Genetics and Molecular Biology, 36, 3, 382-390 (2013)
Copyright © 2013, Sociedade Brasileira de Genética. Printed in Brazil
www.sbg.org.br

Research Article

Genetic structure and a selected core set of Brazilian soybean cultivars

Regina Helena Geribello Priolli, Philip Traldi Wysmierski, Camila Pinto da Cunha, José Baldin Pinheiro and Natal Antonio Vello

Departamento de Genética, Escola Superior de Agricultura “Luiz de Queiroz”, Universidade de São Paulo, Piracicaba, SP, Brazil.

Abstract

Soybean is one of the most valuable and profitable oil crop species and a thorough knowledge of the genetic structure of this crop is necessary for developing the best breeding strategies. In this study, a representative collection of soybean cultivars recommended for farming in all Brazilian regions was genotyped using 27 simple sequence repeat (SSR) loci. A total of 130 alleles were detected, with an average allelic number of 4.81 per locus. These alleles determined the core set that best represented this soybean germplasm. The Bayesian analysis revealed the presence of two clusters or subgroups within the whole collection (435 soybean cultivars) and the core set (31 entries). Cultivars of similar origin (ancestral) were clustered into the same groups in both analyses. The genetic diversity parameters, based on the SSR loci, revealed high similarity between the whole collection and core set. Differences between the two clusters detected in the core set were attributed more to the frequency of their ancestors than to their genetic base. In terms of ancestry, divergent groups were presented and a panel is shown which may foster efficient breeding programs and aid soybean breeders in planning reliable crossings in the development of new varieties.

Keywords: core collection, genetic base, *Glycine max* (L.) Merrill, Bayesian method, SSR markers.

Received: January 22, 2013; Accepted: May 3, 2013.

Introduction

The soybean [*Glycine max* (L.) Merr.] is one of the most valuable and profitable oil crop species consumed worldwide as food and feedstuff. Its cultivation is primarily localized to four countries: USA, Brazil, Argentina and China. Brazil ranks second in crop area and production, with approximately 25 million hectares and 66 million tons, respectively (CONAB, 2012). The genetic breeding programs have been particularly important for the improvement of traits, such as high-yielding, biotic and abiotic stress tolerance, and protein and oil content. In this context, the characterization of the genetic structure of germplasm represents a crucial step to foster efficient breeding strategies and, consequently, the development of new cultivars.

Population structure has been documented in several studies investigating the diversity of elite crop germplasm (Huang *et al.*, 2002, Maccaferri *et al.*, 2005,Van Inghelandt *et al.*, 2010). In soybean populations, traditional estimations of population structure compare the diversity among pre-defined populations based on geographical origins and phenotypes (Cui *et al.*, 2000; Li and Nelson 2001; Abe *et al.*, 2003; Ude *et al.*, 2003; Wang *et al.*, 2006a; Guan *et al.*, 2010). However, in the case of Brazilian soybean germplasm, the recommended cultivars share the same background origin and are cultivated in large geographical areas. Previous studies have shown that the Brazilian soybean has a narrow genetic base, with only five ancestors, representing approximately 60% of the overall genetic base of the soybean (Hiromoto and Vello, 1986; Wysmierski, 2010). Moreover, studies based on genetic distance using molecular markers (Priolli *et al.*, 2002; Bonato *et al.*, 2006) and pedigree information (Miranda *et al.*, 2007; Priolli *et al.*, 2010) have also shown that soybean cultivars cluster according to pedigree.

The Bayesian method applied in the STRUCTURE program (Pritchard *et al.*, 2000; Falush *et al.*, 2003) starts with a predefined number of genetic clusters, before running the algorithm, without any previous information about hypothesized genetic origin, sampling location or phenotype of cultivars (Rosenberg *et al.*, 2002). The program uses a large number of molecular markers, such as simple sequence repeats (SSR) or microsatellite markers. These loci-markers are widely used mainly because of their highly polymorphic and abundantly available loci, which are randomly distributed throughout the genome and codominantly inherited (Powell *et al.*, 1996). STRUCTURE has been successfully applied for genetic structure analyses of Japanese, Chinese and European soybean germplasm (Kuroda *et al.*, 2006; Li *et al.*, 2008; Tavaud-Pirra *et al.*, 2009).

Send correspondence to Regina H.G. Priolli. Departamento de Genética, Escola Superior de Agricultura “Luiz de Queiroz”, Universidade de São Paulo, Av Pádua Dias 11, 13418-900 Piracicaba, SP, Brazil. E-mail: rpriolli@usp.br.

The aim of this study was to systematically survey the genetic structure of cultivated Brazilian soybean germplasm. To this end, a group of soybean cultivars recommended for all regions was genotyped using SSR markers. The core set that best represented this soybean germplasm was determined and used to define divergent clusters in terms of ancestry to aid soybean breeders in the selection of parent-plants for their crossing programs.

Material and Methods

Soybean plant material and SSR genotyping

A collection of 435 soybean elite cultivars, developed and released by public and private institutions, was selected to represent the complete range of cultivars grown in Brazil. The cultivars and their pedigrees are listed in Table S1 in the same order of appearance as depicted in the Bayesian analyses. Ten plants from each soybean cultivar were grown in a greenhouse and leaf tissue samples were collected, frozen in liquid nitrogen and lyophilized for three days. The DNA was isolated from the bulked lyophilized leaf tissue of the plants of each cultivar by mini-prep procedure based on Doyle and Doyle (1990). DNA quality and concentration were evaluated by electrophoresis on agarose gels stained with SYBR Safe (Invitrogen).

Twenty-seven SSR loci with either di- or tri-nucleotide repeats (Table 1) were selected based on their distribution across the soybean genome and amplification quality. All the SSR primer sequences, except for the RGA locus (Priolli *et al.*, 2010), are available in the 2003 USDA con-

Table 1 - Genetic diversity indices for 27 SSR loci between the group of 435 soybean cultivars and the core set, with linkage group (LG) and position (cM - centiMorgan) in the soybean map. N_A, number of alleles; N_{RA}, number of rare alleles; M_{AF}, major allele frequency; *I*, Shannon-Weaver diversity indices; H_E, expected heterozygosity.

