Genetic Diversity in a Soybean Collection

M. Tavaud-Pirra,* P. Sartre, R. Nelson, S. Santoni, N. Texier, and P. Roumet

ABSTRACT

Soybean [*Glycine max* (L.) Merr.] was domesticated in China, and cultivated landraces were initially distributed throughout Asia and more recently extended to Europe and America. Previous studies of genetic diversity suggest a strong genetic bottleneck between Asian and North American soybean genetic pools. However, little is known about the potentially useful genetic diversity present in European soybean germplasm. We evaluated genetic diversity by phenotypic characterization and by comparing nuclear and cytoplasmic microsatellites in 301 genotypes of an INRA soybean collection with those of 31 European breeding lines and 17 ancestors from American cultivars (representing 83% of the North America diversity). We showed that the INRA collection contains 14 and 8% more diversity than the European breeding lines and American ancestors, respectively, based on Nei diversity (He). The genetic structure of this INRA collection does not correlate with either geographic origin or phenotypic differentiation. We generated a core collection composed of 50 accessions from the INRA collection and European breeding lines. This core collection contains 203 of the 226 microsatellite marker alleles available for this germplasm and is also representative of the diversity of morphological traits in temperate regions (maturity 000 to III), which will be useful for future breeding programs.

M. Tavaud-Pirra, P. Sartre, N. Texier, and P. Roumet, INRA-UMR DIAPC, Domaine de Melgueil, 34 130 Mauguio, France; R. Nelson, USDA-ARS, Soybean/Maize Germplasm, Pathology, and Genetics Research Unit, Dep. of Crop Sciences, 1101 W. Peabody Dr., Univ. of Illinois, Urbana, IL 61801; S. Santoni, INRA-UMR DIAPC, 2 place Viala, 34 060 Montpellier Cedex 1, France. Received 15 May 2008. *Corresponding author (Muriel.Tavaud@supagro.inra.fr).

Abbreviations: Fst, F-statistics; He, Nei diversity; PCA, principal component analysis; RAPD, random amplified polymorphic DNA; SSR, simple sequence repeat.

Cultivated soybean [*Glycine max* (L.) Merr.], one of the major crops used for animal feed and human foods, is mainly cultivated in the United States, Brazil, Argentina, and China. The domestication of *G. max* from the wild species (*Glycine soja* Sieb. and Zucc.) occurred in China but there is no consensus about the location within China (Carter et al., 2004). References to soybean appeared in Chinese literature almost as soon as written characters were developed during the period of the Shang dynasty between 1700 and 1100 BC (Qiu et al., 1999). *Glycine max* differs from *G. soja* with smaller pods and seeds, stems that are viny and twining, and nearly complete shattering at plant maturity. Hybridizations between wild and cultivated species produce fertile progenies. These intermediate forms were at one time called *G. gracilis* Skvortsov but Hermann (1962) removed *G. gracilis* from the species rank and incorporated it into *G. max* based on classical taxonomy. Chen and Nelson (2004) showed that semiwild accessions can be distinguished from *G. max* and *G. soja* based on either phenotypic or genotypic data, but do not necessarily support a separate species designation.

Published in Crop Sci. 49:895–902 (2009).
doi: 10.2135/cropsci2008.05.0266
© Crop Science Society of America
677 S. Segoe Rd., Madison, WI 53711 USA

Cultivation of soybean expanded from China to Korea and Japan about 2000 yr ago (Kihara, 1969) and later to other parts of Asia. In the 17th century in Europe, soybean was only known as an exotic plant from the Orient and was first described as a potential food plant in a report by Kaempfer in 1712 (Hymowitz, 1970). Soybean seeds were introduced in Europe from China at the beginning of the 18th century by missionaries and reintroduced on a number of other occasions, but these episodes are not well documented. In 1765, soybean was introduced for the first time into North America from China via London, by S. Bowen (Hymowitz, 2005). From 1765 to 1898, soybean was reintroduced from Europe and Asia to the United States by scientists, seed dealers, merchants, military expeditions, and individuals. In 1898, the Office of Foreign Seed and Plant Introduction was established within the USDA to centralize introduction activities (Hymowitz, 1984).

Given this complex history, we would expect major genetic bottlenecks occurring between wild and cultivated species, and from Asian landraces to European and North American accessions. Indeed, Chen and Nelson (2004), Li and Nelson (2002), and Xu and Gai (2003) observed a loss of genetic diversity from wild to cultivated soybeans: they reported a loss of *G. soja* random amplified polymorphic DNA (RAPD) fragments and a 35% reduction in the Nei diversity (He) index. Characteristic domestication bottlenecks were also detected with chloroplast microsatellites: among the 52 observed haplotypes in the wild species, only eight were found in the cultivated species (Xu et al., 2002). Using DNA sequences from 102 genes, Hyten et al. (2006) reported that diversity present in the wild species was halved in the domestication process and 81% of the rare alleles were lost.

