

## Selection of a core set of RILs from Forrest × Williams 82 to develop a framework map in soybean

Xiaolei Wu · Tri D. Vuong · Jill A. Leroy ·  
J. Grover Shannon · David A. Sleper ·  
Henry T. Nguyen

Received: 27 July 2010 / Accepted: 11 December 2010 / Published online: 19 January 2011  
© The Author(s) 2011. This article is published with open access at Springerlink.com

**Abstract** Soybean BAC-based physical maps provide a useful platform for gene and QTL map-based cloning, EST mapping, marker development, genome sequencing, and comparative genomic research. Soybean physical maps for “Forrest” and “Williams 82” representing the southern and northern US soybean germplasm base, respectively, have been constructed with different fingerprinting methods. These physical maps are complementary for coverage of gaps on the 20 soybean linkage groups. More than 5,000 genetic markers have been anchored onto the Williams 82 physical map, but only a limited number of markers have been anchored to the Forrest physical map. A mapping population of Forrest × Williams 82 made up of 1,025 F<sub>8</sub> recombinant inbred lines (RILs) was used to construct a reference genetic map. A framework map with almost 1,000 genetic markers was constructed using a core set of these RILs. The core set of the population was evaluated with the theoretical population using equality, symmetry and representativeness tests. A high-resolution genetic map will allow integration and utilization of the physical maps

to target QTL regions of interest, and to place a larger number of markers into a map in a more efficient way using a core set of RILs.

### Introduction

Soybean [*Glycine max* (L.) Merr.] represented 56% of world oilseed production in 2007 (<http://www.soystats.com>). Soybeans provide 71% of the edible oils, and have unique benefits in agricultural ecology (e.g. biological nitrogen fixation), documented health benefits (e.g. the anticancer benefits of isoflavones and lunasin) and industrial utilization (e.g. biodiesel). However, soybean production is being challenged due to increasing demands for soybeans in food, feed and value-added products. Soybean breeders around the world are working to improve varieties with better nutritional quality, biotic and abiotic stress tolerance and higher yields. Molecular breeding technologies are increasingly being applied to develop genetic linkage maps and to identify genomic regions influencing traits related to soybean production and seed value, e.g., soybean cyst nematode resistance, high oleic acid, and low linolenic acid.

During the last two decades, both genomic mapping and sequencing methods have advanced significantly to provide tools for scientists to explore genome structure and function in many organisms. Generally speaking, genome mapping relies on genome sequencing to provide location-unique molecular markers to construct a blueprint for understanding genome structure and function. Genetic linkage maps are an essential prerequisite for studying the inheritance of both qualitative and quantitative traits, for molecular breeding and map-based gene cloning, and for genome structure and function studies. Molecular breeding

Communicated by R. Varshney.

**Electronic supplementary material** The online version of this article (doi:10.1007/s00122-010-1522-3) contains supplementary material, which is available to authorized users.

X. Wu · T. D. Vuong · J. A. Leroy · D. A. Sleper ·  
H. T. Nguyen (✉)  
Division of Plant Sciences, National Center for Soybean  
Biotechnology, University of Missouri,  
Columbia, MO 65211, USA  
e-mail: [nguyenhenry@missouri.edu](mailto:nguyenhenry@missouri.edu)

J. Grover Shannon  
Division of Plant Sciences, University of Missouri,  
Delta Center, P.O. Box 160, Portageville, MO 63873, USA

is more effective if the genetic map is densely populated with markers, which would increase the probability of successful trait introgression by transferring defined chromosomal fragments containing target gene(s) and eliminating the linkage drags associated with unfavorable traits. Soybean genome mapping based on the DNA markers began in the early 1990s and numerous genetic linkage maps of soybean have been published in the last decade. The early maps were based primarily on the restriction fragment length polymorphism (RFLP) markers, with more recent maps also including amplified fragment length polymorphism (AFLP) and simple sequence repeats (SSR) and very recently, single nucleotide polymorphism (SNP) markers. In total, several thousand genetic markers (mostly SSR and SNP markers) have been mapped in the past 10 years (Cregan et al. 1999; Wu et al. 2001; Song et al. 2004; Kassem et al. 2006; Choi et al. 2007; Xia et al. 2007; Hisano et al. 2008; Yang et al. 2008). With those markers, more than one thousand quantitative trait loci (QTL) associated with important soybean traits have been identified using different mapping populations (<http://www.soybase.org>).

