2010

Abundance of SSR Motifs and Development of Candidate Polymorphic SSR Markers (BARCSOYSSR_1.0) in Soybean

Qijian Song
USDA-ARS, Soybean Genomics and Improvement Lab., Beltsville, MD, qijian.song@ars.usda.gov

Gaofeng Jia
USDA-ARS, Soybean Genomics and Improvement Lab., Beltsville, MD

Youlin Zhu
Dep. of Bioscience and Biotechnology, Nanchang Univ., Nanchang, 330047, China

David Grant
USDA-ARS, Corn Insects and Crop Genetics Research, Ames, IA

Rex T. Nelson
USDA-ARS, Corn Insects and Crop Genetics Research, Ames, IA

See next page for additional authors

RESEARCH

Abundance of SSR Motifs and Development of Candidate Polymorphic SSR Markers (BARCSOYSSR_1.0) in Soybean

Qijian Song, Gaofeng Jia, Youlin Zhu, David Grant, Rex T. Nelson, Eun-Young Hwang, David L. Hyten, and Perry B. Cregan*

ABSTRACT

Simple sequence repeat (SSR) genetic markers, also referred to as microsatellites, function in map-based cloning and for marker-assisted selection in plant breeding. The objectives of this study were to determine the abundance of SSRs in the soybean genome and to develop and test soybean SSR markers to create a database of locus-specific markers with a high likelihood of polymorphism. A total of 210,990 SSRs with di-, tri-, and tetranucleotide repeats of five or more were identified in the soybean whole genome sequence (WGS) which included 61,458 SSRs consisting of repeat units of di- (≥10), tri- (≥8), and tetranucleotide (≥7). Among the 61,458 SSRs, (AT)n, (ATT)n and (AAAT)n were the most abundant motifs among di-, tri-, and tetranucleotide SSRs, respectively. After screening for a number of factors including locus-specificity using e-PCR, a soybean SSR database (BARCSOYSSR_1.0) with the genome position and primer sequences for 33,065 SSRs was created. To examine the likelihood that primers in the database would function to amplify locus-specific polymorphic products, 1034 primer sets were evaluated by amplifying DNAs of seven diverse *Glycine max* (L.) Merr. and one wild soybean (*Glycine soja* Siebold & Zucc.) genotypes. A total of 978 (94.6%) of the primer sets amplified a single polymerase chain reaction (PCR) product and 798 (77.2%) amplified polymorphic amplicons as determined by 4.5% agarose gel electrophoresis. The BARCSOYSSR1.0 SSR markers can be found in SoyBase (http://soybase.org; verified 21 June 2010) the USDA-ARS Soybean Genome Database.

Q. Song, G. Jia, E. Hwang, D.L. Hyten, and P.B. Cregan, USDA-ARS, Soybean Genomics and Improvement Lab., Beltsville, MD 20705; Q. Song, Dep. of Plant Science and Landscape Architecture, Univ. of Maryland, College Park, MD 20742; Y. Zhu, Dep. of Bioscience and Biotechnology, Nanchang Univ., Nanchang, 330047, China. D. Grant and R.T. Nelson, USDA-ARS, Corn Insects and Crop Genetics Research, Ames, IA 50011-1010 Received 19 Oct. 2009. *Corresponding author (Perry.Cregan@ars.usda.gov).

Abbreviations: EST, expressed sequence tag; PCR, polymerase chain reaction; SSR, simple sequence repeat; WGS, whole genome sequence.

Conventional procedures to develop simple sequence repeat (SSR) markers involve library construction and screening followed by DNA sequence analysis, which are time consuming and expensive. The efficiency of SSR marker development depends on the ease with which the identified repeats can be developed into informative markers. Previous research showed that the success rate of converting a SSR-containing sequence to a locus-specific informative marker varied from 20 to 30% in wheat (*Triticum aestivum* L.), (Roder et al., 1998; Song et al., 2002) and from 11 to 45% in soybean (Shoemaker et al., 2008; Shultz et al., 2007; Song et al., 2004). The low rate is the result of poor polymerase chain reaction (PCR) amplification, multiple amplicons, or the lack of polymorphism among a set of diverse genotypes or the parents of available mapping populations.

The level of polymorphism of SSRs is related to a number of factors including the repeat number and motif. Simple sequence repeat (SSR) polymorphism is positively correlated with number of repeat units (Kayser et al., 2004). Simple sequence repeats (SSRs) with greater repeat numbers are more polymorphic than

Published in Crop Sci. 50:1950–1960 (2010).
doi: 10.2135/cropsci2009.10.0607
Published online 1 July 2010.

those with smaller repeat numbers as reported in humans (Weber, 1990) and confirmed by studies in other organisms, including rice (Ellegren, 2004; Temnykh et al., 2001) and *Medicago truncatula* Gaertn. (Mun et al., 2006). Weber (1990) indicated that in humans, dinucleotide sequences with 10 or fewer repeats were unlikely to be polymorphic. Likewise, Edwards et al. (1991) indicated that in the human populations they investigated, approximately seven tri- or tetrameric repeats were necessary to produce a 50% probability of polymorphism.

The polymorphism rate of GC-rich dinucleotide repeats is low (Morgante et al., 2002; Temnykh et al., 2001; Zhang et al., 2004), as is the polymorphism rate of (CCG)n repeats (<30%) among trinucleotide motifs, whereas (ATT)n, (AAG)n, and (ACT)n have polymorphism rates of greater than 40% in rice (Zhang et al., 2007). To determine which trinucleotide motif(s) would be the most polymorphic and abundant source of trinucleotide SSR markers in wheat, Song et al. (2002) screened four genomic libraries of 'Chinese Spring' with nine trinucleotide probes. The results indicated that (ATT)n-containing SSRs provided the most abundant and polymorphic source of trinucleotide SSR markers in wheat. Further research also indicated that loci containing the (ATT)n motif had a significantly higher polymorphism rate than those with the dinucleotide (CT)n or (CA)n or the tetranucleotide (TAGA)n motifs (Song et al., 2005). It was suggested that GC-rich regions are relatively stable, thus resulting in less replication slippage (Schlotterer and Tautz, 1992) and/or that GC-rich motifs are commonly distributed in exons where polymorphisms occur less frequently (Morgante et al., 2002).