The level of genetic diversity and geographic differentiation in Chinese cultivated soybean have been extensively studied using the coefficient of parentage (Cui et al., 2000), morphological traits (Dong et al., 2004), and simple sequence repeat (SSR) markers (Wang et al., 2006a), showing a clear geographic effect on genetic structure. Wang et al. (2006b) proposed a core collection of 2794 Chinese soybean accessions based on agronomic traits and SSR markers. Genetic differences between Japanese and Chinese germplasm have been assessed using SSR markers (Abe et al., 2003), RAPD (Li and Nelson, 2001), and isozymes (Hirata et al., 1999). Abe et al. (2003) suggested that soybean has been introduced repeatedly and independently from diverse Chinese germplasm pools into Japan, Korea, and South Asia.

Comparisons between the diversity of different samples of Asian soybean landraces and that of North American cultivars have demonstrated a lower level of diversity in the American pools than in the Asian pools, using either phenotypic characterization (Cui et al., 2001) or the coefficient of parentage (Cui et al., 2000). This reduced diversity was confirmed using sequence analyses to show successive genetic bottlenecks between wild and cultivated soybeans and between Asian landraces and North American cultivars (Hyten et al., 2006).

The narrow genetic base of North American soybean cultivars has been confirmed in many studies based on pedigree analysis (Delannay et al., 1983; Gizlice et al., 1994) or molecular markers (Narvel et al., 2000; Thompson et al., 1998). It is now widely accepted that fewer than 20 ancestor lines represent more than 85% of the genetic base of current North American soybean cultivars (Gizlice et al., 1994); in contrast, the USDA Soybean Germplasm Collection holds nearly 17,000 accessions of *G. max*. Other American genetic pools have also been studied (e.g., Brazil [Priolli et al., 2002] and Canada [Fu et al., 2007]).

Although Asian and American soybean diversity has been extensively studied, very little data on genetic diversity in European soybean collections has been published. One previous study analyzed diversity among 19 *G. max* lines from the Czech National Collection using RAPD markers (Baranek et al., 2002). This lack of information is probably due to limited soybean cultivation in Europe. However, given the history of *G. max* cultivation, soybean accessions present in European germplasm collections may be highly diverse, with genotypes adapted for temperate conditions, which may be useful for European breeding. In France at INRA (Institut National de la Recherche Agronomique), a collection of 2000 plant introductions is currently maintained ex situ; it is composed of accessions derived from various scientific and exchange programs with Asia and different countries of Europe. Three hundred one genotypes are not referenced in the USDA germplasm database, and we have labeled these genotypes the "INRA collection." The objectives of this research were to characterize this unknown germplasm using nuclear and chloroplast SSR markers and phenotypic traits, and to use these data to generate and validate a core collection to promote the use of this diversity in the INRA collection and to optimize germplasm management.

MATERIAL AND METHODS

Plant Material

We used 350 cultivated accessions (*G. max*), two wild accessions (*G. soja*), and 17 semiwild accessions (Table 1). The 350 cultivated soybean genotypes were classified into the three main groups described below.

Three hundred one accessions are maintained in the INRA soybean collection at Plant Breeding Station of Montpellier. Among them, 85 accessions are from Asia (including 57 from China) and 185 accessions are from various European countries (including 52 from France). Thirty-one accessions have an unknown origin.

Twelve cultivars (Argenta, Canton, Chandor, Essor, Imari, Kador, Kingsoy, Labrador, Major, Queen, Swift, and Weber) and 20 recent breeding lines were added to this collection, representing the genetic improvement of soybean in Western Europe from 1950 to 2000.

The following 17 accessions are the major contributing ancestors (or first progeny of ancestral lines) of North American cultivars: Arksoy, Capital, CNS, Dunfield, Haberland, Harrow (A.K.), Illini, Korean, Lincoln, Mandarin, Mukden, Ogden, Perry, Ralsoy, Richland, Roanoke, and S-100. These 17 lines represent 83% of genetic diversity in North American progenitors based on an analysis by Gizlice et al. (1994). Seeds of these accessions came from the USDA Soybean Germplasm Collection.

Table 1. Plant material analyzed.

Species	Group	Origin group	Group size
Glycine max	INRA collection	Asian accessions	85
Glycine max	INRA collection	European accessions	185
Glycine max	INRA collection	Other genotypes	31
Glycine max	European breeding lines		32
Glycine max	North American ancestors		17
Semiwild soybean			17
Glycine soja			2
Total			369

Agro-Morphological Traits

Field experiments were conducted over 2 yr (2002–2003) at the INRA Plant Breeding Station of Montpellier (France, 43°34′ N, 3°54′ E). For each accession, raw material was provided by seeds harvested from a single plant. At sowing, seeds were inoculated with *Bradyrhizobium japonicum* (G49 strain). Seeds were sown in single rows (2 m long and 0.45 m between rows) without replication. To control for environmental effects, four commercial lines representing a range of maturity levels (from 0 to II in 2002 and 00 to II the second year) were replicated every 30 rows and used as control lines. R1 and R8 stages were recorded (Fehr and Caviness, 1977). Based on main stem-termination characteristics, growth type ranged from 1 (determinate) to 5 (indeterminate). Plant height and lodging (scored from 0 [no lodging] to 5 [100% lodged plants]) were measured at maturity. Once harvested, seeds were cleaned and dried to approximately 7% moisture. Spectra data were obtained from whole seeds using a near infrared red system (NIRS Model 6500, Foss France, Paris), equipped with a sample transport module over a wavelength range of 400 to 2498 nm (2-nm increments). For each sample, we analyzed the average spectrum from 32 repeat scans; each sample was analyzed twice. Seed protein content was predicted using an equation developed in UMR DIAPC laboratory ($R^2_{\text{calibration}} = 0.970$; standard error of cross-validation = 0.591, sample number = 217).

