Theor Appl Genet (2004) 109:122–128
DOI 10.1007/s00122-004-1602-3

ORIGINAL PAPER

**Q. J. Song · L. F. Marek · R. C. Shoemaker · K. G. Lark · V. C. Concibido · X. Delannay · J. E. Specht · P. B. Cregan**

# A new integrated genetic linkage map of the soybean

Received: 30 October 2003 / Accepted: 8 January 2004 / Published online: 27 February 2004


**Abstract** A total of 391 simple sequence repeat (SSR) markers designed from genomic DNA libraries, 24 derived from existing GenBank genes or ESTs, and five derived from bacterial artificial chromosome (BAC) end sequences were developed. In contrast to SSRs derived from EST sequences, those derived from genomic libraries were a superior source of polymorphic markers, given that the mean number of tandem repeats in the former was significantly less than that of the latter ($P<0.01$). The 420 newly developed SSRs were mapped in one or more of five soybean mapping populations: 'Minsoy' × 'Noir 1', 'Minsoy' × 'Archer', 'Archer' × 'Noir 1', 'Clark' × 'Harosoy', and A81-356022 × PI468916. The JoinMap software package was used to combine the five maps into an integrated genetic map spanning 2,523.6 cM of Kosambi map distance across 20 linkage groups that contained 1,849 markers, including 1,015 SSRs, 709 RFLPs, 73 RAPDs, 24 classical traits, six AFLPs, ten isozymes, and 12 others. The number of new SSR markers added to each linkage group ranged from 12 to 29. In the integrated map, the ratio of SSR marker number to linkage group map distance did not differ among 18 of the 20 linkage groups; however, the SSRs were not uniformly spaced over a linkage group, clusters of SSRs with very limited recombination were frequently present. These clusters of SSRs may be indicative of gene-rich regions of soybean, as has been suggested by a number of recent studies, indicating the significant association of genes and SSRs. Development of SSR markers from map-referenced BAC clones was a very effective means of targeting markers to marker-scarce positions in the genome.

**Electronic Supplementary Material** Supplementary material is available in the online version of this article at http://dx.doi.org/10.1007/s00122-004-1602-3

Communicated by C. Möllers

Q. J. Song · P. B. Cregan (✉)
Soybean Genomics and Improvement Laboratory,
USDA-ARS, Beltsville, MD 20705, USA
e-mail: creganp@ba.ars.usda.gov
Tel.: +1-301-5045070
Fax: +1-301-5045728

L. F. Marek · R. C. Shoemaker
Department of Agronomy, USDA-ARS-CICG,
Iowa State University, Ames, IA 50011, USA

R. C. Shoemaker
Department of Agronomy,
Iowa State University, Ames, IA 50011, USA

K. G. Lark
Department of Biology,
University of Utah,
Salt Lake City, UT 84112, USA

V. C. Concibido · X. Delannay
Monsanto Company,
800 Lindbergh Boulevard, St. Louis, MO 63167, USA

J. E. Specht
Department of Agronomy, University of Nebraska,
Lincoln, NE 68583-0915, USA

## Introduction

The first soybean (*Glycine max* L. Merr.) genetic linkage map of molecular markers was reported by Keim et al. (1990). This map consisted of 26 genetic linkage groups containing a total of 150 restriction fragment length polymorphism (RFLP) loci and was based on a $F_2$ population derived from an interspecific cross of *G. max* (A81-356022) × *G. soja* (PI468916). Lark et al. (1993) subsequently used 132 RFLP, isozyme, and morphological markers to construct a soybean genetic map comprised of 31 linkage groups. Shoemaker and Specht (1995) mapped 110 RFLP, eight random amplified polymorphic DNA (RAPD), seven pigmentation, six morphological, and seven isozyme markers in an $F_2$ population derived from a mating of isolines of the important soybean cultivars 'Clark' and 'Harosoy'.

These early genetic maps were primarily based on RFLP markers. Due to the lack of polymorphism of RFLP loci in soybean and/or the complexity of multiple DNA banding patterns detected with most RFLP probes, simple sequence repeat (SSR) or microsatellite markers were proposed for map development (Akkaya et al. 1992). Most SSRs are single-locus markers, and many SSR loci are multi-allelic. These characteristics make SSRs an ideal marker system not only for creating genetic maps, but also as an unambiguous means of defining linkage group homology across mapping populations. In 1999, Cregan et al. (1999a) reported the development of 606 SSR loci which, together with 689 RFLP, 79 RAPD, 11 AFLP, ten isozyme, and 26 classical loci, were mapped to one or more of three populations: the USDA/Iowa State *G. max* × *G. soja* $F_2$, the University of Utah 'Minsoy' × 'Noir 1' recombinant inbred lines, and the University of Nebraska 'Clark' × 'Harosoy' $F_2$ population. These three separate maps provided useful information relative to the consistency of marker order and genetic distance among the different populations. The Cregan et al. (1999a) report established, for the first time, 20 consensus linkage groups, which were assumed to be the genetic correlates of the 20 soybean chromosomes. In that report, a total of 412 SSR loci were positioned in the 'Minsoy' × 'Noir 1' mapping population of 240 recombinant inbred lines. The resulting map was approximately 2,400 cM in length, but contained 36 intervals of at least 20 cM, and 79 intervals of at least 10 cM, in which no microsatellite loci were positioned. Inversely, there were 67 distinct intervals with less than 0.01 cM of distance between two or more adjacent SSR markers. In some of the 67 intervals, there was no recombination between adjacent SSR loci.

