Theor Appl Genet (2007) 114:1081–1090
DOI 10.1007/s00122-007-0501-9

ORIGINAL PAPER

The development of BAC-end sequence-based microsatellite markers and placement in the physical and genetic maps of soybean

Jeffry L. Shultz · Samreen Kazi · Rabia Bashir · Jawaad A. Afzal · David A. Lightfoot

Received: 20 April 2006 / Accepted: 7 January 2007 / Published online: 8 February 2007

Abstract The composite map of soybean shared among Soybase, LIS and SoyGD (March 2006) contained 3,073 DNA markers in the "Locus" class. Among the markers were 1,019 class I microsatellite markers with 2–3 bp simple sequence repeats (SSRs) of >10 iterations (BARC-SSR markers). However, there were few class II SSRs (2–5 bp repeats with <10 iterations; mostly SIUC-Satt markers). The aims here were to increase the number of classes I and II SSR markers and to integrate bacterial artificial chromosome (BAC) clones onto the soybean physical map using the markers. Used was 10 Mb of BAC-end sequence (BES) derived from 13,473 reads from 7,050 clones constituting minimum tile path 2 of the soybean physical map (http://www.soybeangenome.siu.edu; SoyGD). Identified were 1,053 1–6 bp motif, repeat sequences, 333 from class I (>10 repeats) and 720 from class II (<10 repeats). Potential markers were shown on the MTP_SSR track at Gbrowse. Primers were designed as 20–24 bp oligomers that had Tm of 55 ± 1 C that would generate 100–500 bp amplicons. About 853 useful primer pairs were established. Motifs were not randomly distributed with biases toward AT rich motifs. Strong biases against the GC motif and all tetra-nucleotide repeats were found. The markers discovered were useful. Among the first 135 targeted for use in genetic map improvement about 60% of class II markers and 75% of class I markers were polymorphic among on the parents of four recombinant inbred line (RIL) populations. Many of the BES-based SSRs were located on the soybean genetic map in regions with few BARC-SSR markers. Therefore, BES-based SSRs represent useful tools for genetic map development in soybean. New members of a consortium to map the markers in additional populations are invited.

Keywords Microsatellite · Motif · Repeat · Soybean · Legume · Physical map

Communicated by F. J. Muehlbauer.

Electronic supplementary material The online version of this article (doi:10.1007/s00122-007-0501-9) contains supplementary material, which is available to authorized users.

J. L. Shultz · S. Kazi · R. Bashir · J. A. Afzal · D. A. Lightfoot (✉)
Genomics Core Facility and Center of Excellence in Soybean Research, Teaching and Outreach, and Department of Plant, Soil and Agricultural Systems, Southern Illinois University at Carbondale, Carbondale, IL 62901, USA
e-mail: ga4082@siu.edu

Present Address:
J. L. Shultz
United States Department of Agriculture, Stoneville, MS 38776, USA

Introduction

Microsatellites are 1–6 bp repeat motifs arranged in tandem within DNA (Cregan et al. 1999a; Temnykh et al. 2001). Class I microsatellite markers have more than ten iterations and class II have less than ten iterations of the repeat motif. The repeats exhibit within species polymorphisms at high frequency. The repeats provide an abundant source of simple sequence repeat (SSR) markers. The use of SSRs has been critical to molecular mapping projects in soybean (Demirbas et al. 2001; Iqbal et al. 2001;

Song et al. 2004; Yamanaka et al. 2001; Zhang et al. 2004; Kassem et al. 2006). SSR markers are useful in polyploid or paleopolyploid genomes where they can distiguish among composite genomes and identify homeologous regions in BAC pools (Shultz et al. 2006).

In soybean there were 1,011 BARC_SSR markers listed on Soybase in 2006 (Cregan et al. 1999b; Song et al. 2004). Among them were 605 Satt, 332 Sat, 25 Sct, and 5 Sctt class I markers. Based on a 2,523.6 cM (Kosambi) soybean map and the reported 1,011 SSR markers, the mean distance between SSR marker loci should have been less than 2.48 cM. However, there were 24 gaps that ranged from 10 to 20 cM in the composite map (Song et al. 2004). The gaps account for a total of 375.1 cM, or 14.8% of the current map.

The gaps are targets for marker identification from BAC and contig sequences (Cregan et al. 1999b; Meksem et al. 2000; Triwitayakorn et al. 2005; Ruben et al. 2006). The targeted filling of gaps in the genetic map will require advanced genomic resources (Meksem et al. 2000; Wu et al. 2004; Shultz et al. 2006), particularly large contiguous sets of overlapping BAC clones (hereafter contigs) anchored to reliable markers.

The publicly available physical map of soybean used medium information content fingerprinting (Meksem et al. 2000; Shultz et al. 2003a, 2006; Wu et al. 2004) shown at SoyGD in two forms, build 3 and build 4. Build 3 of the physical map of soybean (Wu et al. 2004) consisted of 2,905 contigs of 491 Kbp mean size. Build 4 of the physical map of soybean (Shultz et al. 2006) consisted of 2,854 contigs of 363 Kbp mean size. However, only 762 contigs could be reliably anchored to the physical map using BARC_SSR markers, RFLPs and BAC DNA pools. Marker amplicon duplication proved to be widespread. Therefore, a new marker system to place contigs on the genetic and physical map and to verify contigs already placed was needed.

The derivation of SSRs from genome encompassing BES has been reported for rice (Temnykh et al. 2001; McCouch et al. 2002), wheat and barley (Rota et al. 2005). In each genome markers of class I and class II were developed that could be mapped within and among genomes. In soybean, clones that made up a minimum tiling path (MTP) of the soybean genome based on build 3 were used to develop 13,473 sequence reads. The objective was to determine whether satellite markers found in soybean BES would be polymorphic and reliable to allow use as integrating markers for the genetic and physical maps.

