Single-Nucleotide Polymorphisms in Soybean

Y. L. Zhu,*,† Q. J. Song,*,‡ D. L. Hyten,* C. P. Van Tassell,§ L. K. Matukumalli,*,** D. R. Grimm,*,[1] S. M. Hyatt,* E. W. Fickus,* N. D. Young†† and P. B. Cregan*,[2]

*Soybean Genomics and Improvement Laboratory, U.S. Department of Agriculture-Agricultural Research Service, Beltsville, Maryland 20705, §Animal Improvement Programs Laboratory and Gene Evaluation and Mapping Laboratory, U.S. Department of Agriculture-Agricultural Research Service, Beltsville, Maryland 20705, ††Department of Plant Pathology, University of Minnesota, St. Paul, Minnesota 55108, †Department of Bioscience and Biotechnology, Nanchang University, Nanchang 330047, People's Republic of China, ‡Agronomy Department, Nanjing Agricultural University, Nanjing 210095, People's Republic of China and **Bioinformatics and Computational Biology, SCS, George Mason University, Fairfax, Virginia 22030

Manuscript received May 7, 2002
Accepted for publication December 12, 2002

ABSTRACT

Single-nucleotide polymorphisms (SNPs) provide an abundant source of DNA polymorphisms in a number of eukaryotic species. Information on the frequency, nature, and distribution of SNPs in plant genomes is limited. Thus, our objectives were (1) to determine SNP frequency in coding and noncoding soybean (*Glycine max* L. Merr.) DNA sequence amplified from genomic DNA using PCR primers designed to complete genes, cDNAs, and random genomic sequence; (2) to characterize haplotype variation in these sequences; and (3) to provide initial estimates of linkage disequilibrium (LD) in soybean. Approximately 28.7 kbp of coding sequence, 37.9 kbp of noncoding perigenic DNA, and 9.7 kbp of random noncoding genomic DNA were sequenced in each of 25 diverse soybean genotypes. Over the >76 kbp, mean nucleotide diversity expressed as Watterson's θ was 0.00097. Nucleotide diversity was 0.00053 and 0.00111 in coding and in noncoding perigenic DNA, respectively, lower than estimates in the autogamous model species *Arabidopsis thaliana*. Haplotype analysis of SNP-containing fragments revealed a deficiency of haplotypes *vs.* the number that would be anticipated at linkage equilibrium. In 49 fragments with three or more SNPs, five haplotypes were present in one fragment while four or less were present in the remaining 48, thereby supporting the suggestion of relatively limited genetic variation in cultivated soybean. Squared allele-frequency correlations (r^2) among haplotypes at 54 loci with two or more SNPs indicated low genome-wide LD. The low level of LD and the limited haplotype diversity suggested that the genome of any given soybean accession is a mosaic of three or four haplotypes. To facilitate SNP discovery and the development of a transcript map, subsets of four to six diverse genotypes, whose sequence analysis would permit the discovery of at least 75% of all SNPs present in the 25 genotypes as well as 90% of the common (frequency >0.10) SNPs, were identified.

SINGLE DNA base differences between homologous DNA fragments plus small insertions and deletions (indels), collectively referred to as single-nucleotide polymorphisms (SNPs), have been shown to be the most abundant source of DNA polymorphisms in humans (Kwok *et al.* 1996; Kruglyak 1997; Collins *et al.* 1998). In humans, these variations were estimated to occur at a frequency of about one per 1000 bp when any two homologous DNA segments were compared (Cooper *et al.* 1985; Kwok *et al.* 1996; Wang *et al.* 1998).

The frequency and nature of SNPs in plants is beginning to receive considerable attention. A number of reports in *Arabidopsis thaliana* (L.) Heynh. and maize (*Zea mays* ssp. *mays* L.) have provided estimates of sequence diversity in these species. In soybean (*Glycine max* L. Merr.), which is an autogamous species, the analysis of DNA sequence variation has been mainly confined to single genes or DNA fragments with the goal of defining gene structure, function, or evolutionary relationships. Scallon *et al.* (1987) compared 3543 bp of the Gy_4 glycinin locus in two genotypes and identified three SNPs. Subsequent work by Xue *et al.* (1992) analyzed a similar region of the Gy_4 gene in another cultivar and found additional sequence variants. Zakharova *et al.* (1989) compared 789 bp of cDNA sequence encoding the A_3B_4 glycinin subunit in three soybean cultivars and found two SNPs. Zhu *et al.* (1995) sequenced a 400-bp fragment of restriction fragment length polymorphism 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), that are normalized for length and adjusted for sample size have been devised. Nucleotide

[1] *Present address:* Abbott Laboratories, Abbott Park, IL 60064.

[2] *Corresponding author:* Soybean Genomics and Improvement Laboratory, Bldg. 006, Rm. 100, USDA, ARS, BARC-West, Beltsville, MD 20705. E-mail: creganp@ba.ars.usda.gov

diversity from the four aforementioned studies of soybean ranges from θ = 0.00085 (Scallon *et al.* 1987) to θ = 0.015 (Zhu *et al.* 1995).

As a consequence of linkage disequilibrium (LD), reduced genetic variability in the form of limited haplotype diversity is a frequent result. SNPs are a useful tool to quantify LD and the analysis of SNP haplotypes has been the focus of recent studies. Patil *et al.* (2001) analyzed SNP haplotype diversity across human chromosome 21 and found an average of fewer than three common haplotypes (those with a frequency ≥0.10). Generally consistent results were reported by Stephens *et al.* (2001) who analyzed SNP haplotype diversity in 313 human genes in 82 unrelated individuals of diverse ancestry. On the basis of frequency of occurrence, Stephens *et al.* (2001) indicated that 80% of the global haplotype diversity detected was defined by three common haplotypes with mean frequencies of 0.5, 0.25, and 0.125.

In plants there are limited data relating to genome-wide haplotype diversity, but haplotype structure can be inferred from analyses of the decline of LD. In maize, Tenaillon *et al.* (2001) reported the rapid decay of LD over only 500 bp, suggesting limited haplotype structure. Similar, but somewhat different conclusions were reached by Remington *et al.* (2001) who found highly variable rates of LD decline in different maize genes. Nordborg *et al.* (2002) indicated that in the autogamous species *A. thaliana* LD generally declines in ∼250 kbp (∼1 cM). Studies of haplotype diversity in soybean are limited. Coryell *et al.* (1999) reported the analysis of two SNP loci separated by 55 bp in 570 soybean genotypes and found only three of the possible four haplotypes that would be anticipated at linkage equilibrium.

In this report we assess the frequency of SNPs and SNP haplotype diversity in 143 DNA fragments. These fragments were derived from coding and noncoding DNA associated with the coding regions, as well as random genomic DNA of soybean, based upon sequence analysis of a group of 25 selected soybean genotypes representative of the genetic base of North American soybean. The rationale for making these determinations was (1) to permit a comparison of SNP frequency and haplotype diversity with other plant and animal species, (2) to provide a preliminary estimate of linkage disequilibrium in soybean, and (3) to develop a strategy for SNP discovery aimed at the development of a SNP-based soybean linkage map.

MATERIALS AND METHODS

Soybean plant material and DNA isolation: *Ancestors of North American soybean cultivars:* On the basis of pedigree analysis, Gizlice *et al.* (1994) identified a group of 35 genotypes from which >95% of the allelic variation present in North American cultivated soybean germplasm was derived. Fourteen of these genotypes, along with a number of others (Table 1), were assayed to determine the frequency and nature of SNPs in soybean. The 14 genotypes were estimated by Gizlice *et al.* (1994) to have contributed 80.5% of the allelic diversity present in North American soybeans. Three additional genotypes that have been used as sources of resistance to the soybean cyst nematode (*Heterodera glycines*) were included, as were a number of additional cultivars that are the parents of available soybean recombinant inbred line mapping populations. Of the 25 genotypes, 17 are reported to be direct introductions or selections from introductions from Asia (Table 1). On the basis of the information available from the Germplasm Resources Information Network (http://www.ars-grin.gov/npgs/acc/acc_queries.html), none of the 17 is derived from a program of hybridization and selection. Seeds of each of the 25 genotypes were obtained from the U.S. Department of Agriculture (USDA) Soybean Germplasm Collection courtesy of Dr. Randall Nelson (USDA-ARS, University of Illinois, Urbana, IL).