To develop microsatellite markers targeted to SSR-free regions as well as to saturate genomic regions of scientific interest, bacterial artificial chromosome (BAC) libraries can be screened by DNA hybridization or by PCR to identify clones from specific regions of the genome. New SSR or other DNA markers can be subsequently developed from those BAC clones, making it feasible to discover new SSRs associated with RFLP or other previously mapped markers. Employing this strategy, Cregan et al. (1999b) successfully developed new SSR markers targeted to two regions of the soybean genome near soybean cyst nematode-resistance loci on linkage groups G and A2. Genetic mapping confirmed that the new SSRs mapped to the correct sites in the genome.

Genetic markers are frequently polymorphic in one population, but monomorphic in another. JoinMap analysis (Stam 1993; Van Ooijen and Voorrips 2001) allows one to combine data from map populations in which not all markers are in common to obtain combined estimates of recombination. This approach not only increases the number of markers on the map, but also increases map precision and resolution.

In the early stages of microsatellite marker development, genomic DNA fragments containing SSRs were isolated from genomic libraries. More recently, EST sequencing projects have resulted in a wealth of sequence DNA information in numerous crop species including soybean. Some ESTs contain di- and trinucleotide-repeat motifs, making EST collections a potential source of microsatellite markers. The use of ESTs as a source of SSRs has been reported in a number of crop species including rice (Cho et al. 2000), grape (Scott et al. 2000), barley (Kota et al. 2001), sugarcane (Cordeiro et al. 2001), and wheat (Eujayl et al. 2002).

The objectives of the work reported here were: (1) to evaluate the potential of soybean ESTs as a source of SSRs for marker development; (2) to assess the success with which the development of SSR markers could be targeted to specific positions in the soybean genome; (3) to develop an additional set of SSR markers to further saturate the soybean linkage map; and (4) to create a consensus linkage map from five commonly used soybean populations using a JoinMap analysis. The creation of a high-density, integrated soybean linkage map with more precisely positioned markers would permit a better overall assessment of the distribution of SSR loci in the soybean genome. Moreover, the map would be useful for map-based cloning efforts and would provide a framework for the positioning of single nucleotide polymorphism (SNP)-based loci that are currently being developed from existing ESTs and other available sources of DNA sequence (Zhu et al. 2003).

## Materials and methods

### Sources of SSR-containing sequences

#### *Random genomic DNA*

The basic procedures of cloning and identification of microsatellite-containing, 500–700-bp genomic clones of 'Williams' soybean DNA were described previously (Cregan et al. 1994; Akkaya et al. 1995). Primer pairs were designed for the flanking regions of repeat motifs that consisted of either ten or more dinucleotide repeat units, or eight or more trinucleotide repeat units.

#### *Targeted SSR-marker development*

BAC clones putatively associated with specific positions in the soybean genome were identified either by hybridization of RFLP probes (Marek and Shoemaker 1997) or via PCR as suggested by Green and Olsen (1990). RFLP probes were used in an attempt to identify BAC clones at genome locations where RFLP loci, but no SSR loci, were present. Conversely, SSRs were used to identify BAC clones in an attempt to develop additional SSR markers targeted to a specific genomic location. The details relating to the use of BAC clones as a source of DNA for targeted-SSR development were described by Cregan et al. (1999b).

#### *SSR-containing repeats from ESTs*

Upon the initiation of this project (December 2000), 136,800 soybean ESTs were available in GenBank. These ESTs were screened to identify sequences containing ten or more dinucleotide SSRs or eight or more trinucleotide SSRs.

### Primer design and examination

PCR primers were designed to the flanking regions of microsatellites with ten or more dinucleotide and eight or more trinucleotide repeats using the software program Oligo 5.0 (National Biosciences, Plymouth, Minn.). Primers were synthesized by BioServe Biotechnologies (Laurel, Md.). Each primer pair was empirically tested for polymorphism using 'Clark', 'Harosoy', 'Jackson', 'Williams', 'Amsoy', 'Archer', 'Fiskeby', 'Minsoy', 'Noir 1', 'Tokyo', A81-356022 (*G. max*) and PI468916 (*G. Soja*) genomic DNA as templates. The first 10 of the above 12 genotypes were described by Cregan et al. (1999a). These ten genotypes represented a range of diversity within the cultivated soybean species. Primers designed from the ESTs were only tested on 'Minsoy', 'Noir 1', and 'Archer'. The $^{32}$P-labelled PCR products were analyzed on a 6% DNA sequencing gel with 30% formamide, followed by autoradiography.