			Whole collection (435 soybean cultivars)					Core set (31 soybean cultivars)				
Locus	LG	cM Position	N_A	N_{RA}	M_{AF}	*I*	H_E	N_A	N_{RA}	M_{AF}	*I*	H_E
Satt129	D1a	109,668	9	5	0.374	1.519	0.729	9	4	0.283	1.846	0.836
Satt186	D2	105,451	8	4	0.390	1.518	0.744	8	3	0.258	1.822	0.840
Satt005	D1b	75,292	7	3	0.486	1.328	0.664	7	2	0.426	1.568	0.760
Satt308	M	130,756	7	0	0.232	1.831	0.828	7	0	0.321	1.784	0.836
Satt191	G	96,572	7	4	0.454	1.286	0.676	7	4	0.339	1.534	0.769
Satt152	N	22,673	7	4	0.467	1.313	0.675	7	4	0.466	1.443	0.717
Satt294	C1	78,645	6	3	0.462	1.198	0.637	6	2	0.500	1.325	0.681
Satt173	O	58,398	6	3	0.453	1.214	0.650	6	2	0.310	1.528	0.783
Satt102	K	30,283	5	0	0.539	1.302	0.651	5	0	0.548	1.312	0.669
Satt263	E	45,397	5	0	0.388	1.350	0.699	5	0	0.355	1.484	0.774
SOYHSP176	F	68,438	5	2	0.587	1.108	0.583	5	2	0.633	1.068	0.565
Satt154	D2	57,07	5	1	0.377	1.375	0.716	5	1	0.393	1.418	0.756
Satt228	A2	154,114	5	1	0.439	1.225	0.660	5	1	0.417	1.363	0.738
AW734043	C2	4,223	5	3	0.780	0.702	0.364	5	3	0.731	0.893	0.455
Satt045	E	46,646	4	1	0.554	0.952	0.557	4	2	0.613	0.886	0.535
Sct_189	I	113,768	4	0	0.528	1.158	0.632	4	0	0.345	1.321	0.744
Satt070	B2	72,808	4	0	0.410	1.274	0.699	4	0	0.387	1.278	0.721
Satt335	F	77,704	4	1	0.640	0.949	0.528	4	0	0.625	1.036	0.574
BE806387	F	22,965	4	1	0.474	1.132	0.647	4	1	0.474	1.136	0.682
BF008905	O	28,951	4	3	0.955	0.195	0.085	4	2	0.857	0.515	0.262
Satt309	G	4,534	3	1	0.517	0.852	0.543	3	0	0.517	0.886	0.574
Satt302	H	81,04	3	1	0.628	0.668	0.469	3	1	0.500	0.772	0.537
AW781285	D1a	67,777	3	1	0.556	0.705	0.497	3	1	0.652	0.727	0.489
AI794821	C1	122,625	3	0	0.760	0.717	0.395	3	0	0.737	0.753	0.444
AW310961	J	5,187	3	1	0.582	0.707	0.492	3	1	0.482	0.822	0.553
SOYGPATR	C1	10,336	2	0	0.930	0.253	0.130	2	0	0.931	0.251	0.133
RGA	-	-	2	0	0.948	0.203	0.098	2	0	0.935	0.239	0.124
Total			130	43				130	36			
Mean			4.81	1.592	0.552	1.039	0.557	4.81	1.333	0.520	1.149	0.613

Table 2 - Analysis of molecular variance (AMOVA) and the respective probabilities (p) based on the 27 SSR loci data from the groups formed by Bayesian analysis.

Source of variation	*df*	Sum of squares	Variance (component)	% (variation)	p
Whole collection (435 soybean cultivars)					
Among groups	1	63.342	0.275	6.73	$p < 0.0001$
Within groups	433	1649.157	3.808	93.27	$p < 0.0001$
Core set (31 soybean cultivars)					
Among groups	1	23.507	0.981	10.55	$p < 0.0001$
Within groups	29	241.171	8.316	89.45	$p < 0.0001$

sensus map (Soybase http://www.soybase.org/) and in the public soybean genetic map (Song *et al.*, 2004). The PCR and electrophoresis conditions described in Priolli *et al.* (2010) were applied. The SSR alleles were resolved on an ABI-377 sequencer using GENESCAN/GENOTYPER software (Applied Biosystems) and ROX-500 as a size standard.

Development of a soybean core set and population structure analysis

The advanced M strategy using a modified heuristic algorithm was implemented in the POWERCORE 1.0 program (Kim *et al.*, 2007) and used to develop the core set from 435 soybean cultivars using 27 SSR loci. This strategy maximizes the allele richness at each marker locus in the core-collection subset, followed by the Shannon diversity index (Schoen and Brown, 1993).

The model-based program, STRUCTURE 2.3.3 (Pritchard *et al.*, 2000; Falush *et al.*, 2003) was used to infer the population structure in the collection of 435 soybean cultivars and to select a core set using POWERCORE. The following parameters were applied to the soybean analysis without prior population information (Tavaud-Pirra *et al.*, 2009): a haploid, no admixture model and an independent allele frequency model. Following a burn-in period of 50,000, ten independent runs were performed for each K value (number of sub-groups, from 1 to 20) with 500,000 iterations. The best estimate of the number of clusters was determined according to the criterion of Evanno, using a ΔK value based on the rate of change in the log probability of data between successive K values (Evanno *et al.*, 2005). An analysis of molecular variance (AMOVA) was used to detect the population differentiation and was calculated with the ARLEQUIN 3.11 program (Excoffier *et al.*, 2005).