DNA isolation: DNA was extracted from bulked leaf tissue of 30–50 plants of each of the soybean genotypes by the method described by Keim *et al.* (1988).

Selection and testing of PCR primers: *From full-length genes:* A total of 90 full-length soybean genes were selected from GenBank to represent a range of functions (Table 2). Primers were designed using OLIGO primer design software (National Biolabs, St. Paul) with the goal of amplifying fragments of ∼500–600 bp in length containing approximately equal amounts of coding and noncoding DNA. In several instances two sets of primers were selected to genes to obtain additional sequence data.

From cDNAs: A total of 88 soybean cDNAs were selected from GenBank (Table 2). Sequences including the poly(A) tail were preferentially selected so primers could be designed as close to the 3′-end of the transcript as possible. Primers were designed as described above with predicted amplicon lengths of 300–500 bp. The rationale for the shorter predicted amplicon length was based upon the likelihood of an intervening intron(s). Additional information relating to primer sequences, fragment lengths, numbers of bases of coding and noncoding sequences, as well as the presence of SNPs in amplicons derived using PCR primers designed to complete both genes and cDNAs can be found at http://www.genetics.org/supplemental/ as Table S2.

From sequences containing mapped simple sequence repeats: The development of a large number of soybean simple sequence repeat (SSR) loci (Cregan *et al.* 1999a) provided a resource of sequence-tagged sites from which to discover SNPs. Primers were designed as described above using these sequence data with predicted product lengths in the 400- to 600-bp range.

From bacterial artificial chromosome subclones: The simple SSR marker BARC-Satt309 is closely linked to the *rhg*1 locus, which is reported to be the most important gene conditioning resistance to the soybean cyst nematode. Primers to BARC-Satt309 were used to identify a bacterial artificial chromosome (BAC) clone, UMN-K4, as previously described by Cregan *et al.* (1999b). The BAC clone was randomly subcloned into pBluescript as described by Cregan *et al.* (1999c). Genomic inserts were sequenced using the ABI Prism BigDye Terminator cycle sequencing kit (Perkin-Elmer, Norwalk, CT; Applied Biosystems, Foster City, CA), using SK and/or KS primers with sequence analysis on a Perkin-Elmer ABI Prism 377 DNA sequencer. PCR primers were designed from these sequence data as described above with predicted product lengths in the 400- to 600-bp range.

BLAST search of SSR flanking regions and BAC subclones: Each SSR-containing sequence and each BAC subclone were analyzed using BLASTN and TBLASTX against the nonredundant

TABLE 1

Soybean genotypes analyzed for the presence of single-nucleotide polymorphisms

Ancestor or first progeny[a]	Origin: introduction (I) or hybridization/ selection (H/S)	Contribution to North American cultivars[a]	Other genotypes assayed	Origin: introduction (I) or hybridization/ selection (H/S)
Lincoln	H/S	17.9	PI 437654[b]	I
Mandarin (Ottawa)	I	12.2	PI 90763[b]	I
CNS	I	9.4	PI 209332[b]	I
Richland	I	8.2	Tokyo	I
S-100	I	7.5	Minsoy	I
Ogden	H/S	4.9	Noir 1	I
A.K. (Harrow)	I	4.9	Archer	H/S
Dunfield	I	3.6	Evans	H/S
Mukden	I	3.5	Clark	H/S
Jackson	H/S	3.3	Harosoy	H/S
Illini	I	2.2	Essex	H/S
Roanoke	I	2.1		
PI 88788[b]	I	0.5		
Peking[b]	I	0.4		

[a] From Gizlice *et al.* (1994). First progeny refers to genotypes including Lincoln, Ogden, and Jackson whose pedigrees are not known.

[b] Sources of resistance to the soybean cyst nematode.

and the expressed sequence tag [EST (est)] databases to determine if any portions of the BAC subclones or SSR flanking sequence were coding regions.

Initial examination of PCR primers: All PCR primers were used to amplify genomic DNA of one or two soybean genotypes. In most cases the cultivar Lincoln was used, but in some instances either Minsoy or Noir 1 DNA was used as template. Amplification reactions used standard PCR reagents including 30 ng of genomic DNA template, 1.5 mm Mg^{2+}, 0.15 μm of 3′- and 5′-end primers, 100 μm of each nucleotide, 1× PCR buffer (10 mm Tris-HCL pH 8.3, 50 mm KCl), and 2 units of *Taq* DNA polymerase in a total volume of 50 μl. PCR cycling conditions were as follows: 45 sec denaturation at 92°, 45 sec annealing at 50° (or higher depending upon optimal annealing indicated by OLIGO), and 45 sec extension at 68° for 32 cycles on a MJ Tetrad thermocycler (MJ Research, Watertown, MA). The products were analyzed on a 1.5% agarose gel stained with ethidium bromide. Those primer sets that produced what appeared to be a single product were selected for further testing. Those that produced no products or multiple products were further examined using lower annealing temperatures or higher Mg^{2+} (those giving no products) or higher annealing temperature or lower Mg^{2+} (those giving multiple products). After these analyses, the amplicons from those primer sets producing what appeared to be single amplicons were selected for sequence analysis.

Sequence analysis of PCR products amplified from genomic DNA: After the initial determination that a set of PCR primers appeared to produce a single amplicon from genomic DNA, the PCR product was directly sequenced using one of the PCR primers with BigDye Terminator cycle sequencing as described above. The results of this sequence analysis determined if the PCR product was derived from a single locus or if it was the result of amplification from two or more homeologous regions. In those cases in which the sequence traces appeared to be derived from a single locus, analysis with AutoAssembler software (Perkin-Elmer, Applied Biosystems) was used to detect ambiguous base calls that appeared as "heterozygotes" that would not be anticipated from the sequence analysis of a homozygous soybean genotype. In most instances, such "heterozygotes" would indicate the presence of two or more paralogues.

Those primer sets that produced a single amplicon suitable for sequencing were used to amplify genomic DNA from each of the remaining 24 genotypes listed in Table 1. The sequence of each of these products was determined as described above. When necessary, products were sequenced from both ends to assure accurate sequence determination.

Single-nucleotide polymorphism discovery: The sequence data from each amplicon were analyzed with PolyBayes SNP detection software (Marth *et al.* 1999). PolyBayes considers alignment depth, the base calls in each of the sequences, the associated base quality scores, the base composition in the region, and the expected *a priori* polymorphism rate and calculates the probability that sequence variants represent true variations rather than sequencing errors. Putative SNPs were accepted as true sequence variants if the probability (SNP score) that the sequence discrepancies represented true sequence variations, as opposed to sequencing errors, exceeded 0.99. To avoid false negatives of low-frequency or singleton SNPs, the PolyBayes output (Phrap alignment) was visually inspected. Fragments containing visually identified sequence variants with SNP scores that did not exceed ($P \geq 0.99$) were resequenced and reanalyzed with PolyBayes. In no case was any type of tandem repeat variant considered to be a SNP.

Statistical analyses: *Nucleotide diversity* (θ *and* π)*:* Nucleotide diversity was estimated as θ, the number of segregating sites (Watterson 1975), and its standard deviation, $S(\theta)$ as per Halushka *et al.* (1999), and as π, the mean pairwise differences, and its standard deviation $S(\pi)$ (Tajima 1983). To calculate θ and π for synonymous (silent nucleotide substitutions) and nonsynonymous (replacement) sites, the numbers of synonymous and nonsynonymous sites were calculated using DNASP sequence polymorphism software version 3.5 (Rozas and Rozas 1999).