The soybean cultivars “Williams 82” and “Forrest”, representing Northern and Southern germplasm in the United States, respectively, have been used as models for soybean genomic research in the same way as the ecotypes *Col* and *Ler* in *Arabidopsis thaliana* (Lister and Dean 1993) or cultivars Mo17 and B73 in maize (*Zea mays*) (Sharopova et al. 2002). The soybean community in the USA has the majority of soybean genomic tools based on the using the two cultivars, Forrest and Williams 82, which were recently reviewed in the literature (Jackson et al. 2006; Lightfoot 2008). The Williams 82 genome was sequenced by the Department of Energy, Joint Genome Institute (Schmutz et al. 2010), and the 8× scaffold assembly was released to the public (<http://www.phytozome.net/soybean>). Williams 82 and Forrest not only provide us numerous ‘-mics’ data, but also are important ancestors of some modern cultivars because they carry useful genes (Lightfoot 2008). However, there is no genetic map available to date developed from a population of Forrest and Williams 82.

In this study, we report the construction of a framework genetic map using a core set of RILs selected from a large Forrest × Williams 82 mapping population, to serve as a reference mapping population in the soybean genomics community. This genetic map also offers opportunities to link the existing genetic maps to the “Williams 82” and “Forrest” physical maps ([soybase.org](http://www.soybase.org); Wu et al. 2004). The mapped markers from the core set of the population would help with the integration of the two physical maps to the draft sequence as framework for genomic research in soybean.

## Materials and methods

### Plant materials

The mapping population used in this study consisted of 1,025 F<sub>2:7</sub> recombinant inbred lines (RILs) derived from the F<sub>1</sub> seed from a cross between the cultivars Forrest (female parent) and Williams 82 (male parent). Each RIL tracing to a single F<sub>2</sub> plant was advanced to the F<sub>7</sub> generation by the single seed descent (SSD) method. In the F<sub>8</sub> generation, 1 m rows of each F<sub>2:7</sub> RIL were individually bulked. Young leaves from each F<sub>2</sub> plant and ~4–8 plants of each corresponding RIL were collected, freeze-dried and ground for DNA isolation. To evaluate the genetic structure of the population before advancing to late generations, a subset of 760 plants in F<sub>2</sub> generation were genotyped with 295 SSR markers covering the whole genome.

### DNA isolation

The genomic DNA was extracted with an AutoGenprep 965/960 machine (AutoGen, Holliston, MA, USA) using the AGP965/960 Plant DNA Extraction Kit, following the manufacturer’s instructions. The DNA was quantified by a NanoDrop spectrophotometer (NanoDrop Technologies Inc., Centreville, DE, USA) and normalized to 25 ng/μl as PCR working template.

### Marker development

New SSR markers were identified from 25,640 Forrest sequences downloaded from the National Center for Biotechnology Information (NCBI) (<http://www.ncbi.nlm.nih.gov/>). Those sequences were analyzed with microsatellite identification tool—MISA (Thiel et al. 2003) to identify SSRs with at least six unit repeats for di-nucleotide repeats, five unit repeats for tri-nucleotide and above. Primers flanking the SSRs were designed using BatchPrimer 3 (You et al. 2008) in such a way that the amplified PCR products ranged between 150 and 400 bp in length.

New Indel markers were identified by blasting the 25,640 Forrest DNA sequences against the soybean Williams 82 genome sequence (<http://www.phytozome.net>) with a cutoff *E* value at E-50. If Indels were identified with more than 2 bp of size difference, primers flanking these Indel were designed using the same software with the same parameter settings as SSR primer design (You et al. 2008).