Genetic diversity analysis

The number of alleles (N_A), the number of rare alleles (alleles with frequencies less than 5%) (N_{RA}), major allele frequency (M_{AF}) and Shannon and Weaver diversity index (*I*) per locus were calculated using the MSA 4.05 program (Dieringer and Schlotterer, 2003). The analyses of diversity among the cultivars in the core set were based on a modified Roger's distance method using the TFPGA 1.3 program (Miller, 1997) and the consensus Ward dendrogram was obtained using the PAST 2.03 program (Hammer *et al.*, 2001).

Genetic base of the core set

The number of ancestors and their relative genetic contribution (RGC) to each cultivar of the core clusters from STRUCTURE were estimated through the coefficient of parentage between the cultivars and their ancestors. The ancestors were defined as the oldest parent in the pedigree of a cultivar, beyond which no genealogical information was available.

The coefficient of parentage was considered in a range from 0 (individuals with completely different pedigrees) to 1 (individuals with the same genetic constitution) (Cox *et al.*, 1985). It was calculated in Microsoft Excel® using the following equation: $f_{X,Y} = 1/4\ (f_{AC} + f_{AD} + f_{BC} + f_{BD})$, where f is the coefficient of parentage of two individuals; X is individual 1; Y is individual 2; A and B are the parents of X; and C and D are the parents of Y.

The RGC was calculated as the arithmetic mean of all coefficients of parentage between the ancestor and cultivars of each core cluster. The frequency of all parents present in each core cluster was also estimated and represented as the individual contribution of each parent. Additionally, the presence of ancestors and intermediate parents in each core cluster from the Bayesian analysis was calculated. Intermediate parents are those that are present in a pedigree of known genealogy. For each ancestor and intermediate parent, we counted the number of cultivars in which it appeared, at least once, for core clusters A and B (it's frequency in each core cluster). For example, in core cluster A, the ancestor 'Arksoy' appeared in the genealogies of seven of the cultivars; in core cluster B, it appeared in the genealogies of five of the cultivars. The difference between these frequencies (core cluster A - core cluster B) indicates the predominance of this ancestor in either group. Positive values indicate predominance in core cluster A and negative values indicate the predominance in core cluster B.

Results

Analysis of population structure using molecular markers

To distinguish putative clusters among the 435 cultivars we evaluated the entire collection using the Bayesian method with SSR loci-markers. After ten independent runs, the best estimate of delta K for the whole collection was K = 2 (Figure 1). The cultivars from same origin (ancestral) were grouped into the same cluster (Table S1), indicating ancestors that contributed to the formation of each cluster, as shown in Figure 2. ‘Santa Rosa’, ‘IAC 2’, ‘Industrial’, ‘Bragg’, ‘União’, ‘Davis’, ‘IAC 7’ and ‘Hood’ were some of the parents in cluster A (red bar), while ‘UFV1’, ‘FT Cristalina’, ‘FT Estrela’, ‘Lee’, ‘Sharkey’, ‘Forrest’, ‘IAC 8’, ‘Tropical’, ‘IAS 5’, ‘Paraná’, ‘Lancer’, ‘Hampton’ were some of the parents recognized in ancestry cluster B (green bar).

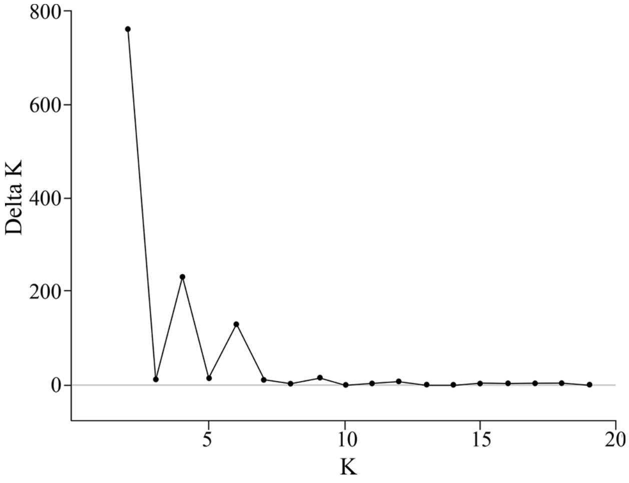

Figure 1 - Values of ΔK (delta K) in 435 soybean cultivars using 27 SSR loci markers, with its modal value detecting a true K of two clusters (K = 2).

Genetic diversity and structure of selected core set

Using the genetic diversity parameters based on SSR loci a marker, a core set was defined, which best represented the genetic diversity in the whole group. Table 1 showed that the total number of alleles and the average genetic diversity value per locus were similar between the core set and the whole collection. The correlation coefficients (r) of the total number of alleles, Shannon-Weaver diversity index and heterozygosity between the whole collection and core set were 1.000, 0.9633 and 0.9706, respectively. To assess whether the same alleles were represented we also compared the allele frequencies of the SSR markers in the core set with the frequencies observed in the whole group. The frequency of alleles and major allele frequency was high between the groups, which were significantly correlated with values of 0.9667 and 0.9277, respectively. Notably, both groups showed approximately 30% rare alleles (alleles with frequency < 0.05 in each locus), which might reflect the microsatellite mutation rate in soybean.

Once the cultivars of the core set were selected, a Bayesian analysis was used to infer the population structure that might be present in these 31 entries. As shown in Figure 3, the best estimate of delta K was again K = 2, and the analyses of diversity among the cultivars also showed the same value with two major core clusters with 16 (green bar) and 15 (red bar) cultivars, respectively. Cultivars from the same ancestral origin still were grouped into the same groups for better observation in the dendrogram (Figure 4). For instance, in the group named “C”, the cultivars ‘BRSMA Acará’, ‘Embrapa 63 (Mirador)’ descended from ‘Santa Rosa’ or from a selection of its progeny, ‘Dourados’. For the same reason, in the group named B, the cultivars ‘Emgopa 302’, ‘FT Eureka’ and ‘BR/EMG 312 (Potiguar)’ descended from ‘Paraná’. Besides the cultivars clearly as-