Analysis of Nuclear and Chloroplast Microsatellites

To analyze genetic diversity, we chose 18 nuclear microsatellite markers with trinucleotide core repeats (ATT), developed by Cregan and colleagues (http://soybase.agron.iastate.edu/resources/ssr.php; Cregan et al., 1999), each on a different linkage group (Table 2). We also analyzed six chloroplast microsatellites developed by Powell et al. (1995) and adapted for soybean studies by Xu et al. (2002) (Table 3). The microsatellite RD19 was monomorphic in our samples and thus was not analyzed further.

Table 2. Description of the 18 nuclear microsatellites used including forward and reverse primers, linkage groups, and range of allele size.

Name	Forward primer	Reverse primer	Linkage group	Allele no.	Range	He[†]	Ho[‡]
					bp		
Satt114	GGGTTATCCTCCCCAATA	ATATGGGATGATAAGGTGAAA	F	9	75–115	0.737	0.011
Satt147	CCATCCCTTCCTCCAAATAGAT	CTTCCACACCCTAGTTTAGTGACAA	D1a	13	170–227	0.723	0.003
Satt156	CGCACCCCTCATCCTATGTA	CCAACTAATCCCAGGGACTTACTT	L	8	200–227	0.734	0.003
Satt168	CGCTTGCCCAAAAATTAATAGTA	CCATTCTCCAACCTCAATCTTATAT	B2	11	178–233	0.735	0.005
Satt172	AGCCTCCGGTATCACAG	CCTCCTTTCTCCCATTTT	D1b	10	206–236	0.751	0.005
Satt177	CGTTTCATTCCCATGCCAATA	CCCGCATCTTTTTCAACCAC	A2	8	98–119	0.676	0.005
Satt186	CGCTTGCCCAAAAATTAATAGTA	CCATTCTCCAACCTCAATCTTATAT	D2	14	200–245	0.764	0.011
Satt197	CACTGCTTTTTCCCCTCTCT	AAGATACCCCCAACATTATTTGTAA	B1	10	130–189	0.775	0.014
Satt268	CTACGAGAACTCATAGAATAGAACA	TGCTAGTGGAAGCCATTTAT	E	12	201–252	0.758	0.008
Satt281	GCGTACACCTCTTTTGATGAC	GCGAGTAACATGAAGTCTACGATAACA	C2	28	156–253	0.843	0.011
Satt300	TGGAGTAAACCATCAATTAATTGTGTG	ATTATGCGTTGATGCGACTGTTA	A1	9	232–262	0.688	0.003
Satt324	GTTCCCAGGTCCCACCATCTATG	GCGTTTCTTTTATACCTTCAAG	G	10	198–243	0.598	0.003
Satt354	CAAATAAAAATGGACACCAAAAGTA	AATTGCCAAAAATAGCCACAC	I	27	178–277	0.852	0.006
Satt414	GCGTATTCCTAGTCACATGCTATTTCA	GCGTCATAATAATGCCTAGAACATAAA	J	16	262–330	0.824	0.003
Satt434	GCGTTCCGATATACTATATAATCCTAAT	GCGGGGTTAGTCTTTTTATTTAACTTAA	H	19	262–360	0.591	0.017
Satt441	AAACCCACCCTCAAAAATAAAAA	AAATGCACCCATCAATCACA	K	20	239–315	0.848	0.008
Satt590	GCGCGCATTTTTTAAGTTAATGTTCT	GCGCGAGTTAGCGAATTATTTGTC	M	23	256–344	0.872	0.011
Satt592	GCGAAGATTGGTCTTTTATGTCAAATG	GCGGAGGAATACAAGTCTCTATTCAA	O	6	230–270	0.654	0.003
All				253		0.746	0.007

[†] Nei diversity.

[‡] Observed heterozygosity.

Table 3. Haplotypes defined using five chloroplast microsatellites and types of genotypes in each haplotype.

Haplotypes	Chloroplast microsatellites					Class size	Group composition
	gmcp1	gmcp2	gmcp3	gmcp4	soycp		
H1	121	138	107	131	91	336	7 semiwild soybean, 282 *G. max* from collection, all U.S. ancestral lines, 30 European breeding lines
H2	120	137	106	132	91	16	9 semiwild soybean, 7 *G. max* from collection
H3	121	138	106	131	91	5	3 *G. max* from collection, 2 European breeding lines
H4	121	138	107	130	91	4	4 *G. max* from collection (Asia)
H5	120	137	107	132	90	4	4 *G max* from collection (Poland and Japan)
H6	121	138	107	132	91	1	1 *G. max* from collection (Canada)
H7	120	137	106	131	90	1	1 semiwild soybean
H8	120	137	107	131	91	1	1 *G. soja*
H9	120	137	103	131	91	1	1 *G. soja*

