Chapter 2
Soybean Molecular Genetic Diversity

Perry B. Cregan

Introduction

The cultivated soybean [*Glycine max* (L.) Merr.] and the wild soybean (*Glycine soja* Seib. et Zucc.) are annuals and the two members of the *Glycine* subgenus. *G. soja* grows wild in China, Japan, Korea, Russia and Taiwan (Hymowitz 2004). It is generally accepted that cultivated soybean was domesticated 3000–5000 years ago on the Chinese mainland from the wild soybean (Hymowitz and Newell 1981). Cultivated soybean exhibits wide phenotypic variability in terms of seed shape, size, color, and chemical composition; plant morphology and maturity, as well as resistance to a broad range of biotic and abiotic stresses. This genetic diversity and the underlying genetic control of numerous specific traits were described in works such as the recent Third Edition of Soybeans: Improvement, Production and Uses (Boerma and Specht 2004). In particular, Carter et al. (2004) thoroughly documented genetic diversity in terms of the formation, collection, evaluation and utilization of diversity by soybean geneticists and breeders in North American and Asia over 70 years and the impacts of their work on genetic diversity. It is the intent of this review to specifically focus on molecular genetic diversity of the nuclear genome and the multitude of research that was directed at the assessment of molecular diversity of cultivated and wild soybean. This research employed a number of different molecular genetic tools beginning with the analysis of isozyme variation followed by a range of DNA marker types and ultimately variation in DNA sequence. The literature relating to the assessment of isozyme variability in *G. max* and *G. soja* recently received a thorough review by Palmer et al. (2004) and will not be considered here.

The first reports of the assessment of genome-wide molecular genetic diversity of the soybean nuclear genome began in the 1980s with the application of restriction fragment length polymorphism technology (RFLP) (Roth and Lark 1984; Apuya et al. 1988). Subsequent analyses employed RFLP, random amplified polymorphic

P.B. Cregan
Soybean Genomics and Improvement Laboratory, U.S. Department of Agriculture, Agricultural Research Service, Beltsville, Maryland
e-mail: creganp@ba.ars.usda.gov

G. Stacey (ed.), *Genetics and Genomics of Soybean*,

DNA (RAPD) or arbitrary primer PCR, amplified fragment length polymorphism (AFLP), microsatellite or simple sequence repeat (SSR), and DNA sequence analysis for the quantification of genetic diversity in both cultivated and wild soybean. This research had a number of different objectives including (1) the assessment of particular DNA marker systems for appropriately distinguishing and grouping cultivated and wild genotypes, (2) the quantification and comparison of diversity within and among various groups of cultivated and/or wild soybean genotypes (3) the use of genetic diversity estimates as tools in soybean breeding for increasing useful genetic variation, (4) the development of unique DNA fingerprints for genotype and cultivar identification and (5) the assessment of linkage disequilibrium.

Applicability of DNA Marker Types in Soybean

Restriction Fragment Length Polymorphism (RFLP)

Apuya et al. (1988) analyzed 300 RFLP probes selected as low-copy clones in Southern hybridizations to genomic DNA of the genetically distinct soybean cultivars Minsoy and Noir 1. Genomic DNAs were digested with a number of different restriction endonucleases in order to detect RFLP. Of the 300 probes examined only one in five was polymorphic. Despite the low level of polymorphism, 27 loci were analyzed in a population of F_2 plants derived from Minsoy × Noir 1. All loci segregated in a Mendelian fashion and 11 of the 27 loci were contained in four linkage groups. Keim et al. (1989) conducted a survey of RFLP via the analysis of 48 cultivated, eight wild and two *G. gracilis* genotypes using 17 probes to assess the allelic structure of RFLP markers and to identify diverse genotypes that would maximize variability in a resulting mapping population. The *G. gracilis* genotypes were previously joined with *G. max* (Hermann 1962) but were included to maximize morphological diversity in the sampling of genotypes. Extremely low levels of RFLP were recorded despite the diversity of the germplasm analyzed. Two of the 17 probes detected three alleles per locus while the remaining 15 detected only two. The *G. max* genotype A81-356022 and *G. soja* PI 468916 were identified as being particularly diverse with a high level of RFLP that was approximately two-fold higher than that of the Minsoy×Noir 1 cross identified by Apuya et al. (1988). Based upon these data, as well as previous analysis of these two genotypes, a mapping population was created from the interspecific cross of A81-356022 × PI 468916. In a subsequent report, Keim et al. (1992) analyzed 132 RFLP probes in 18 ancestors of U.S. cultivars (ancestral cultivars) as well as 20 adapted cultivars. One objective was to estimate the usefulness of the probes in revealing variation in adapted germplasm. Only one in five markers were informative in any pair of adapted soybean genotypes, again suggesting the relatively low level of RFLP particularly among adapted soybean genotypes.

Skorupska et al. (1993) assessed the feasibility of using the markers from the A81-356022 × PI 468916 RFLP map in the distinct subpopulation of soybean

genotypes with maturities adapted to the Southern U.S. A total of 108 genotypes, including older as well as elite cultivars and breeding lines, were analyzed with 83 RFLP probes. Fifty-four percent of the probes were non-informative while 35% had gene diversity values (the probability of detecting polymorphism between any two randomly selected genotypes) of ≥ 0.3. Despite the low levels of molecular diversity in the Southern germplasm pool, the authors indicated that polymorphic probes would serve as a core set for the genetic mapping of agronomic traits in Southern U.S. soybean germplasm. Lorenzen et al. (1995) analyzed 64 soybean ancestral and "milestone" cultivars at 217 RFLP loci to identify a core set of markers that would be useful for pedigree-based analyses of elite soybean cultivars. A set of 97 polymorphic loci were defined that could be used to trace genomic regions contributed by parents to their progeny. Of the 97 loci, 67 had gene diversity scores ≥ 0.30 in the set of 64 cultivars.