Expressed sequence tag (EST)-sequence data provide a convenient source of SSR-containing sequences. However, SSR markers isolated from EST libraries generally contain fewer repeat units as well as a smaller range of allele sizes than markers derived from genomic libraries. In soybean, of 6920 primer sets designed to EST sequences containing di-, tri-, or tetra-SSRs of total nucleotide length ≥13, only 680 SSR markers (<10%) were informative and useful for mapping and allelic screening (Hisano et al., 2007). A similarly low rate of polymorphism was observed by Song et al. (2004) in soybean. Of 133 primer sets designed to EST-derived SSRs with (AT)n and (ATT)n motifs, the estimated polymorphism rate was 18%. The striking difference of polymorphism between the SSRs derived from EST versus genomic sequence has also been reported in rice (Cho et al., 2000), sugarcane (Cordeiro et al., 2001), tomato (Areshchenkova and Ganal, 2002), wheat (Eujayl et al., 2002), and barley (Thiel et al., 2003). The expansion or contraction of dinucleotide repeat number in exons is likely suppressed due to the deleterious nature of the resulting frame-shift mutation. Other factors such as selection against an alteration in coding sequence likely constrain SSR expansion or contraction in genic sequence (Metzgar et al., 2000). Thus, even in species with large EST collections, the development of large numbers of informative SSR markers using the EST-derived SSRs has proven difficult.

The availability of the recently completed soybean whole genome sequence (WGS) (http://www.phytozome.net/index.php; verified 28 May 2010) (Schmutz et al., 2010) provides information not only for the identification of genes, the study of the evolution of the species, and genome structure but also is an ideal resource for the genome-wide identification of SSRs *in silico* and the development of locus-specific SSR markers. Thus, the objectives of this study were to determine the abundance of SSRs in the soybean genome and to use stringent screening and filtering to develop a database containing candidate polymorphic SSRs which have a high likelihood of serving as useful markers.

MATERIALS AND METHODS

Screening of SSRs in the Soybean Genome

The soybean WGS (Glyma1.01), which is publicly available at www.phytozome.net (verified 28 May 2010), was used as a source of SSRs with di-, tri-, and tetranucleotide repeats. The Perl script MISA (Thiel et al., 2003) was used to screen the soybean genome sequences for the desired SSRs with the parameter (maximum_difference_for_two_SSRs) setting of 100. Di-, tri-, and tetranucleotide SSRs with repeat number of five or greater were identified.

Position of Previously Developed Soybean SSRs in the Soybean Genome Sequence

Of 1122 SSR markers released by the USDA-ARS, Beltsville, MD (Cregan et al., 1999; Song et al., 2004), a total of 1015 loci have been mapped in various soybean populations and are widely used for quantitative trait locus (QTL) mapping, marker-assisted selection or association studies, etc. To position these loci in the Glyma1.01 soybean genome sequence, source sequences of the 1122 SSR loci were aligned using standalone Megablast software (http://www.ncbi.nlm.nih.gov/blast/megablast.shtml; verified 28 May 2010) to the Glyma1.01 soy sequence with W = 50, cutoff percentage of alignment = 98 and low complexity filtered, and then the primer sequences of the SSR loci were mapped to the genome sequence using the standalone software e-PCR (http://www.ncbi.nlm.nih.gov/sutils/e-pcr/; verified 21 June 2010). Simple sequence repeat (SSR) positions were definitively determined if both the source sequences and primer sequences of the SSRs aligned to the same region of the genome sequence with expected e-PCR amplicon length and with the expected SSR motif between the two primer sequences. Because the WGS was developed from the Williams 82 cultivar and the previously developed SSRs were from the cultivar Williams, which was the recurrent backcross parent used to create Williams 82 and is predicted to have its genome 99.7% identical by descent to Williams 82, high stringency of alignment (gap = 0, number of mismatch = 0) of e-PCR of primer sequences to the genome sequence was used to map the primer sequences to the soybean genome

sequence after eliminating GCG clamps added to the 3' end of some primer sets when the primers were originally designed.

Extraction and Filtering of SSRs from the Whole Genome Sequence

Based on our experience in the development of SSR markers in soybean (Cregan et al., 1999; Song et al., 2004) and wheat (Song et al., 2002, 2005) as well as the suggestions in previous reports (Edwards et al., 1991; Subramanian et al., 2003; Weber, 1990), we further screened the SSRs identified using MISA (Thiel et al., 2003). First, a total of 500 bases of sequence flanking each identified SSR was extracted from the Glyma1.01 soybean genome sequence. Only SSRs with perfect repeat units from 10 to 35, 8 to 35, and 7 to 35 for di-, tri-, and tetranucleotide SSR motifs, respectively, were retained. In the case of loci with compound repeats, only the sequences from loci with total compound SSR length less than 120 nucleotides were retained. Simple sequence repeat (SSR)-containing sequences were further filtered using the following procedures: The flanking sequences of extracted SSR-containing loci were aligned to the repetitive genome sequence (http://www.soymap.org/data/misc/soy_repeats.fasta; verified 28 May 2010), the mitochondrial sequence released with the soybean genome sequence Glyma1.01 (ftp://ftp.jgi-psf.org/pub/JGI_data/Glycine_max/Glyma1/assembly/sequences/; verified 28 May 2010) and the chloroplast sequence (NC_007942) at a stringency of W = 50, p = 95 using megablast. Any SSR-containing sequence aligned to these elements was eliminated.

During the development of SSR markers for soybean at the USDA-ARS, Beltsville, MD (Cregan et al., 1999; Song et al., 2004), primers were designed to a total of 2601 SSR-containing sequences, of which a total of 1479 sequences were not converted to informative markers as a result of poor PCR amplification (3.9%), the amplification of multiple products (84.4%), or the lack of polymorphism (11.7%) among a panel of diverse genotypes. These 1479 sequences were aligned to the extracted SSR-containing sequences from the WGS and the corresponding sequences were eliminated from the set of extracted SSR-containing sequences.

Assembly of Expressed Sequence Tag Sequence and Extraction of SSR-Containing Sequences from the Whole Genome Sequence that Aligned to the SSR-Containing Expressed Sequence Tags

In July 2009, a total of 1490,273 soybean ESTs were available in GenBank. The sequences were downloaded and then assembled using the software CAP3 (Huang and Madan, 1999) with the default parameters (except for the cutoff percentage of alignments which was increased to 95%). Assembled EST sequences containing SSRs with di-, tri-, and tetranucleotide repeat units $\geq$5 were retrieved and analyzed with blastn to the extracted SSR-containing genome sequences (–e 20 and low complexity filtered). Simple sequence repeat (SSR)-containing sequences from the WGS that aligned to these SSR-containing ESTs were kept and so indicated in the database. Although the SSRs in the ESTs may have a lower probability of polymorphism versus SSRs of similar length in genomic sequence (Areshchenkova and Ganal, 2002; Cho et al., 2000; Cordeiro et al., 2001; Eujayl et al., 2002; Thiel et al., 2003), these SSRs were nonetheless retained.

Design of Primer Sets Flanking Extracted and Filtered SSRs and *in-silico* PCR for the Selection of Primer Sets with Unique Genome Positions

Primers were designed to the final set of extracted SSR-containing sequences using standalone Primer 3 (http://primer3.sourceforge.net/releases.php; verified 28 May 2010). The targeted PCR product length ranged from 80 to 300 base pairs, the annealing temperature from 55 to 62°C and the primer length from 18 to 27 nucleotides.

