2007

A Soybean Transcript Map: Gene Distribution, Haplotype and Single-Nucleotide Polymorphism Analysis

Ik-Young Choi
Soybean Genomics and Improvement Laboratory, USDA, ARS, Beltsville, Maryland 20705

D. L. Hyten
Soybean Genomics and Improvement Laboratory, USDA Agricultural Research Service, Beltsville, Maryland, david.hyten@unl.edu

Lakshmi K. Matukumalli
Bovine Functional Genomics Laboratory, USDA, ARS, Beltsville, Maryland 20705

Qijian Song
University of Maryland - College Park, qijian.song@ars.usda.gov

Julian M. Chaky
University of Nebraska-Lincoln

Authors

Ik-Young Choi, D. L. Hyten, Lakshmi K. Matukumalli, Qijian Song, Julian M. Chaky, Charles V. Quigley, Kevin Chase, K. Gordon Lark, Robert S. Reiter, Mun-Sup Yoon, Eun-Young Hwang, Seung-In Yi, Nevin D. Young, Randy C. Shoemaker, Curtis P. van Tassell, James E. Specht, and P. B. Cregan

A Soybean Transcript Map: Gene Distribution, Haplotype and Single-Nucleotide Polymorphism Analysis

Ik-Young Choi,*,[1] David L. Hyten,* Lakshmi K. Matukumalli,[†] Qijian Song,*,[‡]
Julian M. Chaky,[§] Charles V. Quigley,* Kevin Chase,** K. Gordon Lark,**
Robert S. Reiter,[††] Mun-Sup Yoon,*,[‡‡] Eun-Young Hwang,*,[‡] Seung-In Yi,*,[§§]
Nevin D. Young,*** Randy C. Shoemaker,[†††] Curtis P. van Tassell,[†]
James E. Specht[§] and Perry B. Cregan*,[2]

*Soybean Genomics and Improvement Laboratory and [†]Bovine Functional Genomics Laboratory, USDA, ARS, Beltsville, Maryland 20705,
[‡]Department of Natural Resources and Landscape Architecture, University of Maryland, College Park, Maryland 20742, [§]Department of
Agronomy, University of Nebraska, Lincoln, Nebraska 68583-0915, **Department of Biology, University of Utah, Salt Lake City,
Utah 84112, [††]Monsanto, St. Louis, Missouri 63167, [‡‡]Genetic Resources Division, National Institute of Agricultural
Biotechnology, Rural Development Administration, Suwon, 441-707, Republic of Korea, [§§]National Seed
Management Office, Suwon, 442-400, Republic of Korea, ***Department of Plant Pathology,
University of Minnesota, St. Paul, Minnesota 55108 and [†††]Department of Agronomy,
USDA–ARS–CICG, Iowa State University, Ames, Iowa 50011

ABSTRACT

The first genetic transcript map of the soybean genome was created by mapping one SNP in each of 1141 genes in one or more of three recombinant inbred line mapping populations, thus providing a picture of the distribution of genic sequences across the mapped portion of the genome. Single-nucleotide polymorphisms (SNPs) were discovered via the resequencing of sequence-tagged sites (STSs) developed from expressed sequence tag (EST) sequence. From an initial set of 9459 polymerase chain reaction primer sets designed to a diverse set of genes, 4240 STSs were amplified and sequenced in each of six diverse soybean genotypes. In the resulting 2.44 Mbp of aligned sequence, a total of 5551 SNPs were discovered, including 4712 single-base changes and 839 indels for an average nucleotide diversity of $\theta = 0.000997$. The analysis of the observed genetic distances between adjacent genes *vs.* the theoretical distribution based upon the assumption of a random distribution of genes across the 20 soybean linkage groups clearly indicated that genes were clustered. Of the 1141 genes, 291 mapped to 72 of the 112 gaps of 5–10 cM in the preexisting simple sequence repeat (SSR)-based map, while 111 genes mapped in 19 of the 26 gaps >10 cM. The addition of 1141 sequence-based genic markers to the soybean genome map will provide an important resource to soybean geneticists for quantitative trait locus discovery and map-based cloning, as well as to soybean breeders who increasingly depend upon marker-assisted selection in cultivar improvement.

THE gene space of soybean [*Glycine max* (L.) Merr.] is as yet undefined. MUDGE *et al.* (2004) estimated that most genes in soybean are clustered in ~25% of the genome (275 Mbp). One suggestion for providing information relating to gene space was to place 2000–3000 cDNA sequences onto the physical map (STACEY *et al.* 2004). Another approach would be to genetically map coding sequences onto the existing simple sequence repeat (SSR)-based map (SONG *et al.* 2004). The resulting genetic map would not only indicate the positions of coding sequences, but also answer questions about relationships of coding sequences with existing SSR

and RFLP markers. These relationships are important in map-based cloning, quantitative trait locus discovery, and marker-assisted plant improvement. While a fairly extensive set of >1000 genetically mapped SSR markers is available to soybean breeders and geneticists (CREGAN *et al.* 1999; SONG *et al.* 2004) the current map has 138 gaps of >5 cM in which no SSR marker is present. Twenty-six of these 138 gaps are >10 cM, which is troubling if these regions of low SSR marker density are also gene rich.

The presence of SSRs in expressed sequence tag (EST) sequence provides one means for the genetic mapping of the EST; however, the number of soybean ESTs that contain polymorphic SSRs appears rather limited. SONG *et al.* (2004) were successful in developing and mapping only 24 polymorphic SSR markers from >136,000 soybean ESTs. Alternatively, discovery of single-nucleotide polymorphisms (SNPs, which include

[1]*Present address:* NICEM, CALS, Seoul National University, San 56-1, Sillim-9-dong, Gwanak-gu, Seoul, 151-921, South Korea.

[2]*Corresponding author:* Soybean Genomics and Improvement Laboratory, Bldg. 006, Room 100, USDA, ARS, Beltsville, MD 20705.
E-mail: perry.cregan@ars.usda.gov

single-base changes and insertion/deletions) in genic sequence would provide a source of markers that can be analyzed via PCR and other approaches. Because SNPs are more abundant than SSRs, they improve the odds of success in a diversity of applications including positional cloning, association analysis, QTL mapping, and the determination of genetic relationships among individuals. As of December 2006, almost 28 million human SNPs had been cataloged by the National Center for Biotechnology Information SNP database, dbSNP, of which >5.6 million had been validated. The discovery and application of SNPs in plant species has lagged behind that in humans. However, the pace of SNP discovery and application has increased in species including *Arabidopsis thaliana* (JANDER *et al.* 2002; SCHMID *et al.* 2003, 2005; NORDBORG *et al.* 2005), maize (*Zea mays* L.) (TENAILLON *et al.* 2001; CHING *et al.* 2002), rice (*Oryza sativa* L.) (NASU *et al.* 2002; FELTUS *et al.* 2004), barley (ROSTOKS *et al.* 2005), poplar (*Populus trichocarpa* Torr. & Gray) (TUSKAN *et al.* 2006), and a number of other species.

Two important factors influencing SNP discovery are the frequency of sequence variants and the presence of genome duplications. The frequency of sequence variants in soybean is low due to historical genetic bottlenecks and low sequence diversity in soybean's wild ancestor *G. soja* (Sieb and Zucc.) (HYTEN *et al.* 2006). ZHU *et al.* (2003) reported that SNP frequency as measured by nucleotide diversity (θ) was ~0.00053 in 28.7 kbp of coding sequence analyzed in a diverse set of 25 soybean genotypes. There was more than a twofold higher nucleotide diversity in untranslated regions (UTRs), introns, and genomic DNA in close proximity to coding sequence ($\theta = 0.00111$). These values are similar to those reported in humans ($\pi = 0.000751$) from 1.2 billion bp of aligned human sequence (SACHIDANANDAM *et al.* 2001), which indicates that despite the low level of sequence variation, successful SNP discovery is still possible. In addition, >356,000 soybean EST sequences are available in GenBank, suggesting that *in silico* SNP discovery should be successful in soybean. However, *in silico* SNP discovery in soybean is impeded by the nature of the soybean genome and limited genotypic diversity of available EST data. NELSON *et al.* (2005) suggested the *in silico* analysis of EST FASTA data as an approach to SNP discovery in an inbred species such as soybean. However, the duplicated nature of the soybean genome and the paucity of a large set of EST data from alternative genotypes as well as sequencing errors inherent in FASTA data make *in silico* SNP discovery difficult. As a result of the duplicated nature of the soybean genome, a large proportion of soybean ESTs are members of paralogous sets of expressed genes. The soybean is likely an ancient tetraploid with a chromosome number of $2n = 40$ (HYMOWITZ 2004). Recent analyses suggest that the soybean genome has undergone two major duplication events (BLANC and WOLFE 2004;

SCHLUETER *et al.* 2004). This complicates development of sequence-tagged sites (STSs) required for resequencing to either validate *in silico*-derived SNPs or for *de novo* SNP discovery. The analysis of EST data to create the set of unique soybean genes or unigenes might reduce the problems associated with genome duplications and expedite selection of robust STSs. VODKIN *et al.* (2004) compiled an initial set of 61,127 soybean unigenes on the basis of the alignment of 5′ ESTs. This was followed by 3′-end sequencing of one cDNA from each of 27,513 unigenes. These sequence data provide an important resource for soybean SNP discovery.