Materials and methods

Source of sequences

About 16,128 sequence reads (Genbank. CG812653 to CG826126) were attempted from 8,064 the MTP clones identified from build 3 of the physical map (Shultz et al. 2006). The MTP2 clones were hand picked from the *Bam*HI and *Hin*dIII BAC libraries to provide genome coverage with about 25% overlap among neighboring clones in the tile. There were 13,473 useful BES reads. Mean read length was about 736 bp. The total amount of sequence generated from MTP2 was 9.9 Mbp (about 1%) of the soybean genome. There were 5,555 paired, forward and reverse reads.

Frequency of microsatellites

A JAVA program was written to identify possible simple sequence repeats (SSRs) within the ~10 Mb of soybean BES reported on Genbank (Shultz et al. 2003b). All the BES reads were placed into a single text file, with one line per sequence in FASTA format. Each record was sequentially tested for 2, 3, 4, 5 and 6 base repeated DNAs. A requirement that each group of repeats must be at least 15 base pairs in length (3–7 repeats depending on motif size) was included in the program. No attempt was made to filter out single nucleotide sequences. Once detected, the microsatellite region was surrounded with "[ ]" delimiters to aid the automated primer selection. Also, once an acceptable match was found, the program output the data and proceeded to the next sequence record making no attempt to find any other repeats or to increase the length of the accepted match.

The JAVA output file was transferred to the Primer 3 program located at http://www.frodo.wi.mit.edu/cgi-bin/primer3/primer3_www.cgi/. Settings used for primer design were T_m 55° ± 1°C, amplicon 100–500 bp, primer length 20 ± 2 bp. No constraints on GC% were set to avoid potential bias against the AT rich regions of the soybean genome. Equally, repeated DNA amplicons (minisatellites, transposons etc.) were not filtered out by Blast searching.

Polymorphism detection

The parents of four recombinant inbred line (RIL) populations; 'Flyer' × 'Hartwig' (F × H; Yuan et al. 2002; Kazi et al. 2006), 'Essex' × 'Forrest' (E × F; Iqbal et al. 2001; Kassem et al. 2006; Lightfoot et al. 2005), 'Pyramid' × 'Douglas' (P × D; Njiti et al. 2002) and 'Minsoy' × 'Noir 1' (M × N; Cregan et al. 1999b; Lark

et al. 1993; Song et al. 2004) were tested for polymorphism using 96 of the BES-derived microsatellite primers (See Supplementary Table for sequences). Seed were obtained from ARC seed store maintained by Dr. Lightfoot at SIUC except Minoy and Noir that were obtained from Dr. Orf, (University of Minnesota). DNA was extracted as described previously (Iqbal et al. 2001). Primers (Supplementary Table) were obtained from Sigma Genosys (Woodlands, TX, USA).

Polymerase chain reaction (PCR) was performed in a PE 9700 (Boston, MA, USA). An initial 95°C denaturation for 5 min was followed by 30 cycles of 95° for 30 s, 55° for 30 s, and 72° for 30 s. After PCR was complete, gel electrophoresis was performed in a 4% (w/v) agarose gel stained with ethidium bromide. Polymorphism was documented using a BioRad GelDoc (Hercules, CA, USA) system.

Annotation and Map representation

All potential microsatellites were named with the SIUC_ suffix (at first mention) followed by S, the repeat motif and the BAC of origin. In contrast, earlier SSR markers were assigned a sequential number (Cregan et al. 1999b; Meksem et al. 2000; Song et al. 2004; Triwitayakorn et al. 2005). The altered naming convention used here was designed to aid users find the clone of origin in the physical map. All potential markers were shown at SoyGD in the BES_SSR track. Markers that have been located in the genetic map by scored RIL DNA polymorphism will also be shown on the locus track.

Results

Frequency of microsatellite types and motifs

A total of 1,053 among 13,473 BESs contained 2–6 bp repeats of at least 15 bp, that represented 7.8% of the sequences tested. There were 333 from class I and 720 from class II. Among the 333 class I motifs 153 were mononucleotide repeats. Only 180 were composed of di- and tri-nucleotide repeats like BARC-SSRs.

Of the 1,053 repeat sequence regions, 853 were flanked by sequence capable of PCR amplification using the parameters indicated in methods and materials. A complete list of the primers, with expected amplicon size, contig number and SSR repeat is included in the supplemental table. The rejected 200 contained very AT rich regions on one or both sides of the amplicon that prevented primers being designed. The initial AT composition of the 10 Mbp of BES was 63.7%; the 1,053 BES that contained satellites were 68.2% AT. Of the 853 BES selected for amplification the total AT composition was 66.3%; and the amplicons of about 200 bp were 65.7% AT.

Table 1 showed each microsatellite type by motif and the frequency observed. Based on the canonical set of SSR motifs expected were the two singlets (A or T, and G or C runs) and 4 di-nucleotides (GC, AC, AG and AT). Among the singlet repeats A(x) was significantly more abundant ($n = 137$) than the G(x) motif ($n = 17$), reflecting the AT rich nature of the soybean genome. Among the singlets (x) was 15 or greater and all markers were class 1 SSRs. Among the di-nucleotides AT was significantly more abundant ($n = 196$), GC was absent ($n = 0$) and AG ($n = 46$) and AC ($n = 40$) were less abundant. Among the 12 tri-nucleotide motifs a similar pattern was observed with AAT (reverse complement ATT) motifs most common ($n = 66$), but AAG nearly as common ($n = 57$). The AAC tri-nucleotide was significantly more common ($n = 36$) than all

