented an independent sampling of genotypes. A total of 71 alleles was detected among the 36 cultivars at the 13 loci (Table 6), versus 228 alleles among the 35 N. American ancestors and 66 elite lines (Table 5). This is not a surprising result given the fact that the 36 elite cultivars were from only four maturity groups and that the genotypes within each group were purposely selected for the homogeneity of their morphological and pigmentation descriptors. What was surprising was the presence of 10 alleles that had not been detected among the 35 N. American ancestors and 66 elite cultivars (Table 6). Gizlice *et al.* (1994) indicated that the 35 ancestors of N. American soybeans they identified should account for 95% of the alleles present in the N. American germplasm pool. Thus, three or four new additional alleles might be anticipated in the 36 elite cultivars that were not present in the 35 ancestors. However, most of these additional alleles should have been present in the 66 elite cultivars. It is our assumption that most of the 10 'new' alleles detected in the 36 MG I, II, IV, and VI currently grown elite cultivars arose from mutation events. Diwan & Cregan (1997) suggested that mutation contributed significantly to the rate of new SSR allele development. From limited observation of mapping populations they indicated new allele formation at a rate of 1 per 5,000 or 2×10^{-4} meioses. Furthermore, Diwan & Cregan (1997) indicated that this mutation rate was similar to that observed in humans and should not be an obstacle in the use of SSR markers for genotype identification. They suggested that description of soybean cultivars for purposes of PVP should be determined using bulked DNA of 30–50 plants of a cultivar rather than from DNA of a single plant. The suggestion that a bulked sample of 30 to 50 plants of a cultivar be used as the source of leaf tissue for DNA extraction and SSR allele size determination is predicated upon the assumption that if one of 30 to 50 plants carries a 'new' allele that has arisen as a result of mutation that this allele will not be detected. In this way, mutations in single plants would not alter the SSR allelic constitution of a cultivar. In contrast, when a cultivar of an inbreeding species does carry a

Table 7. Mean, maximum, and miminum of simple similarity coefficients calculated between cultivars within 10 Maturity Group (MG) I, seven MG II, 10 MG IV, and nine MG VI elite cultivars based upon allele size data at each of 13 selected simple sequence repeat loci

Maturity group	*Simple similarity coefficient*		
	Mean	*Minimum*	*Maximum*
I	0.46	0.19	0.81
II	0.51	0.31	0.77
IV	0.39	0.15	0.73
VI	0.33	0.15	0.62
Mean	0.42	0.20	0.73

mixture of two alleles at a locus, as frequently occurs when new cultivars are derived from F_4 plants of a segregating population, the sampling of 30–50 plants should ensure that both alleles are detected. The sampling of one or a few plants might not detect both alleles.

The SSR allele size data were used to calculate similarity coefficients between the elite cultivars within each of the MG I, II, IV, and VI groups. The average similarity among all cultivars was 0.42 (Table 7). Within each of the maturity groups, the most dissimilar cultivars had common alleles at only one in five SSR loci. For purposes of distinguishing cultivars, the most important comparison is between cultivars that are the most similar. In the case of the MG I, the cultivars DSR138 and DSR189 were the most similar, having a similarity coefficient of 0.81 (Table 7). Thus, of the 26 comparisons between these two cultivars (13 loci × 2 possible alleles), they were identical at 21 of the 26 comparisons. For MG II, the most similar cultivars were HS2812 and CM274, with a similarity coefficient of 0.77. In MG IV and VI, the most similar cultivars had similarity coefficients of 0.73 and 0.62, respectively. The relatively high similarity of cultivars within maturity groups is indicative of the high degree of genetic similarity among some highly bred soybean germplasm. Nonetheless, the group of 13 SSR loci were adequate to distinguish all cultivars from one another.

The research described herein has identified a set of SSR loci that should provide unambiguous allele size profiles of soybean genotypes. Using the 13 loci identified as effective for SSR allele profiling, it was possible to distinguish cultivars that were selected based upon their identical maturity, morphological, pigmentation, and other characteristics. Despite the high information content of SSR loci, one pair of cultivars within these groups was 81% similar. However, because approximately 500 trinucleotide SSR loci are available in soybean (Cregan *et al.* 1999), it is a simple matter to replace or add additional loci if the current set of 13 is inadequate to distinguish any pair of genotypes. At the very least, the use of this set of 13 SSR loci provides a system to substantially reduce the number of cultivars in the PVP database from which a candidate cultivar is indistinguishable, thereby reducing the number of entries to which the candidate need be compared in grow-out trials. Using an ABI Prism 373 or 377 DNA Sequencer or a PE Applied Biosystems 3700 DNA Analyzer, allele size determinations at 13 loci can be made very rapidly. Furthermore, any number of other effective systems exist for SSR allele sizing. Thus, we are proposing this set of loci as a standard set that can be used to support applications for Plant Variety Protection of new soybean cultivars.