Molecular Protocols

Total DNA was extracted from soybean young leaves according to Dneasy Plant Mini Kit (Qiagen, Courtaboeuf, France) with the following modification: 1% of polyvinylpyrrolidone (PVP 40,000) was added to buffer AP1. Polymerase chain reaction was performed in a final volume of 20 μL containing 5 ng of DNA, 200 μM of each dNTP, 1 U of *Taq* polymerase (Qiagen), 4 pmol of the fluorescent forward primer, and 1 pmol of reverse primer in Qiagen buffer (1×). Polymerase chain reaction conditions were an initial denaturation step at 95°C for 4 min, followed by 35 amplification cycles of template denaturation at 95°C for 30 s, primer annealing at 47°C for 30 s, primer extension at 72°C for 30 s, and a final elongation at 72°C for 5 min. Electrophoresis was performed using ABIPRISM 3100 sequencer and allele sizes were determined using GENESCAN and GENOTYPER softwares (Applied Biosystems, Foster City, CA). We used four reference genotypes present in each molecular analysis to validate these results; the reproducibility between analyses was 100%.

The whole molecular data are included in Supplementary Table 1.

Data Analysis

Principal component analysis (PCA) was performed using supplier's software (Winisi software, Foss France) to reduce the size of the spectrum matrix and to generate independent variables. Broad-sense heritability was computed on agro-morphological data using Proc Varcomp procedure (SAS Institute, 1999).

Alleles were scored by the length of the polymerase chain reaction product. Mean allele number, He indices, and F-statistics (Fst) were estimated using the software GENETIX (available at http://www.univ-montp2.fr/~genetix/genetix/genetix.htm; Belkhir et al., 1996–2004). We calculated the frequency of alleles specific to each group. To compare genetic diversity between different groups, we standardized allelic richness in a fixed sample size (e.g., size of North American ancestral group was 17) using the rarefaction method, described by Petit et al. (1998). Sign tests were used to compare standardized allelic richness in different groups.

The genetic structure in the INRA collection and European breeding lines (i.e., 333 genotypes) was inferred using the software STRUCTURE developed by Pritchard et al. (2000) with 18 nuclear microsatellites and six haplotypes based on chloroplast variation. We used burn-in periods of 50,000 followed by 1,200,000 Markov chain Monte Carlo repeats. Haploid data were used because of the very high level of homozygosity. Markers used mapped to different linkage groups and could therefore be assumed to be independent; a no-admixture model was performed. We calculated the most probable number of populations (K), ranging between one and eight, with four runs for each K. We determined the optimal value for K based on likelihood variance values obtained for each run. We then used optimal run to define the STRUCTURE clusters. Nei diversity indices and cluster differentiation (Fst) were estimated using the GENETIX software, as mentioned above.

We constructed a core collection and analyzed diversity captured using MSTRAT software (http://www.montpellier.inra.fr/gap/MSTRAT/mstratno.htm, Gouesnard et al., 2001) with 50 replicates and 10 iterations. We performed sampling based on the maximization strategy, which maximizes first the allele richness at each marker locus in the core-collection subset and second the Shannon diversity index (Schoen and Brown, 1993). Sampling quality was assessed by the proportion of allele classes represented in the core sample. For each quantitative trait, the number of classes was defined by the ratio of the difference between the maximum and minimum trait values to twice the residual standard deviation obtained from ANOVA analysis.

RESULTS

Agro-Morphological Traits

For each trait, no significant differences between subsets of control lines were observed over the 2 yr (data not shown). Thus, experiments were considered as uniform and data comparisons did not require correction.

Most accessions were indeterminate types (>73%) and plants with gray pubescence and white flowers represented 40 and 33%, respectively, of the total. Each quantitative trait was highly variable; accession effect accounted for most of this variation (at least seven times residual standard deviation; Table 4). Broad-sense heritability of different traits varied between 0.41 (seed protein concentration) and 0.92 (R1). As expected, highly significant correlations ($0.39 < r < 0.81$, $P < 0.0001$) were observed between phenological traits (R1, R8, reproductive phase length), lodging score, and plant height, whereas seed protein concentration was only weakly correlated with lodging ($r = 0.12$, $P < 0.05$).

Spectra Data

We performed PCA on nonaveraged spectra from the 2 yr. Ten principal component axes were obtained, each representing between 0.2 and 68% of the total spectra variance (Table 5). Heritability of each PCA axis was computed and reported: coefficients ranged between 0.125 (Axis 7) and 0.982 (Axis 1). The three highest heritability coefficients observed (Axes 1, 3, and 6) were mainly built on spectra from the visible region of the spectra (400– 800 nm) (data not shown); thus, these high heritability values may due to the relationship between these spectra and seed color.

Table 4. Agro-morphological trait variation among soybean accessions in the INRA collection.

Traits	Range	Average	Accession effect (% total sum of squares)	Residual SE	Heritability
R1†	30–97	44.6	95.8	2.68	0.92
R8†	85–170	118.5	89.7	6.63	0.78
Reproductive phase duration, d	46–99	73.8	75.0	6.38	0.48
Plant height, cm	20–180	84.3	67.7	12.87	0.68
Lodging score	1–5	1.92	74.9	0.66	0.66
Seed protein concentration, g kg^{-1} dry matter	27.2–52.7	43.6	64.1	2.39	0.41

†Days after sowing.