Table 1 Simple sequence repeat motifs found in BAC end sequences

Motif	*N*	Motif	*N*	Motif	*N*
A	137	AACCT	1	AC	40
AAAAAC	2	AACGC	1	ACAGAG	1
AAAAAG	8	AACGT	1	ACAGGC	1
AAAAAT	7	AACT	1	ACAGT	1
AAAAC	7	AACTTC	1	ACC	4
AAAACT	3	AAG	57	ACCATC	1
AAAAG	17	AAGAG	2	ACG	2
AAAAGG	1	AAGAGG	1	ACT	2
AAAAT	55	AAGAT	5	AG	46
AAAATC	2	AAGGAG	1	AGAGC	1
AAAATT	3	AAGGTG	1	AGAT	1
AAACCC	1	AAGTAC	1	AGC	5
AAACT	3	AAT	66	AGG	18
AAACTT	3	AATAC	3	AGGGG	1
AAAG	1	AATAG	6	AGGT	1
AAAGAG	2	AATAT	3	AGGTGG	1
AAAGC	1	AATATT	1	AT	196
AAAGG	1	AATC	1	ATC	9
AAAGTG	1	AATCC	1	ATCCT	1
AAAT	7	AATCCC	1	ATCTGC	1
AAATAT	1	AATCG	1	ATG	9
AAATC	1	AATCT	2	ATGAT	1
AAATG	2	AATGAC	1	ATGT	1
AAATT	4	AATGGT	1	CC	17
AAATTC	3	AATGT	2	CCG	1
AAATTT	1	AATT	1	CGG	1
AAC	36	AATTAT	1		
AACAAG	6	AATTC	1		
AACAC	4	AATTG	1		
AACAG	3	AATTT	3		

Each motif listed was present with more than four repeats in at least one BES. The number of examples found (*N*) among the 853 amplicons developed from soybean BAC-end sequences are listed

tri-nucleotides with two Gs, Cs or a GC ($n = 29$). Among the latter AGG tri-nucleotides were the most common ($n = 18$). There were only two BESs with repeats composed entirely of a mix of Gs and Cs.

Among the 33 different tetra-nucleotide repeat motif classes potentially present only eight were observed and only 15 examples were recorded. Only AAAT ($n = 7$) was present in multiple copies. Among the 102 potential different penta-nucleotide motifs 31 were observed. AAAAT was most abundant ($n = 55$). AAAAG ($n = 17$) was more common than AAAAC ($n = 7$), but the remaining 28 motifs were rare ($1 < n < 5$). Among the hundreds of potential sexta-nucleotides 31 were observed. The AAAAAN motif was most common ($n = 15$) and AACAAG ($n = 6$) was also abundant. The other 29 motifs were rare.

Polymorphism

Thirty six primer pairs for class II SSR markers among the first 96 tested were polymorphic among the 8 parents (38%; Table 2). Typically amplicons were single bands of about equal intensity on agarose gels, two BAC-derived SSRs and two BARC-Satt amplicons are shown in Fig. 1. Among the markers tested 5 were polymorphic in three populations, 15 were polymorphic

Table 2 Polymorphism of BAC-end sequence based SSR markers within four soybean mapping population parents. The BAC of origin of the BES is shown along with the size of the core motif and its sequence, the build 4 contig if any. For the 14 markers that were mapped shown are the map position as left and right flanking markers and the linkage group. Gap indicates the markers were not linked to any marker in the existing map and so cannot be placed

Primers	Motif		Build3			Build 4			Polymorphic populations	Genetic map locations		
	Size (n)	Sequence	Contig no.	LG	Anchor	Contig no.	LG	Anchor		Map Interval		L.G.
B15B14	2	AA	59	A1	Satt248		ND		3			Gap
B17K12	2	AA	830	D1a	Satt370		ND		2			ND
B16A10	2	AA	366	ND		253	ND		2			ND
B04I02	2	AA	1806	ND			ND		2			
B17P05	2	AA	283	ND			ND		1			
B10O02	2	AA	1604	O	Satt466		ND		1			
B14G13	2	AC	1772	ND			ND		1			
B13M19	2	AC	217	G	Satt533	28	ND		1			
B09L01	2	AC	760	D2	A257	1023	D1a	Satt507	1	Sct_10	Satt324	G
B02K20	2	AG	173	ND			ND		3	Satt510		F
H100B10[a]	2	AG	1793	A2	A117	214	A2	Sat_400	1	Blt65	Satt162	A2
B15LO6	2	AT	107	H	Sat_122	ND			3		Gap	
B15L05	2	AT	123	O	Satt633	628	K	Satt588	3	Satt76	Satt252	F
B17E19	2	AT	132	ND		158	ND		2			Gap
B16L10	2	AT	1357	A1A2	A110		ND		2			
B15P23	2	AT	168	ND		244	H	Sat_206	2			
B15I12	2	AT	1344	A1	A236		ND		2			
B08G14	2	AT	1590	ND			ND		2	Satt129	Satt408	D1a
B08D14	2	AT	1118	ND			ND		2			Gap
B01I14	2	AT	171	A1	Satt200		ND		2			
B17C16	2	AT	42	H	Satt253		ND		1			
B14O11	2	AT	573	D1a	Satt572	8198	A1D1a	Satt507	1			
B12B12	2	AT	1555	ND			ND		1			
B03P01	2	AT	1123	D1b	Satt459		ND		1			
B10P12	3	AAC	426	K	Satt046		ND		2			Gap
B14B13	3	AAC	1216	ND			ND		1			
B13G15	3	AAG	1507	ND			ND		1			Gap
B02B24	3	ATT	791	A1	Satt225		ND		2	Satt009	Satt152	N
B15A19	5	AAAAC	1417	ND		1632	ND		1			
B10M21	5	AAAAC	2365	ND			ND		1			
B15C01	5	AAATA	1848	ND			ND		2			
B12C13	5	AACAC	775	ND			ND		2			Gap
B15J11	5	AACCT	791	A1	Satt225		ND		1			Gap
B11006	5	AGAAC	791	A1	Satt225		ND		2			
B13L17	6	AAAAAG	2082	ND		9092	ND		3			Gap
B16I17	6	AAAAAG	67	ND		1771	ND		1			