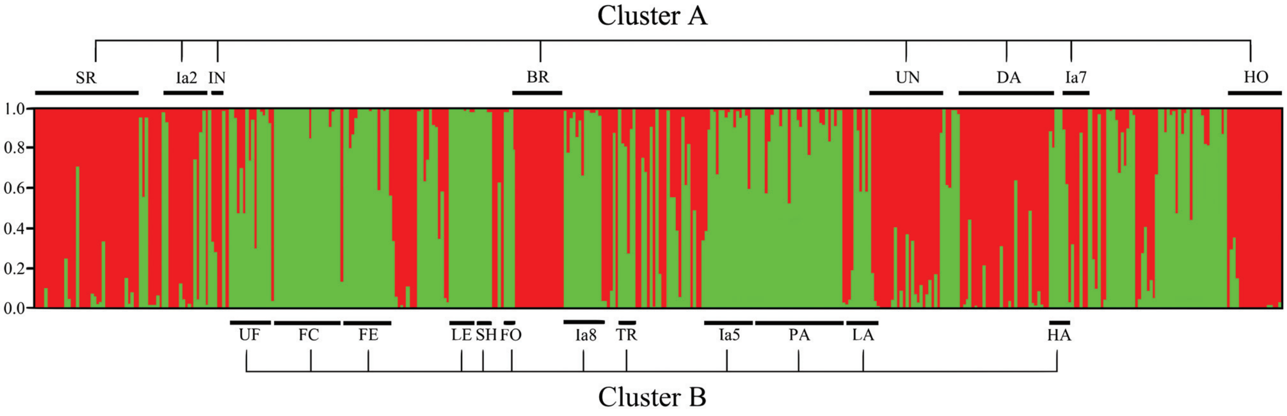

Figure 2 - Two germplasm clusters A (red) and B (green) based on Bayesian analyses for the 435 soybean cultivars analyzed using the 27 SSR markers. Each individual is represented by a single vertical line (red/ green), with lengths proportional to each of the two inferred clusters. The names of some of the ancestors identified are abbreviated, representing: ‘Santa Rosa’ (SR), ‘IAC 2’ (Ia2), ‘Industrial’ (IN), ‘Bragg’ (BR), ‘União’ (UN), ‘Davis’ (DA), ‘IAC 7’ (Ia7), ‘Hood’ (HO), ‘UFV1’ (UF), ‘FT Cristalina’ (FC), ‘FT Estrela’ (FE), ‘Lee’ (LE), ‘Sharkey’ (SH), ‘Forrest’ (FO), ‘IAC 8’ (Ia8), ‘Tropical’ (TR), ‘IAS 5’ (Ia5), ‘Paraná’ (PA), ‘Lancer’ (LA), ‘Hampton’ (HA).

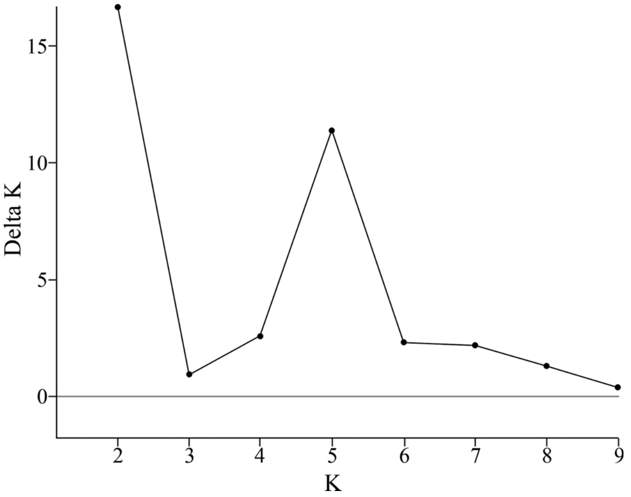

Figure 3 - Values of ΔK (delta K) in 31 soybean cultivars using the 27 SSR loci markers, with its modal value detecting a true K of two core clusters (K = 2).

signed to one core cluster, some cultivars were assigned to two core clusters. For example, 'BRSMA Aracá' and 'Embrapa 63', although belonging to the red core cluster (core cluster A), also were assigned to the green core cluster (cluster B), probably because of their pedigree, which included 'Ocepar 9 (SS-1)', a mutation of 'Paraná', widely present in the pedigree of the red core cluster.The analysis of molecular variance (Table 2) was used to partition the SSR variation among and within the clusters from the whole collection (435 soybean cultivars) and core set (31 soybean cultivars). Both collections had most of the variation (93.27% in the whole collection and 89.45% in the core set) within clusters, while only a small but significant portion of the variation (0.0673 and 0.1055%, $p < 0.0001$, to whole collection and core set, respectively) was attributed to variation among clusters. These results indicate that significant genotypic differentiation exists within the clusters revealed using Bayesian analysis.

Genetic base of selected core set

To fully reveal the differences between the cultivars of each core cluster we defined the number of ancestors and their relative genetic contributions. Table 3 presents the ancestors of each core cluster. The same five main ancestors ('CNS', 'Nanking', 'Tokyo', 'PI 54610' and 'S-100') contributed with 46.23 and 45.15% to core clusters A and B, respectively. Among the 30 ancestors identified, 27 and 23 were present in core clusters A and B, respectively. Neither of these ancestors changed the feature of the core clusters because they had almost the same frequency. However, three and seven of the ancestors were exclusive to core clusters A and B, respectively.