### Genotyping

For SSR genotyping, an M13 primer sequence was added to the forward primer to allow detection with a common fluorescently labeled (FAM, PEP, NED or HEX) M13 primer as

previously described (Schuelke 2000). The polymerase chain reactions (PCR) were carried out in a final volume of 15  $\mu$ l on Eppendorf thermocycler with a thermal profile consisting of a 3-min initial denaturation step at 95°C followed by 35 cycles of 30 s at 94°C, 30 s at 52°C and 45 s at 72°C, and a final 7-min extension step at 72°C. Reactions were conducted in 10 mM Tris–HCl (pH 8.4 at 25°C); 50 mM KCl; 0.1% (v/v); 2.5 mM MgCl<sub>2</sub>; 1% PVP-40; 200  $\mu$ M dNTPs in the presence of 1 U *Taq* polymerase; 50-ng genomic DNA; 0.5  $\mu$ M marker-specific reverse primer; 0.033  $\mu$ M marker-specific M13-tailed forward primer and 0.5  $\mu$ M FAM, PEP, NED or HEX-labeled M13 primer. PCR products were 8-plexly pooled into a Whatman 384-well DNA-binding plate for clean-up following the instructions of the manufacturer. The PCR products were eluted with 10  $\mu$ l of formamide into 384-well plate and run on an ABI 3100 or 3730xl along with LIZ-labeled 500 bp size standard. Genotyping was performed by GeneMapper (3.7) with a 0.5–1.5 ratio as heterozygous allele.

For SNP genotyping, a universal soybean linkage panel (USLP 1.0) containing 1,536 SNP marker (Hyten et al. 2010) was employed to genotype the F<sub>2:7</sub> RIL mapping population following an Illumina GoldenGate assay as described by Fan et al. (2006). In brief, after the completion of a multiple-step procedure, such as genomic DNA activation, PCR amplification, hybridization, and washing, array imaging was performed using the Illumina BeadStation (Illumina, San Diego, CA, USA) to generate intensity data. The allele calling for each SNP locus was subsequently conducted with the BeadStudio 3.0 software (Illumina, San Diego, CA, USA). The clusters of homozygous and heterozygous genotypes for each SNP were manually checked for polymorphisms between the two parental lines. The polymorphic SNP loci were then employed for genetic linkage map construction and QTL analysis as earlier described (Vuong et al. 2010).

#### Mapping population evaluation

A set of tests, including equality test, symmetry test, and representativeness test, using chi-squared test ( $\chi^2$ ) as the criteria were used to test the probability that the parameters of the core set of RILs matched the theoretical RIL population, supposed that the distortion of the population was mainly due to the shifting environments during various seasons of generation advancement rather than the skewed selection of gametes and zygotes during meiosis (Gai et al. 2007). The theoretical values of a RIL population were estimated from 200 simulated populations with the software QGENE (Nelson 1997). The parameter settings for the simulated map and populations were adapted with the real values: population size = 376, chromosome number = 20, map length = 2,600 cM, proportion of missing

genotypes = 2.5%, proportion of parent A (Forrest) allele = 45.7%, proportion of parent B (Williams 82) allele = 47.1%, and proportion of heterozygous allele = 4.7%. Two scenarios of simulations were performed under fixed chromosome length (150 cM) and variable length with statistical distribution based on parameter settings.

#### Linkage map construction

Linkage group (LG) marker order and map distance were calculated using the software JoinMap (version 4.0). The segregated markers were grouped in LGs on the basis of an LOD (logarithm of the odds ratio for linkage) score of  $\geq 4.0$  and referral to previously reported LGs of the public SSR marker loci. Marker order of known SSR markers in the soybean genome was used for fixed order to construct an LG. Markers were tested for deviation from expected Mendelian segregation by the chi-squared test performed with the JoinMap software under the ‘Locus Genotypic Frequency’ command. Linkage between markers, recombination rate ( $Q$ ), and map distances were calculated using the Kosambi mapping function.

## Results

#### BES-derived markers for map integration

About 25,640 Forrest sequence entries, including 25,463 BAC end sequences (BES) from the minimum tile path (MTP) clones of the Forrest physical map (Shultz et al. 2006) and 146 Forrest EST sequences posted in NCBI (<http://www.ncbi.nlm.nih.gov/>) were used to search SSR motif repeats. The whole set of Forrest sequences was also used to identify Indel markers with more than two continuous nucleotide deletions or insertions after filtering out the repetitive elements using RepeatMasker (<http://www.repeatmasker.org>). The repeat-masked sequences were then used to blast the Williams 82 genome sequence 8 $\times$  scaffold assembly to in silico map these BESs onto the Williams 82 genome (<http://www.phytozome.net/soybean>). We extracted all aligned sequences with  $\geq 2$  bp difference of continuous nucleotides between Forrest BES and Williams 82 genome sequence. A total of 3,272 putative polymorphic markers, including 581 SSRs with 2–7 bp motif repeats and 2,691 Indels, were identified (Additional File 1). Of the 3,272 markers, 3,015 markers were anchored onto the Forrest physical map (Shultz et al. 2006). Based on the locations of the new polymorphic markers in the Williams 82 genome and the polymorphism test on public SSRs (Song et al. 2004), primer pairs were designed for 175 putative BES-derived markers that could fill the gaps of the genetic map (Additional File 2). Of the