The objective of this study was to develop STSs using EST and 3′-unigene sequence and then to use the STSs for the discovery of SNPs via the resequencing of six diverse soybean genotypes. SNPs were discovered in >2000 STSs, and >1000 SNPs derived from genic sequence were genetically mapped to create a transcript map that provides a map-based definition of the soybean gene space and will play a key role in the assembly of the whole-genome shotgun sequence.

MATERIALS AND METHODS

Plant materials: ZHU *et al.* (2003) indicated that the sequence analysis of six diverse soybean genotypes Archer, Minsoy, Noir 1, Evans, PI 209332, and Peking would permit discovery of 93% of the common SNPs (frequency >0.10) discovered in a diverse group of 25 genotypes. Among these six genotypes are the parents of four recombinant inbred line (RIL) mapping populations used in our laboratories. DNA of the six genotypes was isolated from bulked leaf tissue as described by KEIM *et al.* (1988).

PCR primer selection in EST clusters: All ESTs available as of August 2003 generated from cultivars Williams and Williams 82 were clustered using the assembly program CAP3 with the requirement that all ESTs within a cluster should have >95% similarity. Each of the clusters was then individually analyzed for the presence of high Phred quality polymorphisms. All like sequences within a cluster were further grouped into smaller sets of similar sequences that were likely to be paralogs. Paralog sequence variants (PSVs) were identified and primers were designed to include the PSVs at the 3′ end of the primers with the intention of providing paralog-specific amplification. Primers were selected using Primer3 (Whitehead Institute, Massachusetts Institute of Technology, MIT).

PCR primer selection in 3′ unigene sequence: PCR primers were designed to 8587 3′-end soybean unigene sequences deposited in GenBank as a result of work reported by VODKIN *et al.* (2004) with a range of predicted fragments lengths between 300 and 800 bp. A total of 5798 primers were designed with OLIGO (National Biolabs, St. Paul) and Array Designer 2 (Premier Biosoft International, Palo Alto, CA) using only FASTA data while 2789 primer sets were designed with Primer 3 (Whitehead Institute, MIT) using FASTA data and Phred quality scores. When quality scores were used, all bases within primers were required to have a minimum quality score of 20. In an attempt to provide specificity, the reverse primer was positioned as far to the 3′ end of the cDNA sequence as possible with the intention of maximizing the likelihood of priming from the 3′ UTR.

Preliminary analysis of PCR primers: Each primer pair was used to amplify genomic DNA of Archer soybean. Amplification reactions were performed with 20 ng of DNA, 0.1 μM of forward and reverse primers, 1× FailSafe PCR PreMix B buffer (Epicentre Technologies, Madison, WI) or a buffer consisting of 20 mM Tris–HCl (pH 8.0), 50 mM KCl, 200 μM each dNTP, 1.0% glycerol, 1.5 mM MgCl₂, 2.0 ng/μl single-stranded DNA-binding protein, and Taq DNA polymerase in a 10-μl reaction volume (45 sec at 92°, denaturation; 45 sec at 58°, annealing; and 45 sec 68°, extension) for 40 cycles. PCR products were resolved on a 2.5% agarose gel stained with ethidium bromide. Reactions that gave no product or multiple products were reamplified with either a lower (no product) or a higher (multiple products) annealing temperature. The primer pairs that amplified a single discrete product were selected for further analysis.

The amplicon from each selected primer pair was prepared for sequence analysis by treatment with 4 units of shrimp alkaline phosphatase (SAP) and 4 units of exonuclease I incubated at 37° for 1 hr followed by 72° for 15 min to deactivate the enzymes. Labeling reactions were performed with 1 μl of PCR product, 0.5 μl of BigDye Terminators, version 1.1 or 3.1 (Applied Biosystems, Foster City, CA), 0.3 μM of one of the original PCR primers, 1× Taq DNA polymerase buffer (Promega, Madison, WI), and 1.75 mM of MgCl₂ in a 5-μl reaction volume (10 sec at 90°, denaturation; 5 sec at 50°, annealing; 60 sec at 60°, extension) for 40 cycles. The PCR products were labeled from both ends and the resulting termination products were analyzed on an ABI 3730 DNA analyzer. The two resulting sequence traces derived from opposite ends of each amplicon were analyzed and aligned with standard DNA analysis software Phred (Ewing and Green 1998) and Phrap (http://www.phrap.org/). Resulting alignments and trace data were visually inspected in the Consed viewer (Gordon et al. 1998) to distinguish those amplicons that were locus specific and those that apparently resulted from amplification of two or more paralogous loci. Those primer sets that produced what appeared to be single-locus amplicons were used for PCR amplification of genomic DNA of the other five soybean genotypes Minsoy, Noir 1, Evans, PI 209332, and Peking. Resulting amplicons were treated with SAP and exonuclease I as described above, followed by sequence analysis on the ABI 3730. Forward and reverse sequence traces from the five genotypes as well as that from Archer were analyzed and aligned as described above. SNP discovery was carried out in the sequence alignments with a machine learning algorithm based on PolyBayes SNP discovery software (Marth et al. 1999; Matukumalli et al. 2006a).

Sequence analysis for STS verification and SNP discovery: Sequence traces for each putative SNP identified were visually inspected to verify sequence polymorphisms. Single-nucleotide changes and indels present in each alignment as well as the haplotypes present among the six genotypes were recorded as described by Matukumalli et al. (2006b).

Analysis of introns: Consensus sequence obtained from the Phrap alignment of the six soybean genotypes was aligned with the EST sequence to which primers were designed with the software alignment program bl2seq (http://www.ncbi.nlm.nih.gov) to determine the presence and length of introns.

Genetic mapping of SNP-containing loci: One SNP was mapped from SNP-containing STSs derived from the EST clusters and 3′ unigene sequence, as well as from the sequence of soybean genes deposited in GenBank. The allele present at each SNP locus was determined using single-base extension on either the Sequenom MassARRAY platform or the Luminex flow cytometer. The MassARRAY system is based on single-base primer extension technology. The MassARRAY technology uses matrix-assisted laser desorption ionization time-of-flight

(MALDI-TOF) mass spectrometry to measure directly the mass of the extension product(s) and then correlates the detected mass with a specific genotype (Griffin et al. 1999). Details of the protocol are available on the Sequenom web site: http://www.sequenom.com/applications/hme_assay.php. Single-base extension technology was also used to map SNP-containing loci on the Luminex flow cytometer as described by Chen et al. (2000). SNPs were mapped in three different mapping populations including the University of Utah Minsoy × Noir 1 (M × N) and Minsoy × Archer (M × A) as well as the Evans × PI 209332 (E × PI) (Concibido et al. 1996) RIL mapping populations. The first two populations were used in creating the current integrated linkage map (Song et al. 2004). A total of 89 RILs from the M × A population were assayed using the MassARRAY technology and the Luminex flow cytometry system. In the case of the M × N population, 89 RILs were assayed using MassARRAY genotyping and 75 RILs were genotyped using the Luminex flow cytometer. A total of 77 F₆-derived RILs from E × PI were used for genetic mapping. SNPs in 500 genes in the M × N and 501 genes in the M × A populations, respectively, were analyzed using the MassARRAY system. SNPs in 337 genes in the M × N and 128 genes in the M × A populations, respectively, were analyzed using the Luminex flow cytometer. One hundred forty-six genes were mapped in the E × PI population using the Luminex flow cytometer. A total of 233 SNPs in the M × N and 115 SNPs in the M × A populations were genotyped using both analysis platforms. In those cases in which the same SNP was mapped using both the MassARRAY and Luminex systems, genotyping data with the least amount of missing data were used in map construction. To provide a set of markers common to the three mapping populations to expedite the JoinMap linkage analysis, 77 RILs of the Evans × PI 209332 were genotyped with 155 SSR loci using the analysis system described by Wang et al. (2003).