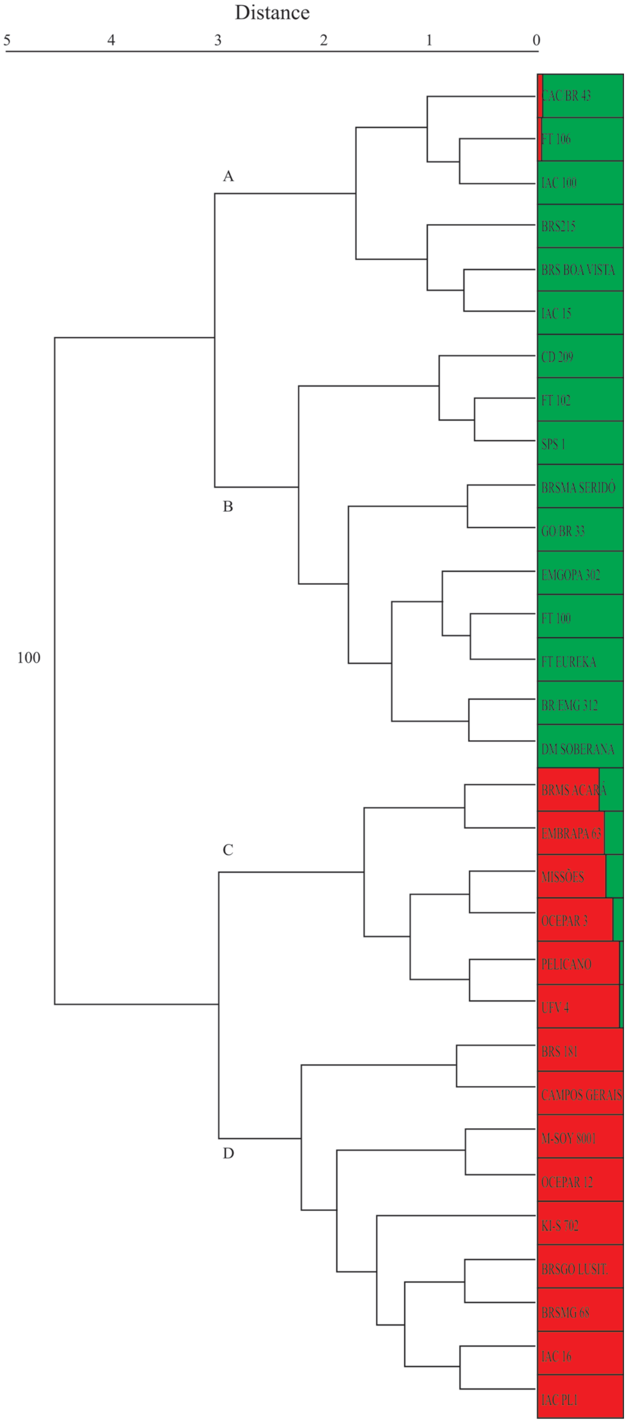

Figure 4 - Ward tree based on a Roger's genetic distance matrix between (left) the 31 selected entries of the core set based on the polymorphism of 27 simple sequence repeat markers and their correspondent Q matrix of Bayesian analysis (right). Red and green bar-plots represent the two core clusters and their respective identifications.

To assess whether the exclusive ancestors and intermediate parents could cause differences between core clusters we again turned to the pedigree of the 31 cultivars in the core set and listed all of the ancestors that participated in their genealogies, up to the oldest ancestor. The ancestors identified were sorted in descending order, with positive values for the parental frequency in core cluster A and negative values in core cluster B (Supplementary Material Table S2). Among the 147 ancestors identified, 48 (32.65%)

Table 3 - Genetic base with relative genetic contribution (RGC) and frequency of oldest identified ancestor of core cluster A and B. Both core clusters were detected from Bayesian analysis in core set of 31 soybean cultivars.

Core cluster A				Core cluster B			
Accessions A	RGC A	FreqA	%FreqA	Accessions B	RGC B	Freq B	%Freq B
CNS (PI 71569)	0.1218	13	86.67	CNS (PI 71569)	0.1231	12	75.00
Nanking (Roanoke)	0.1062	12	80.00	Nanking (Roanoke)	0.0857	12	75.00
Tokyo (PI 8424)	0.0823	12	80.00	Tokyo (PI 8424)	0.0739	13	81.25
PI 54610	0.0771	12	80.00	PI 54610	0.0661	13	81.25
S-100 (Illini; A.K.)	0.0747	11	73.33	S-100 (Illini; A.K.)	0.1026	12	75.00
PI 60406	0.0570	8	53.33	PI 60406	0.0325	7	43.75
Arksoy (PI 37335)	0.0374	8	53.33	Arksoy (PI 37335)	0.0634	10	62.50
Biloxi (PI 23211)	0.0298	9	60.00	Biloxi (PI 23211)	0.0162	7	43.75
Tanner	0.0285	8	53.33	Tanner	0.0162	7	43.75
Dunfield (PI 36846)	0.0206	10	66.67	Dunfield (PI 36846)	0.0337	10	62.50
Haberlandt(PI 6396)	0.0206	10	66.67	Haberlandt (PI 6396)	0.0318	9	56.25
Nanda (PI 95727)	0.0166	2	13.33	Nanda (PI 95727)	0.0000	0	0.00
Palmetto(PI 71587)	0.0137	4	26.67	Palmetto(PI 71587)	0.0014	2	12.50
Bilomi 3(PI 240664)	0.0125	2	13.33	Bilomi 3(PI 240664)	0.0119	3	18.75
Mandarin (Ottawa)	0.0124	1	6.67	Mandarin (Ottawa)	0.0000	0	0.00
Laredo (PI 40658)	0.0109	4	26.67	Laredo (PI 40658)	0.0039	2	12.50
Mammoth Yellow	0.0109	4	26.67	Mammoth Yellow	0.0039	2	12.50
Mandarin(PI 36653)	0.0103	2	13.33	Mandarin(PI 36653)	0.0354	2	12.50
Manchu (PI 30593)	0.0103	2	13.33	Manchu (PI 30593)	0.0041	1	6.25
Pine Dell Perfection	0.0099	3	20.00	Pine Dell Perfection	0.0019	1	6.25
Blyvoor (PI221713)	0.0083	1	6.67	Blyvoor (PI 221713)	0.0000	0	0.00
Mogiana	0.0062	2	13.33	Mogiana	0.0039	1	6.25
Richland(PI 70502)	0.0042	1	6.67	Richland (PI 70502)	0.0024	1	6.25
FC 31745	0.0041	1	6.67	FC 31745	0.0000	0	0.00
PI 171442	0.0041	1	6.67	PI 171442	0.0000	0	0.00
Peking (PI 17852B)	0.0002	2	13.33	Peking (PI 17852B)	0.0000	0	0.00
Mukden (PI 50523)	0.0001	1	6.67	Mukden (PI 50523)	0.0000	0	0.00
PI 274454	0.0000	0	0.00	PI 274454	0.0156	1	6.25
PI 346304	0.0000	0	0.00	PI 346304	0.0156	1	6.25
PI 229358	0.0000	0	0.00	PI 229358	0.0039	1	6.25