Tajima's D: Tajima's (1989) *D* test for the frequency distributions of nucleotide polymorphisms was calculated for each functional region [coding regions, untranslated regions (UTRs), introns, etc.].

Gene diversity: Gene diversity (Weir 1990) was calculated as

TABLE 2

GenBank accessions used in the analysis of sequence diversity

GenBank accession no.	Description	GenBank accession no.	Description
Complete genes			
AB003680	A3B4 glycinin	L28831	Ribosomal protein S11
AB003908	Phosphoenolpyruvate carboxylase	L34842	Chloroplast phytochrome A (phyA)
AB004062	A5A4B3 glycinin	L34843	Phytochrome B (phyB)
AB007127	Acidic chitinase	L42814	Acetyl coA carboxylase (ACCase-A)
AB018378	Early nodulin	M10594	Uricase I I
AB029159	GmMYB29A1	M10595	Peribacteroid membrane protein
AB030491	Thiamin biosynthetic enzyme	M11317	Low-MW heat-shock protein[a]
AF049106	Actin 4 (Sac4)	M13759	α′-type β-conglycinin storage protein
AF055369	Nitrate reductase (nr2)	M16772	Urease
AF061564	Glyceraldehyde-3-phosphate dehydrogenase 1	M16884	Cytochrome oxidase subunit I
AF079058	Alcohol dehydrogenase (Adh-1)	M21296	β-Tubulin (S-β-1)
AF083880	Alternative oxidase precursor (Aox 1)	M76980	Vegetative storage protein (vspB)
AF105199	Glutathione reductase (GR-5)	M76981	vspA
AF162283	Acetyl-CoA carboxylase (accB-1)	M97285	Seed maturation protein
AF180335	Malate dehydrogenase (Mdh1)	M98871	Chalcone synthase (chs7)
AF195819	Isoflavone synthase 2 (ifs2)	U31648	Ferritin
AJ223037	Leginsulin	U41323	β-1,3-Glucanase (SGN1)
AJ239127	Major latex protein homolog	U47143	Nonsymbiotic hemoglobin
AJ276407	Pre-pro-subtilisin	U60500	Actin (Soy57)
D13999	Lipoxygenase L-4	U87999	Phosphoribosylpyrophosphate amidotransferase
D16107	Basic 7s globulin	V00452	Leghemoglobin
D16248	Ubiquitin	V00458	Ribulose-1,5-bisphosphate carboxylase
D26092	Ubiquitin	X05024	Nodulin 22
D64115	Cysteine proteinase inhibitor	X07675	NADH dehydrogenase and rps7
E00532	Heat-shock protein	X16875	Ngm-75
E13668	DNA-binding protein	X52863	Glycinin
E01433	Leghemoglobin c3	X63198	Low-MW heat-shock protein[a]
E03629	Lipoxygenase	X68707	Proteinase inhibitor D-II
J01297	Actin 3 (Sac 3)	X71083	Coproporphyrinogen oxidase
J02746	Proline-rich protein	X78548	Epoxide hydrolase
K00821	Lectin (Le1)	Z11980	Cytochrome oxidase subunit 2
L00921	Maturation protein (MAT 1)	Z12021	Catalase
L20310	Nodulin (nod-20)		
cDNAs			
AB025102	Protoporphyrinogen IX oxidase	L27265	Phosphatidylinositol 3-kinase
AB029441	Trypsin inhibitor p20	L27417	GTP-binding protein (STGA1)
AB030493	Thiamin biosynthetic enzyme	L28005	TGACG-motif binding protein
AB040040	Nonclathrin coat protein	M64267	Iron superoxide dismutase (FeSOD)
AF005030	2S albumin prepropeptide	M80664	Late embryogenesis abundant (LEA) protein
AF007211	Peroxidase precursor (GMIPER1)	M94012	Maturation-associated protein (MAT9)
AF022462	Cytochrome P450 monooxygenase	U04525	δ-Aminolevulinic acid dehydrase (Alad)
AF089850	Urate-degrading peroxidase (PP1)	U12150	Protease inhibitor
AF117885	Seed maturation protein PM31 (PM31)	U26457	Lipoxygenase (vlxC)
AF124148	Trehalase 1 GMTRE1	U32185	Guanine nucleotide regulalory protein
AF127110	GO8 ripening related protein	U35367	Arginine decarboxylase
AF127112	Norvegicus 2-oxoglutarate carrier protein	U63725	Metalloproteinase
AF128443	SNF-1-like serine/threonine protein kinase	U63726	γ glutamyl hydrolase
AF141602	Cystathionine-γ-synthase precursor	U66836	RecA/Rad51/DMC1-like protein
AF167556	Dihydroflavonol-4-reductase DFR1	U82810	Early light-induced protein
D13505	Early nodulin	X60043	Stress-induced gene (SAM22)
D13949	Lipoxygenase-2 (lox2)	X63565	Seed maturation polypeptide
D31700	Cysteine proteinase inhibitor	X67304	Lipoxygenase 1
D45857	Mg chelatase	X68702	Alternative oxidase
D50866	β-Amylase	X69639	Auxin downregulated gene (ADR6)
D78510	β-Glucan-elicitor receptor	X78547	Epoxide hydrolase
L01433	Calmodulin (SCaM-4)	Z32795	Cysteine endopeptidase
L01447	G-box binding factor (GBF1)	Z46951	Heat-shock transcription factor 29
L10292	Ascorbate peroxidase	Z46953	Heat-shock transcription factor 34
L19359	Calmodulin (ScaM-5)	Z46954	Heat-shock transcription factor 33

[a] MW, molecular weight.

$1 - \Sigma P_{ij}^2$, where P_{ij} is the frequency of the *j*th allele for *i*th locus summed across all alleles in the locus. In the case of the SNPs reported here, there were only two alleles at a locus.

Distribution of SNPs in coding and noncoding DNA: To determine if SNPs were evenly distributed in the fragments assayed, theoretical SNP cumulative frequency distributions were calculated for SNPs in both coding and noncoding DNA on the basis of the assumption of uniform distribution. In the case of SNPs in coding regions, the cumulative frequency distribution of coding SNPs (cSNPs) was calculated assuming a uniform distribution of SNPs in these fragments. This distribution was compared with the actual cumulative frequency distribution in these fragments using a Kolmogorov-Smirnov (KS) test (Gibbons 1976). The KS test assessed the degree of agreement between a sample of empirically gathered values and a target theoretical distribution. Cumulative frequency distributions also were calculated for the SNPs discovered in noncoding DNA fragments (from GenBank genes, cDNAs, and random genomic sequence), and the KS test was used to test for deviation from a uniform distribution in these fragments.

SNP haplotype frequencies in sequenced fragments: The number of haplotypes among the 25 genotypes in fragments containing two or more SNPs was determined by visual inspection. A permutation algorithm based on Churchill and Doerge (1994) was designed to assess the likelihood that the limited numbers of observed SNP haplotypes present in each fragment could have occurred by chance. This algorithm involves a number of discrete steps: (1) Allele frequency at each SNP locus within each fragment was calculated on the basis of the sample of 25 genotypes; (2) alleles at each locus within each fragment were randomly shuffled to create 25 hypothetical genotypes, and the order of the shuffling was independently generated for each SNP locus within a fragment; and (3) the number of haplotypes present in the 25 hypothetical genotypes in the permuted data was determined. For these analyses 10,000,000 permutations of the observed data were generated for each fragment containing three or more SNPs to characterize the distribution of the number of haplotypes that would be observed if the loci within a fragment were independent. The probability of observing the number of haplotypes present in the original data or fewer was determined on the basis of the permuted data.

LD in introductions from the Far East: For purposes of calculating LD only genotypes reported to be direct introductions from the Far East were analyzed. This was done to eliminate genotypes that would be anticipated to have reduced LD as a result of hybridization and subsequent recombination. LD was analyzed on the following three subsets of data using haplotypes determined from fragments containing two or more SNPs:

Subclones derived from BAC clone UMN-K4: The haplotypes of BAC subclones were used in the calculation of squared allele-frequency correlations (r^2; Weir 1996) with the multiple-allele option of Tassel 0.2 (http://statgen.ncsu.edu/buckler/software/TASSEL/TASSEL.htm; Remington *et al.* 2001). The significance of r^2 ($P < 0.05$) was determined via permutation analysis using 1000 permutations.