Statistical analyses: *Nucleotide diversity:* Nucleotide diversity (θ) was estimated as per Halushka et al. (1999),

$$\theta = K/aL$$

$$a = \sum_{i=2}^{n} 1/(i-1),$$

where *K* is the number of SNPs identified in an alignment of *n* genotypes, *L* bp in length. Nucleotide diversity was determined in both the intron and the exon sequence.

Haplotype diversity: Haplotype diversity was calculated in a manner identical to the calculation of gene diversity (Weir 1996) as $1 - \sum P_{ij}^2$, where P_{ij} is the frequency of the *j*th haplotype for the *i*th locus summed across all haplotypes in the locus.

Distribution of SNPs in STSs: To determine if SNPs were evenly distributed in the fragments assayed, the theoretical SNP cumulative frequency distribution for SNPs was calculated on the basis of the assumption of uniform distribution. 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.

Consensus map construction: The data sets for JoinMap analysis consisted of 1023 A81-356022 × PI 468916, 1610 M × N, 1072 M × A, and 374 E × PI markers. Of 2807 unique markers in the data sets, a total of 533, 212, and 36 markers were common to two, three, and four populations, respectively. The markers in each of the four populations were grouped separately on the basis of their LOD scores and then integrated using JoinMap (Van Ooijen and Voorrips 2001). The soybean linkage groups were identified using the

alphanumeric codes described by CREGAN *et al.* (1999) and SONG *et al.* (2004). Recombination values were converted to genetic distances using the Kosambi mapping function.

Distribution of genic sequences—test of goodness-of-fit of theoretical and observed numbers of newly mapped genes to each linkage group: The chi-square statistic is commonly used to test for goodness-of-fit of an observed to a theoretical distribution. However, the test of each individual linkage group is problematic due to null degrees of freedom. Therefore, a permutation algorithm was designed to assess the significance level. For each permutation, a total of 1141 SNP markers were randomly assigned to any position in a total length of 2388.61 cM (the total length of the 20 linkage groups) at 0.01-cM intervals. The theoretical number of genes assigned to each linkage group was compared to the observed number of genes mapped to each linkage group for each of 5000 permutations. The probability of goodness-of-fit was measured by the proportion of times that the observed number of genes in each linkage group was larger or smaller than expectation.

Distribution of genes within linkage groups: The theoretical distribution of map distances between adjacent genic sequences within linkage groups was estimated on the basis of the assumption of a random distribution of markers over the total length of the linkage map. The goodness-of-fit between observed and theoretical distributions was tested using chi-square analysis.

Distribution of genes vs. SSR and RFLP loci: To estimate the relative proximities of genes on the newly developed transcript map with the preexisting SSR and RFLP loci, the genetic distance between each SSR and RFLP locus to the two flanking genic loci was determined. Genic loci included both the loci mapped in this study as well as classical genes that were previously mapped and 37 genes that were mapped by virtue of an SSR they contained. The SSRs in these 37 genes were not included in the analysis. The proportions of genetic distances that fell in classes of 0.0–0.1 cM, 0.1–0.2 cM, 0.2–0.3 cM, etc., intervals were calculated both for SSRs *vs.* genes and for RFLPs *vs.* genes. The proportions of genetic distances

in the various distance classes allowed a comparison of the relative genetic proximities of SSR and RFLP loci to adjacent genes.

RESULTS

SNPs in EST clusters: A total of 160,000 ESTs from cultivars Williams and Williams 82 were clustered using CAP3 and then further grouped into subsets believed to correspond to paralogs. Primers were then designed to high Phred quality polymorphisms with the intention of providing paralog specificity. Of 872 primers tested, one-third failed to produce an amplicon and 516 (59.2%) produced a single band on agarose (Table 1). Upon sequence analysis of these amplicons in the genotype Archer, 367 (42.1%) produced generally high-quality sequence data allowing alignment of the 5′ and 3′ traces. In a number of instances, the sequence traces contained what appeared to be "heterozygous" positions indicative of either haplotype variation or variation between very similar homeologous loci. The 367 primers pairs that produced good-quality sequence data were used to amplify the corresponding fragment from the five additional genotypes Minsoy, Noir 1, Evans, Peking, and PI 209332 followed by sequence analysis from both ends. Alignment and analysis of these sequence traces with those from Archer soybean indicated that 46 fragments that contained heterozygous positions in Archer soybean also had identical heterozygous bases at the same positions in each of the five other genotypes. A typical example of this phenomenon is illustrated

TABLE 1

Number of PCR primers designed and results of PCR and sequence analysis in six soybean genotypes

	Primers designed to EST clusters		Primers designed to 3′ unigene FASTA data		Primers designed to 3′ unigene sequence, all positions with Phred >20		Total	
	No.	% of total	No.	% of total	No.	% of total	No.	% of total
Primers designed	872		5,798		2,789		9,459	
Agarose gel analysis								
Primers producing no product	291	33.4	1,404	24.2	603	21.6	2,298	24.3
Primers producing multiple bands	65	7.5	561	9.7	245	8.8	871	9.2
Primers producing a single band	516	59.2	3,833	66.1	1,941	69.6	6,290	66.5
Sequence analysis								
Multiple amplicons	81	9.3	390	6.7	332	11.9	803	8.5
Paralogous sequence	46	5.3	321	5.5	106	3.8	473	5.0
Single amplicon (STS)	321	36.8	2,548	43.9	1,371	49.2	4,240	44.8
No sequence data	68	7.8	574	9.9	132	4.7	774	8.2
Fragments with at least 1 intron	181	20.7	1,161	20.0	657	23.6	1,999	21.1
Fragments with at least 1 SNP	225	25.8	1,222	21.1	585	21.0	2,032	21.5
Length of aligned sequence								
Total (bp)	214,223		1,444,053		778,992		2,437,268	
Mean STS length	667		567		568		575	
Total SNPs	683		3,312		1,556		5,551	
θ ($\times$ 1000)	1.396		1.004		0.875		0.997	

in supplemental Figure 1 at http://www.genetics.org/supplemental/ and suggested that the variation was PSV rather than haplotype variation. The analysis of the remaining 321 alignments indicated that they were the result of amplification from a single locus, *i.e.*, an STS. One hundred eighty-one of these STSs contained one or more introns and 225 (25.8% of 872 primer pairs analyzed) contained at least one single-base change or indel, as detected by PolyBayes and verified by visual inspection of the alignments. The 321 STSs contained 214.2 kbp of aligned sequence and a nucleotide diversity of $\theta = 0.001396$.

SNP discovery in primers designed to 3′ unigene sequence data: In an attempt to increase primer specificity, 3′ unigene sequence data were used in primer design such that one primer was positioned as far to the 3′ end of the unigene as possible. Of those primers designed to FASTA data, 24.2% did not produce a PCR product, which is slightly higher than the 21.6% that did not produce a PCR product when quality scores were used in primer design (Table 1). The proportion of primer sets that amplified multiple or paralogous loci was slightly different between the two groups. However, the proportion of primer pairs that resulted in an STS when primers were designed to sequence with a Phred score of ≥ 20 was 49.2%, which was greater than the 43.9% of successful STSs produced when primers were designed to FASTA sequence.