Simple Sequence Repeat (SSR) or Microsatellite Markers

The high level of variability and Mendelian inheritance of SSR DNA markers in plants was first reported in soybean by Akkaya et al. (1992) and Morgante and Olivieri (1993). Akkaya et al. (1992) assessed SSR allelic variation at two $(AT)_n$ and one $(ATT)_n$ SSR loci and reported from six to eight alleles among a group of 38 diverse *G. max* and five *G. soja* genotypes and concluded that SSRs would serve as an abundant source of highly polymorphic PCR-based genetic markers in soybean. Morgante and Olivieri (1993) reached similar conclusions regarding the utility and abundance of SSR loci in soybean. Subsequent reports by Rongwen et al. (1995) and Maughan et al. (1995) provided further demonstrations of the usefulness of SSRs for the assessment of genetic diversity in soybean. Rongwen et al. (1995) determined allelic variation at seven SSR loci in a diverse set of 96 soybean genotypes that included N. American cultivars, N. American ancestral cultivars, landraces from the USDA Soybean Germplasm Collection and from China as well as five *G. soja* accessions. From 11 to 26 alleles were found at the seven loci. Gene diversities ranged from 0.71 to 0.95 for the complete set of 96 genotypes and from 0.52 to 0.88 in the set of 28 N. American cultivars. It was concluded that SSRs would be an excellent complement to RFLP loci that were being used by soybean molecular geneticists at the time. Maughan et al. (1995) analyzed a similar set of genotypes including 62 *G. max* lines (landraces, ancestral cultivars, and adapted cultivars) and 32 wild soybeans from diverse Asian origins. From 5 to 21 alleles were detected at five SSR loci with gene diversities ranging from 0.55 to 0.81 in the complete set of 94 genotypes and from 0.29 to 0.62 among the 62 cultivated soybean lines. They suggested that SSRs were the "marker of choice" for a species such as soybean in which molecular genetic diversity is relatively limited. Similar conclusions regarding the usefulness of SSR markers to quantify molecular genetic variation were reached by Song et al. (1998) who analyzed 59 Korean landraces with eight SSR loci.

Amplified Fragment Length Polymorphism (AFLP) Markers

AFLP markers (Vos et al. 1995) are PCR based and permit the multiplex amplification of as many as 50 loci without prior knowledge of DNA sequence. The relatively low level of sequence variation in soybean would make AFLP an attractive alternative to RFLP. Maughan et al. (1996) assessed the use of AFLP in soybean via the analysis of 12 *G. max* and 11 *G. soja* genotypes with 15 AFLP primer pairs. A total of 759 fragments were amplified and 274 (36%) were polymorphic. The number of polymorphic fragments per primer pair varied from 9 to 27 with an average of 18.3. It was concluded that the capacity to rapidly detect thousands of genetic loci at relatively low cost made AFLP an ideal marker for a wide array of genetic investigations in soybean.

Random Amplified Polymorphic DNA or Arbitrary Primer PCR

RAPD (Williams et al. 1990) or AP-PCR markers (Welsh and McClelland 1990) are PCR based, require no prior knowledge of DNA sequence and are analyzed simply as the presence or absence of an amplicon via agarose gel electrophoresis. In order to identify a particularly informative set of RAPD primers, Thompson and Nelson (1998b) analyzed 125 random 10-base primers in 35 soybean genotypes that included 18 ancestral cultivars and 17 maturity group (MG) I-III landraces from the USDA Soybean Germplasm Collection. A total of 281 polymorphic RAPD fragments were identified of which 120 fragments from 64 primers were highly reproducible. A principal-components analysis was used to identify a core set of 35 primers that were critical to the analysis of the 35 genotypes. Thompson and Nelson (1998b) indicated that the correlation of pairwise distances between the 35 genotypes analyzed with the 35 selected primers was highly correlated with those based upon the complete set of RAPD fragment data. This set of 35 RAPD primers was subsequently used in a number of studies to assess molecular genetic variation in cultivated and wild soybean.

Variation in DNA Sequence

The ultimate measure of molecular genetic diversity is the direct comparison of DNA sequence. An important advantage of diversity estimates based upon variation in DNA sequence is the ability to compare across species. Initial estimates of DNA sequence variation were confined to single genes or DNA fragments with the goal of defining gene structure, function, or evolutionary relationships. Scallon et al. (1987) discovered three single nucleotide polymorphisms (SNPs) via the comparison of the 3543 bp sequence of the Gy_4 glycinin locus in the two cultivars Dare and Raiden. Two SNPs were discovered by Zakharova et al. (1989) in the 789 bp of cDNA sequence encoding the A_3B_4 glycinin subunit in the soybean cultivars Mandarin, Mukden and Rannaya-10. Zhu et al. (1995) sequenced a 400 bp fragment of RFLP probe A-199a in three diverse soybean genotypes and found a total of

nine SNPs. To compare SNP frequency among DNA fragments of varying length and between populations that vary in size, measures of nucleotide diversity including π (Tajima 1983) and θ (Watterson 1975) were devised, which are normalized for length and adjusted for sample size. Nucleotide diversity from the three aforementioned studies of soybean ranged from $\theta = 0.00085$ (Scallon et al. 1987) to $\theta = 0.015$ (Zhu et al. 1995). This translates into an average of 0.85–10.5 SNPs per kilobase of sequence. In order to provide an estimate of sequence variation in the soybean genome based upon a more extensive analysis of DNA sequence in a large sampling of genotypes, Zhu et al. (2003) analyzed more than 76 kbp of sequence in each of 25 diverse soybean genotypes. The 76 kbp included approximately 28.7 kbp of coding sequence, 37.9 kbp of non-coding perigenic DNA (introns, UTRs and associated genomic DNA), and 9.7 kbp of random non-coding genomic DNA. The mean nucleotide diversity expressed as θ was 0.00097 (an average of slightly less than one SNP per kilobase between any two genotypes in the set of 25 genotypes). Nucleotide diversity was 0.00053, 0.00114, and 0.00179 in coding, non-coding perigenic DNA, and random genomic DNA, respectively. Recent work by Choi et al. (2007) reported the discovery of SNPs in 4240 sequence tagged sites derived from amplicons produced with primers designed to soybean unigenes. In a total of 2.44 mbp of aligned sequence of six diverse genotypes, 4712 single base changes and 839 insertion–deletions (indels) were discovered. This translates to a nucleotide diversity of $\theta = 0.000997$.

The aforementioned reports of nucleotide diversity indicate that as compared to other species, diversity in soybean is relatively low. For example, in rice, Feltus et al. (2004) analyzed 358 mbp of draft sequences of the rice subspecies *Oryza sativa* ssp. *indica* and *japonica* and reported 1.7 single base changes plus 0.11 indels per kbp, which is the equivalent of a nucleotide diversity of $\theta = 0.00181$. Likewise, a calculation of nucleotide diversity in 21.3 kbp of sequence analyzed in five diverse barley cultivars by Kanazin et al. (2002) indicated $\theta = 0.0025$. In sorghum (*Sorghum bicolor*), Hamblin et al. (2004) reported nucleotide diversity of $\theta = 0.0023$, which is more than twice that of soybean. Likewise, Wright et al. (2005) reported nucleotide diversity of $\theta = 0.00627$ in modern maize (*Zea mays* L.) inbreds, while in sugarbeet (*Beta vulgaris* L.) a similarly high nucleotide diversity of $\theta = 0.0077$ was reported in a comparison of two genotypes by Schneider et al. (2001). While the level of DNA sequence variation in soybean is relatively low, it can nonetheless provide an excellent means to compare molecular genetic variability as suggested by the recent report of Hyten et al. (2006) in which nucleotide diversity in *G. soja* was compared with three distinct *G. max* populations.