To reduce the likelihood of nonspecific annealing of primers, all primer sequences were examined with e-PCR software with the parameters of maximum number of mismatching nucleotides in one primer sequence of N = 3 and maximum number of gaps of one primer sequence of G = 1. Primer sets that were not uniquely aligned to the genome sequence under this stringency were discarded and reselected. If four sets of primer pairs designed to flank the same SSR locus failed, the corresponding SSR-containing sequence was discarded. Those extracted SSR-containing sequences to which primers were designed which met the e-PCR criteria were included in the BARCSOYSSR_1.0 database.

Statistics and Analysis

The average and the standard deviation of SSR repeat unit number of each motif were calculated as:

$$\bar{x} = \sum f_i X_i;\ \text{sd} = \sqrt{\left(\sum f_i X_i^2 - \left(\sum f_i X_i\right)^2 / n\right) / \left(\sum f_i - 1\right)}$$

where $\bar{x}$ and sd are mean and standard deviation, respectively, f_i is the frequency of the motif with the ith repeat unit, X_i is the number of the ith repeat unit; n is the total number of SSR occurrences for the motif. A Chi-squared test was used to test the difference of the densities of di- and trinucleotide SSRs among the 20 chromosomes based on the assumption that the numbers of SSRs are directly proportional to the total length of the chromosome:

$$\chi^2 = \sum (O_i - E_i)^2 / E_i,$$

where O_i is the number of SSRs observed in each chromosome, E_i is the number of SSRs expected based on the chromosome length, i is the ith chromosome.

The Polymorphism Rate of the SSRs with Various Motifs in Previous SSR Development

Of the 2601 primer sets previously designed to SSR-containing sequences by the USDA-ARS, Beltsville, MD, in the process of SSR development (Cregan et al., 1999; Song et al., 2004), a total of 2196 primer sets were designed to SSRs with various perfect repeats (i.e., with one motif only). The polymorphism rates of these primer sets were calculated and the overall rate was used as a reference for comparison with the polymorphism rate based on BARCSOYSSR_1.0 database. In addition, the polymorphism rates of the SSRs with various motifs were also compared.

Evaluation of PCR Amplification and Polymorphism of SSRs Using Primer Sets Randomly Selected from the BARCSOYSSR_1.0

To provide an estimate of the value of the SSRs in BARCSOYSSR_1.0, a total of 413 primer sets were randomly selected for testing and an additional 621 primer sets designed to SSRs residing in the intervals and at the ends of chromosomes with more than 1 Mb of sequence without a previously mapped SSR were randomly selected from the BARCSOYSSR_1.0 database. Genomic DNA from eight diverse soybean genotypes, including Williams 82, 'Noir 1', 'Minsoy', 'Archer', 'Evans', 'Peking', 'Essex' and the wild soybean (*Glycine soja* Siebold & Zucc.), PI 468916, were used as template for PCR amplifications. These genotypes are the parents of commonly used soybean mapping populations in the United States. The length polymorphisms of the amplicons from the selected primer sets were evaluated among these genotypes. Polymerase chain reaction (PCR) mixes contained 30 ng of genomic DNA, 0.15 μM of forward and reverse primers, 200 μM of each nucleotide, 1X PCR Buffer containing 50 mM KCl, 10 mM Tris HCl pH 9.0, 0.1% Triton X 100, and *Taq* DNA polymerase in a total volume of 10 μL. The reaction mixes were heated for 3 min at 94°C before PCR cycling, the PCR cycling consisted of 40 sec denaturation at 94°C, 30 sec annealing at 55 to 62°C, and 40 sec extension at 72°C for 35 cycles, followed by an additional 7 min extension at 72°C. Polymerase chain reaction (PCR) products were analyzed on a 4.5% high resolution agarose gel (Agarose SFR, Amresco, Vernon Hills, IL).

RESULTS

The Abundance of SSRs in the Soybean Genome

A total of 210,990 SSRs with di-, tri-, and tetranucleotide repeat units equal to or greater than 5 were identified in the soybean genome (Table 1). Among these, 168,625, 38,411 and 3954 were di-, tri-, and tetranucleotide SSRs, respectively. Of the dinucleotide motifs, (AT)n is the most abundant (58.5%), followed by (AG)n (24.2%) and (AC)n (17.2%). The (GC)n motif is the least frequent (0.1%) dinucleotide in the genome. Of the trinucleotide motifs, (ATT)n is the most abundant (44.8%), followed by (AAG)n (23.6%), (AAC)n (13.1%), (AGG)n (4.1%), (ACT)n (3.9%), (AGT)n (3.8%), (ACC)n (3.8%), (ACG)n (1.0%), (AGC)n (1.0%), and (CCG)n (0.9%). Of the tetranucleotide SSRs, (AAAT)n is the most abundant (52.3%), followed by (ACAT)n (13.7%), (AAAG)n (7.9%), (AATT)n (7.7%), (AGAT)n (4.1%), (AAAC)n (2.8%), (AATC)n (2.3%), (AACT)n (1.9%), and (ACTC)n (1.9%). The remaining tetranucleotide motifs were either absent (such as motifs with only G or C) or were present in less than 1% of the total.

A total of 30,057 SSR-containing sequences had two or more adjacent repeat motifs. Among these compound repeats, the most common were (CA)n(AT)n (45.3%), (CT)n(AT)n (24.8%), (CT)n(CA)n (6.0%) and (AT)n(ATT)n (2.3%). The frequency of other types of compound repeats was less than 2.0%. The frequency of two different motifs adjacent to each other in the compound SSRs was not random ($p < 0.01$) and deviated from their expected probabilities based on the frequencies of two individual repeat motifs in the soybean genome.

Among dinucleotide SSRs, the average repeat number of (AT)n SSRs was 13.6, followed by (AC)n (7.9), (AG)n (6.8), and (CG)n (6.6). Among the trinucleotide SSRs, the average number of (ATT)n repeat units was 8.8 while the remaining trinucleotide SSRs had less than 6.4 repeat units. The average repeat units among the tetranucleotide SSRs were less than that among the di- and trinucleotide SSRs. The (TGAT)n motif with an average of 6.6 repeats had the greatest average number of repeat units among the tetranucleotide SSRs.

A Chi-square test showed that the densities of di- and trinucleotide SSRs were similar among the 20 chromosomes (Table 2), however, the density of the tetranucleotide SSRs among the chromosomes differed significantly. The average SSR density in the genome was 222 Mbp^{-1} and di-, tri-, and tetranucleotide SSRs densities were 178 Mbp^{-1}, 40 Mbp^{-1}, and 4 Mbp^{-1}, respectively.Of the 210,990 SSRs identified in the genome sequence, 61,458 contained repeat units equal to or greater than 10, 8, and 7 for di-, tri-, and tetranucleotide repeats, respectively. Due to the difficulty of PCR amplification of sequences containing very long repeats, SSRs with repeat numbers above 35 were eliminated. After eliminating these sequences, there remained a total of 50,586 SSRs with repeat units of 10 to 35, 8 to 35, and 7 to 35 for di-, tri-, and tetranucleotide SSRs, respectively or with compound repeats of a total length of $\leq$120 nt.