ND is not determined

[a] marker mapped to A2 was b, AY858570

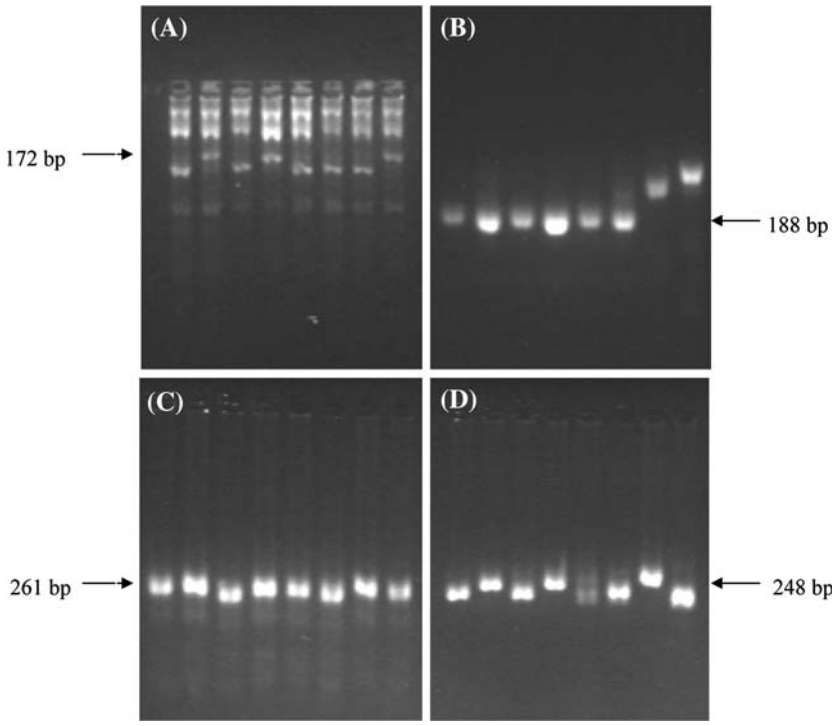

Fig. 1 Agarose gel electrophoresis of simple sequence repeat markers. **a** SIU-Sct_B02K20, **b** SIU-Sat_B15L06, **c** BARC-Satt232 and **d** BARC-Satt 237. *Lane* order is (*left to right*) Flyer, Hartwig, Essex, Forrest, Pyramid, Douglas, Minsoy and Noir1

in two populations and 16 were polymorphic in one population. Within populations 15% of markers tested were polymorphic in F × H, 18% in E × F, 14% in P × D, and 15% in M × N. Regardless of whether polymorphism was detected in these populations, BES-SSR were displayed at SoyGD (Fig. 2).

Among mononucleotide repeat motifs, 6 of 11 markers (55%) were polymorphic. Among di-nucleotide motifs, 3 of 7 AC markers, 14 of 20 AT markers and 1 of 3 AG markers (60%) were polymorphic. Among tri-nucleotide repeats, 2 of 12 were polymorphic (17%). Among tetra-nucleotide repeats, only 1 was tested and was not polymorphic. Among penta-nucleotide repeats, 6 of 17 were polymorphic (35%). Among sexta-nucleotide repeats, 2 of 7 were polymorphic (31%).

Map locations

Genetic map locations were sought for the 14 SSR markers polymorphic in the E × F population and 17 SSRs polymorphic in the F × H population. Jointly the markers provided 20 unique markers. The remaining 18 markers of the set of 38 were only polymorphic in M × N or P × D and were not prioritized for map placement.

Among the mapped markers 6 were placed on E × F linkage groups (Table 2) and one in the F × H map by associations (LOD > 3.0) with markers in the existing maps (430 for E × F and 151 for F × H). The markers placed in the E × F map were; SIUC-SacB09L01 that mapped between Sct_10 and Satt324 on L.G. G (Fig. 2); SIUC-SagB02K20 that was linked to Satt510 on L.G. F; SIUC-SagH100B10b (AY858570) that mapped between Blt65 and Satt162 on L.G. A2; SIUC-SatB15L05 that mapped between Satt76 and Satt252 on L.G. F; SIUC-SatB08G14 that mapped between Satt129 and Satt408 on L.G. D1a; and SIUC-SattB02B24 that mapped between Satt009 and Satt152 on L.G. N. The marker placed in the F × H map but not the E × F map was Sag_B02K20 linked to Satt510 on L.G. F.

The marker Sac_B09L01 shows the power of the method in that four contigs were anchored by Satt507, one of which is on L.G. D1a and the other three are homeologs from other locations. The marker suggested that ctg 8092 be moved to its correct location on L.G. G. Further, a homeologous relationship between these regions of L.G.s D1a and G is inferred. However, markers H100B10a (L.G. A1) and H100B10b (L.G.A2) mapped to different linkage groups and identified homeologous regions with conserved gene orders (Campbell et al. unpublished). Therefore, contig assignments will not always be correct by this method.