175 tested markers, 127 markers were validated to be polymorphic by genotyping parental lines, showing 71.8% polymorphism. A total of 114 markers derived from Forrest BES or EST were mapped in this study. These 114 markers could serve as new anchors for Forrest physical map integration with the Williams 82 genome sequence assembly through the Forrest  $\times$  Williams 82 genetic map. An example is shown in Fig. 1.

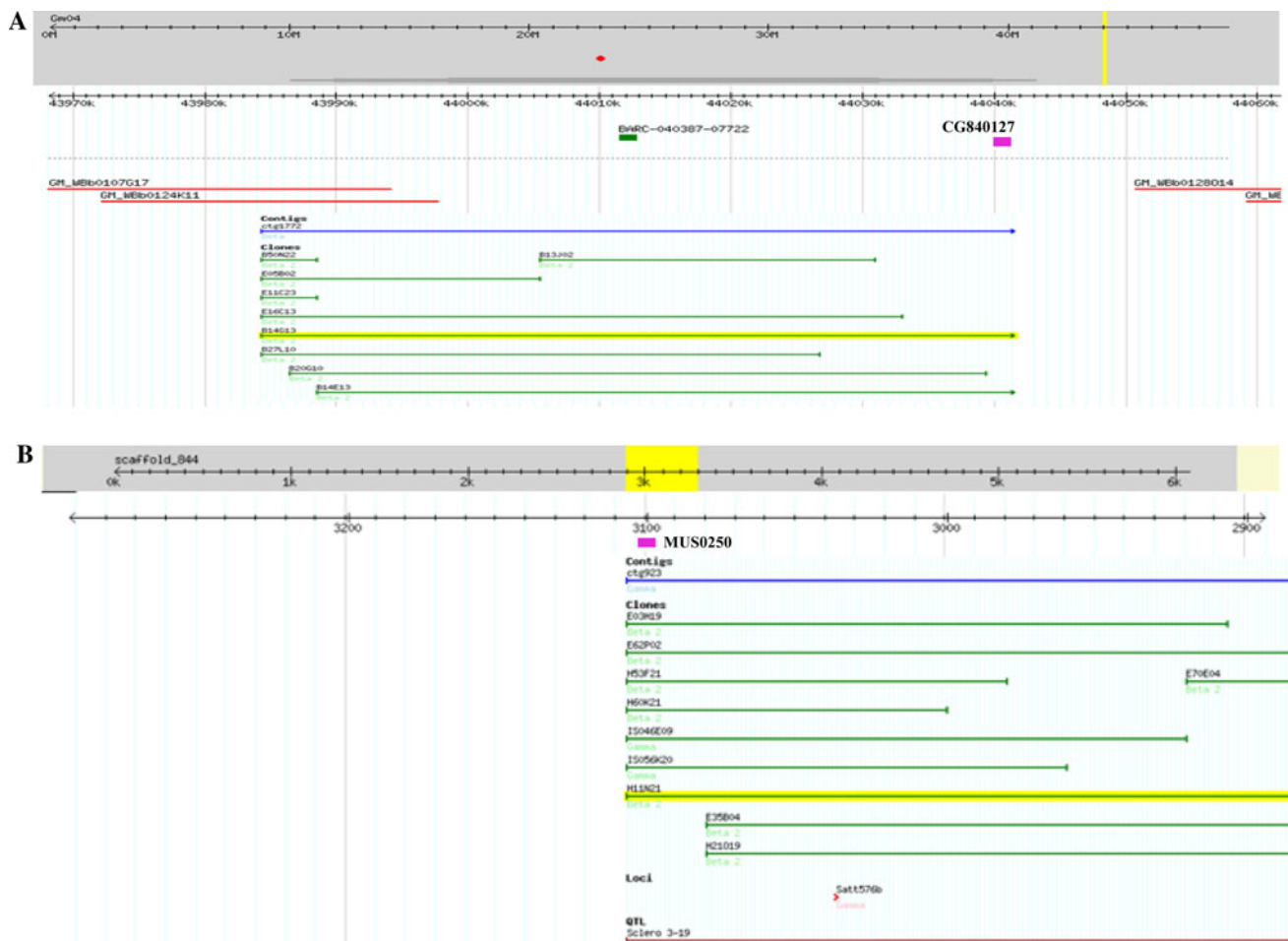
#### Selection of a core set population

A total of 1,025  $F_2$  families were advanced by SSD until  $F_7$  generation. The core set comprising 376 RILs was selected according to a framework map with 295 mapped SSR markers which were evenly distributed across linkage groups in the composite genetic map (Choi et al. 2007). The procedures for selecting a core set of RILs were as

follows: all markers were sorted in same order as in the genetic map, and the number of crossovers and valid markers in each line was calculated to get the exchange frequency rate of the genome. The lines were sorted and selected in order based on the number of the crossovers, the number of valid markers and the exchange frequency rate. Lines that provided unique crossover breakpoints were selected to be part of the core set. The core set includes all of the crossover breakpoints detected by the 295 SSR markers across the whole genome in order to adequately represent the original large mapping population.

#### Mapping population evaluation

A total of 1,029 polymorphic markers segregated in this population, including 532 SNP markers from the Universal 1,536 Soy Linkage Panel (Hyten et al. 2010) and 497 SSR



**Fig. 1** Examples of Forrest BAC contig integration to Williams 82 genome assembly via genetic markers derived from BESs. **a** Forrest FPC contig1772 was anchored by the new genetic marker CG840127 to a 53 kb gap of the Williams 82 physical map integrated on the sequence assembly, **b** Forrest FPC contig923 was anchored by marker Satt576 onto the LG O (chromosome 10) covering a QTL for

*Sclerotinia* stem rot resistance. This contig anchoring on the LG was confirmed by the new SSR marker MUS0250, but the BES sequence that was used to develop the MUS0250 marker was aligned to scaffold\_844, which indicated that scaffold\_844 might belong to chromosome 10