Differences of Frequencies and Numbers of Repeat Units of SSRs from Expressed Sequence Tags versus Genomic Sequences in Soybean

The 1,490,273 EST sequences were assembled to 37,890 contigs and 396,480 singletons. A total of 33,237 SSRs were identified from the assembled EST sequences. Of these, a total of 20,161, 12,440, and 636 contained di-tri-, and tetranucleotide repeats with repeat units $\geq$5, respectively. The most common motifs in ESTs were (AG)n (11,689), (AT)n (5076), (AAG)n (3115), (AC)n (3305), (AAC)n (1606), (ATT)n (1493), (ACC)n (1489), and (CCG)n (1225) (Table 3). Except in the case of the (AG)n, (AAG)n, and (ACC)n motifs, the average repeat unit numbers of the remaining five motifs were significantly lower in ESTs than in genomic sequences ($p < 0.01$) and the standard deviation of all motifs but (AG)n and (ACC)n were significantly lower in ESTs than in genomic sequences ($p < 0.01$). The percentage of SSRs with units

Table 1. Numbers of simple sequence repeats (SSRs) with di-, tri-, and tetranucleotide motifs in different repeat unit classes, the total number of SSRs with each motif, the standard deviation and mean of SSR repeat units and the number of di-, tri-, and tetranucleotide SSRs with the number of repeat units equal to or greater than 10, 8, and 7, respectively in the whole soybean genome sequence (Glyma1.01).

	No. of repeats												
Motifs	5_7	8_10	11_15	16_20	21_25	26_30	31_35	36_40	> 40	Total	SD of SSR repeat length	Avg. repeat unit	No. of SSRs with units of di-, tri-, and tetranucleotide ≥10, 8, and 7, respectively
(AC)n	20,567	4818	1888	631	325	232	153	94	266	28,974	6.59	7.9	4549
(AG)n	32,280	5037	2149	760	254	84	44	33	100	40,741	4.22	6.8	4439
(AT)n	46,896	11,079	9819	7575	7240	6208	4645	2657	2565	98,684	11.74	13.6	43,878
(CG)n	191	13	14	4	4	0	0	0	0	226	3.23	6.6	25
(AAC)n	4420	500	96	19	6	0	0	0	0	5041	1.72	6	623
(AAG)n	8239	562	172	46	17	6	5	2	4	9053	2.25	5.8	818
(AAT)n	11,206	2014	1867	1214	549	222	73	33	48	17,226	6.37	8.8	6022
(ACC)n	1358	88	8	1	0	0	0	0	0	1455	1.10	5.6	97
(ACG)n	396	15	1	0	0	0	0	0	0	412	0.87	5.5	16
(ACT)n	1370	81	17	14	6	2	1	1	1	1493	2.66	5.9	123
(AGC)n	338	25	1	0	0	0	0	0	0	364	1.06	5.6	27
(AGG)n	1297	206	54	7	0	0	0	0	0	1564	1.78	6.3	267
(AGT)n	1337	95	22	3	2	0	1	0	1	1461	1.96	5.8	124
(CCG)n	332	9	1	0	0	0	0	0	0	342	0.97	5.6	11
(AAAC)n	108	2	0	0	0	0	0	0	0	110	0.58	5.2	4
(AAAG)n	299	10	2	0	0	0	0	0	0	311	0.95	5.4	29
(AAAT)n	2030	34	4	0	0	0	0	0	0	2068	0.72	5.3	106
(AACC)n	11		0	0	0	0	0	0	0	11	0.30	5.1	
(AACG)n	2		0	0	0	0	0	0	0	2	0.00	5	
(AACT)n	81	1	0	0	0	0	0	0	0	82	0.66	5.4	6
(AAGC)n	3		0	0	0	0	0	0	0	3	0.00	5	
(AAGG)n	27		0	0	0	0	0	0	0	27	0.36	5.1	
(AAGT)n	37		0	0	0	0	0	0	0	37	0.46	5.2	1
(AATC)n	80	4	0	0	0	0	0	0	0	84	0.81	5.4	6
(AATG)n	28		0	0	0	0	0	0	0	28	0.59	5.3	2
(AATT)n	299	4	1	0	0	0	0	0	0	304	0.75	5.3	21
(ACAG)n	7		0	0	0	0	0	0	0	7	0.38	5.1	
(ACAT)n	419	88	28	3	0	0	0	0	0	538	2.17	6.5	180
(ACCC)n	11		0	0	0	0	0	0	0	11	0.60	5.2	1
(ACCT)n	7		0	0	0	0	0	0	0	7	0.49	5.3	
(ACGC)n	13	2	3	0	0	0	0	0	0	18	2.46	6.9	9
(ACGT)n	16	1	0	0	0	0	0	0	0	17	0.75	5.2	1
(ACTC)n	68	6	0	0	0	0	0	0	0	74	1.04	5.6	13
(AGAT)n	125	25	11	1	0	0	0	0	0	162	2.27	6.6	59
(AGCG)n	2		0	0	0	0	0	0	0	2	0.00	5	
(AGCT)n	12		0	0	0	0	0	0	0	12	0.29	5.1	
(AGGC)n	1		0	0	0	0	0	0	0	1		5	
(AGGG)n	35		0	0	0	0	0	0	0	35	0.55	5.4	1
(AGGT)n	3		0	0	0	0	0	0	0	3	0.58	5.3	

of di-, tri-, and tetranucleotide SSR repeats ≥10, 8, and 7, respectively, from the EST sequences was 13.4% (4441 of 33,237). This is in contrast to 29.10% (61,458 of the 210,990) calculated based on SSRs in genomic sequence.

The 4441 EST sequences were aligned to the 50,586 SSRs with repeat lengths in a range that would allow a reasonable probability of conversion to a useful SSR marker (10 to 35, 8 to 35, and 7 to 35 for di-, tri-, and tetranucleotide SSR motifs, respectively) which had been identified in the WGS. A total of 4326 of the 4441 EST-derived SSRs aligned to the SSRs with desirable repeat lengths identified in the WGS. The failure of all 4441 EST-derived SSRs to be present among the 50,586 SSRs from the Williams 82 WGS was a result of a small number of EST-derived SSRs with desirable repeat lengths in the genotype used as a source of mRNA at loci at which Williams 82 did not contain an SSR which met the repeat unit length requirement.

Table 2. Numbers, frequencies and Chi-square tests of the density of total simple sequence repeats (SSRs) and di-, tri-, and tetranucleotide SSRs with repeat unit numbers greater than five on each of the 20 soybean chromosomes.