Molecular Genetic Diversity Within and Among Various Groups of Soybean Genotypes

Numerous studies using a variety of DNA marker types reported on the levels of molecular genetic diversity within and between populations of both cultivated and wild soybean germplasm. These reports included comparisons of (1) adapted

cultivars, ancestral cultivars, and landraces, (2) cultivars and landraces from various Asian origins and (3) cultivated versus wild soybean genotypes. A number of studies in each of these categories are briefly summarized.

North American Adapted Cultivars, Ancestral Cultivars and Landraces

Based upon the analysis of 17 RFLP loci in 48 cultivated soybean genotypes including cultivars, ancestral cultivars and landraces, Keim et al. (1989) calculated Euclidean distances as a measure of diversity between individuals in the three groups. The average diversity among the landraces (0.37) was greater than that among the ancestral cultivars (0.26) which were in turn greater than that among the cultivars (0.16). Kisha et al. (1998) analyzed "gene pools" of cultivated soybean genotypes including 53 northern elite, 50 southern elite, 20 N. American ancestral cultivars, as well as 28 southern landraces (MG V-VIII) and 14 northern landraces (MG 0-IV) with 53 RFLP probes. A cluster analysis of these data generally grouped genotypes based upon their gene pool of origin although the ancestral cultivars were dispersed among the clusters. Based upon the average percent heterozygosity across all loci for each pool, the ancestral cultivar pool was determined to be the most diverse while the southern elite cultivars were the least diverse. It was concluded that more diversity was present between the northern elite, southern elite, northern landrace and southern landrace pools than within them. In the AFLP analysis by Maughan et al. (1996) a diverse set of 16 ancestral and adapted cultivars were analyzed along with 11 wild soybean genotypes. In this analysis, the ancestral and adapted cultivars clustered tightly together and separately from the wild soybean accessions. In another study involving ancestral cultivars of the N. American as well as the Chinese soybean germplasm pools, Li et al. (2001) compared the genetic diversity of 18 N. American and 32 Chinese ancestral cultivars using RAPD markers to establish the genetic relationships between the two groups of ancestors. Based upon mean genetic distance among cultivars within the N. American and Chinese ancestral groups, the N. American ancestors were determined to have a slightly lower level of genetic diversity. Cluster analyses generally separated the two gene pools. In particular, large differences were detected between the ancestors of northern U.S. and Canadian soybeans and the Chinese ancestors.

Diwan and Cregan (1997) assayed allelic variation in 35 N. American ancestral cultivars that represented 95% of the allelic variation present in North American cultivated soybean germplasm as determined via pedigree analysis (Gizlice et al. 1994). Twenty SSR loci were analyzed and an average of 10.1 alleles was detected per locus (range 5–17) with a mean gene diversity of 0.80 (range: 0.50–0.87). In an extensive study of SSR allelic diversity, Narvel et al. (2000) analyzed 39 adapted cultivars and 40 MG I-IV landraces which were selected for their yield potential in a replicated field trial. Each genotype was analyzed with 74 SSR loci distributed across the 20 consensus linkage groups and a total of 397 alleles were detected. There were 138

alleles specific to the landraces and only 32 alleles specific to the cultivars. Average gene diversity among the landraces was 0.56 and ranged from 0.0 to 0.84 while gene diversity among the adapted lines was 0.50 and ranged from 0.0 to 0.79. As would be anticipated, genetic similarity estimates based on simple matching coefficients revealed more genetic diversity among the landraces than among the cultivars.

Cultivated Soybean with Different Asian Origins

Li and Nelson (2001) used RAPD analysis with the objective of comparing genetic variation within and among 120 cultivated soybean accessions from eight Chinese and three South Korean provinces and three Japanese districts in an attempt to relate patterns of diversity to geographic origin. Of 115 polymorphic RAPD fragments, all were present in the Chinese accessions while only eight were not present in the S. Korean or Japanese accessions. Thus, divergence among the three national genepools was mainly a function of fragment frequencies. Genetic distances among genotypes ranged from 0.14 to 0.55 with a mean of 0.42. The highest genetic distances were between accessions from China versus those from Japan and S. Korea. The accessions from Japan and S. Korea had similar but much lower genetic distances than those from China. Cluster analyses generally put the Korean and Japanese genotypes together and separate from the Chinese accessions. It was concluded from these data that the S. Korean and Japanese gene pools were probably derived from a relatively few introductions from China.

In an extensive analysis of 131 *G. max* landraces and/or pureline selections from 14 Asian countries, Abe et al. (2003) assayed allelic variation at 20 SSR loci, one each from the 20 consensus linkage groups defined by Cregan et al. (1999). The landraces were primarily from China and Japan and were selected to represent the diversity of geographic regions in these two nations. Germplasm from southeast and south central Asia was also included. An extremely high level of allelic diversity was detected with an average of 11.9 alleles per locus and a mean gene diversity of 0.782. A cluster analysis separated the Japanese from the Chinese accessions and suggested their origins from different germplasm pools. Korean accessions clustered in both the Chinese and Japanese groups while the southeast and south central accessions clustered with the Chinese lines. It was concluded that the soybeans from southeast and south central Asia were derivatives of the diverse Chinese germplasm pool.

In an analysis of cultivated soybean of the seven primary ecotypes from the three Chinese production regions (Northern, Yellow River and Southern), Wang et al. (2006) analyzed 122 landraces and seven cultivars selected to represent the range of phenotypic diversity for 14 agronomic and morphological traits. Allelic diversity was determined at 60 SSR loci that were uniformly distributed across the 20 soybean linkage groups. An average of 12.2 alleles per locus was detected and gene diversity ranged from 0.5 to 0.92 with a mean of 0.78. A cluster analysis yielded five major groups, two that contained primarily Northern ecotypes, one

Yellow River ecotypes, one Southern ecotypes and one that contained both Northern and Yellow River ecotypes. The Yellow River ecotypes had the greatest allelic diversity and were present in each of the five clusters supporting the suggestion that the Yellow River is the center of diversity of Chinese soybean.