A total 8587 primer sets were designed to 3′ unigene sequence of which 4346 produced high-quality sequence data from the analysis of the STSs amplified from genomic DNA of Archer soybean. As was the case with primers designed to the EST clusters, a number of the sequence traces contained heterozygous positions. The subsequent analysis of the 4346 amplicons from the five genotypes Minsoy, Noir 1, Evans, Peking, and PI 209332 indicated that in 427 alignments of traces from the six genotypes each of the genotypes had "heterozygotes" at the same positions, indicating that variation was between paralogs as described earlier and illustrated in supplemental Figure 1 at http://www.genetics.org/supplemental/. A total of 3919 of the primers designed to 3′ unigene sequence resulted in what appeared to be a robust STS as indicated by the sequence analysis of the six genotypes and 1807 of these contained at least one

SNP. Thus, 21% of the primer pairs tested resulted in an SNP-containing amplicon.

Multiple amplicons: A total of 22.7% of the primers designed and tested produced (1) multiple products on an agarose gel (9.2%), (2) multiple amplicons as determined by sequence analysis (8.5%), or (3) evidence of PSV as depicted in supplemental Figure 1 at http://www.genetics.org/supplemental/ (5.0%).

Heterogeneity of nucleotide diversity among genes: The analysis of 9459 primer pairs resulted in the development of 4240 sequence-tagged sites with an average length of 575 bp and a mean of 1.31 SNPs per fragment. The nucleotide diversity in the 2.44 Mbp of aligned genic sequence was $\theta = 0.000997$. More than half of the gene fragments had no sequence variation, which suggested an uneven distribution of sequence variation among the fragments. The Kolmogorov–Smirnov test was conducted to compare the observed cumulative frequency distributions of SNPs in fragments with the theoretical distributions based upon the assumption of mutations being evenly distributed across the 4240 fragments. 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 gene fragments analyzed in this study.

Characteristics of SNPs in exons and introns: The aligned sequence of the 4240 STSs resulted in the discovery of 5551 SNPs (Table 2). The analysis of the aligned sequence indicated that 91.6 kbp (37.6%) was intron sequence with a mean of 1.64 introns per gene fragment and an average intron length of 279 bp. The number of introns per gene fragment ranged from 1 to 8 and there were 2.66 SNPs/kbp in introns *vs.* 2.04 SNPs/kbp in exon sequence. Thus, the intron nucleotide diversity ($\theta = 0.001168$) was somewhat greater than that in the exons ($\theta = 0.000895$). Approximately 85% of the SNPs were single-base changes of which 55.7% were transitions and 44.3% transversions. In five cases, triallelic single-base changes were discovered in the six genotypes. The remaining 15% of the SNPs were indels that ranged in length from 1 to 104 bp, of which 14.3% were >5 bp in length and 4.8% were >10 bp in length. Fifty-one percent of the 827 indels were 1 bp in length. The distribution of indel lengths in exons and introns indicated that the proportion of indels in different

TABLE 2

Characteristics of SNPs in soybean genic sequence

	Aligned sequence length (bp)	Single-base changes	Transitions	Transversions	Others[a]	Indels	SNPs	SNP frequency (SNPs/kbp)	Nucleotide diversity ($\theta \times 10^3$)
Exon	1,520,855	2,718	1,559	1,159	3	390	3,108	2.04	0.895 ± 0.47
Intron	916,413	1,994	1,064	930	2	449	2,443	2.66	1.168 ± 0.62
Total	2,437,268	4,712	2,623	2,089	5	839	5,551	2.28	0.997 ± 0.53

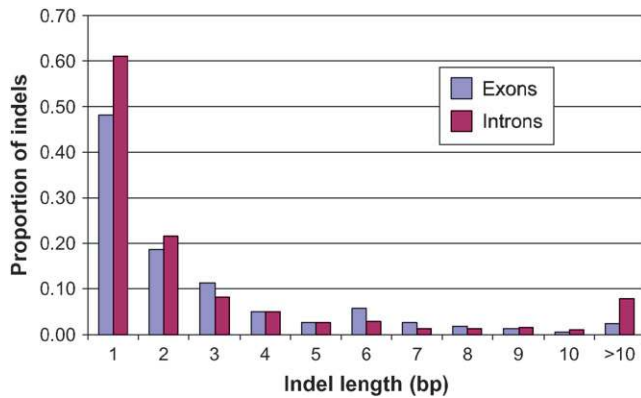[a] Triallelic single-nucleotide changes.

Figure 1.—The distribution of indel lengths among 839 indels discovered in 4240 sequence-tagged sites comprising 2.44 Mbp of aligned genic sequence.

length classes was fairly similar in exons and introns (Figure 1). There was a slightly higher proportion of indels of 1, 2, and >10 bp in introns while the proportion of 3- and 6-bp indels was greater in exons. Information relating to each of the 2032 SNP-containing STSs and the 5551 SNPs discovered as a result of the resequencing analysis is provided at http://bfgl.anri. barc.usda.gov/soybean/ (files: 2032-SNP-containing STS.xls and 5551-SNPs-2032 STS.xls).

Haplotype diversity: Among the six genotypes analyzed, just two haplotypes were found in 1502 (73.9%) of the 2032 gene fragments that contained a sequence variant (Table 3). Only two haplotypes were found in 837 fragments with a single sequence variant, as well as many with multiple SNPs. Four or more haplotypes were observed in <5% of the gene fragments. There was a mean of 2.31 haplotypes in the 2032 SNP-containing genes. Haplotype diversity in gene fragments that contained two or more SNPs was 0.48 ± 0.13.

Molecular function of genes: Since genes were chosen randomly, it was anticipated that a diverse set of genes would be included in this study. This anticipation was realized as is suggested by Figure 2. There were small differences among the proportions of genes in the various gene prediction classes for which primers were designed *vs.* STSs developed as opposed to those in which sequence diversity was discovered to permit genetic mapping. For example, ~12% of genes to which primers were designed were in the category of catalytic activity whereas 14% of loci that were mapped were in this category. Similarly, 1.3% of genes to which primers were designed were predicted to function in signal transduction whereas only 0.52% of these genes were mapped. In general, for most gene prediction categories, such differences were relatively minor.

Genetic Mapping of STSs: A total of 1141 genes were placed on the genetic map by virtue of a segregating SNP mapped in one or more of the three RIL mapping populations. In the case of the M × A population, 115 SNPs were mapped using both the Sequenom Mass Spectrometer MassARRAY technology and the Luminex