### Mapping populations

Five widely used soybean mapping populations were used for microsatellite positioning; three of these, the USDA/Iowa State University A81-356022 × PI468916 (MS) population, the University of Nebraska, 'Clark' isoline × 'Harosoy' isoline (CH) population, and University of Utah 'Minsoy' × 'Noir 1' (MN) population, were previously described by Cregan et al. (1999a). The University of Utah 'Minsoy' × 'Archer' (MA) and 'Archer' × 'Noir 1' (AN) RIL populations were described by Mansur et al. (1995, 1996). Newly developed SSRs were mapped to MA, MN, and/or NA populations, and then JoinMap analysis was used on the five populations.

### DNA, isozyme, and classical genetic markers

A data set containing 1,019 SSR, 749 RFLP, 13 AFLP, 90 RAPD, ten isozyme, 24 classical, and 12 other markers that mapped in at least one of the five populations CH, MS, MA, MN, and/or AN was used for map integration.

### Statistical analysis

#### *Linkage map construction using JoinMap analysis*

Linkage maps of the five mapping populations were integrated based on the principle described by Stam (1993) using the JoinMap 3.0 (Van Ooijen and Voorrips 2001) program. The initial step involved calculating the LOD scores and pairwise recombination frequencies between markers. A LOD of 5.0 was used to create linkage groups in the MS, MA, and AN populations, whereas a LOD 4.0 was used in the MN and CH populations. The five maps of each linkage group were then integrated. Recombination values were converted to genetic distances using the Kosambi mapping function. The resulting 20 linkage groups were identified using the alphanumeric codes described in Cregan et al. (1999a).

#### *SSR marker distribution*

The theoretical distribution of map distance between adjacent SSR markers was estimated based on the assumption of random distribution of markers over the total length of the linkage map. The goodness of fit between the observed and theoretical distribution was tested using the Monte Carlo estimate of chi-square in Proc-StatXact 5 of SAS (Mehta and Patel 2002). The Monte Carlo estimate of the exact *P*-value was based on a Monte Carlo sample of size 10,000. To avoid the bias, markers developed from targeted isolation of BACs were excluded from this analysis.

## Results

### Development of SSR markers from EST and genomic DNA sequences

Dinucleotide and trinucleotide SSRs were identified in EST and BAC end sequences from GenBank, BAC subclones, and from clones of genomic libraries. The minimum length criteria were ten or more repeat units for dinucleotide repeats and eight or more for trinucleotide repeats. A total of 420 new SSR loci were developed to add to the 606 SSR loci published by Cregan et al. (1999a). Among these 420 SSRs, 24 were developed from EST sequences, five from GenBank BAC end sequences, 127 from DNA of BAC subclone libraries intended to target specific map positions, and 264 from genomic libraries. Primer pairs designed for sequences with an ATT/TAA, AT/TA, CT/GA, and various other repeat motifs, numbered 110, 276, 12, and 22, respectively.

Of the 136,800 soybean EST sequences examined, 75 contained dinucleotide repeats of ten or more, and 58 ESTs contained trinucleotide repeats of eight or more. The average percentage of ESTs containing the minimum number of repeats was thus less than 0.1%. Of the 133 primer sets designed for the EST-derived SSRs, just 24 (18.0%) amplified polymorphic products among the genotypes of 'Minsoy', 'Noir 1', and 'Archer' (Table 1). In contrast, over the course of several years of SSR-marker development in soybean, 824 (43%) primer sets designed for SSRs derived from genomic libraries were polymorphic among these three genotypes. This proportion was significantly higher than the observed polymorphism rate from the EST-derived primer sets ($t$=6.34, $P$<0.01). The mean length of di- and trinucleotide repeats was also significantly shorter ($t$=5.7, $P$<0.01 and $t$=9.3, $P$<0.01 for di- and trinucleotide repeats, respectively) in the EST-derived SSRs compared to the SSRs from genomic DNA sequences (Table 1).

**Table 1** Means and standard deviations (SD) of repeat numbers in simple sequence repeats (*SSRs*) obtained from either ESTs or from genomic DNA sequences

| Motif | Primers designed to SSRs from EST sequences | | | Primers designed to SSRs from genomic DNA sequences | | |
|---|---|---|---|---|---|---|
| | No. of primer pairs | Mean(±SD) repeat length | No. of polymorphic loci | No. of primer pairs | Mean(±SD) repeat length | No. of polymorphic loci |
| Dinucleotide | 75 | 18±6.7 | 14 (19%) | 693 | 24±7.4 | 283 (40%) |
| Trinucleotide | 58 | 11±2.8 | 10 (17%) | 1,211 | 16±5.9 | 541 (45%) |