Cultivated Versus Wild Soybean

The loss of genetic diversity due to domestication of a cultivated species from its wild progenitor is a subject of interest in many crop species. The so-called "genetic bottleneck" of domestication can drastically reduce variability as a result of selection for traits such as non-shattering of seeds, loss of germination inhibition, erect growth habit, seed size and seed composition. In maize, Tenaillon et al. (2004) estimated a 38% loss of genetic variability through the domestication bottleneck from the maize progenitor Teosinte (*Zea mays* ssp. *Parviglumis*) to cultivated maize. Keim et al. (1989) provided one of the first molecular genetic comparisons of cultivated versus wild soybean using 17 RFLP markers. Based upon a Euclidean genetic distance measure they indicated that molecular diversity was least among the 18 cultivars examined and was greatest among a group of 10 *G. max* landraces. Surprisingly, diversity among eight *G. soja* lines was intermediate to the cultivars and landraces.

Powell et al. (1996) used 11 SSR loci to analyze 22 cultivated and 25 wild soybean genotypes. The cultivated genotypes included 10 genotypes, nine of which were ancestral cultivars. These were selected based on a high level of RFLP diversity. An additional 12 accessions were selected from throughout the geographical range of cultivated soybean in Asia. The *G. soja* accessions were similarly selected to represent the geographical range of the wild soybean in Asia. The diversity index, ($\hat{H}$) (Weir 1990) was used to measure variability at each locus within each set of genotypes. $\hat{H}$ was significantly greater among the *G. soja* genotypes for eight of the 11 SSR loci. The mean diversity over loci was $\hat{H} = 0.539$ and $\hat{H} = 0.830$ in *G. max* and *G. soja*, respectively [calculated from (Powell et al. 1996)]. It was concluded that the domestication of *G. max* from *G. soja* was a key factor influencing the lower level of genetic variability in cultivated soybean. Li and Nelson (2002) selected 10 wild soybean genotypes and 10 cultivars from each of four Chinese provinces. With the exception of two of the *G. max* accessions from one province, all of the cultivated soybeans were landraces or primitive varieties. DNA of each of the 80 genotypes was analyzed with a selected set of 35 previously identified RAPD primers (Thompson and Nelson 1998b) along with two additional primers which produced a total of 269 fragments, 172 of which were polymorphic. Euclidean distances (D_{ij}) were calculated between all pairs of lines. The mean distance among the *G soja* lines ($D_{ij} = 0.46$) was significantly greater than that among the *G. max* genotypes ($D_{ij} = 0.40$) indicating greater genetic diversity. RAPD analysis was also used by Xu and Gai (2003) to measure diversity in a set of 27 *G. max* landraces and 21 wild soybean genotypes from China. The cultivated accessions were selected

based upon geographical distribution and seasonal type and the wild soybeans based upon geographical origin. Data were collected from 20 RAPD primers producing 177 bands, of which 66 were polymorphic and none was species specific. The mean gene diversity was 0.188 in the cultivated and 0.285 in the wild soybean indicating significantly greater genetic diversity in *G. soja*.

Kuroda et al. (2006) used one randomly selected SSR locus from each of the 20 consensus soybean linkage groups to analyze 77 *G. soja* accessions collected from across Japan as well as 53 currently grown soybean cultivars. A total of 405 alleles were detected in the wild, and 109 in the cultivated genotypes. Mean gene diversity across the 20 loci was 0.870 in the wild accessions and 0.496 in the cultivated genotypes, which was indicative of significantly less genetic variability in cultivated versus wild soybean. Another estimate of genetic variability in cultivated versus wild soybean based on DNA sequence variation in 102 randomly chosen gene fragments was reported by Hyten et al. (2006). More than 55 kbp of sequence from each of 26 wild soybean genotypes was compared with that of 52 Asian landraces. Both sets of genotypes were selected to maximize diversity based upon geographic origin. Hyten et al. (2006) reported nucleotide diversity values of $\theta = 0.00115$ and $\theta = 0.00235$ in cultivated and wild soybean, respectively.

The molecular diversity values from Powell et al. (1996), Li and Nelson (2002), Xu and Gai (2003), Kuroda et al. (2006) and Hyten et al. (2006) each provide estimates of the loss of diversity through the domestication bottleneck (Table 2.1). Estimates of the proportion of diversity retained after domestication range from 0.49 (Hyten et al. 2006) to 0.87 (Li and Nelson 2002) with a mean of 0.65. However, the 37 RAPD loci used by Li and Nelson (2002) included 35 loci that were carefully selected for high levels of molecular diversity and thus probably do not provide unbiased estimates of diversity. In the remaining studies, the loci appeared to be selected at random, although this is not completely clear in the case of Xu and Gai (2003). If the Li and Nelson (2002) estimate is excluded, the proportion of diversity retained after domestication is 0.59 which is very close to the estimates in maize.

Table 2.1 Molecular genetic diversity values reported in studies comparing cultivated and wild soybean genotypes and the ratio of genetic diversity values in *G. max* versus *G. soja* as an estimate of diversity retained through the genetic bottleneck of domestication

	Genetic diversity estimate		Proportion of diversity retained
Data source	*G. max*	*G. soja*	*G. max/G.soja*
Powell et al. (1996)	0.538	0.830	0.65
Li and Nelson (2002)	0.40	0.46	0.87
Xu and Gai (2003)	0.188	0.285	0.66
Hyten et al. (2006)	0.00115	0.00235	0.49
Kuroda et al. (2006)	0.496	0.870	0.57
Mean			0.65

The Search for Increased Genetic Diversity in Soybean Breeding

Plant breeders are continuously searching for new sources of useful genetic variation that will positively impact their breeding programs. Concerns of the lack of genetic variation in soybean resulting from the narrow genetic base of Asian introductions and many years of intense selection (National Research Council 1972) led to a number of assessments of molecular genetic variability aimed at the discovery of unique variation that could facilitate continued gains in soybean yields. Sneller et al. (1997) evaluated molecular diversity of landraces with maturities adapted to the southern U.S., elite southern cultivars and the elite northern parents of north × south elite cultivar crosses at 60 RFLP loci. Agronomic evaluations were also conducted of the landraces and the progeny from the north × south crosses. The RFLP analysis indicated that the landraces and the northern elite cultivars were genetically divergent from the southern elite cultivars and from each other. While the agronomic characteristics of many of the landraces were inferior, some genetically diverse lines (based upon genetic distances calculated from the RFLP data) with better agronomic potential were identified that might serve as sources of useful genetic variability in breeding. Likewise, the genetically divergent northern cultivars were suggested as another source of genetic variation for use in southern U.S. breeding. The similar report by Kisha et al. (1998), described earlier, used 53 RFLP loci to study genetic relationships among northern elite cultivars, southern elite cultivars, ancestral cultivars and landraces with maturities adapted to both the northern and southern U.S. Much like Sneller et al. (1997), it was concluded that northern elite cultivars would be particularly useful for providing increased diversity to the southern U.S. germplasm pool. The southern landraces were also suggested as sources of diversity that would be useful in both the northern and southern germplasm pools.