markers (349 public SSRs and 148 new SSR and Indel markers developed from this research) were used to evaluate the 376 core set RILs. The RILs can be classified as three categories according their genotypes: families with same parental allele frequency, families with dominant maternal allele frequency ( $p > 0.5$ ), and families with dominant paternal allele frequency ( $q > 0.5$ ). Based on the characterizations of the RIL population—equal distribution of two parental alleles, a set of statistical tests, including equality, symmetry, and representativeness tests, were used to examine the coincidence and derivation of the core set of population against the theoretical RIL population (Gai et al. 2007). The equality test examined whether the ratio of the total number of two parental alleles fit the expected theoretical value 1:1. The symmetry test examined the fitness of the 1:1 ratio for the total number of families with maternal allele frequency  $>0.5$  to the number of the families with paternal allele frequency  $>0.5$ . The representativeness test examined whether each family as well as the whole experimental sample was a random sample from the corresponding theoretical population, by comparing the rates of families with extreme-biased segregation or of markers with extreme-biased distortion to the ones with the expected rate obtained from the simulated populations (Gai et al. 2007). The results of the tests are shown in Table 1. The equality test showed a larger  $\chi_c^2$  value, indicating unequal genetic contribution from both parents. But the symmetry test showed the different categories of families in the population were balanced since the  $\chi_c^2$  values were  $<3.84$  (5% significance threshold value, when degree of freedom is 1), indicating the numbers of families basically fit a 1:1 ratio. The representativeness test of markers showed some markers had serious distortion so that the extreme-biased rate was larger than the expected value of the simulated population; but the rate of extreme-biased families was 40.4%, lower than the expected values of the simulated population in scenario 1 (45.4%) and scenario 2 (41.0%).

#### Framework map construction

Based on the theoretical requirements for genetic population analysis and the three tests as mentioned above, we removed 16 families with the highest  $\chi_c^2$  or with more than 20% heterozygous alleles and 18 markers with the most extreme-biased distortion or with more than 40% missing data. Finally, a total of 986 markers, which included 471 SSRs and 515 SNPs were used to construct a framework map in which a total of 145 newly developed SSR or Indel markers were integrated (Additional File 2). Except for chromosomes 5 and 16 which had two unmerged linkage groups, the other 18 chromosomes constituted one LG for each chromosome (Table 2). Thus, the 986 markers were

assembled to 22 LGs with a total genetic map length of 2,723.8 cM (Fig. 2).