**Table 2** Number of markers mapped to each linkage group and linkage group length in Kosambi mapping distance

| Linkage group | No. SSR | | No. RFLP | | No. RAPD | No. AFLP | Other | Total | Length (cM) |
|---|---|---|---|---|---|---|---|---|---|
| | Previously mapped | New | Previously mapped | New | | | | | |
| A1 | 27 | 23 | 36 | 1 | - | - | - | 87 | 102.3 |
| A2 | 37 | 27 | 44 | 2 | 2 | - | 4 | 116 | 165.7 |
| B1 | 19 | 16 | 32 | 1 | 2 | - | 1 | 71 | 131.8 |
| B2 | 24 | 12 | 38 | 4 | 6 | 2 | 2 | 88 | 120.9 |
| C1 | 21 | 22 | 19 | 4 | - | | 4 | 70 | 135.6 |
| C2 | 35 | 18 | 41 | 3 | 2 | - | 1 | 100 | 157.9 |
| D1a | 39 | 14 | 33 | 4 | 5 | - | 6 | 101 | 121.0 |
| D1b | 30 | 29 | 18 | 1 | 1 | 1 | 2 | 82 | 138.0 |
| D2 | 39 | 21 | 18 | 1 | 4 | 1 | 3 | 87 | 133.9 |
| E | 28 | 15 | 42 | 5 | 11 | - | 2 | 103 | 71.3 |
| F | 40 | 24 | 37 | 4 | 4 | 1 | 3 | 113 | 151.0 |
| G | 36 | 27 | 50 | 3 | 12 | - | 1 | 129 | 116.8 |
| H | 21 | 17 | 34 | 7 | 3 | - | 2 | 84 | 124.0 |
| I | 21 | 19 | 30 | 2 | 2 | - | 2 | 76 | 125.2 |
| J | 22 | 28 | 31 | 12 | 5 | - | - | 98 | 91.0 |
| K | 40 | 19 | 22 | 2 | 4 | 1 | 4 | 92 | 117.0 |
| L | 31 | 21 | 41 | 2 | 2 | - | 2 | 99 | 115.1 |
| M | 25 | 26 | 22 | 2 | 2 | - | 1 | 78 | 142.2 |
| N | 24 | 21 | 25 | 4 | 4 | - | 4 | 82 | 116.7 |
| O | 36 | 21 | 30 | 2 | 2 | - | 2 | 93 | 146.4 |
| Total | 595 | 420 | 643 | 66 | 73 | 6 | 46 | 1849 | 2,523.6 |

## Targeted SSR marker development

Of the 127 SSR markers developed from BAC subclone libraries, 91 originated from BAC clones identified by existing RFLP probes and 36 from BAC clones identified via existing SSR markers. However, only 36 of the 91 (39.6%) compared to 23 of 36 (64%) markers subsequently mapped to the genomic regions to which they were targeted.

## Mapping of the SSR markers

A JoinMap analysis of the 1,019 SSR, 749 RFLP, 13 AFLP, 90 RAPD, ten isozyme, and 30 other markers that segregated in at least one of the five populations produced a genetic map comprised of 20 consensus linkage groups that spanned 2,523.6 cM of Kosambi map distance. A total of 1,849 markers, including 1,015 SSRs, 709 RFLPs, 73 RAPDs, 24 classical traits, six AFLPs, ten isozymes, and 12 others, were integrated to form the current map (Table 2). Four SSRs remained unlinked, as did 40 of the RFLP loci. Among the 1,849 markers, a total of 420 SSR and 66 RFLP have been added to the map since the report by Cregan et al. (1999a). The numbers of SSRs mapped per linkage group averaged about 51, but varied from 35 to 64. The average length of the interval between any two adjacent SSR markers was 2.5 cM. The primer sequences for all SSR loci, as well as genetic maps of each of the 20 consensus linkage groups, are available on the SoyBase Web site of the USDA, ARS Soybean Genome Database (http://soybase.agron.iastate.edu/). Additional details can be found on the corresponding author's Web site http://bldg6.arsusda.gov/~pooley/soy/cregan/soy_map1.html.

## Distribution of SSR markers among and within linkage maps

The MN map presented in Cregan et al. (1999a) had 36 intervals of greater than 20 cM in which there was no SSR locus. With JoinMap integration and SSR markers from the other four populations, the gaps in the MN map were filled with 76 markers previously mapped by Cregan et al. (1999a) in the MS and CH populations. In the current study, 90 of the 420 new SSR loci developed either randomly or by targeting mapped to 30 of the 36 intervals. Six of the 36 intervals, C2 Satt202-Satt371; D1a Satt531-Satt368; D1b Satt542-Satt412; H Satt353-Satt192; I top-Satt571; and O Sat_109-Scaa001 still contain no new SSR markers. The number of markers mapped to the remaining 30 intervals varied from one to eight (Table 3).

Chi-square tests of the number of markers mapped to each linkage group indicated a significant deviation from that anticipated based upon linkage group length ($\chi^2$=36.7, $P$<0.05). However, this deviation was mainly due to fewer and greater numbers of new SSRs mapping to linkage groups B1 and G (Fig. 1). Indeed, a recalculation of the chi-square, with the G and B1 linkage groups excluded from the analysis, indicated similar SSR marker density among the 18 remaining linkage groups.