Thompson et al. (1998) used data derived from 125 RAPD primers to compare 18 ancestral cultivars with 17 maturity group I–III Asian landraces that had produced high yielding progeny in crosses with adapted N. American cultivars. Genetic distances were calculated based upon the RAPD data and the pairwise distances ranged from 0.26 to 0.67 with an average of 0.56. The averages and ranges for the two groups were similar, indicating approximately the same diversity in the two groups of genotypes. However, in cluster analyses the landraces clustered apart from the ancestral cultivars suggesting that they may be a useful source of genetic diversity to be exploited in soybean breeding. Thompson and Nelson (1998a) reported that lines derived from crosses of seven of the diverse landraces identified by Thompson et al. (1998) with adapted cultivars produced progeny that out-yielded their adapted parent. This result indicated that exotic germplasm could contribute genes to enhance yield. Brown-Guedira et al. (2000) compared patterns of genetic diversity in the same set of 18 ancestral cultivars used by Thompson et al. (1998) with 87 lines that included MG 00-IV landraces that had produced progeny with high yields. A few U.S. cultivars with uncertain parentage were also included in this group of 87 genotypes. A total of 46 RAPD markers and 3 SSRs from different linkage groups were used to characterize molecular genetic variation. Genetic distances ranged from 0.08 to 0.76 with a mean of 0.52. Cluster analyses of the distance matrix identified 11 clusters, three of which were composed almost exclusively of

landraces that were distinct from the ancestral base of U.S. soybean cultivars. These accessions were considered to be of particular interest as sources of useful genetic variation for yield improvement of northern U.S. cultivars.

An AFLP analysis was used to compare the level of genetic diversity within and between 59 modern cultivars from China, 30 from Japan, 66 from N. America along with and 35 N. American ancestral cultivars (Ude et al. 2004). Genetic distance (GD) between pairs of genotypes was calculated on the basis of the similarity indices determined by the 332 AFLP fragments, 90 of which were polymorphic. Within each of the cultivar groups, the average GD between pairs of genotypes was 6.3% among the 30 Japanese cultivars, 7.1% among the 66 N. American cultivars, 7.3% among the 35 N. American ancestral cultivars and 7.5% among the 59 Chinese cultivars. The average GD between the N. American cultivars and the Chinese and Japanese cultivars was 8.5% and 8.9%, respectively. None of these distances was significantly different; however, the greater genetic distances between the N. American cultivars and those from China and Japan versus the distances among the N. American cultivars indicated that the Asian cultivars may be a useful source of genetic variation for cultivar improvement in N. America. A cluster analysis indicated that the Japanese cultivars were more removed from the N. American cultivars than were the Chinese cultivars and would probably be the better of the two sets of Asian cultivars as a source of genetic diversity for yield improvement in N. American breeding.

Molecular Genetic Diversity for Unambiguous DNA Fingerprinting

The first report of the use of DNA markers to distinguish soybean cultivars used RFLP loci. In their evaluation of RFLP loci, Apuya et al. (1988) discovered multiple locus RFLP probes that distinguished a set of five cultivars. Lorenzen and Shoemaker (1996) used between 37 and 50 RFLP loci to distinguish members of 17 "cultivar groups", where groups were defined as all accessions with a similar common name and the selections made from these cultivars. This analysis successfully identified cultivar groups that were originally a heterogeneous seed mixture such as A.K. (*A*ll *K*inds) and Manchu. In cases where no phenotypic diversity was reported between two members of a group, molecular genetic diversity was detected 40 of 44 times. These results clearly suggested the power of molecular marker technology to detect genetic differences for purposes of distinguishing phenotypically similar genotypes.

SSR Markers for DNA Fingerprinting

Diwan and Cregan (1997) examined the use of SSR loci to distinguish sets of cultivars that were phenotypically indistinguishable. A total of 10 MG I, seven MG II, 10 MG IV and 9 MG VI cultivars were identified such that cultivars within each

group were indistinguishable based upon eight morphological and pigmentation traits. The cultivars within the four groups were readily distinguishable using the 20 SSR loci. Keim et al. (1989) previously identified seven cultivars that could not be distinguished using 17 RFLP probes. The seven cultivars were readily distinguished using the 20 SSR loci. Cregan and Diwan (1997) found a mean of 2.95 alleles among the seven cultivars at the 20 loci. Subsequent research by Song et al. (1999) analyzed 48 SSR loci on the set of 35 N. American ancestral cultivars used by Diwan and Cregan (1997) along with a diverse set of 66 elite N. American cultivars. Only loci in which adjacent alleles differed by at least three basepairs were maintained for further statistical analysis via a clustering procedure. A final set of 13 loci from 12 different linkage groups was identified which easily produced unique SSR allele size profiles for each of the 66 elite cultivars. The 13 loci also readily distinguished the members of the four sets of cultivars examined by Diwan and Cregan (1997) that were phenotypically identical. The set of 13 loci was proposed by Song et al. (1999) as a standard set for use in DNA profiling of soybean cultivars for purposes of Plant Variety Protection.

SNP Markers for DNA Fingerprinting

The availability of genetically mapped SNP markers in soybean (Choi et al. 2007) provides an alternative source of DNA markers for genetic fingerprinting. Yoon et al. (2007) determined the allele present at each of 58 mapped SNP loci selected from across the 20 soybean consensus linkage groups. Each was analyzed in a set of cultivars that included 16 N. American ancestral cultivars, 59 elite N. American cultivars, 21 elite Korean cultivars, as well as the same four sets of MG I, II, IV and VI cultivars examined by Diwan and Cregan (1997) that were phenotypically indistinguishable. Based upon a clustering procedure, a set of 23 informative SNPs loci was identified. The 23 loci were spread across 19 of the 20 soybean consensus linkage groups. The 23 loci very efficiently distinguished the N. American ancestral, Korean and N. American cultivars as well as the cultivars within the four sets of phenotypically identical cultivars. This set of SNP markers provide an alternative to SSR markers for the DNA fingerprinting of cultivars or other germplasm.