Loci on soybean linkage group G: The SNP haplotypes of SSR flanking regions on soybean linkage group G (Cregan *et al.* 1999a) were used in the calculation of squared allele-frequency correlations as described above. The r^2 values were plotted against the known genetic distances between loci to examine the relationship of genetic distance and LD.

Remaining loci with undefined genome positions (genome-wide LD): Twelve of the 65 fragments that contained two or more SNPs were located in proximity to each other on linkage group G. Because of this known linkage relationship, only one of these fragments (flanking regions of SSR locus BARC-Satt309) was included in the analysis of genome-wide LD. Squared allele-frequency correlations were calculated as described above using haplotypes of the 54 remaining loci.

RESULTS

Nature and frequency of SNPs in soybean: Sequence data for 25 soybean genotypes were obtained from fragments amplified using PCR primers derived from 66 complete GenBank genes and 50 cDNAs. In addition to the genes and cDNAs, sequence data were obtained from 13 BAC subclones and 15 SSR flanking regions. The BLASTN analysis indicated that one of the BAC subclones was homologous to a *G. max* aspartokinase-homoserine dehydrogenase gene (GenBank accession no. AF049708). The remaining BAC and SSR flanking sequences were tentatively classified as noncoding. In total, ~28.7 kbp of coding sequence, 9.3 kbp of 5′- and 3′-UTR, 22.9 kbp of intron, and 5.8 kbp perigenic genomic sequence as well as 9.7 kbp of random noncoding genomic sequence data were obtained for each of the 25 genotypes (Table 3). A total of 280 SNPs including 233 single-base changes and 47 indels were identified in 143 amplicons totaling ~76.3 kbp of sequence. The mean frequency of the least common allele at the 280 loci was 0.23 with a mean gene diversity of 0.35. The distribution of the allele frequencies of the least common allele are shown in Table 4. A total of 212 SNPs (76%) could be considered common (frequency >0.10). Of the 233 single-base changes, transitions accounted for 112 (48%) and transversions for 121 (52%). Moriyama and Powell (1996) found that transversions accounted for 54% of single-base changes in *Drosophila melanogaster*, which is very similar to the 52% in soybean. In contrast, a 2:1 ratio of transitions to transversions has been reported in humans (Wang *et al.* 1998) and mice (Lindblad-Toh *et al.* 2000).

The mean nucleotide diversity (θ) in the 76.3 kbp of sequence analyzed was 0.00097 (Table 3). The estimate of nucleotide diversity in coding DNA ($\theta = 0.00053$) was less than half that in noncoding sequence associated with genes ($\theta = 0.00111$). Nucleotide diversities in the UTRs, introns, and genomic sequence adjacent to genes were similar, ranging from $\theta = 0.00087$–0.00126. In random noncoding genomic sequence from BAC clones and SSR flanking regions nucleotide diversity of $\theta = 0.00179$ was numerically, but nonsignificantly, higher than that of the genomic DNA associated with genes.

Tajima's *D* was determined across loci in the various functional regions of genes and in genomic DNA (Table 3) to provide information on population structure such as genetic bottlenecks and expanding population size that would be anticipated to affect the entire genome rather than specific genes. Tajima's *D* values were generally positive although none was significant. *D* was sig-

TABLE 3

Nucleotide diversity in soybean genes and genomic sequence

Source of sequence	Primers designed	Successful sequences	Coding regions and associated noncoding sequence										Random noncoding genomic sequence		Total	
					Noncoding regions											
			Coding regions		5′- and 3′-UTR		Introns		Associated genomic sequence		Total noncoding					
			Bases	SNPs	Bases	SNPs	Bases	SNPs	Bases	SNPs	Bases	SNPs	Bases	SNPs	Bases	SNPs
GenBank genes	90	65	17,913	38	2,325	13	15,641	71	5,781	19	23,747	103			41,660	141
cDNAs	88	50	10,630	19	6,933	31	7,092	25			14,025	56			24,655	75
BAC subclones		13	150	0			169	0			169		4,200	37	4,519	37
SSR flanking regions		15											5,534	27	5,534	27
Totals			28,693	57	9,258	44	22,902	96	5,781	19	37,941	159	9,734	64	76,368	280
Synonymous			6,601	25												
Nonsynonymous			22,092	32												
θ (×10^3)			0.53 + 0.19		1.26 ± 0.46		1.11 ± 0.39		0.87 ± 0.35		1.11 ± 0.38		1.79 ± 0.62		0.97 ± 0.33	
Synonymous			1.00 ± 0.39													
Nonsynonymous			0.38 ± 0.15													
π (×10^3)			0.54 ± 0.26		1.75 ± 0.74		1.36 ± 0.55		1.14 ± 0.54		1.45 ± 0.56		2.64 ± 1.01		1.25 ± 0.60	
Synonymous			0.79 ± 0.49													
Nonsynonymous			0.41 ± 0.19													
Tajima's *D*			0.10 NS		1.49 NS		0.90 NS		1.05 NS		1.09 NS		2.00 ($P < 0.10$)		1.08 NS	
Synonymous			−0.80 NS													
Nonsynonymous			0.22 NS													

Shown are the number of bases sequenced and measures of nucleotide diversity (θ and π); Tajima's *D* in coding, noncoding, and associated genomic DNA of fragments amplified from genomic DNA using PCR primers designed to GenBank-derived genes and cDNAs; and random genomic sequences from BAC subclones and SSR flanking regions determined from the sequence analysis of 25 diverse soybean genotypes. NS, not significant.

TABLE 4

Distribution of allele frequencies of the least common allele of SNPs discovered in 25 diverse soybean genotypes

	Frequency of the least common allele				
	0–0.10	0.11–0.20	0.21–0.30	0.31–0.40	0.41–0.50
No. of SNPs (proportion)	68 (0.24)	55 (0.20)	57 (0.20)	72 (0.26)	28 (0.10)

nificant at $P < 0.10$ in the 9.7 kbp of random noncoding genomic sequence derived from BAC clones and SSR flanking regions.

Polymorphisms in coding regions: In the 28.7 kbp of coding sequence analyzed, the 57 cSNPs included 51 single-base changes and six indels. Of the 51 single-base changes, 13, 8, and 30 were detected in the first, second, and third codon positions, respectively. A total of 25 were synonymous (no alteration in amino acid) while 26 were nonsynonymous or replacement SNPs that included a single-base change in the third position in the start codon of the glycinin gene (GenBank accession no. X52863) of PI 88788 that had been reported previously by Scallon *et al.* (1987). The indels included two events in accession no. AF167556, which is a dihydroflavonol-4-reductase (DFR1) gene, where two separate insertions, one of 5 and one of 4 bases, and a single-base change in the third position of the stop codon were detected in eight genotypes with a predicted alteration of the last 10 amino acids to 6 new amino acids. A single-base deletion in accession no. L10292, an ascorbate peroxidase gene, resulted in the change of the last 4 amino acids to 11 new amino acids in the genotypes Roanoke, PI209332, and Tokyo. In GenBank accession no. M94012, a 3-base deletion of GCT covering two codons was discovered: G(GC T)AC → GAC (Gly Tyr → Asp). The remaining two indels, one insertion of a codon CCA and one deletion of two codons CGACCA, were found in GenBank accession nos. X63198 and M13759, respectively.

The DNASP analysis indicated that of the 28.7 kbp of coding sequence 22.1 kbp (77%) were nonsynonymous. Thus, about three-quarters of randomly occurring single-base changes in coding DNA would be anticipated to result in an amino acid alteration. However, of the 57 cSNPs, 32 were nonsynonymous, which included 26 single-base changes and 6 indels, while 25 were synonymous. The nucleotide diversity of synonymous changes, $\theta = 0.00100$, while not significantly greater than that of nonsynonymous changes, $\theta = 0.00038$, was 2.6-fold greater. The higher frequency of synonymous cSNPs suggests selection against mutations that result in an amino acid replacement.