		SSR numbers				SSR Frequency				Chi-square for SSR Frequency			
Chromosome	Size (bp)	Total SSRs	Dinucleotide SSRs	Trinucleotide SSRs	Tetranucleotide SSRs	Total SSRs Mbp^{-1}	Dinucleotide SSRs Mbp^{-1}	Trinucleotide SSRs Mbp^{-1}	Tetranucleotide SSRs Mbp^{-1}	Total SSRs	Dinucleotide SSRs	Trinucleotide SSRs	Tetranucleotide SSRs
Gm1	55,915,595	11,552	9193	2148	211	207	164	38	4	0.1	0.2	0.7	6.5
Gm2	51,656,713	11,611	9344	2041	226	225	181	40	4	0	0.1	0.2	2.3
Gm3	47781,076	10,261	8129	1944	188	215	170	41	4	0	0.1	0.3	2.6
Gm4	49,243,852	10,516	8502	1828	186	214	173	37	4	0.1	0.1	0.3	3.2
Gm5	41,936,504	8963	7208	1586	169	214	172	38	4	0.1	0.1	0.3	3.1
Gm6	50,722,821	11,205	8913	2073	219	221	176	41	4	0	0	0	0.4
Gm7	44,683,157	10,138	8115	1837	186	227	182	41	4	0.1	0.1	0.3	3.3
Gm8	46,995,532	10,985	8810	1973	202	234	187	42	4	0.1	0.2	0.7	6.6
Gm9	46,843,750	10,547	8376	1976	195	225	179	42	4	0	0.1	0.3	2.4
Gm10	50,969,635	11,788	9380	2162	246	231	184	42	4	0.1	0.1	0.6	5.4
Gm11	39,172,790	8932	7149	1606	177	228	182	41	5	0.1	0.1	0.4	3.8
Gm12	40,113,140	8748	6982	1609	157	218	174	40	4	0	0	0.1	1
Gm13	44,408,971	10,548	8530	1827	191	238	192	41	4	0.2	0.2	0.9	8.4
Gm14	49,711,204	10,604	8430	1968	206	213	170	40	4	0.1	0.1	0.3	3.3
Gm15	50,939,160	10,633	8469	1959	205	209	166	38	4	0.1	0.1	0.6	5.5
Gm16	37,397,385	8367	6695	1525	147	224	179	41	4	0	0	0.2	1.8
Gm17	41,906,774	9791	7767	1866	158	234	185	45	4	0.1	0.2	0.7	6.6
Gm18	62,308,140	13,091	10,496	2309	286	210	168	37	5	0.1	0.1	0.5	4.8
Gm19	50,589,441	10,835	8666	1972	197	214	171	39	4	0.1	0.1	0.3	2.9
Gm20	46,773,167	10,006	7907	1928	171	214	169	41	4	0.1	0.1	0.3	3
Total	9.50E+08	209,121	167,061	38,137	3923					1.5	2.1	7.9	76.9**

**Indicates a significant deviation at the 0.01 level of the density of SSRs on an individual chromosome from the average density in the entire genome.

Screening of SSR Loci and the Development of the BARCSOYSSR_1.0 SSR Database

The alignment of the 50,586 SSR-containing sequences with the SSRs from non-nuclear sequences and the SSR and primer sequences of previously examined SSR loci which had failed to convert to robust sequence tagged site (STS) indicated that a total of 9748 SSR loci were aligned with these elements. These loci were excluded from primer selection.

Of the primers sets designed to the remaining 40,838 loci, a total of 33,065 primer sets (81%) were identified as being locus-specific after e-PCR. The retention rate of the SSR-containing loci after stringent filtering was 65% (33,065/50,586). Proportions of loci with primers designed to perfect (AT)n, (ATT)n, (AG)n, and (AC)n motifs in the current database were 0.78, 0.09, 0.06, and 0.03, respectively (Supplementary Table 1).

Anchoring Previously Mapped or Released SSRs to the Soybean Genome Sequence

Of the 1122 SSR markers released by Cregan et al. (1999) and Song et al. (2004), a total of 874 were positioned on the Glyma1.01 genome sequence (Table 4, Supplementary Table 1). Similar to the distribution of previously mapped SSRs on the genetic map, these SSR markers are not uniformly distributed within chromosomes based on their position on the Glyma1.01 genome sequence. We observed 287 intervals and chromosome ends each spanning from 1.0 to 15.0 Mbp with no previously mapped SSRs. The number of such intervals or ends varied from 7 to 21 among the 20 soybean chromosomes. In the current BARCSOYSSR_1.0 database, a total of 21,206 SSR primer sets were designed to SSRs from these regions with no previously mapped SSRs. The numbers ranged from 5 to 259 SSR loci per interval or chromosome end. The average density of SSR loci in the whole genome to which primers were successfully designed was 35 Mbp^{-1} or one SSR marker per 0.072 cM as calculated based on a total of 2383.8 cM of soybean genetic map (Choi et al., 2007). The total number of intervals or chromosome ends in

Table 3. Numbers of simple sequence repeats (SSRs) with repeat units ≥5, mean and standard deviations of SSR repeat unit numbers, and numbers and proportions of SSRs with units of di-, tri-, and tetranucleotide repeats ≥10, 8, and 7, respectively in soybean expressed sequence tags (ESTs).

Repeats	No. of SSRs with repeat units ≥5	Avg. number of repeat units	SD of repeat units	No. of SSRs with units of di-, tri-, and tetranucleotide repeats ≥10, 8, and 7, respectively	Proportion of SSRs in ESTs with units of di-, tri-, and tetranucleotide repeats ≥10, 8, and 7, respectively
(AC)n	3305	5.8	1.88	127	0.03
(AG)n	11,689	8.0	6.91	1945	0.16
(AT)n	5076	8.3	6.02	1039	0.20
(CG)n	91	5.9	3.39	4	0.04
(AAC)n	1606	5.8	1.28	138	0.08
(AAG)n	3115	6.1	1.81	429	0.13
(AAT)n	1493	6.4	2.51	280	0.18
(ACC)n	1489	5.7	1.09	111	0.07
(ACG)n	533	5.6	1.05	42	0.07
(ACT)n	835	5.7	1.52	65	0.07
(AGC)n	545	5.8	1.17	47	0.08
(AGG)n	799	5.7	0.97	34	0.04
(AGT)n	800	5.9	1.58	86	0.10
(CCG)n	1225	5.1	0.56	14	0.01
(AAAC)n	22	5.2	0.39		
(AAAG)n	200	5.8	2.19	32	0.16
(AAAT)n	139	5.3	0.59	8	0.05
(AACC)n	19	5.1	0.23		
(AACG)n	5	6.0	1.00	2	0.40
(AACT)n	7	5.9	0.38		
(AAGC)n	1	5.0			
(AAGG)n	19	5.4	0.50		
(AAGT)n	21	5.9	0.96	6	0.28
(AATC)n	31	6.0	0.68	7	0.22
(AATG)n	6	5.7	0.82	1	0.16
(AATT)n	31	5.4	0.62	2	0.06
(ACAG)n	16	5.8	1.34	4	0.25
(ACAT)n	9	6.2	3.31	1	0.11
(ACCC)n	2	5.0			
(ACCT)n	1	6.0			
(ACGC)n	2	5.0			
(ACGT)n	10	5.0			
(ACTC)n	4	5.0			
(ACTG)n	31	5.6	1.05	6	0.19
(AGAT)n	1	8.0		1	
(AGCC)n	23	7.0	4.06	7	0.30
(AGCG)n	9	5.0			
(AGCT)n	13	5.0			
(AGGC)n	1	5.0			
(AGGG)n	1	5.0			
(AGGT)n	6	5.2	0.41		
(AGTC)n	6	6.5	1.52	3	
	33,237			4441	