Linkage Disequilibrium

Linkage disequilibrium (LD) is the non-random association of alleles at two or more loci and is affected by a number of factors. These include the (1) rate of recombination with higher recombination lowering LD, (2) population subdivision or admixture which increase LD, (3) selection which increases LD in the vicinity of selected loci, and (4) mutation rate with high mutation rate decreasing overall LD but increasing LD in proximity to newly mutated loci (Rafalski and Morgante 2004). Mating system is also an important determinant of LD. In selfing species such as

soybean, with high homozygosity, recombination occurs between identical haplotypes and thus does not reduce LD. LD is the basis of genetic association analysis for the discovery and fine mapping of genes or quantitative trait loci (QTL) in natural populations (Risch and Merikangas 1996). Genetic association analysis measures correlations between allelic variants and phenotypic differences in naturally occurring populations and depends on historical LD for the detection of significant associations (Flint-Garcia et al. 2003).

There are only two published estimates of LD in soybean. Zhu et al. (2003) indicated that LD significantly decayed at distances of 2.0–2.5 centiMorgans (cM) (roughly equivalent to 1.0–1.25 mbp) across a 12.5 cM region of linkage group G. In a more extensive examination of LD in four sets of soybean germplasm accessions including 26 *G. soja*, 52 landraces, 17 N. American ancestral cultivars and 25 modern cultivars, Hyten et al. (2007a) analyzed LD decay across three genome regions ranging in length from 336 to 574 kbp. In *G. soja*, LD was the least extensive and did not extend past 100 kbp; however, in the three cultivated soybean populations LD extended from 90 kb to 574 kbp. The extent of LD in the three *G. max* populations varied greatly between the three genome regions. The structure of LD was described using haplotype blocks that are consecutive loci in high LD flanked by blocks demonstrating historical recombination (Altshuler et al. 2005; Daly et al. 2001; Gabriel et al. 2002). Using common methods to define haplotype blocks (Barrett et al. 2005; Gabriel et al. 2002; Wang et al. 2002), Hyten et al. (2007a) determined that *G. soja* had haplotype blocks with an average block length of 4.8 kb/block. The largest haplotype block spanned 25 kb with the majority of blocks spanning <1 kb. The 52 landraces and the ancestral cultivar populations had similar size haplotype blocks, which were on average much larger than those of *G. soja.* The average block size in the modern cultivar population was more than twice that of any of the other populations as estimated by each of the three methods of haplotype block determination and there were only a few blocks that were <1 kb in length. Tag SNPs are defined as a subset of SNPs that capture a large fraction of the allelic variation of all SNP loci (Altshuler et al. 2005). Therefore, based upon allelic variation in the 52 landraces Hyten et al. (2007a) calculated the number of tag SNPs needed to capture 100% of alleles in the three genome regions. The estimate ranged from a SNP every 9 kb to a SNP every 51 kb. (Hymowitz 2004) estimated that the euchromatic DNA represented about 64% of the genome or approximately 705 Mb. Thus, to fully capture allelic variation in the euchromatic portion of the genome would require from 13,800 to 78,300 SNPs depending upon which of the three genomic regions is the most representative of the soybean genome.

While sequenced tagged sites containing only a few thousand SNPs have been mapped in soybean to date (Choi et al. 2007), the imminent availability of the Department of Energy, Joint Genome Institute whole genome sequence, as well as the sequence of alternative genotypes will greatly accelerate SNP discovery. In addition, the availability of high throughput SNP detection assays will expedite genotyping. A report at the Plant Animal Genome XV meeting in San Diego, CA (Hyten et al. 2007b) indicated that the Illumina Inc. GoldenGate SNP detection assay (http://www.illumina.com/pages.ilmn?ID=11) functions extremely well

in soybean despite its highly duplicated genome. It is likely that the Illumina Infinium assay (http://www.illumina.com/downloads/INFINWKFLOW.pdf) which is capable of the analysis of more than 100,000 SNPs in parallel will also function in soybean. Such analysis platforms have the potential for the rapid characterization of the entire genomes of thousands of diverse soybean genotypes.

Conclusions

The characterization of the molecular genetic diversity of various sets of wild and cultivated soybean genotypes began more than 20 years ago. A number of different DNA marker systems were used in these analyses. The technical advances from Southern hybridization-based analysis to various types of PCR-based markers have increased the speed and reduced the cost of data acquisition. These analyses were undertaken to meet wide ranging objectives from simply testing the usefulness of a particular marker system to identifying exotic germplasm accessions to expand the genetic diversity of the elite germplasm pool in order to permit genetic improvement for increased soybean yield. Recent advances in high throughput DNA sequencing technology for inexpensive SNP discovery and tools for the detection of tens of thousands of SNP DNA markers in parallel suggest that in the near future there will be few limits to our ability to characterize genetic diversity in very fine detail. In the next few years it is likely that the entire USDA Soybean Germplasm Collection of more than 17,000 accessions will be characterized at each of 100,000 or more loci and that this level of characterization will define the entire haplotype variation of each accession. At that point, the major question facing soybean genomicists will be how to most effectively mine this large dataset for the genetic improvement of soybean. Furthermore, the availability of such a dataset increases the need for rapid and accurate phenotypic analysis. The analysis of molecular diversity is, and will remain important, but until extensive and accurate phenotypic data are available with which to associate genotypic diversity we will not realize the genetic progress that appears to be possible.

References

Abe, J., Xu, D.H., Suzuki, Y., Kanazawa, A., and Shimamoto, Y. (2003) Soybean germplasm pools in Asia revealed by nuclear SSRs. Theor. Appl. Genet. 106, 445–53.

Akkaya, M.S., Bhagwat, A.A., and Cregan, P.B. (1992) Length polymorphisms of simple sequence repeat DNA in soybean. Genetics 132, 1131–9.

Altshuler, D., Brooks, L.D., Chakravarti, A., Collins, F.S., Daly, M.J., and Donnelly, P. (2005) A haplotype map of the human genome. Nature 437, 1299–320.

Apuya, N.R., Frazier, B.L., Keim, P., Roth, E.J., and Lark, K.G. (1988) Restriction fragment length polymorphisms as genetic markers in soybean, *Glycine max* (L.) Merrill. Theoretical and Applied Genetics. 75, 889–901.

Barrett, J.C., Fry, B., Maller, J., and Daly, M.J. (2005) Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics 21, 263–5.

Boerma, H.R., and Specht, J.E. (2004) *Soybeans: Improvement, Production, and Uses*. 3rd ed. American Society of Agronomy, Crop Science Society of America, Soil Science Society of America, Madison, WI.