Failure to obtain sequence data from genes and cDNAs: High-quality sequence data were obtained from 65 of the 90 complete genes to which PCR primers were designed. The failure to obtain data from the remaining 25 resulted from either the failure of primers to amplify (4 cases) or the amplification of two or more products as determined via agarose gel electrophoresis (3 cases). The failure to obtain data from the remaining 18 genes was the result of sequence analyses that indicated heterogeneous template, as would be anticipated if members of a gene family or homeologous loci were amplified. In the case of cDNAs, high-quality sequence data for all 25 genotypes were obtained for only 50 of the 88 cDNAs for which primers were designed. The failure to obtain data from the remaining 38 was the result of failure to amplify a product (5 cases), the amplification of multiple products as determined by agarose gel electrophoresis (12 cases), and poor quality sequence data from what appeared to be a single PCR product on agarose (21 cases). As with the complete genes, this latter outcome generally appeared to result from multiple sequencing templates**.**

Heterogeneity of nucleotide diversity among DNA fragments: The average length of the 143 amplicons analyzed for the presence of SNPs was 534 bp with a mean of 1.95 SNPs per amplicon. There was no sequence variation in 47 of the fragments while only 1 SNP was discovered in 30 of the 143 (Table 5), suggesting an uneven distribution of sequence variation in this sample of amplicons. The Kolmogorov-Smirnov test was done to compare the observed cumulative frequency distributions of SNPs in fragments with the theoretical distributions on the basis of the assumption of mutations being

TABLE 5

Numbers of SNP haplotypes observed in SNP-containing DNA fragments

SNPs/ fragment	Potential haplotypes	SNP haplotype no.				
		2	3	4	5	Total
1	2	30				30
2	4	8	11			19
3	8	4	5	5		14
4	16	2	6	5		13
5	25		5	5		10
6	25	1	0	2		3
7	25		2			2
8	25		1			1
9	25		1	1		2
10	25				1	1

evenly distributed across the 143 fragments. In both coding and noncoding sequences, the observed and theoretical frequency distributions were determined to be significantly different ($P < 0.01$), indicating that there was heterogeneity in the nucleotide diversity of both the coding and noncoding DNA fragments included in this study.

SNP haplotypes and haplotype frequency: The number of SNP haplotypes present among the 25 soybean genotypes was determined in each of the 66 fragments that contained two or more SNPs (Table 5). Gene diversity based upon haplotypes was 0.52. In only one case were more than four haplotypes observed among the 25 genotypes. In this instance five haplotypes were found. The permutation analyses of the 49 fragments with three or more SNPs indicated that 44 of the 49 fragments had an empirical probability of the limited number of haplotypes observed of ≤ 0.001. A total of 30 of the 49 fragments never had a single permutation randomly generated with as few haplotypes present as that observed in the original data. In the five fragments where probability did not exceed 0.001, allele frequencies at one or more SNP loci were maximally asymmetric. The analysis indicated a shortage of haplotypes in relation to the number that would be anticipated at linkage equilibrium.

LD in introductions from the Far East: *Among subclones of BAC UMN-K4:* Squared allele-frequency correlations (r^2) were calculated among each of the seven subclones with two or more SNPs discovered in 16 genotypes that were direct introductions from Asia (Table 1). The ancestral cultivar Illini was eliminated from the analysis because it was determined to be identical at all SNP loci to A.K. (Harrow). Both Illini and A.K. (Harrow) were selected from the older cultivar A.K. and were anticipated to be similar. The mean r^2 value for the 21 estimates of LD was 0.36 and 18 of the 21 estimates of r^2 indicated significant LD ($P < 0.05$). The positions of the subclones in the 110-kbp BAC clone UMN-K4 BAC are not known; however, a simulation analysis using 1000 permutations indicated that seven 550-bp fragments drawn at random from a 110-kb BAC would span a region of at least 53.9 kbp ($P > 0.95$). Thus, significant r^2 values among most subclones suggest that LD exists over a distance of ~50 kbp in this region of the soybean genome.

Among loci on soybean linkage group G: The SNP haplotypes of flanking regions of seven SSR loci in soybean linkage group G were used to provide an initial estimate of the relationship of LD with genetic map distance. The loci cover a map distance of 12.5 cM. The mean r^2 of the 21 pairwise estimates of LD was 0.14 and only four of the r^2 values were significant ($P < 0.05$). The trend line developed from the plot of r^2 against genetic map distance is presented in Figure 1. Although these data are limited, it appears that LD has significantly decayed at distances of 2.0–2.5 cM, which is roughly equivalent to 1.0–1.5 mbp.

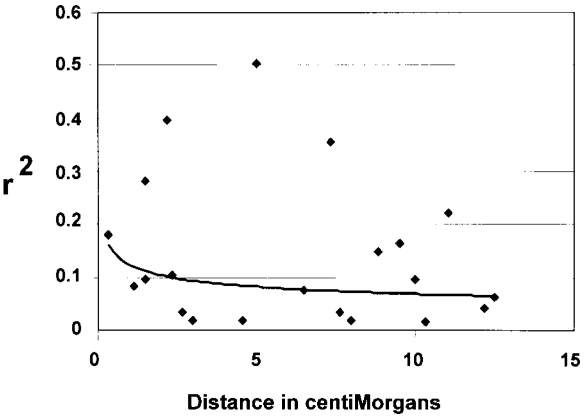

FIGURE 1.—Linkage disequilibrium as a function of genetic map distance based upon squared allele-frequency correlations (r^2) among haplotypes at seven loci in a 12.5-cM region on soybean linkage group G.

Genome-wide LD: Squared allele-frequency correlations were calculated among the haplotypes of 54 loci with two or more SNPs. The mean r^2 over all pairwise estimates was 0.091 with only 8.9% of the r^2 values significant at $P < 0.05$. This result indicated a low level of genome-wide LD in the set of 16 soybean accessions used in the analysis.

A subset of genotypes with maximum SNP diversity: If the 25 genotypes analyzed here are representative of North American cultivated soybean, the limited number of haplotypes suggested that SNP discovery might proceed with the sequence analysis of a relatively small selected set of genotypes. The three genotypes whose analysis would detect the largest proportion (71%) of all 280 SNPs and 83% of the 212 common SNPs discovered among all 25 genotypes were Peking, PI 209332, and Tokyo (Table 6). The addition of the fourth genotype, Noir 1, brought these figures to 78 and 91%, respectively. Two sets of 8 genotypes would permit the discovery of >90% of the total SNPs and 98% of the common SNPs. A total of 89% of the total and 99% of the common SNPs were polymorphic in the 14 soybean genotypes that contributed 80.5% of the allelic diversity present in North American soybeans. The 6 genotypes Minsoy, Noir 1, Archer, Peking, Evans, and PI 209332 represent the parents of recombinant inbred line (RIL) mapping populations available in our laboratory (University of Utah Minsoy × Noir 1 and Archer × Minsoy and University of Minnesota Evans × Peking and Evans × PI 209332). The SNPs detected in these 6 genotypes included 85% of the total and 93% of the common SNPs. Totals of 83 and 89% of the total and common SNPs could be mapped in at least one of the aforementioned RIL populations, respectively.