and 39 (26.53%) were exclusive to A and B, respectively, while 60 appeared in both groups. Among these ancestors, 27 and 9 had higher frequencies (> 1 and < -1) in core clusters A and B, respectively. It is worthy of note that soybean cultivars located at the end of the list might be considered more divergent among the core clusters, *i.e.*, 'Ocepar 9-SS1' (core cluster A) in relation to 'Viçoja' or 'UFV 1' (selection of 'Viçoja') (core cluster B).

Discussion

Genetic structure of the whole group

This study is the first to evaluate the genetic diversity in a large group of Brazilian soybean cultivars using the Bayesian method. The model-based structural analysis used here revealed the presence of two stable clusters (A and B) within the whole group of cultivars, with discrimination of parent background participation in the pedigree. In Asian and European soybean germplasms, the number of clusters were seven and three, respectively (Li *et al.*, 2008; Tavaud-Pirra *et al.*, 2009), which is another indication of the narrow genetic base of Brazilian germplasm. Many of the ancestral genotypes mentioned in clusters A and B presented some degree of parentage. 'Santa Rosa', 'IAC 2' and 'Industrial', for instance, are descendants of 'La 41-1219 (Pelican)'. 'UFV1' and 'FT Cristalina' also share a common parent, 'Viçoja', or a selection of its ancestors 'D4924-91', such as 'FT Estrela'. As can be verified in Table S1, 'Paraná' and 'IAS 5' also have the same ancestry. Cultivars belonging to different clusters (A and B) should have differing genetic constitutions. Therefore, crosses between cultivars from different clusters (for example, between cultivars 'Dourados', 'União', 'MG/BR 46 Conquista' from cluster A and cultivars 'Paraná', 'IAS-5' and 'IAC 8' from cluster B) could provide a higher genetic variability for breeding programs to explore and also broaden the soybean genetic base.

Consistent with our results, current studies based on genetic distances and traditional cluster analysis, have identified similar groups in relation to their pedigrees. Polymorphisms in AFLP markers and SSR loci were consistent with the cultivar pedigree information (Priolli *et al.*, 2002, 2010; Bonato *et al.*, 2006). Nonetheless, in relation to the structure, the method for determining the number of populations in STRUCTURE frequently fails in germplasm data sets for various reasons, including isolation by distance and inbreeding (Falush *et al.*, 2007). The presence of dominant groups in the evaluated germplasm can overshadow minor subdivisions and sequential detections (Yan and Ye, 2007). A traditional cluster analysis using UPGMA or the Ward method occasionally provides the best way for determining the genetic structure in germplasm collections (Odong *et al.*, 2011). We compared dendrograms generated using different methods, such as SSRs (Priolli *et al.*, 2002) and the Malécot coefficient of parentage (Miranda *et al.*, 2007), which were coincident. In fact, both methods successfully distinguished similar groups (Priolli *et al.*, 2010). The current study complements a growing body of work focused on the genetic structure of Brazilian soybeans for choosing hybridizations and controlled crosses.

Structure of core set

We first ran Structure on a subsample representing the genetic diversity of the whole collection and removed families of related accessions. Once the genetic structure of the subsample had been assessed, the admixture proportions of the additional individuals could be calculated, assuming that the population allelic frequencies were equal to the ones previously estimated (Camus-Kulandaivelu *et al.*, 2006). Maximizing genetic diversity and reducing the number of entries from 435 to a core set of 31 provided a working collection of Brazilian soybean germplasm containing all 130 alleles of the whole collection (Table 1), this fostering the study of economically important traits in soybean breeding programs. There are several methods for developing a core set in plants, but all of these aim at representing the maximum genetic diversity with the fewest possible number of entries. Wang *et al.* (2006b) indicated the efficiency of a core collection in capturing the genetic diversity of agronomic traits using only 2% of the total accessions to represent approximately 70% of the genetic diversity from a whole soybean sample set. Based on SSR information and the maximization method, 50 soybean accessions (15% of the total sampling) captured more than 90% of the global allelic richness available in European soybean germplasm (Tavaud-Pirra *et al.*, 2009).

The microsatellite mutation rate in soybean has been estimated at 10^{-5} to 10^{-4} per generation (Diwan and Cregan 1997), which can explain the presence of low frequency alleles in some of the SSR loci (30% rare alleles). In a study of soybean cultivar identification, ten new alleles in 66 soybean cultivars that were not present in the 35 ancestral lines were identified (Song *et al.*, 1999). Additionally, 32 alleles specific to elite cultivars within a total of 397 alleles were identified in another study of 79 soybean genotypes (Narvel *et al.*, 2000).

The population structure identified in the core set revealed the presence of two core clusters, which were basically consistent with the grouping analysis based on genetic distance. To increase the genetic variability available in the breeding program, soybean breeders may choose cultivars belonging to different core clusters, for example 'IAC100', 'BRSMA Seridó', BR/Emgopa 312' from the green core cluster (core cluster B) and 'UFV4', BRS181','KI-S 702' from red core cluster (core cluster A). Although some clusters presented common ancestors, it was not possible to identify origin-related sites for the core clusters formed, as observed in previous studies of soybean germplasm from China (Cui *et al.*, 2000; Li and Nelson 2001; Wang *et al.*, 2006a; Li *et al.*, 2008; Guan *et al.*, 2010) or from China, Japan, Korea and USA (Abe *et al.*, 2003; Ude *et al.*, 2003). The genetic structure of the soybean European collection also did not correlate with geographical origin (Tavaud-Pirra *et al.*, 2009), probably because the plant material represented only elite soybean cultivars, such as those included in this study. Modern breeding limits gene flow and can result in a large amount of variation attributed to differences within the groups, rather than between the two inferred groups. The AMOVA showed a greater difference within the clusters in both the whole (93.27%, 435 soybean cultivars) and core set analyses (89.45%; 31 soybean cultivars), and only a low but significant portion of the variation was attributed to variation among clusters (6.73 and 10.55%, $p < 0.0001$) in both analyses.