The randomness of SSR-marker distribution within linkage groups was also examined. Observed and theoretical distributions of map distances between adjacent SSR markers were not completely congruent; the Monte Carlo estimate of the exact $P$-value based on a Monte Carlo sample of size 10,000 is less than 0.01. As indicated in Fig. 2, there were large differences in the observed and expected frequencies of the cases in which adjacent SSR markers were separated by 0.5 cM or by 1.0 cM. The observed and the expected numbers were

**Table 3** Number of new SSR loci mapped to genomic intervals of at least 20 cM that contained no SSR markers in the soybean genetic map reported by Cregan et al. (1999a)

| Linkage group | Flanking SSR loci or linkage-group end | No. of previously existing markers positioned via JoinMap analysis | No. of new markers |
|---|---|---|---|
| A1 | Satt050-Satt385 | 0 | 4 |
| | Satt424-Sat_115 | 2 | 1 |
| B1 | Top-Satt509 | 1 | 4 |
| | Satt197-Satt298 | 2 | 2 |
| | Sat_123-Satt453 | 2 | 1 |
| B2 | Satt577-Satt126 | 2 | 2 |
| | Satt126-Sct_034 | 0 | 2 |
| | Satt534-Satt560 | 2 | 1 |
| C1 | Soygpatr-Satt578 | 1 | 4 |
| | Sat_042-Satt524 | 1 | 6 |
| C2 | Sat_130-Sat_062 | 4 | 3 |
| | Satt291-Satt170 | 4 | 2 |
| | Satt202-Satt371 | 1 | 0 |
| D1a | Satt531-Satt368 | 0 | 0 |
| | Sat_036-Satt071 | 2 | 2 |
| D1b | Sat_096-Satt095 | 1 | 3 |
| | Satt542-Satt412 | 3 | 0 |
| | Sat_069-Satt459 | 2 | 1 |
| D2 | Satt301-Sat_086 | 5 | 2 |
| E | Satt384-Satt598 | 9 | 7 |
| F | Satt522-Sat_074 | 1 | 3 |
| G | Satt288-Satt472 | 0 | 3 |
| H | Satt353-Satt192 | 3 | 0 |
| I | Top-Satt571 | 0 | 0 |
| J | Sct046-Satt456 | 6 | 8 |
| | Satt215-Satt244 | 3 | 3 |
| K | Sat_043-Satt475 | 1 | 1 |
| | Satt260-Sat_020 | 0 | 4 |
| L | Satt462-Satt481 | 0 | 4 |
| M | Satt150-Satt567 | 0 | 1 |
| N | Top-Satt159 | 1 | 3 |
| | Satt387-Satt521 | 1 | 1 |
| O | Satt445-Satt259 | 0 | 5 |
| | Satt347-Satt262 | 10 | 5 |
| | Satt123-Satt243 | 6 | 2 |
| | Sat_109-Scaa001 | 0 | 0 |

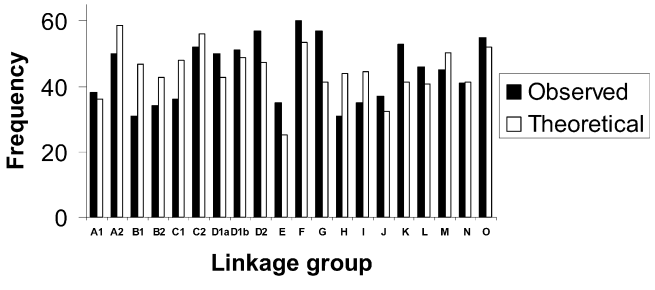


**Fig. 1** Observed and theoretical distribution of simple sequence repeat (SSR) markers in linkage groups based on the ratio of mapped SSR markers to linkage group length (cM)

187 versus 133 and 148 versus 113, respectively. These data indicated that more SSRs than average are closely linked, thus suggesting some degree of SSR-marker clustering.

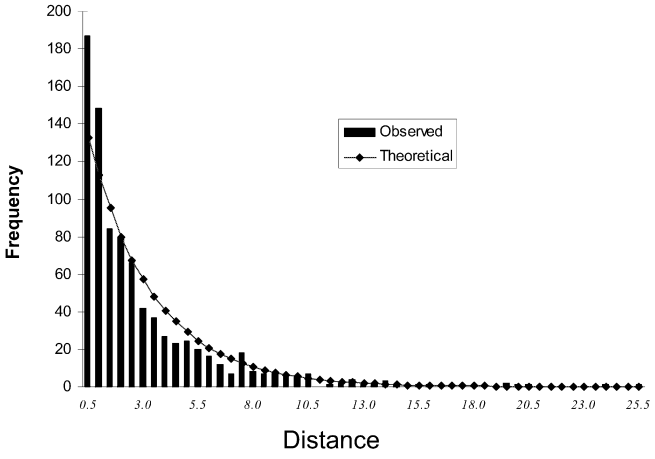


**Fig. 2** Theoretical and observed distribution of Kosambi map distance between adjacent SSR markers (summarized over all linkage groups)