DISCUSSION

Nucleotide diversity: The results of this survey provide the first extensive sampling of DNA sequence diversity

TABLE 6

The proportion of total SNPs and common SNPs (frequency >0.10) discovered in 25 soybean genotypes by the analysis of selected subsets of genotypes

Genotypes included in subset	Proportion of total SNPs discovered	Proportion of common SNPs discovered
Peking, PI 209332, Tokyo	0.71	0.83
Peking, PI 209332, Tokyo, Noir 1	0.78	0.91
Peking, PI 209332, Tokyo, Noir 1, S-100	0.83	0.94
Peking, PI 209332, Tokyo, Noir 1, S-100, Minsoy, Archer, CNS	0.92	0.98
Peking, PI 209332, Tokyo, Noir 1, S-100, Minsoy, CNS, Richland	0.93	0.98
Fourteen ancestors or first progeny (Table 1)	0.89	0.99
Minsoy, Noir 1, Archer	0.58	0.68
Minsoy, Noir 1, Archer, Peking	0.73	0.84
Minsoy, Noir 1, Archer, Peking, Evans, PI 209332	0.85	0.93
Polymorphic in Minsoy × Noir 1 and/or Archer × Minsoy and/or Evans × Peking and/or Evans × PI 209332	0.83	0.89

in cultivated soybean. The initial estimate suggests that mean nucleotide diversity is much lower in soybean ($\theta = 0.00097$) than in the wild plant *A. thaliana*. Numerous reports of sequence variation in individual Arabidopsis genes (Kawabe and Miyashita 1999; Purugganan and Suddith 1999; Kawabe *et al.* 2000; Kuittinen and Aguade 2000) have indicated a level of nucleotide diversity 5- to 8-fold higher than what we have detected in domesticated soybean. Likewise, data from maize (Tenaillon *et al.* 2001) indicated diversity ($\theta = 0.0096$) 10-fold greater than that in soybean. This calculation was based on >14 kbp of sequence from each of 25 inbreds and exotic landraces. The level of sequence diversity in an inbreeding species is expected to be lower than that in an outcrossing species because of smaller effective population size (Pollak 1987) and as a result of additional effects of background selection (Nordborg *et al.* 1996). Nonetheless, nucleotide diversity in the soybean germplasm included in our analysis is lower than some of the lowest values reported in Arabidopsis, the model selfing species. For example, Olsen *et al.* (2002) noted the unusually low nucleotide variation in the *TERMINAL FLOWER1* ($\theta = 0.0017$) and *LEAFY* ($\theta = 0.0033$; calculated from Olsen *et al.* 2002) loci in Arabidopsis. The relatively low level of sequence diversity we have observed in soybean supports concerns of the narrow genetic base of North American soybean (Committee on Genetic Vulnerability of Major Crops 1972; Gizlice *et al.* 1994).

The ratio of synonymous to nonsynonymous changes in soybean (2.6) was somewhat lower than the ratio of 4.8 reported in maize (Tenaillon *et al.* 2001) and much lower than the 8.7 ratio reported in *D. melanogaster* (Moriyama and Powell 1996). In Arabidopsis, Olsen *et al.* (2002) analyzed six genes in the floral development pathway and found ratios of synonymous/nonsynonymous nucleotide diversity ranging from 0.5 to 9.5 (mean = 2.9). Low diversity at nonsynonymous sites is the result of selection against deleterious mutations. Outcrossing species are generally more effective at purging deleterious mutations as a result of large effective population size. Soybean, like Arabidopsis, has a low ratio of synonymous to nonsynonymous mutation, suggesting the presence of a relatively high level of slightly deleterious mutations.

A notable difference between sequence diversity in soybean *vs.* reports in humans was the relative levels of nucleotide diversity in coding and noncoding sequences. In the reports of Halushka *et al.* (1999) and Cargill *et al.* (1999) sequence polymorphism in humans was essentially identical in coding and closely associated noncoding DNA. Cargill *et al.* (1999) suggested that the similar nucleotide diversity might be indicative of regulatory or splicing function of noncoding perigenic sequence. In soybean, nucleotide diversity was 2.2 times greater in noncoding DNA closely associated with coding sequence and in *D. melanogaster* it was 2.6 times greater (Moriyama and Powell 1996). Data derived from three studies in Arabidopsis (Kawabe and Miyashita 1999; Kawabe *et al.* 2000; Kuittinen and Aguade 2000) indicated that perigenic sequences had levels of nucleotide diversity that were 2.6 times greater than that of the associated coding sequence. Apparently the level of functional constraint on perigenic sequence in soybean, Arabidopsis, and *D. melanogaster* is less than that in humans.

Limited haplotype diversity: The small number of haplotypes observed in our data suggests that the genetic base of cultivated soybean is built upon a small group of progenitor genotypes. This may be the result of a small number of domestication events from the wild relative *G. soja*. Alternatively, the limited haplotype diversity observed here may be only the result of the narrow genetic base of North American soybean germplasm or of limited variability in *G. soja*. A number of reports have documented the small group of progenitor genotypes that form the genetic base of North American soybean germplasm (Committee on Genetic Vulner-

ABILITY OF MAJOR CROPS 1972; DELANNAY *et al.* 1983). A comparison of sequence and haplotype diversity of North American and Asian *G. max* genotypes along with a representative sample of *G. soja* genotypes would provide useful information to serve as a guide in the development of strategies aimed at enhancing the genetic variability available for soybean improvement.

Linkage disequilibrium in soybean: Our data indicated that over relatively short distances of perhaps 50 kbp there is little decay in LD in soybean. This conclusion is based upon limited data from a set of subclones derived from one BAC clone that is 110 kbp in length. This finding is in marked contrast to reports in maize indicating that LD, as estimated by r^2, decayed to values <0.10 within 1500 bp (REMINGTON *et al.* 2001). An even more rapid rate of LD decay was noted by TENAILLON *et al.* (2001) in the analysis of genes on maize chromosome 1. Because of its autogamous nature, LD decay in Arabidopsis is likely to be more comparable to that of soybean. Reports by HANFSTINGL *et al.* (1994) and AGUADE (2001) studied individual Arabidopsis genes and found limited recombination over distances of 1.2–2.6 kbp, indicative of high levels of LD over short distances as would be anticipated in soybean.

The second estimate of LD reported here was also based upon limited data derived from the SNPs discovered in seven SSR flanking regions on soybean linkage group G. These data are quite variable as evidenced by large deviations from the trend line developed from the plot of the squared allele-frequency correlations on genetic map distance (Figure 1). Nonetheless, LD as estimated by r^2 decays to <0.10 at genetic map distances >2.5 cM. Recent work by NORDBORG *et al.* (2002) indicated that in Arabidopsis LD dissipates over distances of 1 cM, which corresponds to ~250 kbp. These authors found a generally similar level of LD decay across the Arabidopsis genome. Reports from other species have noted wide variation in LD decay in different genome regions. For example, highly variable rates of LD decline were observed among different maize genes (REMINGTON *et al.* 2001). Likewise, in humans wide variation in LD decay has been observed among genes (STEPHENS *et al.* 2001) and among genomic regions (REICH *et al.* 2001). Additional estimates of LD decline in soybean will be necessary to determine if the LD decay observed in the 12.5-cM region on linkage group G is typical of the soybean genome.

Our assessment of genome-wide LD used haplotypes at 54 loci that we assumed were distributed across the soybean genome. There was no reason to suggest that randomly selected genes and cDNAs would derive from only one or a few linkage groups. This analysis, like those of localized LD, used the subset of 16 Asian soybean introductions that were not derivatives of modern breeding programs and therefore artificial hybridization and recombination had not contributed to the dissipation of genome-wide LD. Among the 66 loci with two or more SNPs there was an average of 3.1 haplotypes. The lack of genome-wide LD coupled with the limited haplotype diversity suggests that the cultivated soybean genome is a mosaic of a limited number of haplotypes that may be the result of recombination among three or four ancestral haplotypes. Some natural outcrossing does occur in soybean despite its autogamous nature. The progeny of these rare outcrosses might have one or more distinctive features that would cause them to be the target of selection. Such cycles of outcrossing and selection could result in substantial recombination over a period of >3000 years since the estimated time of the domestication of the soybean, which probably took place in China during the Shang Dynasty (*ca.* 1700–1100 BC) or earlier (HYMOWITZ 1990). Another alternative is that the haplotype structure of the soybean genome predates domestication. Whatever the origins of haplotype structure in the soybean genome it is important to further define that structure. NORDBORG *et al.* (2002) concluded that the extensive haplotype structure of *A. thaliana* indicated that the development of a linkage disequilibrium map of Arabidopsis is feasible. Our limited data relating to the decay of LD in soybean suggest that the haplotype structure may be somewhat more extensive than that in Arabidopsis. However, a systematic assessment of genome-wide LD in soybean is clearly needed. Such an analysis would permit an appraisal of the likelihood that association analysis (CARDON and BELL 2001) could be successfully applied for gene discovery in soybean.