Genetic base and difference of core clusters

Our findings show that the two core clusters detected by using Bayesian analysis have the same genetic base with 10% (3 in 30) and 23% (7 in 30) absent ancestors in core clusters A and B, respectively. Previous studies had determined that the genetic base of Brazilian soybean cultivars was narrow and comprised five ancestors ('CNS', 'S-100', 'Nanking', 'Tokyo' and 'PI 54610'), representing 57.49% of the genetic base (Hiromoto and Vello, 1986) or 63.84% (Wysmierski, 2010). These same five ancestors are those that contributed most to the genetic base of the core set (Table 3). A comparison between the genetic base of core clusters A and B of the core set showed that they are quantitatively similar to one another, as denoted by the accumulated RGC among their five main ancestors, at 46.23% and 45.16%, for A and B, respectively. These values are lower than those reported by the same authors, but this result was expected, as the genetic base was calculated for a collection of elite cultivars with no selection for diversity. In contrast, in our present study, the core set was selected to maximize genetic diversity, thereby reducing the accumulated RGC to the main ancestors to some extent.

Our data also showed that there are some qualitative differences between the genetic bases of the two core clusters in relation to their ancestral composition, but these differences are usually concentrated on ancestors with low contributions, typically less than 2%. For example, ‘PI 274454’ was found to contribute exclusively to group B, but only with approximately 1.56%. These low-contributing exclusive ancestors are probably due to their restricted use in cultivar development to transfer certain specific characteristics, as is the case of ‘PI 274454’ mentioned above, which is reported to be an insect-resistant cultivar that contributed to the pedigree of ‘IAC-100’, a moderately insect-resistant Brazilian cultivar (Fernandes *et al.*, 1994; de Godoi and Pinheiro 2009).

Most of the accessions have little tendency toward either group, but 27 were more predominant in group A and 9 were predominant in group B. For example, some of the agronomic traits observed in ‘Ocepar 9 -SS1’ and ‘Viçoja’ indicate that the cycle and the consequent maturity range are divergent traits among the two groups. The first was developed and recommended for southern and southeastern areas, and the second, for the central area of the Brazilian Cerrado. In germplasm with a restricted genetic base, such as the Brazilian soybean germplasm, this association must be further tested, but to the breeders, all such information is important because the divergence among parents is obtained by their pedigree analyses and, on many occasions, it is the only information available in cultivar development.

In conclusion, the analysis of population structure based on SSR markers showed the existence of genetic diversity and structure in the studied plant material. The Bayesian analysis revealed the presence of two clusters in the whole collection (435 soybean cultivars) and in the core set (31 soybean cultivars). Here, we show that individuals of the core set maintained all the alleles of the large group. We speculate that by using this framework of genetically defined populations, it may be possible to exploit the soybean germplasm and suggests that the use of these SSR loci resulting in the panel presented may allow breeders to perform reliable crossings or to strategically plan their breeding programs.

Acknowledgments

This work was supported by the São Paulo Research Foundation (FAPESP). We thank Lizandra L. Catelli and Carlos A.A. Arias (Embrapa soybean) for the genotypes. CNPq and CAPES provided scholarships.

References

Abe J, Xu DH, Suzuki Y, Kanazawa A and Shimamoto Y (2003) Soybean germplasm pools in Asia revealed by nuclear SSRs. Theor Appl Genet 106:445-453.

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

Camus-Kulandaivelu L, Veyrieras JB, Madur D, Combes V, Fourmann M, Barraud S, Dubreuil P, Gouesnard B, Manicacci D and Charcosset A (2006) Maize adaptation to temperate climate: Relationship between population structure and polymorphism in the Dwarf8 gene. Genetics 172:2449-2463.

Cox TS, Kiang YT, Gorman MB and Rodgers DM (1985) Relationship between coefficient of parentage and genetic similarity indexes in the soybean. Crop Sci 25:529-532.

Cui ZL, Carter TE and Burton JW (2000) Genetic diversity patterns in Chinese soybean cultivars based on coefficient of parentage. Crop Sci 40:1780-1793.

de Godoi CRC and Pinheiro JB (2009) Genetic parameters and selection strategies for soybean genotypes resistant to the stink bug-complex. Genet Mol Biol 32:328-336.

Dieringer D and Schlotterer C (2003) Microsatellite analyser (MSA): A platform independent analysis tool for large microsatellite data sets. Mol Ecol Notes 3:167-169.

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.

Doyle JJ and Doyle JL (1990) Isolation of plant DNA from fresh tissue. BRL Focus 12:13-15.

Evanno G, Regnaut S and Goudet J (2005) Detecting the number of clusters of individuals using the software STRUCTURE: A simulation study. Mol Ecol 14:2611-2620.

Excoffier L, Laval G and Schneider S (2005) Arlequin ver. 3.0: An integrated software package for population genetics data analysis. Evol Bioinform 1:47-50.

Falush D, Stephens M and Pritchard JK (2003) Inference of population structure using multilocus genotype data: Linked loci and correlated allele frequencies. Genetics 164:1567-1587.

Falush D, Stephens M and Pritchard JK (2007) Inference of population structure using multilocus genotype data: Dominant markers and null alleles. Mol Ecol Notes 7:574-578.

Fernandes FM, Athayde MLF and Lara FM (1994) Performance of soybean cultivars in the field under stink bug attack. Pesq Agr Brasil 29:363-367.

Guan RX, Chang RZ, Li YH, Wang LX, Liu ZX and Qiu LJ (2010) Genetic diversity comparison between Chinese and Japanese soybeans (*Glycine max* (L.) Merr.) revealed by nuclear SSRs. Genet Resour Crop Evol 57:229-242.