Eight markers were not linked to any other marker in the E × F map and so were inferred to be located in gaps (Kassem et al. 2006). The markers in gaps were; SIUC-SaaB15B14 from ctg59 anchored to L.G. A1 by Satt248 in build 3; SIUC-SatB15LO6 from ctg107 anchored to L.G. H by Sat_122; SIUC-SatB17E19

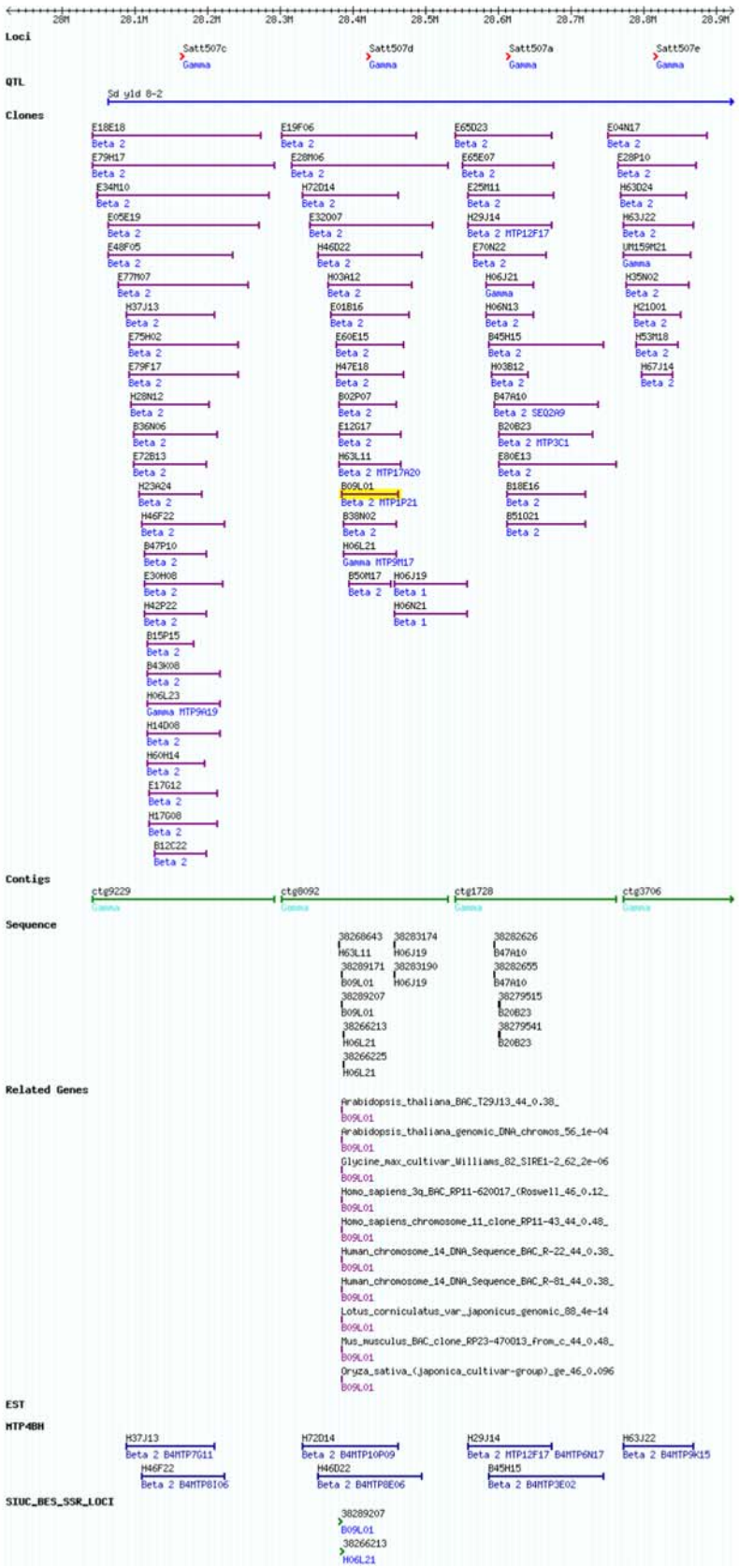

Fig. 2 Gbrowse representation of the MTP clones in a portion of the soybean genome showing build 4 linkage group D from 28 to 29 Mbp. A 1 Mbp region with loci, QTL, clones, contigs, sequences and gene models are shown. Loci, or genetic map DNA markers, are shown as *red arrowheads*. QTL in the region are shown as *blue bars*. BAC clones are shown as the *coalesced purple bar*. Contigs are shown as *green bars*. Polyploid region contigs have ctg numbers greater than 8,000. Sequences from MTP BAC ends are shown as *black lines*. Related gene annotations are shown as *purple lines* (the five most probable Blastx hits at $P < \text{e-}5$ are listed). Clicking on MTP clones brings up the gene index number. MTP4 clones are annotated below the bar with MTP and the MTP plate address. MTP2 clones can be identified as they have BES and EST hits shown. BES-SSR markers are shown as *green arrowheads* below the MTP clones. Highlighted in *yellow* is the clone B09L01 that contains the marker Sag_B09L01 mapped to linkage group G suggesting this contig (one of four homeologs) was misplaced in the physical map

from ctg132 and ctg 158 unplaced in builds 3 and 4; SIUC-SatB08D14 from unplaced ctg1118 (build 3); SIUC-SaacB10P12 from ctg426 anchored to L.G. K by Satt046 in build 3 but a singleton in build 4; SIUC-SaagB13G15 from ctg1507 unplaced in build 3 and a singleton in build 4; SIUC-SaacacB12C13 from ctg775 unplaced in build 3 and a singleton in build 4; SaacctB15J11 from ctg 791 anchored to L.G. A1 by Satt225 in build 3 but a singleton in build 4; and SIUC-SaaaaagB13L17 from ctg2082 and ctg 9092 unplaced in builds 3 and 4.