A soybean transcript map: One of the objectives of this research was to develop a strategy for SNP discovery aimed at the development of a SNP-based soybean linkage map. The large amount of soybean EST sequence data is a resource that may be useful for *in silico* SNP discovery as well as for the design of sequence-tagged sites (STSs) for SNP discovery via resequencing. The mapping of these SNPs would create a transcript map with candidate genes to associate with quantitative trait loci. Information on nucleotide diversity, the rate of success with which STSs can be developed from EST and genomic sequence, and SNP distribution allow an estimate of the feasibility of creating such a map. Of the 178 primer sets designed to complete genes and cDNAs, 115 yielded a sequence-tagged site from which sequence data were obtained. To a great extent, the failure to convert primer sets into STSs was the result of the amplification of multiple sites. Previous reports of genome duplication in soybean suggest the occurrence of tetraploidization as well as other duplication events (SHOEMAKER *et al.* 1996). SHOEMAKER *et al.* (1996) estimated that, on average, a given chromosomal segment was duplicated 2.55 times in the soybean genome. Because a STS is required both for SNP discovery via resequencing and for most methods of SNP detection, soybean genome duplication will no doubt reduce the efficiency and increase the cost of SNP discovery.

A total of 216 SNPs were discovered in the 66,634 bases of DNA analyzed in the 116 gene-derived STSs

(GenBank genes + cDNAs + 1 BAC subclone; Table 3) or a rate of 3.24 SNPs/kbp in these 25 soybean genotypes. The average length of the 116 gene fragments was 574 bp and at the rate of 3.24 SNPs/kbp, one would anticipate 1.86 SNPs per fragment. Under these circumstances, a rough estimate of the probability of finding at least one SNP in a 574-bp fragment is $[1 - (0.99676)^{574}] = 0.776$ and one would anticipate 90 of the 116 STSs to contain at least 1 SNP. However, at least 1 SNP was discovered in only 74 of the gene or perigenic DNA fragments analyzed. The heterogeneous distribution of SNPs was detected by the KS tests and is a phenomenon that is not unique to soybean. Similar evidence of wide differences in nucleotide diversity of genes and gene fragments has been reported in maize (Tenaillon *et al.* 2001), Arabidopsis (Olsen *et al.* 2002), *D. melanogaster* (Moriyama and Powell 1996), and humans (Cargill *et al.* 1999; Halushka *et al.* 1999). Thus, heterogeneity of nucleotide diversity is likely to reduce the success rate at which a SNP can be discovered in any given gene or gene fragment.

A strategy for SNP discovery in soybean: While genome duplication and heterogeneity of nucleotide diversity across fragments will negatively impact the likelihood of successfully discovering sequence variation in a particular DNA fragment, the knowledge that nucleotide diversity is more than twofold greater in noncoding perigenic DNA than in coding sequence suggests that SNP discovery should focus on these noncoding regions. The increasing availability of 3′-UTR data in soybean will be useful in this regard. Another approach to maximizing the usefulness of the large soybean EST database is an intron prediction protocol being used in SNP discovery in cattle (*Bos taurus*; Stone *et al.* 2002). Introns are predicted on the basis of comparison of cattle EST sequence with homologous human genomic sequence so that primers can be designed to the exon sequence around predicted intron-exon splice sites. The resulting amplification product from genomic DNA is composed mainly of intron sequence and would therefore be anticipated to have a greater likelihood of sequence variation. In the case of soybean, the availability of an essentially complete genome sequence of *A. thaliana* should permit the use of a similar approach for intron prediction in combination with soybean EST data. The genomic sequence of the model legume *Medicago truncatula* should be more useful in this type of analysis. In addition, while we assayed only ~10 kbp of random genomic sequence, the level of sequence diversity was higher, although nonsignificantly higher than that in the noncoding DNA associated with genes. This suggests that the BAC-end sequence will be a good source of data for SNP discovery as will SSR flanking regions.

The data reported here from a diverse set of soybean genotypes indicate the feasibility of large-scale SNP discovery in soybean and also provide guidance for such an effort. Discovery needs to focus, when possible, on primer design to noncoding sequence, where nucleotide diversity is expected to be greatest. Primer testing may be expedited by heteroduplex analysis of a homozygous genotype to eliminate those primer sets that amplify multiple (and heterogeneous) amplicons. Putative locus-specific primer sets can then be used to amplify genomic DNA of a pool of diverse genotypes followed by heteroduplex analysis to identify SNP-containing fragments. Heteroduplex analysis for SNP discovery using denaturing HPLC is well established (Jin *et al.* 1995). When heteroduplexes are detected, the individual genotypes can be sequenced or the amplicon derived from the pooled genotypes can be sequenced as suggested by Taillon-Miller *et al.* (1999).

The identification of a small set of soybean genotypes in which sequence diversity is maximized will enhance the efficiency of SNP discovery. Sequence analysis of the six genotypes Minsoy, Noir 1, Archer, Evans, Peking, and PI 209332 (Table 6) is likely to identify most sequence variants in North American soybean germplasm. Likewise, most of these variants will segregate in at least one of the RIL mapping populations available in our laboratory. These readily available populations are well characterized and have well-developed molecular genetic maps. An important and unanswered question is the utility of SNPs discovered in North American germplasm in a wider range of cultivated and wild soybean germplasm. If North American genotypes represent a unique soybean subpopulation in terms of sequence and haplotype diversity, then the strategy suggested here for SNP discovery will need to be modified. If we are to successfully mine germplasm using the power of association analysis it is important to have at least a basic understanding of the variability of the target germplasm for which these analyses are intended.

The authors thank Joann Mudge for her assistance in the identification of BAC clone UMN-K4. The authors thank the United Soybean Board (USB grants 9222 and 1243) for support of this research.

LITERATURE CITED

Aguade, M., 2001 Nucleotide sequence variation at two genes of the phenylpropanoid pathway, the *FAH1* and *F3H* genes, in *Arabidopsis thaliana.* Mol. Biol. Evol. **18:** 1–9.

Cardon, L. R., and J. I. Bell, 2001 Association study designs for complex diseases. Nat. Rev. Genet. **2:** 91–99.

Cargill, M., D. Altshuler, J. Ireland, P. Sklar, K. Ardlie *et al.*, 1999 Characterization of single-nucleotide polymorphisms in coding regions of human genes. Nat. Genet. **22:** 231–238.

Churchill, G. A., and R. W. Doerge, 1994 Empirical threshold values for quantitative trait mapping. Genetics **138:** 963–971.

Collins, F. S., L. D. Brooks and A. Charkravarti, 1998 A DNA polymorphism discovery resource for research on human genetic variation. Genome Res. **8:** 1229–1231.

Committee on Genetic Vulnerability of Major Crops, 1972 *Genetic Vulnerability of Major Crops.* National Academy of Science, Washington, DC.

Cooper, D. N., B. A. Smith, H. J. Cooke, S. Niemann and J. Schmidtke, 1985 An estimate of unique DNA sequence heterozygosity in the human genome. Hum. Genet. **69:** 201–205.

Coryell, V. H., H. Jessen, J. M. Schupp, D. Webb and P. Keim, 1999 Allele-specific hybridization markers for soybean. Theor. Appl. Genet. **98:** 690–696.

Cregan, P. B., T. Jarvik, A. L. Bush, R. C. Shoemaker, K. G. Lark *et al.*, 1999a An integrated genetic linkage map of the soybean genome. Crop Sci. **39:** 1464–1490.