the 500 kbp to 1.0 Mbp size range to which a SSR was not successfully designed was 75. Of the 1.0 to 1.5 Mb intervals only eight did not contain a SSR to which primers were designed. The total number of SSR markers on the 20 chromosomes in BARCSOYSSR_1.0 varied from 1327 to 1963, and the average density ranged from 28 to 41 Mbp^{-1} (Table 4).

Evaluation of the Quality and Polymorphism of Markers in the Database

To provide a prediction of the rate at which the primer sets in the BARCSOYSSR_1.0 database would be predicted to amplify a single discrete PCR product with length polymorphism among a set of diverse genotypes, a total of 1034 primer sets were evaluated by amplifying genomic

Table 4. Number of candidate simple sequence repeats (SSRs) (di- tri-, and tetranucleotide SSRs with repeat units of 10 to 35, 8 to 35, and 7 to 35 for di-, tri-, and tetranucleotide SSRs, respectively or with compound repeats of a total length of ≤120 nt), number of previously mapped SSRs and the number and frequency of candidate plus previously developed SSRs on each of the 20 soybean chromosomes to which primers were successfully designed and which are included in the BARCSOYSSR 1.0 database.

Chromosome	Linkage group	Size (bp)	No. of candidate SSRs	No. of previously developed SSRs	Total no. of potential SSR markers	Potential SSR markers Mbp^{-1}
Gm1	D1a	55,915,595	1674	49	1723	31
Gm2	D1b	51,656,713	1781	52	1833	35
Gm3	N	47,781,076	1666	41	1707	36
Gm4	C1	49,243,852	1484	36	1520	31
Gm5	A1	41,936,504	1388	43	1431	34
Gm6	C2	50,722,821	1807	48	1855	37
Gm7	M	44,683,157	1612	43	1655	37
Gm8	A2	46,995,532	1863	56	1919	41
Gm9	K	46,843,750	1694	51	1745	37
Gm10	O	50,969,635	1742	48	1790	35
Gm11	B1	39,172,790	1482	26	1508	38
Gm12	H	40,113,140	1380	33	1413	35
Gm13	F	44,408,971	1906	57	1963	44
Gm14	B2	49,711,204	1501	30	1531	31
Gm15	E	50,939,160	1561	38	1599	31
Gm16	J	37,397,385	1291	36	1327	35
Gm17	D2	41,906,774	1614	51	1665	40
Gm18	G	62,308,140	1895	55	1950	31
Gm19	L	50,589,441	1552	50	1602	32
Gm20	I	46,773,167	1298	31	1329	28
Total		9.50E+08	32,191	874	33,065	35

DNAs of seven *Glycine max* (L.) Merr. and one *G. soja* genotypes. Of these primer sets, 978 (94.6%) amplified a single discrete PCR product and 798 (77.2%) amplified polymorphic amplicons as evaluated on a 4.5% agarose gel. When the *G. soja* genotype is excluded, the polymorphism rate of loci among the seven *G. max* genotypes was 63.6%. Details relating to the 1034 primers sets tested including the allele size classes for the 798 polymorphic markers in the eight genotypes are given in Supplementary Table 2.

Polymorphism of SSRs with Various Motifs and Repeat Unit Numbers in Previous Soybean SSR Marker Development

Of the 2196 primer sets designed to perfect repeats in previous work at the USDA, Beltsville, MD (Cregan et al., 1999; Song et al., 2004), the polymorphism rates of primer sets designed to SSRs with perfect (AT)n and (ATT)n were 42.1 and 45.4% and were higher than those designed to (CT)n (25.0%) and (CTT)n (21.9%), (CA)n (18.2%), and (CAA)n (10.0%) repeats, respectively. The overall polymorphism rate of primer sets designed to di- and trinucleotide motifs were 40.6 and 44.5%, respectively, and the polymorphism rate of the 2196 primer sets was 43.0% (Table 5).

DISCUSSION

Efficacy of the Database BARCSOYSSR_1.0

From more than 210,000 SSRs identified in the soybean genome sequence Glyma1.01 and from the assembly of available EST sequences, 33,065 SSRs were identified with unique genome positions based on screening to eliminate repetitive sequence in SSR flanking regions, via analysis with e-PCR and by only maintaining those SSRs with repeat unit numbers ranging from 10 to 35, 8 to 35, and 7 to 35 for di-, tri-, and tetranucleotide SSR motifs, respectively. These ranges of repeat numbers were selected to maximize the probability of polymorphism in soybean germplasm. Polymerase chain reaction (PCR) primer sequences to these loci and information relative to their position on the 20 soybean chromosomes is available in the SSR database BARCSOYSSR_1.0. The database is effective in providing informative SSRs at essentially any genome position at which fine mapping to discern the position of a causative gene is required. The availability of inexpensive and versatile markers will help to facilitate map-based cloning. In addition, BARCSOYSSR_1.0 provides a source of candidate markers that can be used in marker-assisted selection. In wheat, less than 67% of the primer pairs designed from wheat SSR sequences amplified PCR products with expected size, and the polymorphism rate of the resulting SSR markers was lower than 30% (Bryan et al., 1997; Roder et al., 1995; Song et al., 2002, 2005). In rice, percentages

Table 5. The rate of success simple sequence repeat (SSR) marker development (single locus amplicon and polymorphism among a set of diverse soybean genotypes) of perfect SSRs (i.e., exclusion of the SSRs with compound motifs) derived from different sources and with different motifs based on 2196 primer sets previously evaluated by Cregan et al. (1999) and Song et al. (2004).