TABLE 3

Haplotype numbers in 2032 SNP-containing gene fragments

| SNPs/fragment	Potential haplotypes	SNP haplotype no.					
		2	3	4	5	6	Total
1	2	837					837
2	4	336	171	1			508
3	6	153	92	16			261
4	6	68	69	14			151
5	6	36	30	12	2		80
6	6	18	20	8	2		48
7	6	15	15	5	3	1	39
8	6	9	14	3	1		27
9	6	9	8	5			22
10	6	8	5	3			16
11	6	1	4	2			7
12	6	2	3		1		6
13	6		4	1			5
14	6	2	1	1			4
15	6	2		1			3
>15	6	6	10	1		1	18
Total fragments		1502	446	73	9	2	2032
Total haplotypes		3004	1338	292	45	12	4691
Proportion with haplotype no.		0.64	0.29	0.06	0.01	0.003	
Mean haplotype no./fragment							2.31
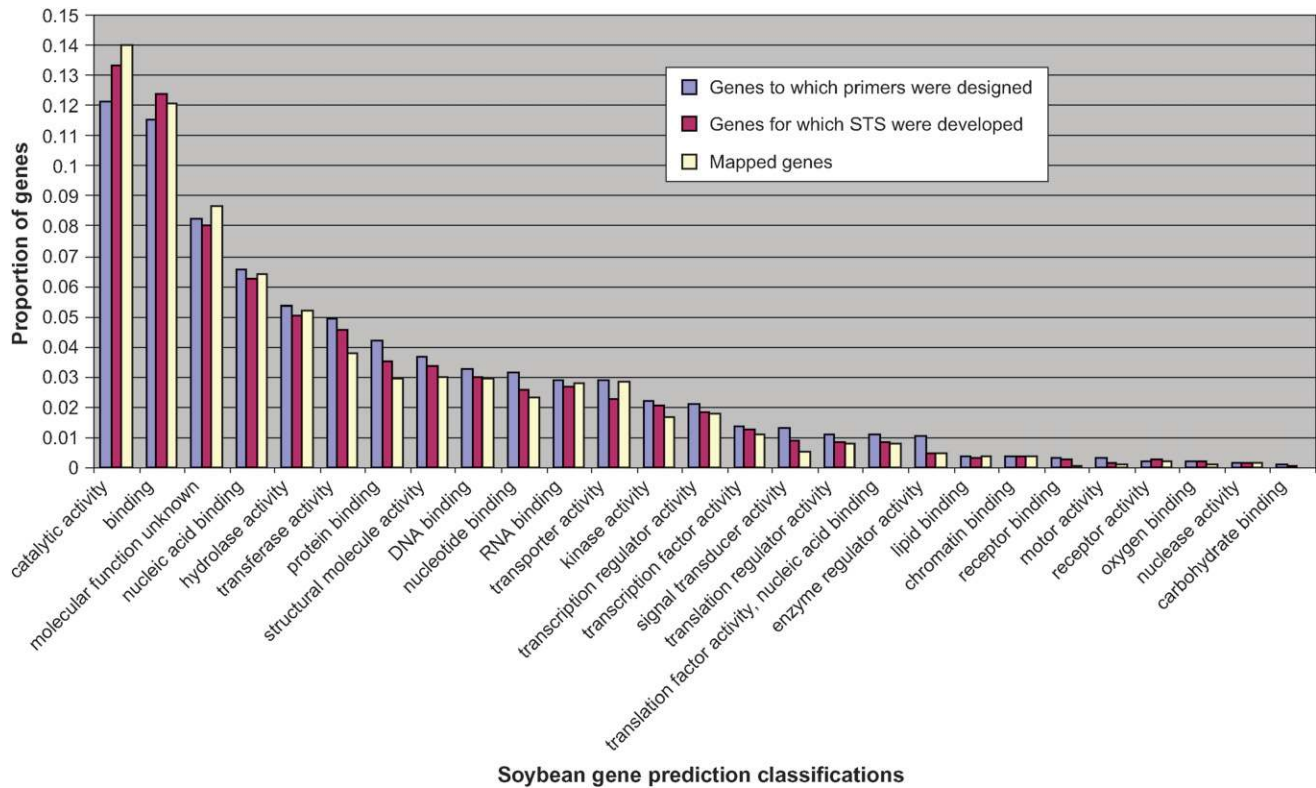
FIGURE 2.—Predicted molecular functions of genes analyzed.

flow cytometer on the same set of 89 RILs. Greater than 98.5% of RIL genotype calls were identical using both SNP detection platforms. Loci were placed on the existing framework map, which consists of 1015 SSR and 709 RFLP loci (SONG *et al.* 2004). The genetic map with SSR, RFLP, SNP, and other loci is provided in supplemental data file 1 at http://www.genetics.org/supplemental/. On the basis of the lengths of the 20 linkage groups in the framework map, the predicted number of genes mapped per linkage group was determined assuming a random distribution of genes. A significantly greater than predicted gene density occurred in 3 of the 20 linkage groups (E, J, and K) (Table 4), while a smaller than expected number of genes mapped to linkage group M. An additional analysis was undertaken to determine the nature of the distribution of genes within linkage groups. The analysis of the theoretical distribution of map distances between adjacent genic sequences clearly indicated that genes were clustered (Figure 3).

Distribution of genes *vs.* SSR and RFLP loci: To estimate the relative proximities of genes on the newly developed transcript map to the preexisting SSR and RFLP loci, genetic distances of each SSR and RFLP locus to the two closest flanking genic loci were determined. The graph of the proportion of SSR and RFLP loci that fall at various intervals from the closest two flanking genes suggested little difference between the proximities of SSR and RFLP loci to genes (Figure 4). However, a higher proportion of SSRs appeared to be in very close

proximity (0.0–0.5 cM) to genes than were RFLP loci. In addition, correlations of the number of genes with SSRs and genes with RFLPs on the 20 soybean linkage groups indicated a moderately strong relationship between SSR density and gene density. The correlation of SSR numbers per linkage group with genes per linkage group was $r = 0.58$ ($P = 0.0075$), whereas that between the numbers of RFLP loci and genic loci was substantially lower, $r = 0.25$ ($P = 0.29$). Together these data suggested an apparent association between SSRs and genic sequence.

Gaps in the linkage map now populated with SNP markers: In the SONG *et al.* (2004) linkage map, a total of 112 and 26 intervals between adjacent SSR markers are >5 and 10 cM, respectively. In the new map, a total of 291 genes were mapped in 72 of the 112 gaps with distances between 5 and 10 cM and 111 genes were mapped in 19 of 26 gaps >10 cM. Thus, sequence-based markers effectively filled many of the gaps between the preexisting sequence-based markers in the linkage map.

A soybean SNP database: To provide web-based access to the mapped SNP markers generated in this study, a database was created that can be accessed at http://bfgl.anri.barc.usda.gov/soybean/. Information includes descriptive data for each SNP-containing STS, STS information including primer sequences, and SNP positions in the STSs as well as the allele present in each of the six genotypes. The complete integrated map with positions of all SSR, RFLP, and SNP loci is also available.

TABLE 4

Existing linkage group lengths, number of SSRs, RFLPs, and other marker types including RAPD, AFLP, and classical and isozyme loci in the Song *et al.* (2004) consensus genetic linkage map and newly mapped genes

| Linkage group | Linkage group length and preexisting marker nos. | | | | | Newly mapped genes | | |
	Length (cM)	SSR	RFLP	Other marker types	Total markers	Observed	Predicted based on LG length	χ^2
A1	102.3	50	37	0	87	53	46	1.07
A2	165.7	64	46	6	116	78	75	0.12
B1	131.8	35	33	3	71	47	65	4.98
B2	125.0	36	42	10	88	47	51	0.31
C1	136.1	43	23	4	70	51	36	6.25
C2	157.9	53	44	3	100	53	70	4.13
D1a	120.9	53	37	11	101	45	52	0.94
D1b	138.0	58	19	4	81	57	69	2.09
D2	140.9	60	19	8	87	77	61	4.20
E	71.3	43	47	13	103	70	50	8.00*
F	151.4	64	41	8	113	72	69	0.13
G	126.0	63	53	13	129	68	53	4.25
H	124.0	38	41	5	84	48	57	1.42
I	120.9	40	32	4	76	54	59	0.42
J	91.2	50	43	5	98	74	44	20.45*
K	120.0	59	24	9	92	69	49	8.16*
L	117.0	52	43	4	99	41	53	2.72
M	146.3	51	24	3	78	39	64	9.77*
N	117.2	45	29	8	82	48	51	0.18
O	146.4	57	32	4	93	51	67	3.82
Total	2550.3	1014	709	125	1848	1141		

*Significant deviation ($P \leq 0.01$) from the predicted number of genes based on linkage group length.

In addition, data related to SNP detection using (1) the Sequenom MassARRAY system including the assay ID, PCR, and single-base extension primer sequences and (2) the Luminex flow cytometer including PCR and single-base extension primer sequences are available in the database.
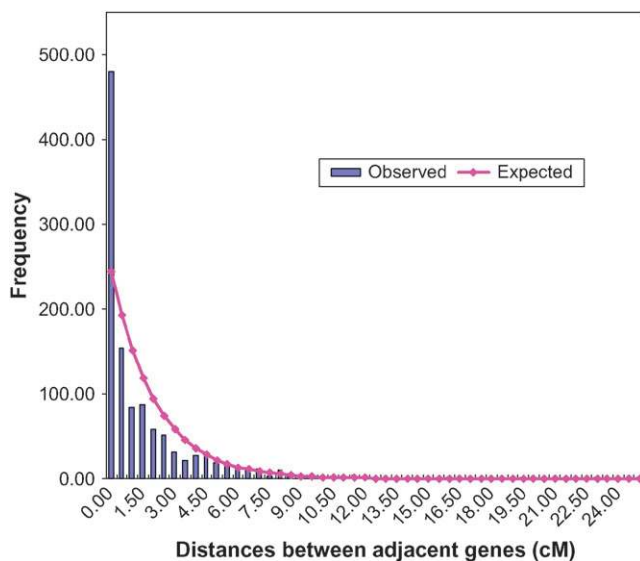

FIGURE 3.—The observed distribution of map distances between adjacent genic sequences within linkage groups and the theoretical distribution based on the assumption of random distribution of markers over the total length of the linkage map.