Hammer O, Harper DAT and Ryan PD (2001) PAST: Paleontological Statistics Software Package for Education and Data Analysis. Palaeontol Electronica, 9 pp.

Hiromoto DM and Vello NA (1986) The genetic base of Brazilian soybean (*Glycine max* (L) Merrill) cultivars. Braz J Genet 9:295-306.

Huang XQ, Borner A, Roder MS and Ganal MW (2002) Assessing genetic diversity of wheat (*Triticum aestivum* L.) germplasm using microsatellite markers. Theor Appl Genet 105:699-707.

Kim KW, Chung HK, Cho GT, Ma KH, Chandrabalan D, Gwag JG, Kim TS, Cho EG and Park YJ (2007) PowerCore: A program applying the advanced M strategy with a heuristic search for establishing core sets. Bioinformatics 23:2155-2162.

Kuroda Y, Kaga A, Tomooka N and Vaughan DA (2006) Population genetic structure of Japanese wild soybean (*Glycine soja*) based on microsatellite variation. Mol Ecol 15:959-974.

Li YH, Guan RX, Liu ZX, Ma YS, Wang LX, Li LH, Lin FY, Luan WJ, Chen PY, Yan Z, *et al.* (2008) Genetic structure and diversity of cultivated soybean (*Glycine max* (L.) Merr.) landraces in China. Theor Appl Genet 117:857-871.

Li ZL and Nelson RL (2001) Genetic diversity among soybean accessions from three countries measured by RAPDs. Crop Sci 41:1337-1347.

Maccaferri M, Sanguineti MC, Noli E and Tuberosa R (2005) Population structure and long-range linkage disequilibrium in a durum wheat elite collection. Mol Breed 15:271-289.

Miller MP (1997) TFPGA: Tools for population genetic analyses ver. 1.3. Flagstaff, Northern Arizona University.

Miranda ZFS, Arias CAA, Prete CEC, Kiihl RAS, Almeida LA, Toledo JFF and Destro D (2007) Genetic characterization of ninety elite soybean cultivars using coefficient of parentage. Pesq Agr Brasil 42:363-369.

Narvel JM, Fehr WR, Chu WC, Grant D and Shoemaker RC (2000) Simple sequence repeat diversity among soybean plant introductions and elite genotypes. Crop Sci 40:1452-1458.

Odong D, van Heerwaarden J, Jansen J, van Hintum T and van Eeuwijk F (2011) Determination of genetic structure of germplasm collections are traditional hierarchical clustering methods appropriate for molecular marker data? Theor Appl Genet. 123:195-205.

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

Priolli RHG, Mendes-Junior CT, Arantes NE and Contel EPB (2002) Characterization of Brazilian soybean cultivars using microsatellite markers. Genet Mol Biol 25:185-193.

Priolli RHG, Pinheiro JB, Zucchi MI, Bajay MM and Vello NA (2010) Genetic diversity among Brazilian soybean cultivars based on SSR loci and pedigree data. Braz Arch Biol Techn 53:519-531.

Pritchard JK, Stephens M and Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155:945-959.

Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA and Feldman MW (2002) Genetic structure of human populations. Science 298:2381-2385.

Schoen DJ and Brown AHD (1993) Conservation of allelic richness in wild crop relatives is aided by assessment of genetic-markers. Proc Natl Acad Sci USA 90:10623-10627.

Song QJ, Quigley CV, Nelson RL, Carter TE, Boerma HR, Strachan JL and Cregan PB (1999) A selected set of trinucleotide simple sequence repeat markers for soybean cultivar identification. Plant Var Seeds 12:207-220.

Song QJ, Marek LF, Shoemaker RC, Lark KG, Concibido VC, Delannay X, Specht JE and Cregan PB (2004) A new integrated genetic linkage map of the soybean. Theor Appl Genet 109:122-128.

Tavaud-Pirra M, Sartre P, Nelson R, Santoni S, Texier N and Roumet P (2009) Genetic diversity in a soybean collection. Crop Sci 49:895-902.

Ude GN, Kenworthy WJ, Costa JM, Cregan PB and Alvernaz J (2003) Genetic diversity of soybean cultivars from China, Japan, North America, and North American ancestral lines determined by amplified fragment length polymorphism. Crop Sci 43:858-1867.

Van Inghelandt D, Melchinger AE, Lebreton C and Stich B (2010) Population structure and genetic diversity in a commercial maize breeding program assessed with SSR and SNP markers. Theor Appl Genet 120:1289-1299.

Wang LX, Guan RX, Liu ZX, Chang RZ and Qiu LJ (2006a) Genetic diversity of chinese cultivated soybean revealed by SSR markers. Crop Sci 46:1032-1038.

Wang LX, Guan Y, Guan RX, Li YH, Ma YS, Dong ZM, Liu X, Zhang HY, Zhang YQ, Liu ZX, *et al.* (2006b) Establishment of Chinese soybean *Glycine max* core collections with agronomic traits and SSR markers. Euphytica 151:215-223.

Yan M and Ye K (2007) Determining the number of clusters using the weighted gap statistic. Biometrics 63:1031-1037.

Internet Resources

CONAB (2012) National Company of Food Supply, http://www.conab.gov.br/ Brazilian Grain Crops 2012/2013 (November 12, 2012).

Wysmierski PT (2010) Genetic contribution of soybean ancestors to Brazilian soybean cultivars. Dissertation, University of Sao Paulo, São Paulo, Brasil http://www.teses.usp.br/teses/disponiveis/11/11137/tde-11022011-105217.

Supplementary Material

The following online material is available for this article:

Table S1 - Soybean genotypes and their pedigrees.

Table S2 - Frequency of the oldest ancestors and intermediate parents of each subgroup resulting from Bayesian analysis.

This material is available as part of the online article from http://www.scielo.br/gmb.

Associate Editor: Everaldo Gonçalves de Barros