ACKNOWLEDGEMENT

The authors wish to thank the United Soybean Board (USB Grants # 6027 and 9211) for the support of this research.

REFERENCES

Akkaya M.S., Bhagwat A.A. & Cregan P.B. (1992) Length polymorphism of simple sequence repeat DNA in soybean. *Genetics* **132**, 1131–1139.

Alford R.L., Hammond H.A., Coto I. & Caskey C.T. (1994) Rapid efficient resolution of parentage by amplification of short tandem repeats. *American Journal of Human Genetics* **55**, 190–195.

Cregan P.B., Bhagwat A.A., Akkaya M.S. & Jiang Rongwen (1994) Microsatellite fingerprinting and mapping of soybean. *Methods of Molecular Cell Biology* **5**, 49–61.

Cregan P.B., Jarvik T., Bush A.L., Shoemaker R.C., Lark K.G., Kahler A.L., Kaya N., VanToai T.T., Lohnes D.G., Chung J. & Specht J.E. (1999) An integrated genetic linkage map of the soybean genome. *Crop Science* **39**, 1464–1490

Diwan N. & Cregan P.B. (1997) Automated sizing of fluorescent labelled simple sequence repeat markers to assay genetic variation in soybean. *Theoretical and Applied Genetics* **95**, 723–733.

Edwards A., Hammond H.A., Jin L., Caskey C.T. & Chakraborty R. (1992) Genetic variation at five trimeric and tetrameric tandem repeat loci in four human population groups. *Genomics* **12**, 241–253.

Gizlice Z., Carter T.E. & Burton J.W. (1994) Genetic base for North American public soybean cultivars released between 1947 and 1988. *Crop Science* **34**, 1143–1151.

Hammond H.A., Jin L., Zhong Y., Caskey C.T. & Chakraborty R. (1994) Evaluation of 13 short tandem repeat loci for use in personal identification applications. *American Journal of Human Genetics* **55**, 175–189.

Hokanson S.C., Szewc-McFadden A.K., Lamboy W.F. & McFerson J.R. (1998) Microsatellite (SSR) markers reveal genetic identities, genetic diversity, and relationships in a *Malus* x *domestica* Borkh. core subset collection. *Theoretical and Applied Genetics* **97**, 671–683.

Keim, P., Olson, T. & Shoemaker, R. (1988) A rapid protocol for isolating soybean DNA. *Soybean Genetics Newsletter* **15**, 150–152.

Kresovich, S., Szewc-McFadden, A.K. & Bliek, S.M. (1995) Abundance and characterization of simple-sequence repeats (SSRs) isolated from a size-fractionated genomic library of *Brassica napus* L. (Rapeseed). *Theoretical and Applied Genetics* **91**, 206–211.

Nei, M. & Li, W.H. (1997) Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proceedings of the National Academy of Sciences, USA* **76**, 5269–5273.

Maughan, P.J., Saghai Maroof, M.A. & Buss, R.G. (1995) Microsatellite and amplified sequence length polymorphisms in cultivated and wild soybean. *Genome* **38**, 715–723.

Morgante, M. & Olivieri, A.M. (1993) PCR-amplified microsatellites as markers in plant genetics. *Plant Journal* **3**, 175–182.

Rohlf, F.J. (1992) *NTSYS-pc Numerical Taxonomy and Multivariate Analysis System version 1.7.* Owner's manual.

Rongwen, J., Akkaya, M.S., Lavi, U. & Cregan, P.B. (1995) The use of microsatellite DNA markers for soybean genotype identification. *Theoretical and Applied Genetics* **90**, 43–48.

SAS Institute (1989) *SAS/STAT User's Guide. Version 6. 4th edition.* SAS Institute, Inc., Cary, NC.

Thomas, M.R. & Scott, N.S. (1993) Microsatellite repeats in grapevine reveal DNA polymorphisms when analyzed as sequence-tagged sites (STSs). *Theoretical and Applied Genetics* **86**, 985–990.

Weir, B.S. (1990) *Genetic data analysis methods for discrete genetic data.* Sinauer Association, Sunderland, Mass.