DISCUSSION

SNP discovery in soybean: Of a total of 9459 primer sets designed to EST sequence and examined for locus-specific PCR amplification followed by resequencing for SNP discovery, 21.5% were determined to contain a sequence variant. The low rate of SNP discovery reported here was, first, the result of the inability to develop robust STSs, and, second, the result of the low level of sequence variation in cultivated soybean. The difficulty of STS development stems from a number of factors. To facilitate specificity, the 872 primers designed to ESTs clustered using CAP3 frequently contained a PSV at their 3′ end. In some instances, the PSV bases may have been reverse transcriptase errors that would result in poor or absent PCR amplification. Reverse transcriptase errors have been documented to occur at frequencies of 1/1000–4000 bases in cDNA synthesis in Drosophila (STAPLETON *et al.* 2002). Primer design across intron–exon splice sites would also reduce the success of PCR amplification of primers designed to EST sequence. Likewise, the high level of duplication in the soybean genome complicated the development of STSs. More
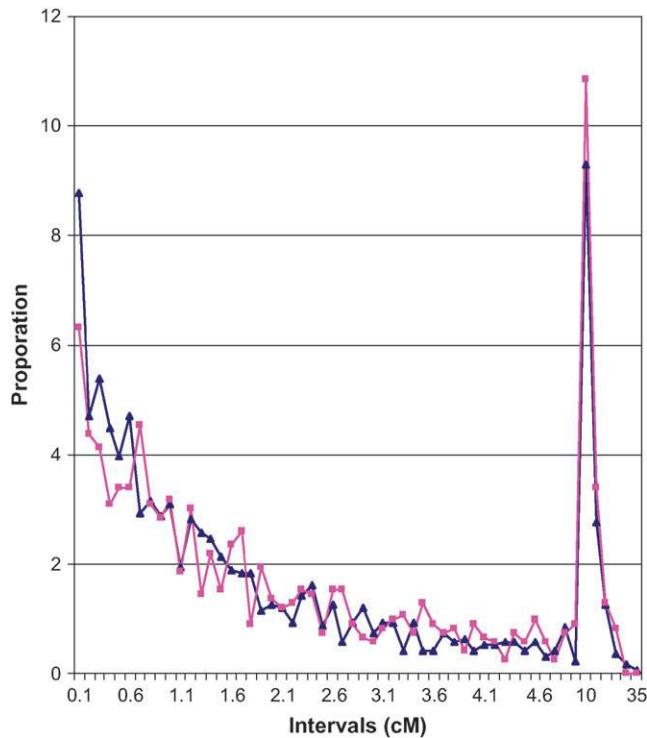
FIGURE 4.—Proportion of intervals between SSR loci (blue lines and triangles) and RFLP loci (purple lines and boxes) and the closest flanking genic sequences that fall within centimorgan distances of 0–0.1 cM, 0.1–0.2 cM, 0.2–0.3 cM, etc., to 35 cM.

than 20% of primers designed produced either multiple amplicons as determined by agarose gel analysis or single-band amplicons that yielded traces indicative of two or more amplicons. This would be anticipated in light of reports that, on average, a given chromosomal segment is duplicated 2.55 times in the soybean genome and that some segments are present as many as six times (SHOEMAKER *et al.* 1996).

The relatively low nucleotide diversity of cultivated soybean also contributed to the difficulty of SNP discovery. In the 2.44 Mbp of aligned sequence, nucleotide diversity ($\theta = 0.000997$) was similar to that reported by ZHU *et al.* (2003) ($\theta = 0.00086$) in 66.6 kbp of coding sequence and associated introns, untranslated regions, and perigenic sequence. The level of sequence diversity in cultivated soybean in relation to other cultivated species is relatively low. For example, in rice, FELTUS *et al.* (2004) reported 1.7 single-base changes plus 0.11 indels/kbp in 358 Mbp of low-copy DNA sequence on the basis of a comparison of draft sequences of the two rice subspecies *O. sativa* ssp. *indica* and *japonica*. This is equivalent to a nucleotide diversity of $\theta = 0.00181$. A calculation from KANAZIN *et al.* (2002) indicated nucleotide diversity of $\theta = 0.0025$ in 21.3 kbp of sequence analyzed in five diverse barley cultivars. Similarly, nucleotide diversity, $\theta = 0.0023$, in sorghum (*Sorghum bicolor*) (HAMBLIN *et al.* 2004) is more than twice that of

soybean. In modern maize (*Z. mays* L.) inbreds, WRIGHT *et al.* (2005) reported nucleotide diversity of $\theta = 0.00627$, which is more than sixfold that of soybean. A similarly high nucleotide diversity of $\theta = 0.0077$ was reported in a comparison of two sugar beet genotypes by SCHNEIDER *et al.* (2001). Along with low sequence diversity, the heterogeneity of nucleotide diversity across the 2.44 Mbp of sequence analyzed in this study also reduced the number of SNP-containing genic fragments discovered *vs.* what would have been anticipated if polymorphisms had been distributed evenly across the 4240 STSs. This heterogeneity is likely the product of selective sweeps that occurred during soybean domestication, which resulted in regions of the cultivated soybean genome in which little or no sequence variation is present. HYTEN *et al.* (2006) reported that while only 6.8% of genes assayed in 26 wild soybeans [*G. soja* (Sieb and Zucc.)] contained no sequence variants, 24.5% showed no variation in a set of 52 exotic *G. max* germ-plasm accessions. This suggests that nearly one-fifth of the cultivated soybean genome is genetically invariant as a result of selective sweeps associated with domestication.

The proportion of indels to total sequence variants in genic and perigenic sequence in soybean (15%) is quite similar to that reported in Arabidopsis. SCHMID *et al.* (2003) determined that 14% of sequence polymorphisms they detected via the resequencing of 12 genotypes were indels. In cultivated barley the comparable proportion is somewhat lower (8%) on the basis of an analysis of sequence polymorphisms in five diverse barley cultivars (KANAZIN *et al.* 2002). In contrast, the frequency of indels in and around maize genes greatly exceeds that reported to date for soybean and other plant species. BHATTRAMAKKI *et al.* (2002) discovered 655 indels via the sequence analysis of 502 genic loci (180,618 bp of aligned sequence) by the resequencing of eight diverse maize inbreds. This was an average of one indel every 276 bp. In this study, one indel was found every 2905 bp. While only six genotypes were analyzed in the case of soybean *vs.* eight in the work of BHATTRAMAKKI *et al.* (2002), it is clear that the indel frequency is substantially lower in soybean genic sequence than in maize.

Haplotype diversity: The mean number of haplotypes present among the 6 soybean genotypes in the 2032 SNP-containing STSs was 2.31, which was only a little less than the 2.74 haplotypes found in 96 SNP-containing genic fragments analyzed in 25 diverse soybean genotypes by ZHU *et al.* (2003). Thus, using the 6 genotypes identified by ZHU *et al.* (2003) as a subset that would identify 93% of the common SNPs (frequency ≥ 0.10) was apparently successful. The limited haplotype diversity in soybean is comparable to that reported in other crop species. For example, KILIAN *et al.* (2006) determined haplotype diversity among 20 domesticated barley lines in 7 barley genes and reported an average of 2.4 haplotypes per locus. Another report

from barley found an average of 2.81 haplotypes per STS via the analysis of 309 STSs with an average length of 466 bp in 7 cultivated and 1 wild barley (*Hordeum vulgare* ssp. *spontaneum*) genotype (ROSTOKS *et al.* 2005). CHING *et al.* (2002) analyzed fragments of 18 genes in 36 maize inbreds and found from 2 to 8 haplotypes with an average of only 4.4 haplotypes per gene. In cultivated soybean, the relatively limited haplotype diversity is suggested to be mainly a result of limited diversity in the wild progenitor compounded by a further loss of diversity as a result of domestication (HYTEN *et al.* 2006).