Cregan, P. B., J. Mudge, E. W. Fickus, L. F. Marek, D. Danesh *et al.*, 1999b Targeted isolation of simple sequence repeat markers through the use of bacterial artificial chromosomes. Theor. Appl. Genet. **98:** 919–928.

Cregan, P. B., J. Mudge, E. W. Fickus, D. Danesh, R. Denny *et al.*, 1999c Two simple sequence repeat markers to select for soybean cyst nematode resistance conditioned by the *rhg1* locus. Theor. Appl. Genet. **99:** 811–818.

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

Gibbons, J. D., 1976 *Nonparametric Methods for Quantitative Analysis*, pp. 56–77. Holt, Rinehart & Winston, New York.

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

Halushka, M. K., J. B. Fan, K. Bentley, L. Hsie, N. Shen *et al.*, 1999 Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. Nat. Genet. **22:** 239–247.

Hanfstingl, U., A. Berry, E. A. Kellogg, J. T. Costa, W. Rudiger *et al.*, 1994 Haplotypic divergence coupled with lack of diversity at the *Arabidopsis thaliana* alcohol dehydrogenase locus: Roles for both balancing and directional selection? Genetics **138:** 811–828.

Hymowitz, T., 1990 Soybeans: the success story, pp.159–163 in *Advances in New Crops*, edited by J. Janick and J. Simon. Timber Press, Portland, OR.

Jin, L., P. A. Underhill, P. J. Oefner and L. L. Cavalli-Sforza, 1995 Systematic search for polymorphisms in the human genome using denaturing high-performance liquid chromatography (DHPLC). Am. J. Hum. Genet. **57** (Suppl.): A26.

Kawabe, A., and N. T. Miyashita, 1999 DNA variation in the basic chitinase locus (*ChiB*) region of the wild plant *Arabidopsis thaliana*. Genetics **153:** 1445–1453.

Kawabe, A., K. Yamane and N. T. Miyashita, 2000 DNA polymorphism at the cytosolic phosphoglucose isomerase (*PgiC*) locus of the wild plant *Arabidopsis thaliana*. Genetics **156:** 1339–1347.

Keim, P., T. Olson and R. Shoemaker, 1988 A rapid protocol for isolating soybean DNA. Soybean Genet. Newsl. **15:** 150–152.

Kruglyak, L., 1997 The use of a genetic map of biallelic markers in linkage studies. Nat. Genet. **17:** 21–24.

Kuittinen, H., and M. Aguade, 2000 Nucleotide variation at the *CHALCONE ISOMERASE* locus in *Arabidopsis thaliana*. Genetics **155:** 863–872.

Kwok, P.-Y., Q. Deng, H. Zakeri and D. A. Nickerson, 1996 Increasing the information content of STS-based genome maps: identifying polymorphisms in mapped STSs. Genomics **31:** 123–126.

Lindblad-Toh, K., E. Winchester, M. J. Daly, D. G. Wang, J. N. Hirschhorn *et al.*, 2000 Large-scale discovery and genotyping of single-nucleotide polymorphisms in the mouse. Nat. Genet. **24:** 381–385.

Marth, G. T., I. Korf, M. D. Yandell, R. T. Yeh, Z. Gu *et al.*, 1999 A general approach to single-nucleotide polymorphism discovery. Nat. Genet. **23:** 452–456

Moriyama, E. N., and J. R. Powell, 1996 Intraspecific nuclear DNA variation in Drosophila. Mol. Biol. Evol. **13:** 261–277.

Nordborg, M., B. Charlesworth and D. Charlesworth, 1996 Increased levels of polymorphism surrounding selectively maintained sites in highly selfing species. Proc. R. Soc. Lond. Ser. B **263:** 1033–1039.

Nordborg, M., J. O. Borevitz, J. Bergelson, C. C. Berry, J. Chory *et al.*, 2002 The extent of linkage disequilibrium in *Arabidopsis thaliana*. Nat. Genet. **30:** 190–193.

Olsen, K. M., A. Womack, A. R. Garrett, J. I. Suddith and M. D. Purugganan, 2002 Contrasting evolutionary forces in the *Arabidopsis thaliana* floral developmental pathway. Genetics **160:** 1641–1650.

Patil, N., A. J. Berno, D. A. Hinds, W. A. Barrett, J. M. Doshi *et al.*, 2001 Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. Science **294:** 1669–1670.

Pollak, E., 1987 On the theory of partially inbreeding finite populations. I. Partial selfing. Genetics **117:** 353–360.

Purugganan, M. D., and J. I. Suddith, 1999 Molecular population genetics of floral homeotic loci: departures from the equilibrium-neutral model at the *APETALA3* and *PISTILLATA* genes of *Arabidopsis thaliana*. Genetics **151:** 839–848.

Reich, D. E., M. Cargill, S. Bolk, J. Ireland, P. C. Sabeti *et al.*, 2001 Linkage disequilibrium in the human genome. Nature **411:** 199–204.

Remington, D. L, J. M. Thornsberry, Y. Matsuoka, L. M. Wilson, S. R. Whitt *et al.*, 2001 Structure of linkage disequilibrium and phenotypic associations in the maize genome. Proc. Natl. Acad. Sci. USA **98:** 11479–11484.

Rozas, J., and R. Rozas, 1999 DnaSP version 3: an integrated program for molecular population genetics and molecular evolution analysis. Bioinformatics **15:** 174–175.

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

Shoemaker, R. C., K. Polzin, J. Labate, J. Specht, E. C. Brummer *et al.*, 1996 Genome duplication in soybean (*Glycine* subgenus *soja*). Genetics **144:** 329–338.

Stephens, J. C., J. A. Schneider, D. A. Tanguay, J. Choi, T. Acharya *et al.*, 2001 Haplotype variation and linkage disequilibrium in 313 human genes. Science **293:** 489–493.

Stone, R. T., W. M. Grosse, E. Casas, T. P. Smith, J. W. Keele *et al.*, 2002 Use of bovine EST data and human genomic sequences to map 100 gene-specific bovine markers. Mamm. Genome **13:** 211–215.

Taillon-Miller, P., E. E. Piernot and P.-Y. Kwok, 1999 Efficient approach to unique single-nucleotide polymorphism discovery. Genome Res. **9:** 499–505.

Tajima, F., 1983 Evolutionary relationship of DNA sequences in finite populations. Genetics **105:** 437–460.

Tajima, F., 1989 Statistical nethods for testing the neutral mutation hypothesis by DNA polymorphism. Genetics **123:** 585–595.

Tenaillon, M. I., M. C. Sawkins, A. D. Long, R. L. Gaut, J. F. Doebley *et al.*, 2001 Patterns of DNA sequence polymorphism along chromosome 1 of maize (*Zea mays* ssp. *mays* L.). Proc. Natl. Acad. Sci. USA **98:** 9161–9166.

Wang, D. G., J. B. Fan, C. J. Siao, A. Berno, P. Young *et al.*, 1998 Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. Science **280:** 1077–1082.

Watterson, G. A., 1975 On the number of segregating sites in genetical models without recombination. Theor. Popul. Biol. **7:** 256–276.

Weir, B. S., 1990 *Genetic Data Analysis Methods for Discrete Genetic Data*. Sinauer Associates, Sunderland, MA.

Weir, B. S., 1996 *Genetic Data Analysis II*. Sinauer Associates, Sunderland, MA.

Xue, Z. T., M. L. Xu, W. Shen, N. L. Zhuang, W. M. Hu *et al.*, 1992 Characterization of a Gy_4 glycinin gene from soybean *Glycine max* cv. Forrest. Plant Mol. Biol. **18:** 897–908.

Zakharova, E. S., S. M. Epishin and Y. P. Vinetski, 1989 An attempt to elucidate the origin of cultivated soybean via comparison of nucleotide sequences encoding glycinin B_4 polypeptide of cultivated soybean, *Glycine max*, and its presumed wild progenitor, *Glycine soja*. Theor. Appl. Genet. **78:** 852–856.

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

Communicating editor: A. H. D. Brown