Source	Type	No. of primers designed	No. of named SSRs	Success rate (%)
Genomic clones or BAC ends	AT	704	307	43.6
Genomic clones or BAC ends	ATT	1291	591	45.7
Genomic clones or BAC ends	CT	25	9	36.0
Genomic clones or BAC ends	CA	3	0	0
Genomic clones or BAC ends	CTT	13	5	38.5
ESTs or genes	AT	57	14	24.5
ESTs or genes	ATT	22	6	27.3
ESTs or genes	CT	35	6	17.1
ESTs or genes	CA	8	2	25.0
ESTs or genes	CTT	28	4	14.2
ESTs or genes	CAA	10	1	10.0
Total		2196	945	43.0

of locus-specific primer sets and polymorphic SSRs were 63 and 39%, respectively, as calculated based on the data reported by Temnykh et al. (2000). In soybean, the overall polymorphism rates of the previously designed 2196 primer sets were 40.6 and 44.5% for di- and trinucleotide motifs, respectively, and the overall average polymorphism rate was 43.0%. Most of those primer sets were tested on twelve diverse genotypes: 'Clark', 'Harosoy', 'Jackson', Williams 82, 'Amsoy', Archer, 'Fiskeby V', Minsoy, Noir1, 'Tokyo', A81-356022, and PI 468916. Although difficult to accurately assess, the diversity of these twelve genotypes is similar to or greater than the eight genotypes used in the current study for the screening of polymorphism. Based on the analysis of a random selection of 1034 primer sets from the BARCSOYSSR_1.0 database, the percentage of locus specific primer sets and the rate of polymorphism were 94.6 and 77.4%, respectively. This polymorphism rate is much higher than the 43% rate obtained in the development of soybean SSR markers by Cregan et al. (1999) and Song et al. (2004). The polymorphism rate was 63.9% even among the seven *G. max* genotypes which is also substantially higher than the 38% polymorphism rate reported by Shultz et al. (2007) who only analyzed *G. max* genotypes. The higher success rate is likely due to a number of reasons including the increase of specificity of primer design as a result of the availability of the whole genome sequence and the use of e-PCR as well as the elimination of non-nuclear sequences or of SSRs that resided in repetitive sequence.

As an example to demonstrate the impact of the WGS on the success of SSR marker development, the primer sequences from the 1479 primer sets which failed to amplify locus-specific PCR products or that were not polymorphic in previous SSR marker development research at the USDA, Beltsville, MD, were analyzed with e-PCR using the whole soybean genome sequence Glyma1.01. Of the 1479 sets of primer sequences, a total of 276 (18.7%) were aligned to duplicated regions. Thus, avoiding primer design to such regions reduces the failure rate of SSR marker development and further testifies to the value of the WGS.

Polymorphism of SSRs with Various Motifs and Repeat Unit Numbers

The polymorphism rates of primer sets designed to SSRs with perfect (AT)n and (ATT)n were higher than those with (CT)n and (CTT)n, respectively (Table 5). The average and standard deviation of repeat unit numbers of SSRs with (AT)n and (ATT)n motifs were also higher than those of SSRs with other motifs (Table 1). As suggested by a number of previous reports (Ellegren, 2004; Kayser et al., 2004; Mun et al., 2006; Temnykh et al., 2001; Weber, 1990), SSRs with a greater number of repeat units have a higher likelihood of polymorphism. We observed the association of repeat unit number versus polymorphism rate within the (ATT)n motif, to which most of the previous soybean SSR primer sets were designed. The correlation of the number of repeat units (calculated based on the repeat units from 8 to 35) versus the polymorphism rate of the (ATT)n motifs was 0.677 ($p < 0.001$). More primer sets are needed to verify this association within other motifs. Calculations based on data from *Oryza sativa* L. subsp. *indica* Kato rice (Zhang et al., 2007) showed a similar relationship of polymorphism rates versus SSR average length and versus standard deviation of the unit number of the motif (both coefficients were significant at $\alpha = 0.01$). Thus, users of the BARCSOYSSR_1.0 database would be advised to select candidate SSR markers with greater repeat unit numbers to maximize the probability of polymorphism in their application. Our data indicate that in all likelihood the (ATT)n and (AT)n SSRs would be the first candidates chosen as a result of their greater average repeat unit numbers.

The position of SSRs in the BARCSOYSSR_1.0 database is based on the soybean genome sequence Glyma1.01. As additional improvements of scaffold

anchoring and orientation of the genome sequence continues, the actual positions of the SSRs in the genome may in some cases be different from the positions listed in the current database. To facilitate the use of the database, the physical positions of previously published SSR markers (Cregan et al., 1999; Hisano et al., 2007; Shoemaker et al., 2008; Shultz et al., 2007; Song et al., 2004; Xia et al., 2007) were included in the database (Supplementary Table 1). These were included only if the primer sequences of loci and the motifs were identical to the flanking sequences of loci and the motifs of the SSRs in the database, respectively. The positions and information related to the 33,065 BARCSOYSSR_1.0 markers are available in SoyBase the USDA-ARS Soybean Genome Database (http://soybase.org; verified 21 June 2010). The BARCSOYSSR_1.0 database can be downloaded as an Microsoft Excel file at http://bldg6.arsusda.gov/~pooley/soy/cregan/BARCSOYSSR_1.0.html (verified 28 May 2010). With updates of the WGS in the future, candidate SSR markers can still be easily identified from specific regions by referring to the positions of mapped SSRs in the database. These updates can be obtained at SoyBase and at http://bldg6.arsusda.gov/~pooley/soy/cregan/BARCSOYSSR_1.0.html (verified 28 May 2010).

Acknowledgments

These whole soybean genome sequence and sequence assembly was produced by the U.S. Department of Energy, Joint Genome Institute http://www.jgi.doe.gov/ (verified 21 June 2010) in collaboration with the user community. This work was partially supported by United Soybean Board Project 7268. The support of the United Soybean Board is greatly appreciated.

References

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

Bryan, G.J., A.J. Collins, P. Stephenson, A. Orry, , J.B. Smith, and M.D. Gale. 1997. Isolation and characterisation of microsatellites from hexaploid bread wheat. Theor. Appl. Genet. 94:557–563.

Cho, Y.G., T. Ishii, S. Temnykh, C. X., L. Lipovich, S.R. McCouch, W.D. Park, N. Ayers, and S. Cartinhour. 2000. Diversity of microsatellites derived from genomic libraries and GenBank sequences in rice (*Oryza sativa* L.). Theor. Appl. Genet. 100:713–722.

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

Cordeiro, G.M., R. Casu, C.L. McIntyre, J.M. Manners, and R.J. Henry. 2001. Microsatellite markers from sugarcane (*Saccharum* spp.) ESTs cross transferable to erianthus and sorghum. Plant Sci. 160:1115–1123.

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

Edwards, A., A. Civitello, H.A. Hammond, and C.T. Caskey. 1991. DNA typing and genetic mapping with trimeric and tetrameric tandem repeats. Am. J. Hum. Genet. 49:746–756.

Ellegren, H. 2004. Microsatellites: Simple sequences with complex evolution. Nat. Rev. Genet. 5:435–445.