A more comprehensive soybean genetic map: The previous version of the soybean genetic map (SONG *et al.* 2004) included 1015 PCR-based markers (SSRs). The addition of 1141 markers based upon gene sequence provides the first transcript map of soybean (supplemental data file 1 at http://www.genetics.org/supplemental/). The diversity of gene function associated with these transcripts offers researchers an opportunity to identify potential candidate genes for >1150 QTL reported thus far in soybean for a variety of traits related to biotic and abiotic stresses, plant growth and morphology, and seed quality (SoyBase: http://soybase.agron.iastate.edu/). This initial transcript map and the additional markers it contains should enhance both applied and basic soybean genetics and genomics research, including QTL discovery, marker-assisted selection, map-based cloning, and the anchoring of the physical to the genetic map. Doubling of the number of genetically mapped sequence-based markers is a step forward in creating the resources that will be needed to assemble the whole-genome shotgun sequence of soybean (JOINT GENOME INSTITUTE 2006).

Distribution of genes and SSR and RFLP markers: Clustering of the mapped genic loci reported here was not unanticipated. MUDGE *et al.* (2004) hybridized RFLP probes derived from *Pst*I digestion of genomic DNA to arrayed BAC clones. Numerous cases of nonhomologous probes hybridizing to common BAC clones indicated that the gene-derived RFLPs were clustered in the genome. These authors concluded that most genes in soybean are clustered in ~275 Mbp of the genome, which is ~25% of the 1100-Mbp genome. Our analysis clearly supported gene clustering on the genetic map, although it would be difficult to draw conclusions about the proportion of the genome that is gene rich *vs.* gene poor. Such an analysis should be forthcoming as the soybean physical map progresses and the genome sequence becomes available.

An analysis of DNA sequence data from *A. thaliana*, rice, soybean, maize, and wheat (*Triticum aestivium*) (MORGANTE *et al.* 2002) concluded that the frequency of SSRs was significantly higher in transcribed regions and that microsatellites (SSRs) are associated with low-copy portions of plant genomes rather than with regions of repetitive DNA. In contrast, empirical evidence from soybean indicated that end sequences of BAC clones identified with *Pst*I-derived RFLP probes were reported to have 50% more gene-like sequences and 45% less repetitive sequence than end sequences of BAC clones identified with microsatellite markers (MAREK *et al.* 2001). These data suggested that in relation to SSRs, RFLPs tended to be more closely associated with gene-rich regions. Thus, the 138 gaps of >5 cM between adjacent PCR-based markers (SSRs) in the previous SSR-based genetic map of soybean (SONG *et al.* 2004) may include regions of interest to soybean genomicists and breeders because these gaps frequently contain one or more RFLP loci. On the basis of the transcript map developed via the positioning of 1141 genes on the preexisting SSR/RFLP map, SSR loci appear to be at least as closely associated with genic sequence as RFLP loci (Figure 4). Thus, whatever the relationship between genic sequence and SSR and RFLP markers, the gene-based SNPs mapped in this study effectively filled many of the 5- and 10-cM gaps in the previous map with at least one new PCR-based marker.

Access to SNP marker technology: The SNP detection used here was conducted using two SNP detection platforms. The Sequenom MassARRAY technology is well established and all the information required including redesigned PCR primers to amplify the SNP-containing fragment as well as the single-base extension primer for detection of the one SNP mapped per STS is available at http://bfgl.anri.barc.usda.gov/soybean/ (file: Sequenom Information.xls). Multiplex assays using the MassARRAY technology can be designed from this information and genotyping services are available commercially. The assays conducted on the Luminex flow cytometer also used single-base extension. Single-base extension primers for the detection of one SNP in each of 502 STSs are available at http://bfgl.anri.barc.usda.gov/soybean/ (file: Luminex Information.xls). The Luminex flow cytometer has been demonstrated to be flexible in that, in addition to single-base extension assays, hybridization-based as well as oligonucleotide ligation assays (OLA) can be used on this platform. A comparison of single-base extension, hybridization, and OLA assays for SNP detection in soybean was reported by LEE *et al.* (2004). The single-base extension assays were reported to be quite robust, but once optimized the hybridization system was actually more rapid and less costly per datapoint, making the latter assay more suitable for high-throughput marker-assisted selection. The 1141 gene fragments mapped in this study contained a total of 2928 SNPs. The data relating to the allele at each SNP locus in each STS will allow the user to design assays on the basis of an ever increasing number of SNP detection assay systems. These have been reviewed by SYVÄNEN (2001, 2005) and NG and LIU (2006). Existing systems plus the promise of new systems suggest an ever-improving SNP detection throughput coupled with decreasing cost per datapoint. While the newest technologies described by SYVÄNEN (2005) may

not be immediately available to soybean geneticists, a number of systems are currently available in addition to the Sequenom MassARRAY technology and the Luminex flow cytometer. An inexpensive and widely used alternative for the detection of a limited number of SNPs is based upon the alteration in a restriction endonuclease site by the presence of the SNP. These so-called cleaved amplified polymorphic sequences (CAPS) markers (GLAZEBROOK *et al.* 1998) have been successfully used for quite some time. Thus, if a QTL is discovered in a particular region of the genome, SNPs in that region can be identified, assayed, and converted to CAPS markers and used in marker-assisted selection. CAPS assays can be conducted on agarose gels that are available in many plant-breeding laboratories. Indeed, a CAPS marker linked to the important *ms2* gene for male sterility at the bottom of linkage group O was obtained in this way (J. M. CHAKY and J. E. SPECHT, unpublished results). As additional SNP markers are placed on the map, more potential SNPs will be available in any given region of the genome from which to design assays for marker-assisted selection.

LITERATURE CITED

BHATTRAMAKKI, D., M. DOLAN, M. HANAFEY, R. WINELAND, D. VASKE *et al.*, 2002   Insertion-deletion polymorphisms in 3′ regions of maize genes occur frequently and can be used as highly informative genetic markers. Plant Mol. Biol. **48:** 539–547.

BLANC, G., and K. H. WOLFE, 2004   Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. Plant Cell **16:** 1667–1678.

CHEN, J., M. A. IANNONE, M. S. LI, J. D. TAYLOR, P. RIVERS *et al.*, 2000   A microsphere-based assay for multiplexed single nucleotide polymorphism analysis using single base chain extension. Genome Res. **10:** 549–557.

CHING, A., K. S. CALDWELL, M. JUNG, M. DOLAN, O. S. SMITH *et al.*, 2002   SNP frequency, haplotype structure and linkage disequilibrium in elite maize inbred lines. BMC Genet. **3:** 19.

CONCIBIDO, V. C., R. L. DENNY, D. A. LANGE, J. H. ORF and N. D. YOUNG, 1996   RFLP mapping and marker-assisted selection of soybean cyst nematode resistance in PI 209332. Crop Sci. **36:** 1643–1650.

CREGAN, P. B., T. JARVIK, A. L. BUSH, R. C. SHOEMAKER, K. G. LARK *et al.*, 1999   An integrated genetic linkage map of the soybean genome. Crop Sci. **39:** 1464–1490.

EWING, B., and P. GREEN, 1998   Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res. **8:** 186–194.

FELTUS, F. A., J. WAN, S. R. SCHULZE, J. C. ESTILL, N. JIANG *et al.*, 2004   An SNP resource for rice genetics and breeding based on subspecies indica and japonica genome alignments. Genome Res. **14:** 1812–1819.

GIBBONS, J. D., 1976   *Nonparametric Methods for Quantitative Analysis.* Holt, Rinehart and Winston, Austin, TX.

GLAZEBROOK, J., E. DRENKARD, D. PREUSS and F. M. AUSUBEL, 1998   Use of cleaved amplified polymorphic sequences (CAPS) as genetic markers in Arabidopsis thaliana. Methods Mol. Biol. **82:** 173–182.

GORDON, D., C. ABAJIAN and P. GREEN, 1998   Consed: a graphical tool for sequence finishing. Genome Res. **8:** 195–202.

GRIFFIN, T. J., J. G. HALL, J. R. PRUDENT and L. M. SMITH, 1999   Direct genetic analysis by matrix-assisted laser desorption/ionization mass spectrometry. Proc. Natl. Acad. Sci. USA **96:** 6301–6306.

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.

HAMBLIN, M. T., S. E. MITCHELL, G. M. WHITE, J. GALLEGO, R. KUKATLA *et al.*, 2004   Comparative population genetics of the panicoid grasses: sequence polymorphism, linkage disequilibrium and selection in a diverse sample of sorghum bicolor. Genetics **167:** 471–483.