Brown-Guedira, G.L., Thompson, J.A., Nelson, R.L., and Warburton, M.L. (2000) Evaluation of genetic diversity of soybean introductions and North American ancestors using RAPD and SSR markers. Crop Sci. 40, 815–823.

Carter, T.E., Nelson, R., Sneller, C.H., and Cui, Z. (2004) Genetic diversity in soybean:In: H. R. Boerma and J. E. Specht (Eds), *Soybeans: Improvement, Production, and Uses*. American Society of Agronomy, Crop Science Society of America, Soil Science Society of America, Madison, WI, pp. 303–416.

Choi, I.-Y., Hyten, D.L., Matukumalli, L.K., Song, Q.-J., Chaky, J.M., Quigley, C.V., Chase, K., Lark, K.G., Reiter, R.S., Yoon, M.-S., Hwang, E.-Y., Yi, S.-I., Young, N.D., Shoemaker, R.C., van Tassell, C.P., Specht, J.E., and Cregan, P.B. (2007) A soybean transcript map: gene distribution, haplotype and SNP analysis. Genetics 176, 685–696.

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., and Chung, J. (1999) An integrated genetic linkage map of the soybean genome. Crop Sci. 39, 1464–1490.

Daly, M.J., Rioux, J.D., Schaffner, S.F., Hudson, T.J., and Lander, E.S. (2001) High-resolution haplotype structure in the human genome. Nat Genet 29, 229–32.

Diwan, N., and Cregan, P.B. (1997) Automated sizing of fluorescent-labeled simple sequence repeat (SSR) markers to assay genetic variation in soybean. Theor. Appl. Genet. 95, 723–733.

Feltus, F.A., Wan, J., Schulze, S.R., Estill, J.C., Jiang, N., and Paterson, A.H. (2004) An SNP resource for rice genetics and breeding based on subspecies indica and japonica genome alignments. Genome Res. 14, 1812–9.

Flint-Garcia, S.A., Thornsberry, J.M., and Buckler, E.S., IV. (2003) Structure of linkage disequilibrium in plants. Annu. Rev. Plant Biol. 54, 357–74.

Gabriel, S.B., Schaffner, S.F., Nguyen, H., Moore, J.M., Roy, J., Blumenstiel, B., Higgins, J., DeFelice, M., Lochner, A., Faggart, M., Liu-Cordero, S.N., Rotimi, C., Adeyemo, A., Cooper, R., Ward, R., Lander, E.S., Daly, M.J., and Altshuler, D. (2002) The structure of haplotype blocks in the human genome. Science 296, 2225–9.

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

Hamblin, M.T., Mitchell, S.E., White, G.M., Gallego, J., Kukatla, R., Wing, R.A., Paterson, A.H., and Kresovich, S. (2004) Comparative population genetics of the panicoid grasses: sequence polymorphism, linkage disequilibrium and selection in a diverse sample of sorghum bicolor. Genetics 167, 471–83.

Hermann, A. 1962. A revision of genus *Glycine* and its immediate allies. USDA Tech. Bull. 1268, 1–79.

Hymowitz, T. (2004) Speciation and cytogenetics:In: H. R. Boerma and J. E. Specht (Eds), *Soybeans: Improvement, Production, and Uses*. American Society of Agronomy, Crop Science Society of America, Soil Science Society of America, Madison, Wis., pp. 97–136.

Hymowitz, T., and Newell, C.A. (1981) Taxonomy of the genus *Glycine*, domestication and uses of soybeans. Economic botany. 35, 272–288.

Hyten, D.L., Choi, I.-Y., Song, Q., Shoemaker, R.C., Nelson, R.L., Costa, J.M., Specht, J.E., and Cregan, P.B. (2007a) Highly variable patterns of linkage disequilibrium in multiple soybean populations. Genetics 175, 1937–1944.

Hyten, D.L., Song, Q., Zhu, Y., Choi, I.Y., Nelson, R.L., Costa, J.M., Specht, J.E., Shoemaker, R.C., and Cregan, P.B. (2006) Impacts of genetic bottlenecks on soybean genome diversity. Proc. Natl. Acad. Sci. U.S.A. 103, 16666–71.

Hyten, D.L., Choi, I.-Y., Yoon, M.-S., Song, Q.-J., Specht, J.E., Nelson, R.L., Chase, K., Young, N.D., Lark, K.G., Shoemaker, R.C., and Cregan, P.B. 2007b. *An Assessment of Genome-wide Linkage Disequilibrium in Soybean*. Plant & Animal Genome XV, San Diego, CA.

Kanazin, V., Talbert, H., See, D., DeCamp, P., Nevo, E., and Blake, T. (2002) Discovery and assay of single-nucleotide polymorphisms in barley (*Hordeum vulgare*). Plant Mol Biol 48, 529–37.

Keim, P., Shoemaker, R.C., and Palmer, R.G. (1989) Restriction fragment length polymorphism diversity in soybean. Theor. Appl. Genet. 77, 786–792.

Keim, P., Beavis, W., Schupp, J., and Freestone, R. (1992) Evaluation of soybean RFLP marker diversity in adapted germ plasm. Theor. Appl. Genet. 85, 205–212.

Kisha, T.J., Diers, B.W., Hoyt, J.M., and Sneller, C.H. (1998) Genetic diversity among soybean plant introductions and North American germplasm. Crop Sci. 38, 1669–1680.

Kuroda, Y., Kaga, A., Tomooka, N., and Vaughan, D.A. (2006) Population genetic structure of Japanese wild soybean (*Glycine soja*) based on microsatellite variation. Mol. Ecol. 15, 959–74.

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

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

Li, Z., Qiu, L., Thompson, J.A., Welsh, M.M., and Nelson, R.L. (2001) Molecular genetic analysis of U.S. and Chinese soybean ancestral lines. Crop Sci. 41, 1330–1336.

Lorenzen, L.L., and Shoemaker, R.C. (1996) Genetic relationships within old U.S. soybean cultivar groups. Crop Sci. 36, 743–752.

Lorenzen, L.L., Boutin, S., Young, N., Specht, J.E., and Shoemaker, R.C. (1995) Soybean pedigree analysis using map-based molecular markers. I. Tracking RFLP markers in cultivars. Crop Science. 35, 1326–1336.

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

Maughan, P.J., Saghai-Maroof, M.A., Buss, G.R., and Huestis, G.M. (1996) Amplified fragment length polymorphism (AFLP) in soybean: species diversity, inheritance, and near-isogenic line analysis. Theor. Appl. Genet. 93, 392–401.

Morgante, M., and Olivieri, A.M. (1993) PCR-amplified microsatellites as markers in plant genetics. Plant J. 3, 175–82.