## Discussion

We designed primers to 133 sequences with microsatellite repeats derived from ESTs, but only 24 (18.0%) of those primer sets produced useful polymorphic markers. In contrast, when genomic DNA sequences were used as the source of SSR-containing sequences, 43.0% yielded markers that were polymorphic with respect to the genotypes of 'Minsoy', 'Noir 1', and 'Archer'. Markers derived from genomic libraries also contained more repeat units as well as a greater range of allele sizes and genetic diversity than markers isolated from EST libraries. The striking difference of polymorphism between the soybean SSRs derived from the two sources is consistent with differences reported in rice (Temnykh et al. 1999; Cho et al. 2000), sugarcane (Cordeiro et al. 2001), tomato (Arshchenkova and Ganal 2002), wheat (Eujayl et al. 2002), and barley (Thiel et al. 2003). For example, Arshchenkova and Ganal (2002) reported that only 20 of 27,000 tomato ESTs contained microsatellites of more than ten repeat units. EST-derived microsatellites were generally shorter (7.3 repeat units) than genomic DNA-derived microsatellites (22.7 repeat units) in barley (Ramsay et al. 2000). The average number of repeats from EST-derived and genomic DNA-derived SSRs was 6.1 versus 13.7 in sugarcane (Cordeiro et al. 2001). The expansion or contraction of dinucleotide repeat length in exons may likely be suppressed due to the deterious nature of the frame-shift mutation that would frequently result in translated regions. Microsatellite markers derived from repeat arrays in genes are reported to be significantly less polymorphic than markers generated from longer arrays (Smulders et al. 1997). Other factors such as selection against large alteration in coding DNA or even a closely associated sequence that may play a role in gene expression could constrain microsatellite expansion or contraction. Such constraints could contribute to the reduced polymorphism of microsatellites in ESTs. To

**Table 4** Position of repeat motif in the sequenced genes from which polymorphic SSR markers were developed

| GenBank accession number | Description | Motif | Repeat position |
|---|---|---|---|
| AB002807 | *Glycine max* DNA for modulin 35 | (AT)14 | Boundary 5′ upstream sequences |
| AF162283 | *Glycine max* acetyl-CoA carboxylase (*accB-1*) gene | (CT)11 | 5′ Untranslated region |
| AF186183 | *Glycine max* retrovirus-like element Calypso2-1 | (ATT)22 | Boundary 5′ upstream sequences |
| X53404 | *Glycine max* glycin A (1a)B(1b) and A(2)B(1a) boundary DNA | (AT)25 | Intragenic |
| X17120 | Soybean actin *SAc7* gene | (CT)16 | Boundary 5′ upstream sequences |
| X16876 | Soybean *ENOD2B* gene | (AT)17 | Intragenic |
| X01425 | Soybean pseudogene for leghemoglobin | A18 | Intron |
| X07159 | Soybean pseudogene for heat shock protein Gmshp17.9-D (classVI) | (AT)9 | Boundary 5′ upstream sequences |
| V00458 | *Glycine max* gene encoding ribulose-1,5-bisphosphate carboxylase small subunit | A20 | Intron |
| X56139 | Soybean *ac514* gene for lipoxygenase | (AT)13 | Boundary 5′ upstream sequences |
| L23833 | Soybean glutamine phosphoribosylpyrophosphate amidotransferase | (CTT)6(CTT)4 | 5′ Untranslated region |
| M11317 | Soybean (*Glycine max*) low MVV heat shock protein gene (Gmhsp17.6-L) | (AT)15 | Boundary 5′ upstream sequences |
| V00452 | *Glycine max* leghemoglobin gene | (AT)26 | Intron |
| M94764 | *Glycine max* nodulin gene | (AT)24 | Intron |
| J02746 | *Glycine max SbPRP1* gene encoding a proline-rich protein | (ATT)20 | Boundary 5′ upstream sequences |

gain further understanding of the position of SSRs in and around functioning genes, the position of SSRs in the 15 genes from which we have developed polymorphic markers was determined (Table 4). In two instances, the SSR was located in 5′ UTR sequence, while in all others, they were located in either 5′ boundary sequence (seven cases), introns (four cases), or in intragenic sequence (two cases). This suggested that even when polymorphic SSRs were discovered in genic or perigenic regions, the SSR-repeat sequence changes occur only infrequently in mRNA. Obviously, EST-sequence data provide a convenient source of SSR-containing sequences that may be easily and inexpensively exploited. However, even in species with large EST collections, relatively few informative SSR loci are likely to result from this source.