Table 5. Principal component analysis axis inertia and heritability.

NIRS† axis no.	Individual inertia (% of total variation)	Heritability	Axis genetic value‡
1	68.01	0.98	66.02
2	12.51	0.67	8.38
3	6.56	0.80	5.27
4	4.11	0.59	2.43
5	3.33	0.14	0.47
6	1.81	0.88	1.60
7	1.69	0.12	0.21
8	0.42	0.33	0.14
9	0.30	0.42	0.13
10	0.22	0.24	0.05

†Near infrared spectrometry.

‡Calculated as the product of individual inertia axes with their heritability.

Microsatellite Polymorphisms

Among the 369 genotypes analyzed, we observed a total of 253 alleles corresponding to the 18 nuclear microsatellites; the allele number for each locus varied between 6 and 28 (Table 2). Of these 253 alleles, 172 (68%) were rare, with a frequency less than 0.05 in the whole sample studied. Nuclear microsatellite analysis revealed a low observed heterozygosity, estimated at 0.007, as expected for autogamous species such as *G. max* and *G. soja*. The mean He was 0.75 (from 0.59 to 0.87 per locus). The frequency of group-specific alleles, despite its small sample size, was higher for *G. soja* (0.49) than for *G. max* or semiwild (0.27 and 0.07, respectively).

For the five chloroplast microsatellites analyzed across the whole data set, 12 alleles defining nine haplotypes were observed (Table 3). Twenty-five percent of chloroplast microsatellite alleles were present at low frequency (<0.05). Gene diversity estimated for five loci across the 369 genotypes was 0.101 and mean allele number was 2.4.

The two wild genotypes (*G. soja*) represent two haplotypes, H8 and H9 (Table 3). Allele 103, corresponding to microsatellite gmcp3, was observed only in the wild species. We detected three haplotypes, H1, H2, and H7, among the 17 semiwild types. Most *G. max* genotypes (282 out of 350 genotypes) belong to the first haplotype, H1. All U.S. ancestral lines and most recent European breeding lines harbored the most frequent haplotype H1. Twelve *G. max* from the INRA collection had four rare haplotypes (H3, H4, H5, and H6); seven *G. max* shared the haplotype H2 with nine semiwild types.

Genetic Diversity in *Glycine max*

Among 350 *G. max* genotypes, we observed 222 alleles for 18 nuclear microsatellites; 142 of these alleles had a frequency less than 0.05. Gene diversity (He) was 0.734, and the mean allele number was 12.3 alleles per microsatellite. Of the 96 alleles observed in North American ancestral lines, 93 were present in INRA collection and 73 were present in European breeding lines. Additionally, we observed 125 and 17 specific alleles in the INRA and European breeding lines, respectively, which were not present in North American ancestral lines (Fig. 1). In the INRA collection, the diversity estimated by the standardized allelic richness in Asian accessions (5.4) and Europe accessions (5.1) was significantly higher than observed in European breeding lines (4.2; Table 6). In *G. max*, there were six haplotypes defined by the chloroplast microsatellites (Tables 3 and 6). We observed five different haplotypes in Asian genotypes, four in the INRA collection, two in European breeding lines, and only one in North American ancestors.

Structure in *G. max* Sample

F-statistics values estimating differentiation between the five geographic or genetic groups (as defined in Table 7) were low, not exceeding 4.2% (with an average of 2%); thus, geographical origin did not strongly influence genetic structure in this collection.

We used STRUCTURE software to determine if an alternative structure could be elucidated with only molecular data. We identified three clusters: Cluster 1 was composed of 178 genotypes, with 35 Asian genotypes, 91 European,

Table 6. Diversity in different groups of the INRA soybean collection, European breeding lines, and North American ancestral lines.

	Sample size	He†	Ho‡	Alleles observed	Mean allele no.	Standardized allele no.	Haplotypes observed§
Accessions from Asia	85	0.73	0.006	167	9.3	5.4	H1, H2, H3, H4, H5
Accessions from Europe	185	0.72	0.009	176	9.8	5.1	H1, H2, H3, H5
Others accessions	31	0.72	0.006	119	6.6	5.1	H1, H2, H9
Entire INRA collection	301	0.74	0.007	218	12.1		
European breeding lines	32	0.65	0.004	90	5.0	4.2	H1, H3
North American ancestors	17	0.68	0.000	96	5.3	4.8	H1

†Nei diversity.

‡Observed heterozygosity.

§Based on chloroplast microsatellite variation.

21 with other origin, and 31 of the 32 European breeding lines; Cluster 2 contained 32 Asian and 55 European genotypes and two of other origin; Cluster 3 was composed of 18 Asian and 39 European genotypes, eight of other origin, and one European cultivar (Major). The mean Fst value for these three clusters was 0.147, with 33 alleles specific to Cluster 1, 28 specific to Cluster 2, and 29 specific to Cluster 3. Fst values, calculated from pairs of clusters, are listed in Table 8; all Fst values were significant and were higher that those obtained previously for different geographic groups.