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

National Research Council. Committee on Genetic Vulnerability of Major Crops. (1972) *Genetic Vulnerability of Major Crops* National Academy of Sciences, Washington.

Palmer, R.G., Pfeiffer, T.W., Buss, G.R., and Kilen, T.C. (2004) Qualitative genetics:In: H. R. Boerma and J. E. Specht (Eds), *Soybeans: Improvement, Production, and Uses*. American Society of Agronomy, Crop Science Society of America, Soil Science Society of America, Madison, WI, pp. 137–233.

Powell, W., Morgante, M., Doyle, J.J., McNicol, J.W., Tingey, S.V., and Rafalski, A.J. (1996) Genepool variation in genus *Glycine* subgenus *Soja* revealed by polymorphic nuclear and chloroplast microsatellites. Genetics 144, 793–803.

Rafalski, A., and Morgante, M. (2004) Corn and humans: recombination and linkage disequilibrium in two genomes of similar size. Trends Genet. 20, 103–11.

Risch, N., and Merikangas, K. (1996) The future of genetic studies of complex human diseases. Science 273, 1516–7.

Rongwen, J., Akkaya, M.S., Bhagwat, A.A., Lavi, U., and Cregan, P.B. (1995) The use of microsatellite DNA markers for soybean genotype identification. Theor. Appl. Genet. 90, 43–48.

Roth, E.J., and Lark, K.G. (1984) Isopropyl-N(3-chlorophenyl) carbamate (CIPC) induced chromosomal loss in soybean: a new tool for plant somatic cell genetics. Theoretical and Applied Genetics. 68, 421–431.

Scallon, B.J., Dickinson, C.D., and Nielsen, N.C. (1987) Characterization of a null-allele for the Gy_4 glycinin gene from soybean. Mol. Gen. Genet. 208, 107–113.

Schneider, K., Weisshaar, B., Borchardt, D.C., and Salamini, F. (2001) SNP frequency and allelic haplotype structure of Beta vulgaris expressed genes. Molecular Breeding: New Strategies in Plant Improvement. 8, 63–74.

Skorupska, H.T., Shoemaker, R.C., Warner, A., Shipe, E.R., and Bridges, W.C. (1993) Restriction fragment length polymorphism in soybean germplasm of the southern USA. Crop Sci. 33, 1169–1176.

Sneller, C.H., Miles, J.W., and Hoyt, J.M. (1997) Agronomic performance of soybean plant introductions and their genetic similarity to elite lines. Crop Sci. 37, 1595–1600.

Song, Q.-J., Choi, I.-Y., Heo, N.-K., and Kim, N.-S. (1998) Genotype fingerprinting, differentiation and association between morphological traits and SSR loci of soybean landraces. Plant Res. 1, 81–91.

Song, Q.J., Quigley, C.V., Nelson, R.L., Carter, T.E., Boerma, H.R., Strachan, J.L., and Cregan, P.B. (1999) A selected set of trinucleotide simple sequence repeat markers for soybean cultivar identification. Plant Varieties and Seeds 12, 207–220.

Tajima, F. (1983) Evolutionary relationship of DNA sequences in finite populations. Genetics 105, 437–60.

Tenaillon, M.I., U'Ren, J., Tenaillon, O., and Gaut, B.S. (2004) Selection versus demography: a multilocus investigation of the domestication process in maize. Mol. Biol. Evol. 21, 1214–25.

Thompson, J.A., and Nelson, R.L. (1998a) Utilization of diverse germplasm for soybean yield improvement. Crop Sci. 38, 1362–1368.

Thompson, J.A., and Nelson, R.L. (1998b) Core set of primers to evaluate genetic diversity in soybean. Crop Sci. 38, 1356–1362.

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

Ude, G.N., Kenworthy, W.J., Costa, J.M., Cregan, P.B., and Alvernaz, J. (2004) Genetic diversity of soybean cultivars from China, Japan, North America, and North American ancestral lines determined by amplified fragment length polymorphism. Crop Sci. 43, 1858–1867.

Vos, P., Hogers, R., Bleeker, M., Reijans, M., van de Lee, T., Hornes, M., Frijters, A., Pot, J., Peleman, J., Kuiper, M., and et al. (1995) AFLP: a new technique for DNA fingerprinting. Nucleic Acids Res 23, 4407–14.

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

Wang, N., Akey, J.M., Zhang, K., Chakraborty, R., and Jin, L. (2002) Distribution of recombination crossovers and the origin of haplotype blocks: the interplay of population history, recombination, and mutation. Am J Hum Genet 71, 1227–34.

Watterson, G.A. (1975) On the number of segregating sites in genetical models without recombination. Theor. Popul. Biol. 7, 256–76.

Weir, B.S. (1990) *Genetic Data Analysis.* Sinauer Associates, Sunderland, MA.

Welsh, J., and McClelland, M. (1990) Fingerprinting genomes using PCR with arbitrary primers. Nucleic Acids Res. 18, 7213–8.

Williams, J.G., Kubelik, A.R., Livak, K.J., Rafalski, J.A., and Tingey, S.V. (1990) DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. Nucleic Acids Res. 18, 6531–5.

Wright, S.I., Bi, I.V., Schroeder, S.G., Yamasaki, M., Doebley, J.F., McMullen, M.D., and Gaut, B.S. (2005) The effects of artificial selection on the maize genome. Science 308, 1310–4.

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

Yoon, M.S., Song, Q.J., Choi, I.Y., Specht, J.E., Hyten, D.L., and Cregan, P.B. (2007) BARC-SoySNP23: a panel of 23 selected SNPs for soybean cultivar identification. Theor. Appl. Genet. 114, 885–99.

Zakharova, E.S., Epishin, S.M., and Vinetski, Y.P. (1989) An attempt to elucidate the origin of cultivated soybean via comparison of nucleotide sequences encoding glycinin B4 polypeptide

of cultivated soybean, *Glycine max*, and its presumed wild progenitor, *Glycine soja*. Theoretical and Applied Genetics. 78 (6), 852–856.

Zhu, T., Shi, L., Doyle, J.J., and Keim, P. (1995) A single nuclear locus phylogeny of soybean based on DNA sequence. Theor. Appl. Genet. 90, 991–999.

Zhu, Y.L., Song, Q.J., Hyten, D.L., Van Tassell, C.P., Matukumalli, L.K., Grimm, D.R., Hyatt, S.M., Fickus, E.W., Young, N.D., and Cregan, P.B. (2003) Single-nucleotide polymorphisms in soybean. Genetics 163, 1123–34.