Three of the markers were linked together in a cluster in the F × H population. The markers that clustered in the genetic map were SIUC-SaaaagB54L24 (build 4 ctg 3682 in Queue; build 3 ctg 360 on L.G. G), SIUC-SaaattB05N04 (build 4 singleton; build 3 ctg331 on L.G. N) and SIUC-Sat_H35G20 (build 4 ctg 9299; build 3 ctg 853 both in Queue). The markers were from different contigs in the physical map. Location of these markers in a genetic map will correctly place the related contigs.

Physical map locations were predicted in build 3 for most of the markers and the BACs of origin by association in a contig since the clones derived from the MTP of that build (Table 2). For example SaaB15B14 was part of ctg59 that was anchored to L.G. A1 by Satt248 in build 3 but was unplaced in Queue in build 4. Genetic linkage placed the marker BAC and associated contig in a gap of the E × F genetic map. Among the mapped markers twelve were from build 3 contigs that were anchored to linkage groups (e.g. SaaB15B14) and 8 were unassigned in Queue. For example SIUC-SaaB16A10 was part of ctg366 in build 3 and ctg253 in build 4, both were assigned to Queue. Ten of the markers were from BACs included in build 4 (e.g. SaaB16A10) and six of them were assigned to linkage groups by association with an anchored contig. For example SacB09L01 was from ctg760 anchored to L.G D2 by RFLP A257 in build 3 but ctg1023 anchored to L.G. D1a by Satt507in build 4 (Fig. 2). The ExF genetic map placed this marker at a third location, between Sct_10 and Satt324 on L.G. G. The latter is the most likely location because soybean SSR amplicons all had homeologs that complicate BAC anchor determinations (Shultz et al. 2006).

Among the 7 markers that were linked to the genetic map, two agreed with the build 3 position for the contig (e.g. SagH100B10b) and 5 suggested the contig may be placed incorrectly or composed of BACs from different genetic locations (e.g. SacB09L01; Fig. 2). The two markers with consistent positions were; SatB14O11 part of ctg573 anchored to L.G. D1a by Satt572 in build 3 and ctg8198 anchored to both L.G.s A1 and D1a by Satt507 in build 4; and SagH100B10b anchored to ctg1793 on L.G A2 by RFLP A117 in build 3 and ctg214 on A2 by Sat_400. The genetic map of ExF placed the BAC between SCAR Blt65 and Satt162, also on L.G. A2.

Discussion

The distribution of satellite motifs in soybean DNA was not random. AT rich motifs were more common in general. Certain AT rich motifs were significantly more common than expected. For example, the AAG motif was much more common than the AAC motif suggesting runs of pyrimidines and purines are more stable motifs in soybean. The selection of the AT and ATT motifs for BARC-SSR and composite map development by Song et al. (2004) was clearly justified (Table 3) by their abundance. Should more random microsatellites be sought the AAG and AAAAT motifs were abundant and might map to gaps frequently.

The tetra-nucleotide repeats were significantly underrepresented as a group compared to the penta- and tri-nucleotide repeat motifs. The reduction in tetra-nucleotide motifs is not seen among grass genomes (Rota et al. 2005) or non-legume dicots (Paniego et al. 2002). Some tetra-nuclotide motifs were common in the

Table 3 Numbers of microsatellite repeats in soybean genomic DNA. Mapped SSRs (BARC) compared to BES Derived SSRs. (A) The unselected set includes all the motifs found in the BES. (B) The selected set are the satellites within amplicons with useful primers

	TAAT	TTAA	CTGA	CAGT	TAAATT	Other tri-N	Quad to sex	Total motifs
(A) Unselected								
SIUC-BES	290	192	34	58	81	109	289	1,053
BARC-SSR	386	0	24	0	598	11	2	1,021

	Mono-N	Di-N	Tri-N	Quad-N	Quint-N	Sext-N	Total motifs
(B) Selected							
SIUC-BES	154	282	210	14	136	59	853
BARC-SSR	0	410	609	0	2	0	1,021
Polymorphism-BES (%)	55	58	17	0	35	31	42
Polymorphic-SSR (%)		68	69				

legume peanut (Genbank search 2006). There is one tetra-nucleotide motif in the composite map (BARC-Staga002; Song et al. 2004).

Apparently, all tetra-nucleotide motifs and the GC dinucleotide repeat have been strongly selected against in the soybean genome. Long class I tetra- and di-nucleotide repeats were reported clustered around centromeres in tomato (Areshchenkova and Ganal 1999) but not in soybean (Walling et al. 2006). Since the BES used here were derived from a MTP, providing a sequence read about each 70 Kbp, clusters of motifs would be expected to be under-represented in the data set in the manner observed for tetra-nucleotide motifs and the GC di-nucleotide repeat.

Alternately, selection for widespread distribution of penta- and sexta-nucleotide repeats in soybean may be inferred. For example, the CAAAA motif was inferred to be selected for in the legume *Trifolium repens* because it may be involved in breakage-reunion mechanism of tandemly repeating arrays (Ansari et al. 2004). Therefore, satellites containing penta-nucleotide motifs might be selected to be more polymorphic than tetra-nucleotides.

Measurement of polymorphism frequencies suggests that the new markers will be frequently polymorphic in mapping populations. The overall 38% polymorphism rate is somewhat less than that reported for the BARC_SSR markers with the same populations (Njiti et al. 2002; Song et al. 2004; Kazi et al. 2006). Within populations 15% of markers tested were polymorphic in F × H compared to 33% of BARC-SSRs (Yuan et al. 2002; Kazi 2005; Kazi et al. 2006) scored on agarose gels. The 18% polymorphic in ExF compared to the 38% reported (Iqbal et al. 2001; Kassem et al. 2006). The 14% in P × D compared to the 39% reported by Njiti et al. (2002), and 15% in M × N compared to the 41% reported by Song et al. (2004). The lower polymorphism frequencies among the SIUC-BES-SSRs compared to BARC-SSR did not seem to be related to the use of sequencing gels compared to agarose gels (Kazi 2005). Probably, the shorter motif lengths cause the lower overall polymorphism frequencies.