## Discussion

### Mapping population

A limitation in QTL mapping is that the markers closely linked to genes controlling important traits are difficult to identify because the meiotic recombination events along chromosomes are limited in a regular mapping population which prevents the map from reaching a high resolution. Even though in theory using a larger mapping population will result in a genetic map with better resolution, in practice, it is not possible to construct a high density genetic map using a relatively large number of plants from a population due to the genotyping cost and mapping efficiency. Usually, a smaller set of RILs in a mapping population is initially used in QTL mapping to identify the genomic regions associated with targeted traits. Once the associated regions are identified, further mapping or map-based cloning can be pursued to identify markers more closely linked to genes. The major disadvantage of mapping genes from a smaller number of individuals is that the associations detected using a smaller sized population sometimes is not representative (Yan et al. 2006). Thus, QTL cannot be confirmed using a larger number of individuals. Thus, to overcome the disadvantages of using a small population, we developed a large RIL population derived from a cross of Forrest and Williams 82 cultivars, which represent southern and northern US soybean germplasm, respectively. We modified the ‘selective strategy’ proposed by Vision et al. (2000) to select an optimal subset (core set) of the population consisting of 376 RILs from the original 1,025 F<sub>2</sub> Forrest  $\times$  Williams 82 families. Using an optimal core set of RILs will reduce the cost in genotyping a large mapping population and increase the mapping efficiency.

Selective mapping strategy has been widely used to improve the efficiency of mapping genetic markers (Vision et al. 2000; Howad et al. 2005; Han et al. 2009; Sargent et al. 2008). It consists of a two-step process in which, first, a subset of highly informative plants is selected based on a framework genetic map of a mapping population, and second, new markers are added to this map using the subset population. In this study, the method used for selection of the core set population was initially used for mapping AFLP markers into a RFLP and SSR map to construct a high density map (Wu et al. 2001), and anchoring chromosome-specific markers to a *Leymus* AFLP map (Wu et al. 2003). Following the same approach, we selected the core set of individuals according to the most breakpoints