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

Hisano, H., S. Sato, S. Isobe, S. Sasamoto, T. Wada, A. Matsuno, T. Fujishiro, M. Yamada, S. Nakayama, Y. Nakamura, S. Watanabe, K. Harada, and S. Tabata. 2007. Characterization of the soybean genome using EST-derived microsatellite markers. DNA Res. 14:271–281.

Huang, X., and A. Madan. 1999. CAP3: A DNA sequence assembly program. Genome Res. 9:868–877.

Kayser, M., R. Kittler, A. Erler, M. Hedman, A.C. Lee, A. Mohyuddin, S.Q. Mehdi, Z. Rosser, M. Stoneking, M.A. Jobling, A. Sajantila, and C. Tyler-Smith. 2004. A comprehensive survey of human Y-chromosomal microsatellites. Am. J. Hum. Genet. 74:1183–1197.

Metzgar, D., J. Bytof, and C. Wills. 2000. Selection against frameshift mutations limits microsatellite expansion in coding DNA. Genome Res. 10:72–80.

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

Mun, J.H., D.J. Kim, H.K. Choi, J. Gish, F. Debelle, J. Mudge, R. Denny, G. Endre, O. Saurat, A.M. Dudez, G.B. Kiss, B. Roe, N.D. Young, and D.R. Cook. 2006. Distribution of microsatellites in the genome of Medicago truncatula: A resource of genetic markers that integrate genetic and physical maps. Genetics 172:2541–2555.

Roder, M.S., V. Korzun, K. Wendehake, J. Plaschke, M.H. Tixier, P. Leroy, and M.W. Ganal. 1998. A microsatellite map of wheat. Genetics 149:2007–2023.

Roder, M.S., J. Plaschke, S.U. Konig, A. Borner, M.E. Sorrells, S.D. Tanksley, and M.W. Ganal. 1995. Abundance, variability and chromosomal location of microsatellites in wheat. Mol. Gen. Genet. 246:327–333.

Schlotterer, C., and D. Tautz. 1992. Slippage synthesis of simple sequence DNA. Nucleic Acids Res. 20:211–215.

Schmutz, J., S. Cannon, J. Schlueter, J. Ma, T. Mitros, W. Nelson, D. Hyten, Q.J. Song, J.J. Thelen, J. Cheng, D. Xu, U. Hellsten, G. May, Y. Yu, T. Sakurai, T. Umezawa, M. Bhattacharyya, D. Sandhu, B. Valliyodan, E. Lindquist, M. Peto, D. Grant, S. Shu, D. Goodstein, K. Barry, M. Futrell-Griggs, J. Du, Z. Tian, L. Zhu, N. Gill, T. Joshi, M. Libault, A. Sethuraman, X.C. Zhang, K. Shinozaki, H.T. Nguyen, R.A. Wing, P.B. Cregan, J. Specht, J. Grimwood, D. Rokhsar, G. Stacey, R.C. Shoemaker, and S.A. Jackson. 2010. A genome sequence of paleopolyploid soybean (*Glycine max*). Nature 463:178–183.

Shoemaker, R.C., D. Grant, T. Olson, W.C. Warren, R. Wing, Y. Yu, H. Kim, P. Cregan, B. Joseph, M. Futrell-Griggs, W. Nelson, J. Davito, J. Walker, J. Wallis, C. Kremitski, D. Scheer, S.W. Clifton, T. Graves, H. Nguyen, X. Wu, M. Luo, J. Dvorak, R. Nelson, S. Cannon, J. Tomkins, J. Schmutz, G. Stacey, and S. Jackson. 2008. Microsatellite discovery from

BAC end sequences and genetic mapping to anchor the soybean physical and genetic maps. Genome 51:294–302.

Shultz, J.L., S. Kazi, R. Bashir, J.A. Afzal, and D.A. Lightfoot. 2007. The development of BAC-end sequence-based microsatellite markers and placement in the physical and genetic maps of soybean. Theor. Appl. Genet. 114:1081–1090.

Song, Q.J., E.W. Fickus, and P.B. Cregan. 2002. Characterization of trinucleotide SSR motifs in wheat. Theor. Appl. Genet. 104:286–293.

Song, Q.J., L.F. Marek, R.C. Shoemaker, K.G. Lark, V.C. Concibido, X. Delannay, J.E. Specht, and P.B. Cregan. 2004. A new integrated genetic linkage map of the soybean. Theor. Appl. Genet. 109:122–128.

Song, Q.J., J.R. Shi, S. Singh, E.W. Fickus, J.M. Costa, J. Lewis, B.S. Gill, R. Ward, and P.B. Cregan. 2005. Development and mapping of microsatellite (SSR) markers in wheat. Theor. Appl. Genet. 110:550–560.

Subramanian, S., V.M. Madgula, R. George, S. Kumar, M.W. Pandit, and L. Singh. 2003. SSRD: Simple sequence repeats database of the human genome. Comp. Funct. Genomics 4:342–345.

Temnykh, S., G. DeClerck, A. Lukashova, L. Lipovich, S. Cartinhour, and S. McCouch. 2001. Computational and experimental analysis of microsatellites in rice (*Oryza sativa* L.): Frequency, length variation, transposon associations, and genetic marker potential. Genome Res. 11:1441–1452.

Temnykh, S., W.D. Park, N. Ayers, S. Cartinhour, N. Hauck, L. Lipovich, Y.G. Cho, T. Ishii, and S.R. McCouch. 2000. Mapping and genome organization of microsatellite sequences in rice (*Oryza sativa* L.). Theor. Appl. Genet. 100:697–712.

Thiel, T., W. Michalek, R.K. Varshney, and A. Graner. 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.

Weber, J.L. 1990. Informativeness of human (dC-dA)n.(dG-dT)n polymorphisms. Genomics 7:524–530.

Xia, Z., Y. Tsubokura, M. Hoshi, M. Hanawa, C. Yano, K. Okamura, T.A. Ahmed, T. Anai, S. Watanabe, M. Hayashi, T. Kawai, K.G. Hossain, H. Masaki, K. Asai, N. Yamanaka, N. Kubo, K. Kadowaki, Y. Nagamura, M. Yano, T. Sasaki, and K. Harada. 2007. An integrated high-density linkage map of soybean with RFLP, SSR, STS, and AFLP markers using A single F2 population. DNA Res. 14:257–269.

Zhang, L., D. Yuan, S. Yu, Z. Li, Y. Cao, Z. Miao, H. Qian, and K. Tang. 2004. Preference of simple sequence repeats in coding and non-coding regions of Arabidopsis thaliana. Bioinformatics 20:1081–1086.

Zhang, Z., Y. Deng, J. Tan, S. Hu, J. Yu, and Q. Xue. 2007. A genome-wide microsatellite polymorphism database for the indica and japonica rice. DNA Res. 14:37–45.