HYMOWITZ, T., 2004   Speciation and cytogenetics, pp. 97–136 in *Soybeans: Improvement, Production, and Uses*, edited by H. R. BOERMA and J. E. SPECHT. American Society of Agronomy, Crop Science Society of America, Soil Science Society of America, Madison, WI.

HYTEN, D. L., Q. SONG, Y. ZHU, I. Y. CHOI, R. L. NELSON *et al.*, 2006   Impacts of genetic bottlenecks on soybean genome diversity. Proc. Natl. Acad. Sci. USA **103:** 16666–16671.

JANDER, G., S. R. NORRIS, S. D. ROUNSLEY, D. F. BUSH, I. M. LEVIN *et al.*, 2002   Arabidopsis map-based cloning in the post-genome era. Plant Physiol. **129:** 440–450.

JOINT GENOME INSTITUTE, 2006   DOE JGI to decode DNA of soybean, a key alternative energy source (http://www.llnl.gov/PAO/news/news_releases/2006/NR-06-01-05.html).

KANAZIN, V., H. TALBERT, D. SEE, P. DECAMP, E. NEVO *et al.*, 2002   Discovery and assay of single-nucleotide polymorphisms in barley (Hordeum vulgare). Plant Mol. Biol. **48:** 529–537.

KEIM, P., T. C. OLSON and R. C. SHOEMAKER, 1988   A rapid protocol for isolating soybean DNA. Soybean Genet. Newsl. **15:** 150–152.

KILIAN, B., H. OZKAN, J. KOHL, A. VON HAESELER, F. BARALE *et al.*, 2006   Haplotype structure at seven barley genes: relevance to gene pool bottlenecks, phylogeny of ear type and site of barley domestication. Mol. Genet. Genomics **276:** 230–241.

LEE, S. H., D. R. WALKER, P. B. CREGAN and H. R. BOERMA, 2004   Comparison of four flow cytometric SNP detection assays and their use in plant improvement. Theor. Appl. Genet. **110:** 167–174.

MAREK, L. F., J. MUDGE, L. DARNIELLE, D. GRANT, N. HANSON *et al.*, 2001   Soybean genomic survey: BAC-end sequences near RFLP and SSR markers. Genome **44:** 572–581.

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.

MATUKUMALLI, L., J. GREFENSTETTE, D. HYTEN, I.-Y. CHOI, P. CREGAN *et al.*, 2006a   Application of machine learning in SNP discovery. BMC Bioinformtics **7:** 4.

MATUKUMALLI, L., J. GREFENSTETTE, D. HYTEN, I.-Y. CHOI, P. CREGAN *et al.*, 2006b   SNP-PHAGE - high throughput SNP discovery pipeline. BMC Bioinformatics **7:** 468.

MORGANTE, M., M. HANAFEY and W. POWELL, 2002   Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes. Nat. Genet. **30:** 194–200.

MUDGE, J., Y. HUIHUANG, R. L. DENNY, D. K. HOWE, D. DANESH *et al.*, 2004   Soybean bacterial artificial chromosome contigs anchored with RFLPs: insights into genome duplication and gene clustering. Genome **47:** 361–372.

NASU, S., J. SUZUKI, R. OHTA, K. HASEGAWA, R. YUI *et al.*, 2002   Search for and analysis of single nucleotide polymorphisms (SNPs) in rice (Oryza sativa, Oryza rufipogon) and establishment of SNP markers. DNA Res. **9:** 163–171.

NELSON, R. T., D. GRANT and R. C. SHOEMAKER, 2005   ESTminer: a suite of programs for gene and allele identification. Bioinformatics **21:** 691–693.

NG, J. K., and W. T. LIU, 2006   Miniaturized platforms for the detection of single-nucleotide polymorphisms. Anal. Bioanal. Chem. **386:** 427–434.

NORDBORG, M., T. T. HU, Y. ISHINO, J. JHAVERI, C. TOOMAJIAN *et al.*, 2005   The pattern of polymorphism in Arabidopsis thaliana. PLoS Biol. **3:** e196.

ROSTOKS, N., S. MUDIE, L. CARDLE, J. RUSSELL, L. RAMSAY *et al.*, 2005   Genome-wide SNP discovery and linkage analysis in barley

based on genes responsive to abiotic stress. Mol. Genet. Genomics **274:** 515–527.

SACHIDANANDAM, R., D. WEISSMAN, S. C. SCHMIDT, J. M. KAKOL, L. D. STEIN *et al.*, 2001 A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature **409:** 928–933.

SCHLUETER, J. A., P. DIXON, C. GRANGER, D. GRANT, L. CLARK *et al.*, 2004 Mining EST databases to resolve evolutionary events in major crop species. Genome **47:** 868–876.

SCHMID, K. J., T. R. SORENSEN, R. STRACKE, O. TORJEK, T. ALTMANN *et al.*, 2003 Large-scale identification and analysis of genomewide single-nucleotide polymorphisms for mapping in Arabidopsis thaliana. Genome Res. **13:** 1250–1257.

SCHMID, K. J., S. RAMOS-ONSINS, H. RINGYS-BECKSTEIN, B. WEISSHAAR and T. MITCHELL-OLDS, 2005 A multilocus sequence survey in *Arabidopsis thaliana* reveals a genomewide departure from a neutral model of DNA sequence polymorphism. Genetics **169:** 1601–1615.

SCHNEIDER, K., B. WEISSHAAR, D. C. BORCHARDT and F. SALAMINI, 2001 SNP frequency and allelic haplotype structure of Beta vulgaris expressed genes. Mol. Breed. **8:** 63–74.

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.

SONG, Q. J., L. F. MAREK, R. C. SHOEMAKER, K. G. LARK, V. C. CONCIBIDO *et al.*, 2004 A new integrated genetic linkage map of the soybean. Theor. Appl. Genet. **109:** 122–128.

STACEY, G., L. VODKIN, W. A. PARROTT and R. C. SHOEMAKER, 2004 National Science Foundation-sponsored workshop report. Draft plan for soybean genomics. Plant Physiol. **135:** 59–70.

STAPLETON, M., J. CARLSON, P. BROKSTEIN, C. YU, M. CHAMPE *et al.*, 2002 A Drosophila full-length cDNA resource. Genome Biol. **3:** RESEARCH0080.

SYVÄNEN, A. C., 2001 Accessing genetic variation: genotyping single nucleotide polymorphisms. Nat. Rev. Genet. **2:** 930–942.

SYVÄNEN, A. C., 2005 Toward genome-wide SNP genotyping. Nat. Genet. **37**(Suppl.): S5–S10.

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.

TUSKAN, G. A., S. DIFAZIO, S. JANSSON, J. BOHLMANN, I. GRIGORIEV *et al.*, 2006 The genome of black cottonwood, Populus trichocarpa (Torr. & Gray). Science **313:** 1596–1604.

VAN OOIJEN, J. W., and R. E. VOORRIPS, 2001 *JoinMap 3.0 Software for the Calculation of Genetic Linkage Maps.* Plant Research International, Wageningen, The Netherlands.

VODKIN, L. O., A. KHANNA, R. SHEALY, S. J. CLOUGH, D. O. GONZALEZ *et al.*, 2004 Microarrays for global expression constructed with a low redundancy set of 27,500 sequenced cDNAs representing an array of developmental stages and physiological conditions of the soybean plant. BMC Genomics **5:** 73.

WANG, D., J. SHI, S. R. CARLSON, P. B. CREGAN, R. W. WARD *et al.*, 2003 A low-cost, high-throughput polyacrylamide gel electrophoresis system for genotyping with microsatellite DNA markers. Crop Sci. **43:** 1828–1832.

WEIR, B. S., 1996 *Genetic Data Analysis II: Methods for Discrete Population Genetic Data.* Sinauer Associates, Sunderland, MA.

WRIGHT, S. I., I. V. BI, S. G. SCHROEDER, M. YAMASAKI, J. F. DOEBLEY *et al.*, 2005 The effects of artificial selection on the maize genome. Science **308:** 1310–1314.

ZHU, Y. L., Q. J. SONG, D. L. HYTEN, C. P. VAN TASSELL, L. K. MATUKUMALLI *et al.*, 2003 Single-nucleotide polymorphisms in soybean. Genetics **163:** 1123–1134.

Communicating editor: J. A. BIRCHLER