The SIUC-BES-SSR class I motif types showed polymorphism frequencies equal to the class I BARC-SSRs of Song et al. (2004). Even among the mononucleotide motifs (all class I), six of 11 markers (55%) were polymorphic suggesting these markers should have been used for BARC-SSR development.

However, some class II markers were highly polymorphic. Among di-nucleotide motifs 3 of 7 AC markers, 14 of 20 AT markers and 1 of 3 AG markers (60%) were polymorphic (3 class I and 15 class II). Motifs composed primarily of As were highly polymorphic. Several factors may contribute to the polymorphisms in poly-A tracts. Mononucleotide tracts cause DNA polymerase to slow down or stall during synthesis; DNA editing is error prone in such regions; cDNA copied from mRNAs of retrovirus' and incorporated into the nuclear genome carry with them long poly-A tracts. The latter could explain the greater polymorphism of poly-A compared to poly G tracts.

Anchoring BES-SSR to the genetic map proved to be an effective, independent method to check on the assignment of BACs to contig and contigs to linkage groups in the physical map. Build 4 proved to be a more accurate physical map than build 2 or 3, as predicted (Shultz et al. 2006). However, build 4 was not completely correct in marker assignments. Every marker was present in build 3 but only some of the source BACs were included in build 4. The opposite will be true of the MTP4 derived markers under development. As noted (Shultz et al. 2006) the homologous regions of the soybean genome cause most BARC-Satt markers to have multiple amplicons at different locations in the BAC pools used to anchor contigs (Fig. 2). The number of potential amplicons increased as pool complexity decreased. However, markers like and Sac_B09L01 were effective is assigning particular contigs from sets that shared a common anchor to a separate chromosomal location. Markers of this type will greatly improve the quality of the anchors on physical maps.

The large number of BES-SSR markers that mapped into gaps in the genetic map suggest the use of an MTP has reduced the tendency of polymorphic markers to cluster. Although the composite map of soybean currently available on Soybase (March, 2006) contained 3,049 molecular markers, only 1,011 were SSRs. In any single population about 35% can be placed (Njiti et al. 2002; Song et al. 2004; Lightfoot et al. 2005). The maps that result are 30–50% smaller than the known soybean genome size. Gaps in marker coverage are inferred. For example, in the E × F map (Kassem et al. 2006) from the first 50 BES-SSR 25 of the markers were placed into 21 gapped regions on 12 chromosomes (Bashir and Lightfoot, unpublished). Many new QTL were identified. Revising existing maps and the derived composite maps by adding a few hundred BES-SSR markers per population will be productive. A community mapping project will be necessary. The results will strengthen both the genetic and the physical maps of soybean.

SSR markers are highly reliable, co-dominant and simple to use. Consequently, the additional markers reported here should improve coverage of the soybean

genome by maps. The polymorphism rates of the BES-SSRs may be increased if sequencing gels are used to detect amplicons. The 10 Mb of BAC-end sequence used in this marker development project has recently been augmented by seven thousand more BAC-ends and hundreds of thousands of EST sequence. The additional resources will be used to add more markers to SoyGD. A consortium of laboratories are mapping the markers in additional populations (Gore et al. 2002; Yamanaka et al. 2004; Guo et al. 2005) and additional members are invited from the readership.

Acknowledgments This research was funded in part by grants from the NSF 9872635, ISA 95–122–04; 98-122-02 and 02-127-03 and USB 2228-6228. The physical map was based upon work supported by the National Science Foundation under Grant No. 9872635. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. The continued support of SIUC, College of Agriculture and Office of the Vice Chancellor for Research to JA, SK and DAL is appreciated. The authors thank Dr. Q. Tao and Dr, H.B. Zhang for assistance with fingerprinting. We thank Dr. C. Town and Dr. C. Foo at TIGR for the BES.

References

Ansari HA, Ellison NW, Griffiths AG, Williams WM (2004) A lineage-specific centromeric satellite sequence in the genus *Trifolium*. Chromosome Res 12:357–367

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

Cregan PB, Mudge J, Fickus EW, Danesh D, Denny R, Young ND (1999a) Two simple sequence repeat markers to select for soybean cyst nematode resistance conditioned by the *rhg1* locus. Theor Appl Genet 99:811–818

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

Demirbas A, Rector BG, Lohnes DG, Fioritto RJ, Graef GL, Cregan PB, Shoemaker RC, Specht JE (2001) Simple sequence repeat markers linked to the soybean *Rps* genes for phytophthora resistance. Crop Sci 41:1220–1227

Gore MA, Hayes AJ, Jeong SC, Yue YG, Buss GR, Maroof S (2002) Mapping tightly linked genes controlling potyvirus infection at the *Rsv1 and Rpv1* region in soybean. Genome 45:592–599

Guo B, Sleper DA, Arelli PR, Shannon JG, Nguyen HT (2005) Identification of QTLs associated with resistance to soybean cyst nematode races 2, 3 and 5 in soybean PI 90763. Theor Appl Genet. 111:965–971

Iqbal MJ, Meksem K, Njiti VN, Kassem M, Lightfoot DA (2001) Microsatellite markers identify three additional quantitative trait loci for resistance to soybean sudden death syndrome (SDS) in Essex x Forrest RILs. Theor Appl Genet 102:187–192

Kassem MA, Shultz J, Meksem K, Wood AJ, Iqbal MJ, Lightfoot DA (2006) An Updated 'Essex' by 'Forrest' Linkage Map and First Composite Interval Map of QTL Underlying Six Soybean Traits. Theor Appl Genet 113:1015–1026