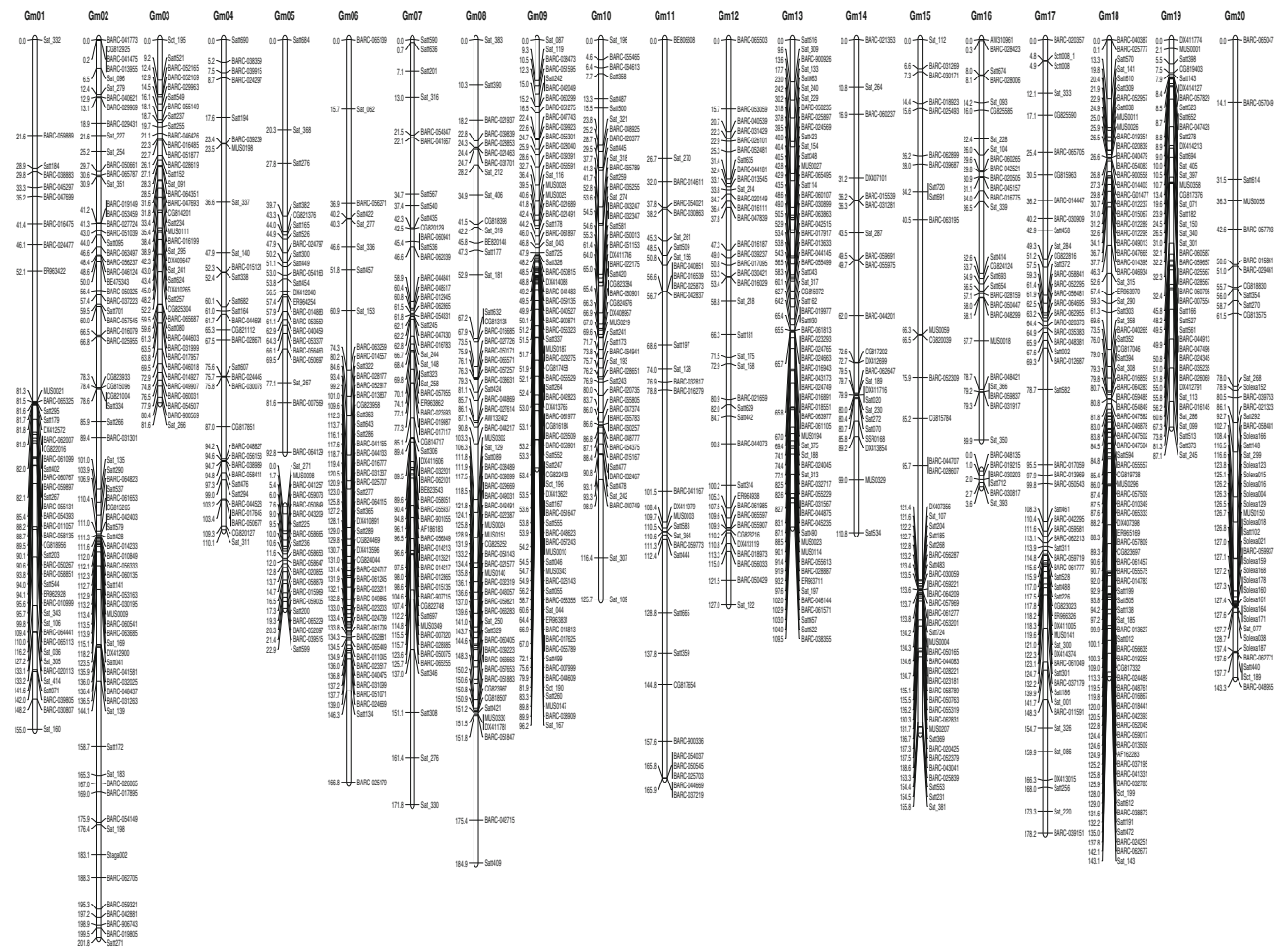


**Table 1** A comparison of core set and original population for equality, symmetry and representativeness tests

Test	Group		Value of core set	Simulated population ( $\alpha = 0.05$ )	
				Scenario 1	Scenario 2
Equality	Total no. of effective loci	Forrest ( $p$ )	172,576	196,890 $\pm$ 1,250	193,485 $\pm$ 6,067
		Williams 82 ( $q$ )	177,728	195,492 $\pm$ 1,109	193,291 $\pm$ 5,798
		$\chi_c^2$	17.75	4.98	15.9
Symmetry	Family number	$p > 0.5, q < 0.5$	170	192 $\pm$ 8	188 $\pm$ 8
		$p < 0.5, q > 0.5$	206	183 $\pm$ 9	187 $\pm$ 8
		$p = q = 0.5$	0	1.2 $\pm$ 1.3	1.3 $\pm$ 0.9
		$\chi_c^2$	3.45	0.11	0.01
		$\chi_c^2 < 3.84$			
	$\chi_c^2 > 3.84$	$p > 0.5, q < 0.5$	73	65.6 $\pm$ 7.1	70.4 $\pm$ 6.4
		$P < 0.5, q > 0.5$	75	61.2 $\pm$ 4.8	70.1 $\pm$ 5.9
		$\chi_c^2$	0.03	1.4	0.0006
		$p > 0.5, q < 0.5$	97	126.2 $\pm$ 6.9	117.7 $\pm$ 7.8
		$p < 0.5, q > 0.5$	131	121.8 $\pm$ 7.1	116.6 $\pm$ 7.3
Representativeness	Marker	$\chi_c^2$	2.54	1.3	0.024
		Number	1,005	1,060	1,044.8 $\pm$ 32.0
		Extreme-biased rate	18.2	6.0	5.3
	Family	Largest $\chi_c^2$	217.5	11.6	17.9
		Number	376	376	376
		Extreme-biased rate	40.4	45.4 $\pm$ 1.4	41.0 $\pm$ 2.3
		Largest $\chi_c^2$	223.1	203.4	157.1

**Table 2** Summary of the genetic map constructed with 990 markers using a core set of RILs selected from a large Forrest  $\times$  Williams 82