A similar trend was observed with phenotypic traits: the phenotypes were highly variable within geographical origins and, consequently, the differences between these groups remained small. The three clusters from STRUCTURE analysis were also more differentiated for these phenotypic traits: the sum of squares (SSE) for the cluster effect represented between 2.5 and 18.3% of the total variation, depending on the observed trait, whereas SSE for geographic origin represented only 1 to 3.7% for the same traits. On average, Cluster 1 accessions matured later than those from other groups (+8 d on average); they were taller (+14 cm); and their grain protein concentration was reduced (−2.1 points). Cluster 3 accessions were closest to the semideterminate growth type, and they were more sensitive to lodging (+0.4 points).

Defining a Core Collection

For efficient use of this genetic diversity, a sample of the INRA collection and European breeding lines was built based on the MSTRAT maximization method. Fifty accessions (15% of the total sampling) identified with this method captured more than 90% of the global allelic richness available (see Supplementary Fig. 1); this was much more efficient than using the random method (60% of global diversity with the same sample size).

Table 7. Differentiation estimated by *F*-statistics values between defined groups.

F-statistics	INRA collection		European inbred lines (*G. max*)	N. American ancestors (*G. max*)
	G. max from Asia	*G. max* not assigned		
G. max from Europe	0.015***	0.002	0.028***	0.029***
G. max from Asia		0.011**	0.042***	0.031**
G. max not assigned			0.014*	0.002
European breeding lines				0.025*

*$P < 0.05$.

**$P < 0.01$.

***$P < 0.001$.

We defined a core collection of 50 genotypes containing 203 alleles out of the 226 available alleles from the total collection and European breeding genotypes. This core collection is composed of four European lines (including the cultivar Queen), and 46 belonging to the INRA collection, including 24 from Europe and 17 from Asia. It contains accessions from the three clusters, with 19 from Cluster 1, 13 from Cluster 2, and 18 from Cluster 3. The core collection is indicated in Supplementary Table 1.

The efficiency of this core collection to capture additional diversity was tested on quantitative traits that had not been used to build the core collection. The core set included about 68% of the diversity for these traits (39 of 57 classes). This capture rate for variation of quantitative traits was twice that obtained from a random sampling of 50 accessions (35% of the whole diversity), highlighting how efficient the MSTRAT maximization method is.

DISCUSSION

Genetic Diversity

This study is the first to evaluate genetic diversity in a European soybean germplasm collection. The INRA germplasm collection appears to be more useful for widening the genetic diversity of soybean breeding programs than the North American ancestors or European breeding lines. The INRA collection contains 127 alleles that are not observed among American progenitors, 130 that are not observed among European inbred lines, and four different cytoplasmic haplotypes that are not present in either of these groups. These allele numbers were compared to the results from Abe et al. (2003) and Narvel et al. (2000) for two and three common loci. As results, genetic diversity in European germplasm was lower than for Asian germplasm: up to six alleles observed in the Asian sampling by Abe et al. (2003) were not observed in the INRA collection; whereas some alleles present in the INRA collection (up to six) were not detected in either the American PIs or elite lines of the sampling by Narvel et al. (2000). These results let us to consider the INRA collection diversity level as being intermediate between Asian and American germplasm.

Despite the small sample collected, we observed substantial specific diversity for *G. soja*, in terms of both chloroplast and nuclear polymorphisms. Using the same chloroplast microsatellites, Xu et al. (2002) also identified alleles specific to *G. soja*. Studies based on molecular markers (Hyten et al., 2006; Xu and Gai, 2003) also indicate that the wild species contains original diversity, consistent with our observations. Despite recurrent introgressions between *G. soja* and *G. max* during the history of these species, the wild pool remains a source of original diversity.

Core Collection

We used molecular markers to build a core collection that is representative of (i) the genetic diversity of the total collection and (ii) the variability of certain quantitative traits. Wang et al. (2006b) demonstrated the efficiency of core collections to capture the genetic diversity of a large collection using 2% of total accessions to represent about 70% of the diversity from a whole soybean sample set.

However, few studies have used a core based only on SSR information to achieve efficient sampling of quantitative trait variability. Some authors (Wang et al., 2006b) have reported little overlap between molecular markers and agro-morphological traits. The close relationship between molecular markers and agro-morphological traits in our study, suggests that they are in linkage disequilibrium; consequently, SSR markers were sufficient to achieve a representative sample independent of the nature of the observed data. The linkage disequilibrium was partly due to the structure within the INRA collection, revealed by STRUCTURE software.

Table 8. Differentiation estimated by F-statistics values between clusters determined by the software STRUCTURE using only molecular data.