Kazi S (2005) Minimum tile derived microsatellite markers improve the physical map of the soybean genome and the Flyer by Hartwig Genetic map. MS. MBMB, SIUC. Pp208

Kazi S, Shultz JL, Bashir R, Afzal J, Njiti VN, Lightfoot DA (2006) Identification of loci underlying resistance to soybean sudden death syndrome in 'Hartwig' by 'Flyer'. Theoret Appl Genet (in review)

Lightfoot DA, Njiti VN, Gibson PT, Kassem MA, Iqbal JM, Meksem K (2005) Registration of the Essex by Forrest recombinant inbred line mapping population. Crop Sci 45:1678–1681

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

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 RJ, 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 9:199–207

Meksem K, Zobrist K, Ruben E, Hyten D, Quanzhou T, Zhang HB, Lightfoot DA (2000) Two large-insert soybean genomic libraries constructed in a binary vector: applications in chromosome walking and genome wide physical mapping. Theor Appl Genet 101:747–755

Meksem K, Pantazopoulos P, Njiti VN, Hyten DL, Arelli PR, Lightfoot DA (2001a) 'Forrest' resistance to the soybean cyst nematode is bigenic: saturation mapping of the *Rhg1* and *Rhg4* loci. Theor Appl Genet 103:710–718

Meksem K, Njiti VN, Banz WJ, Iqbal MJ, Kassem MA, Hyten DL, Yuang J, Winters TA, Lightfoot DA (2001b) Genomic regions that underlie soybean seed isoflavone content. J Biomed Biotechnol 1:38–44

Njiti VN, Meksem K, Iqbal MJ, Johnson JE, Kassem MA, Zobrist KF, Kilo VY, Lightfoot DA (2002) Common loci underlie field resistance to soybean sudden death syndrome in Forrest, Pyramid, Essex, and Douglas. Theor Appl Genet 104: 294–300

Paniego N, Echaide M, Munoz M, Fernandez L, Torales S, Faccio P, Fuxan I, Carrera M, Zandomeni R, Suarez EY, Hopp HE (2002) Microsatellite isolation and characterization in sunflower (*Helianthus annuus* L.). Genome 45:34–43

Rota ML, Kantety RV, Yu JK, Sorrells ME (2005) Nonrandom distribution and frequencies of genomic and EST-derived microsatellite markers in rice, wheat, and barley. BMC Genomics 6:23–32

Ruben E, Jamai A, Afzal J, Njiti VN, Triwitayakorn K, Iqbal MJ, Yaegashi S, Arelli PR, Town CD, Meksem K, Lightfoot DA (2006) Genomic analysis of the 'Peking' *rhg1* locus: Candidate genes that underlie soybean resistance to the cyst nematode. Mol Gen Genome 276:320–330. doi: s00438-006-0150-8

Shultz JL, Meksem K, Lightfoot DA (2003a) Evaluating physical maps by clone location comparison. Genome Lett 2:99–107

Shultz JL, Meksem K, Shetty J, Town CD, Koo H, Potter J, Wakefield K, Zhang HB, Wu C, Lightfoot DA (2003b) End sequencing of BACs comprising a provisional tiling path from a fingerprint physical map of soybean (*Glycine max*) cultivar Forrest. Genbank. CG812653 to CG826126 (13,473 sequences)

Shultz JL, Kurunam DJ, Shopinski KL, Iqbal MJ, Kazi S, Zobrist K, Bashir R, Yaegashi S, Lavu N, Afzal A, Yesudas CR, Kassem MA, Wu C, Zhang HB, Town CD, Meksem K, Lightfoot DA (2006) The soybean genome database (SoyGD): a browser for display of duplicated, polyploid regions and

sequence tagged sites on the integrated physical and genetic maps of *Glycine max*. Nucleic Acids Res 34:D758–D765

Song QJ, Marek LF, Shoemaker RC, Lark KG, Concibido VC, Delannay X, Specht JE, Cregan PB (2004) A new integrated genetic linkage map of the soybean. Theor Appl Genet 109:122–128

Temnykh SG, DeClerck A, Lukashova L, Lipovich S, Cartinhour S, McCouch S (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

Triwitayakorn K, Njiti VN, Iqbal MJ , Yaegashi S, Town CD, Lightfoot DA (2005) Genomic analysis of a region encompassing *QRfs1* and *QRfs2*: genes that underlie soybean resistance to sudden death syndrome. Genome Génome 48:125–138

Walling JG, Shoemaker RC, Young ND, Mudge J, Jackson SA (2006) Chromosome level homeology in paleopolyploid soybean (*Glycine max*) revealed through integration of genetic and chromosome maps. Genetics 172: 1893–1900

Wu CS, Sun P, Nimmakayala P, Santos FA, Meksem K, Springman R, Ding K, Lightfoot DA, Zhang HB (2004) A BAC- and BIBAC-based physical map of the soybean genome. Genome Res 14:319–26

Yuan J, Njiti VN, Meksem K, Iqbal MJ, Triwitayakorn K, Kassem MA, Davis GT, Schmidt ME, Lightfoot DA (2002) Quantitative trait loci in two soybean recombinant inbred line populations segregating for yield and disease resistance. Crop Sci 42:271–277

Yamanaka N, Ninomiya S, Hoshi M, Tsubokura Y, Yano M, Nagamura Y, Sasaki T, Harada K (2001) An informative linkage map of soybean reveals QTLs for flowering time, leaflet morphology and regions of segregation distortion. DNA Res 8:61–72

Zhang WK, Wang YJ, Luo GZ, Zhang JS, He CY, Wu XL, Gai JY, Chen SY (2004). QTL mapping of ten agronomic traits on the soybean (*Glycine max* L. Merr.) genetic map and their association with EST markers. Theor Appl Genet 108:1131–9

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 163:1123–34