Clustering of SSR markers on the soybean map was observed. Similar clustering of SSR markers was also reported in the tomato (Broun and Tanksley 1996; Areshchenkova and Ganal 1999) and rice linkage maps (McCouch et al. 2003). Physical clustering of SSR markers was also reported in the rat radiation hybrid map (Watanabe et al. 1999) and in barley (Cardle et al. 2000). Morgante et al. (2002) and Cardle et al. (2000) indicated that microsatellites are significantly associated with the low-copy fraction of plant genomes based on the estimation of microsatellite density in *Arabidopsis thaliana*, rice, soybean, maize, and wheat. Among these species, the overall frequency of microsatellites was inversely related to genome size and to the proportion of repetitive DNA. This suggests that most microsatellites reside in regions predating the recent genome expansions in many plants. In order to investigate the distribution of SSRs per megabase (Mb) on each of the 12 rice chromosomes, McCouch et al. (2003) divided the total number of SSRs mapped to each chromosome by the total length of genomic sequence available for each. The figures were compared to the number of EST clusters/Mb on each chromosome identified by The Institute for Genomic Research's *Oryza* gene index using the same genomic sequence data. The density of genes was approximately ten times the density of newly developed SSR markers, but there was a significant correlation ($r$=0.45, $P$<0.015) between the number of genes/Mb and the number SSRs/Mb at the level of the chromosome. The clustering of SSR loci we have observed in this report may correspond to gene-rich regions of soybean. However, as a result of the shorter length criteria used to define a microsatellite in studies such as Morgante et al. (2002) versus that used here, this conclusion remains tentative.

Thirty of the 36 intervals in the MN population described by Cregan et al. (1999a) that then contained no SSR marker now have at least one, and often several, SSR markers based on the results of our present study. In many instances, new markers were positioned in these intervals as a result of SSRs obtained from genomic clones, but 127 were developed from BAC clones with the intention of targeting specific genomic intervals. The proportion of SSRs that mapped to the linkage groups to which they were targeted was much higher (64%) when BAC clones were identified using PCR primers to SSR-flanking regions than when the BAC clones were identified with RFLP probes (39.6%). The higher efficiency of targeting when using SSR rather than RFLP probes is likely due to the greater specificity of PCR versus hybridization. The difference is also consistent with the report that RFLP probes hybridize, on average, to 2.55 map positions in the soybean genome (Shoemaker et al. 1996), and that the multiple fragments detected by RFLP frequently occur on different linkage groups (Keim et al. 1990). If only 1 of every 2.55 hybridizations per probe were to the targeted position in the genome, then approximately 39% of the BACs identified would be from the desired position in the genome. Thus, our results suggest that hybridization was about as successful as would be anticipated for the identification of a BAC clone from a specific position in the soybean genome. Although targeting was more

successful when SSRs were used to identify BAC clones, the duplicated nature of the genome still interfered with the efficiency of BAC clone identification despite the greater specificity of PCR. This is likely a reflection of the fact that at least some regions of the soybean genome share very high levels of sequence homology (Zhu et al. 1994; Shoemaker et al. 1996). This may make the development of locus-specific markers to these duplicated regions extremely difficult.

**Acknowledgements** The authors wish to thank Edward Fickus for excellent technical assistance. The financial support of the United Soybean Board (grant nos. 1243 and 2212) and the Monsanto Company is gratefully acknowledged.

## References

Akkaya MS, Bhagwat AA, Cregan PB (1992) Length polymorphisms of simple sequence repeat DNA in soybean. Genetics 132:1131–1139

Akkaya MS, Shoemaker RC, Specht JE, Bhagwat AA, Cregan PB (1995) Integration of simple sequence repeat (SSR) DNA markers into a soybean linkage map. Crop Sci 35:1439–1445.

Areshchenkova, T, Ganal MW (1999) Long tomato microsatellites are predominantly associated with centromeric regions. Genome 42:536–544

Areshchenkova T, Ganal MW (2002) Comparative analysis of polymorphism and chromosomal location of tomato microsatellite markers isolated from different sources. Theor Appl Genet 104:229–235

Broun P, Tanksley SD (1996) Characterization and genetic mapping of simple repeat sequences in the tomato genome. Mol Gen Genet 250:39–49

Cardle L, Ramsay L, Milbourne D, Macaulay M, Marshall D, Waugh (2000) Characterization of physically clustered simple sequence repeats in plants. Genetics 156:847–854

Cho YG, Ishii T, Temnykh S, Chen X, Lipovich L, McCouch SR, Park WD, Ayres N, Cartinhour S (2000) Diversity of microsatellites derived from genomic libraries and GenBank sequences in rice (*Oryza Sativa* L.). Theor Appl Genet 100:713–722

Cordeiro GM, Casu R, McIntyre CL, Manners JM, Henry RJ (2001) Microsatellite markers from sugarcane (Saccharum spp.) ESTs cross transferable to erianthus and sorghum. Plant Sci 160:1115–1123

Cregan PB, Bhagwat AA, Akkaya MS, Jiang R (1994) Microsatellite fingerprinting and mapping of soybean. Methods Mol Cell Biol 5:49–61

Cregan PB, Jarvik T, Bush AL, Shoemaker RC, Lark KG, Kahler AL, Kaya N, VanToai TT, Lohnes DG, Chung J, Specht JE (1999a) An integrated genetic linkage map of the soybean. Crop Sci 39:1464–1490

Cregan PB, Mudge J, Fickus ED, Marek LF, Danesh D, Denny R, Shoemaker RC, Matthews BF, Jarvik T, Young ND (1999b) Targeted isolation of simple sequence repeat markers through the use of bacterial artificial chromosomes. Theor Appl Genet 98:919–928