	Cluster 2	Cluster 3
Cluster 1	0.115***	0.179***
Cluster 2		0.160***

***$P < 0.001$.

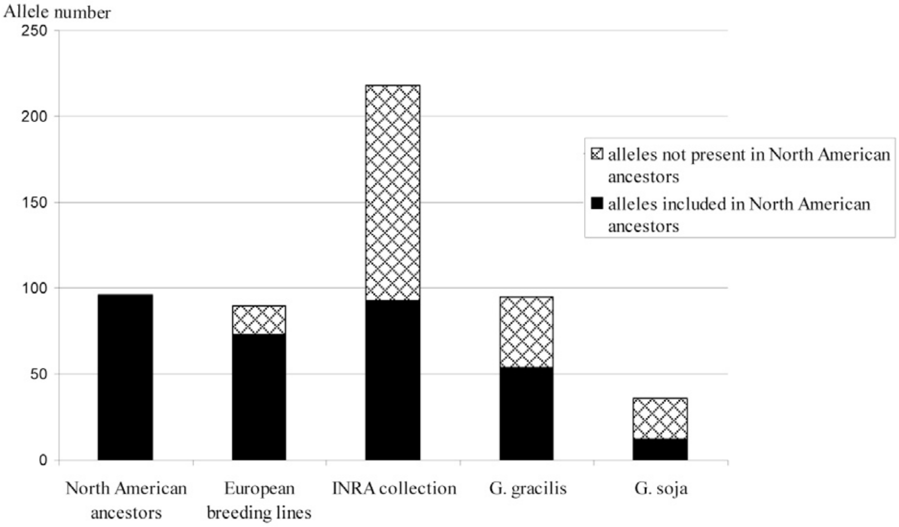

Figure 1. Number of alleles in 32 European breeding lines, 301 INRA collection genotypes, 17 semiwild, and two *Glycine soja*, compared to alleles included in 17 North American ancestors.

Genetic Structure in the INRA Collection

The absence of detailed passport data, such as collect site, introduction date in Europe, and pedigree of raw material, prevented us from performing the subsequent clustering and inferring the main evolutive factors having impacted European germplasm. In addition, analyses based on chloroplast microsatellites did not allow us to identify a clear structure because the H1 haplotype was much more frequent than the eight others. We used an approach based only on molecular data (STRUCTURE software) to define three clusters well differentiated (Fst = 0.147) not only for molecular data but also for agro-morphological traits; this suggests that the two data sets were not independent. The absence of external reliable information related to our sampling prevented interpretation of these clusters or comparison of this structure with regional genetic differentiation, as performed in previous studies (Xu et al., 2002; Xu and Gai, 2003; Wang et al., 2006a). Further complementary analyses of other well-documented collections such as the USDA Soybean Collection or Chinese soybean genetic resources may validate this clustering and allow its proper interpretation.

The tracing of evolutionary events and successive bottlenecks in soybean history requires extensive sampling based both on geographical and historical factors. Indeed, analysis of the genetic diversity available in Asia and the United States is important, as mentioned by Hyten et al. (2006), but genetic diversity in Europe, which received accessions from China and Japan as early as the 17th century, also needs to be taken into account. The core collection generated in this study may be used as a representative sample of European soybean diversity in such analyses. This improved understanding of *G. max* diversity is essential to widen the genetic base in breeding programs, and to promote genetic association programs.

Acknowledgments

The authors would like to thank anonymous reviewers for their comments of the manuscript. Special thanks to the Groupement d'Intérêt Economique des Sélectionneurs de Soja (GIE Soja, France) for the financial support of this work.

References

Abe, J., D. Xu, Y. Suzuki, A. Kanazawa, and Y. Shimamoto. 2003. Soybean germplasm pools in Asia revealed by nuclear SSR. Theor. Appl. Genet. 106:445–453.

Baranek, M., M. Kadlec, J. Raddova, M. Vachun, and M. Pidra. 2002. Evaluation of genetic diversity in 19 *Glycine max* (L.) Merr. accessions included in the Czech National Collection of Soybean Genotypes. Czech J. Genet. Plant Breed. 38:69–74.

Belkhir, K., P. Borsa, L. Chikhi, N. Raufaste, and F. Bonhomme. 1996–2004. GENETIX 4.05, logiciel sous Windows TM pour la génétique des populations. Laboratoire Génome, Populations, Interactions, CNRS UMR 5000, Université de Montpellier II, Montpellier, France.

Carter, T.E., R.L. Nelson, C.H. Sneller, and Z. Cui. 2004. Soybeans: Improvement, production, and uses. p. 303–416. *In* H.R. Boerma and J.E. Specht (ed.) Agron. Mongr. 16. ASA, CSSA, and SSSA, Madison, WI.

Chen, Y.W., and R.L. Nelson. 2004. Genetic variation and relationships among cultivated, wild, and semiwild soybean. Crop Sci. 44:316–325.

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

Cui, Z., T.E. Carter, Jr., and J.W. Burton. 2000. Genetic diversity patterns in Chinese soybean cultivars based on coefficient of parentage. Crop Sci. 40:1780–1793.

Cui, Z., T.E. Carter, Jr., J.W. Burton, and R. Wells. 2001. Phenotypic diversity of modern Chinese and North American soybean cultivars. Crop Sci. 41:1954–1967.

Delannay, X., D.M. Rogers, and R.G. Palmer. 1983. Relative genetic contributions among ancestral lines to North American soybean cultivars. Crop Sci. 23:944–949.

Dong, Y.S., L.M. Zhao, B. Liu, Z.W. Wang, Z.Q. Jin, and H. Sun. 2004. The genetic diversity of cultivated soybean grown in China. Theor. Appl. Genet. 108:931–936.

Fehr, W.R., and C.E. Caviness. 1977. Stages of soybean development. Cooperative Extension Service Special Rep. 80. Iowa State Univ., Ames.

Fu, Y.-B., G.W. Peterson, and M.J. Morrison. 2007. Genetic diversity of Canadian soybean cultivars and exotic germplasm revealed by simple sequence repeat markers. Crop Sci. 47:1947–1954.