Eujayl I, Sorrells ME, Baum M, Wolters P (2002) Isolation of EST-derived microsatellite markers for genotyping the A and B genomes of wheat. Theor Appl Genet 104:399–407

Green ED, Olson MV (1990) Systematic screening of yeast artificial-chromosome libraries by use of the polymerase chain reaction. Proc Nat Acad Sci USA 87:1213–1217

Lark KG, Weisemann JM, Matthews BF, Palmer R, Chase K, Macalma T (1993) A genetic map of soybean (Glycine max L.) using an intraspecific cross of two cultivars: 'Minsoy' and 'Noir 1'. Theor Appl Genet 86:901–906

Kota R, Varshney RK, Thief T, Dehmer KJ, Graner A (2001) Generation and comparison of EST-derived SSRs and SNPs in barley (*Hordeum vulgare* L.). Hereditas 135:145–151

Keim P, Diers BW, Olson TC, Shoemaker RC (1990) RFLP mapping in soybean: association between marker loci and variation in quantitative traits. Genetics 126:735–742

Mansur LM, Orf J (1995) Evaluation of soybean recombinant inbreds for agronomic performance in northern USA and Chile. Crop Sci 35:422–425

Mansur LM, Orf JH, Chase K, Jarvik T, Cregan PB, Lark KG (1996) Genetic mapping of agronomic traits using recombinant inbred lines of soybean. Crop Sci 36:1327–1336

Marek LF, Shoemaker RC (1997) BAC contig development by fingerprint analysis in soybean. Genome 40:420–427

McCouch SR, Teytelman L, Xu Y, Lobos KB, Clare K, Walton M, Fu B, Maghirang R, Li Z, Xing Y, Zhang Q, Kono I, Yano M, Fjellstrom R, DeClerck G, Schneider D, Cartinhour S, Ware D, Stein L (2002) Development and mapping of 2240 new SSR markers for rice (*Oryza sativa* L.). DNA Res 9:257–279

Mehta C, Patel N (1997) Proc-StatXact for SAS users. Cytel, Cambridge, Mass.

Morgante M, Hanafey M, Powell W (2002) Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes. Nat Genet 30:194–200

Ooijen JW van , Voorrips RE (2001) JoinMap 3.0 software for the calculation of genetic linkage maps. Plant Research International, Wageningen, The Netherlands

Ramsay L, Macaulay M, Ivanissivich S, MacLean K, Cardle L, Fuller J, Edwards K, Tuvesson S, Morgante M, Massari A, Maestri E, Marniorlin N, Sjakste T, Ganal M, Powell W, Powell W, Waugh R (2000) A simple sequence repeat-based linkage map of barley. Genetics 156:1997–2005

Scott KD, Eggler P, Seaton G, Rosetto M, Ablett EM, Lee LS, Henry RJ (2000) Analysis of SSRs derived from grape ESTs. Theor Appl Genet 100:723–726

Shoemaker RC, Specht JE (1995) Integration of the soybean molecular and classical genetic linkage groups. Crop Sci 35:436–446

Shoemaker RC, Polzin K, Labate J, Specht J, Brummer EC, Olson T, Young N, Concibido V, Wilcox J, Tamulonis JP, Kochert G, Boerma HR (1996) Genome duplication in soybean (*Glycine* subgenus *soja*). Genetics 144:329–339

Smulders MJM, Bredemeijer G, Rus-Kortekaas W, Arens P, Vosman B (1997) Use of short microsatellites from database sequences to generate polymorphisms among *Lycopersicon esculentum* cultivars and accessions of other *Lycopersicon* species. Theor Appl Genet 94:264–272

Stam P (1993) Construction of integrated genetic linkage maps by means of a new computer package: JoinMap. Plant J 3:739–744

Temnykh S, Park W, Ayres N, Cartinhour S, Hauck N, Lipovich L, Cho YG, Ishii T, McCouch SR (1999) Mapping and genome organization of microsatellites in rice (*Oryza Sativa* L.). Theor Appl Genet 100:698–712

Thiel T, Michalek W, Varshney RK, Graner A (2003) Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (Hordeum vulgare L.). Theor Appl Genet 106:411–422

Watanabe TK, Bihoreau MT, McCarthy LC, Kiguwa SL, Hishigaki H, Tsuji A, Browne J, Yamasaki Y, Mizoguchi-Miyakita A, Oga K, Ono T, Okuno S, Kanemoto N, Takahashi E, Tomita K, Hayashi H, Adachi M, Webber C, Davis M, Kiel S, Knights C, Smith A, Critcher R, Miller J, James MR, et al (1999). A radiation hybrid map of the rat genome containing 5,255 markers. Nat Genet 22:27–36

Zhu T, Schupp JM, Oliphant A, Keim P (1994) Hypomethylated sequences: characterization of the duplicate soybean genome. Mol Gen Genet 244:638–645

Zhu YL, Song QJ, Hyten DL, Van Tassell CP, Matukumalli LK, Grimm DR, Hyatt SM, Fickus EW, Young ND, Cregan PB (2003) Single-nucleotide polymorphisms in soybean. Genetics 163:1123–1134