Gizlice, Z., T.E. Carter, Jr., and J.W. Burton. 1994. Genetic base for North American public soybean cultivars released between 1947 and 1988. Crop Sci. 34:1143–1151.

Gouesnard, B., T.M. Bataillon, G. Decoux, C. Rozale, D.J. Schoen, and J.L. David. 2001. MSTRAT: An algorithm for building germplasm core collections by maximizing allelic or phenotypic richness. J. Hered. 92:93–94.

Hermann, F.J. 1962. A revision of the genus *Glycine* and its immediate allies. USDA Tech. Bull. 1268. Gov. Print. Office, Washington, DC.

Hirata, T., J. Abe, and Y. Shimamoto. 1999. Genetic structure of the Japanese soybean population. Genet. Resour. Crop Evol. 46:441–453.

Hymowitz, T. 1970. On the domestication of the soybean. Econ. Bot. 24:408–421.

Hymowitz, T. 1984. Dorsett–Morse soybean collection trip to East Asia: 50 year retrospective. Econ. Bot. 38:378–388.

Hymowitz, T. 2005. Debunking soybean myths and legends in the historical and popular literature. Crop Sci. 45:473–476.

Hyten, D.L., Q. Song, Y. Zhu, I.Y. Choi, R.L. Nelson, J.M. Costa, J.E. Specht, R.C. Shoemaker, and P.B. Cregan. 2006. Impacts of genetic bottlenecks on soybean genome diversity. Proc. Natl. Acad. Sci. USA 103:16666–16671.

Kihara, H. 1969. History of biology and other sciences in Japan in retrospect. Proc. XII Int. Cong. Genet. 3:49–70.

Li, Z., and R.L. Nelson. 2001. Genetic diversity among soybean accessions from three countries measured by RAPDs. Crop Sci. 41:1337–1347.

Li, Z., and R.L. Nelson. 2002. RAPD marker diversity among cultivated and wild soybean accessions from four Chinese provinces. Crop Sci. 42:1737–1744.

Narvel, J.M., W.R. Fehr, W. Chu, D. Grant, and R.C. Shoemaker. 2000. Simple sequence repeat diversity among soybean plant introductions and elite genotypes. Crop Sci. 40:1452–1458.

Petit, R.J., A. El-Mousadik, and O. Pons. 1998. Identifying populations for conservation on the basis of genetic markers. Conserv. Biol. 12:844–855.

Powell, W., M. Morgante, C. Andre, J.W. McNicol, G.C. Machray, J.J. Doyle, S.V. Tingey, and J.A. Rafalski. 1995. Hypervariable microsatellites provide a general source of polymorphic DNA markers for the chloroplast genome. Curr. Biol. 5:1023–1029.

Priolli, R.H.G., C.T. Mendes, Jr., N.E. Arantes, and E.P.B. Contel. 2002. Characterization of Brazilian soybean cultivars using microsatellite markers. Genet. Mol. Biol. 25:185–193.

Pritchard, J.K., M. Stephens, and P. Donnelly. 2000. Inference of population structure using multilocus genotype data. Genetics 155:945–959.

Qiu, L., R. Chang, J. Sun, X. Li, Z. Cui, and Z. Li. 1999. The history and use of primitive varieties in Chinese soybean breeding. p. 165–172. *In* H.E. Kauffman (ed.) Proc. World Soybean Res. Conf. VI, Chicago, IL. 4–7 Aug. 1999. Superior Print., Champaign, IL.

SAS Institute. 1999. SAS/STAT user's guide. SAS Inst., Cary, NC.

Schoen, D.J., and A.H.D. Brown. 1993. Conservation of allelic richness in wild crop relatives is aided by assessment of genetic markers. Proc. Natl. Acad. Sci. USA 22:10623–10627.

Thompson, J.A., R.L. Nelson, and L.O. Vodkin. 1998. Identification of diverse soybean germplasm using RAPD markers. Crop Sci. 38:1348–1355.

Wang, L., R. Guan, L. Zhangxiong, R. Chang, and L. Qiu. 2006a. Genetic diversity of Chinese cultivated soybean revealed by SSR markers. Crop Sci. 46:1032–1038.

Wang, L., Y. Guan, R. Guan, Y. Li, Y. Ma, Z. Dong, X. Liu, H. Zhang, Y. Zhang, Z. Liu, R. Chang, H. Xu, L. Li, F. Lin, W. Luan, Z. Yan, X. Ning, L. Zhu, Y. Cui, R. Piao, Y. Liu, P. Chen, and L. Qiu. 2006b. Establishment of Chinese soybean (*Glycine max*) core collections with agronomic traits and SSR markers. Euphytica 151:215–223.

Xu, D.H., J. Abe, J.Y. Gai, and Y. Shimamoto. 2002. Diversity of chloroplast DNA SSRs in wild and cultivated soybeans: Evidence for multiple origins of cultivated soybean. Theor. Appl. Genet. 105:645–653.

Xu, D.H., and J.Y. Gai. 2003. Genetic diversity of wild and cultivated soybeans growing in China revealed by RAPD analysis. Plant Breed. 122:503–506.