Chromosome	LG	Number of markers			Length (cM)	Average interval (cM)
		SSR	SNP	Total		
Gm01	D1a	21	23	44	155.0	3.5
Gm02	D1b	31	44	75	201.8	2.7
Gm03	N	19	21	40	81.6	2.0
Gm04	C1	15	16	31	110.1	3.6
Gm05	A1	19	24	43	115.7	2.7
Gm06	C2	19	29	48	166.8	3.5
Gm07	M	27	31	58	171.8	3.0
Gm08	A2	28	33	61	184.9	3.0
Gm09	K	35	37	72	96.2	1.3
Gm10	O	25	24	49	125.7	2.6
Gm11	B1	15	17	32	165.9	5.2
Gm12	H	13	24	37	127.0	3.4
Gm13	F	27	36	63	109.5	1.7
Gm14	B2	18	9	27	110.8	4.1
Gm15	E	19	29	48	155.8	3.2
Gm16	J	16	17	33	93.5	2.8
Gm17	D2	30	25	55	178.2	3.2
Gm18	G	34	52	86	143.1	1.7
Gm19	L	30	14	44	87.1	2.0
Gm20	I	33	11	44	143.3	3.3
Total		474	516	990	2,723.8	
Average		23.7	25.8	49.5	136.2	2.9



**Fig. 2** Genetic linkage map of soybean constructed with SSR, Indel and SNP markers. The linkage map was visualized graphically with MapChart

and unique breakpoints instead of bin size proposed by Vision et al. (2000). Thus, we can capture all informative individuals carrying all breakpoints detected by the framework map, and the core set of individuals is a good representation of the whole mapping population.

**Marker development**

Integration of the Forrest physical map with Williams 82 genome assembly via a high density of genetic map is a very important step for utilizing valuable genomic resources developed in Forrest (Lightfoot 2008). BAC library screening has been used for integration of physical map and genetic maps in soybean (Wu et al. 2008). However, this strategy is time-consuming and can generate false positives because of duplication in the soybean genome. In contrast, genetic markers derived from BESs can anchor corresponding BAC clones or contigs onto a genetic map without BAC library screening. Therefore, development of BES-based markers is a promising tool for constructing

integrated physical and genetic maps. Although BAC-end sequence-based SSR markers have been successfully used to develop genetic maps in cotton (Frelchowski et al. 2006), soybean (Shultz et al. 2007; Shoemaker et al. 2008), and apple (Han et al. 2009), the polymorphism between two parental lines in a given mapping population is very low, 13–15% (Shultz et al. 2007; Shoemaker et al. 2008), which prevents genetic mapping of larger number of BES-based SSR markers in one mapping population. In this study, we took advantage of the available Williams 82 genome sequence to identify the potential polymorphic SSRs and Indels derived from BESs of the minimum tilling path (MTP) of the Forrest physical map. The polymorphism of 175 putative polymorphic SSRs or Indels was 71.8% (Additional File 2). Therefore, this approach has dramatically improved the efficiency of polymorphic marker identification and mapping in the Forrest × Williams 82 population. Moreover, selection of the SSR or Indel markers for mapping was more specific. The unmerged linkage groups in this study indicated that a big gap

existed in chromosomes 5 and 16, respectively, that may reflect either a lack of polymorphic markers in a highly homozygous region or the presence of hot spots of recombination that enlarge the genetic distance corresponding to a short physical distance (Hwang et al. 2009). The new polymorphic markers that could be mapped into gaps in the genetic map were preferentially selected to fill gaps and reduce the tendency of marker clustering. In addition to the identified SSR and Indel markers derived from Forrest BESs, we also predicted SNPs from Forrest BESs by comparing them with the Williams 82 genome sequence. Moreover, using high-throughput Solexa sequencing technology, we have discovered thousands of SNPs between Forrest and Williams 82 (unpublished). These SNPs will be mapped to this population using customized high-throughput SNP genotyping arrays.

### Map integration

The Williams 82 soybean physical map has been integrated with genetic map via BAC clone screening with the known genetic markers (Wu et al. 2008) or genetic mapping SSRs derived from BESs of Williams 82 BAC library (Shoemaker et al. 2008). The precise placement of BAC contigs was hindered by multiple hits of FPC contigs caused by soybean genome duplication or false positives. To build a reliable integrated physical/genetic map, a high-resolution genetic map is needed. The markers especially developed from BES of BAC clones consisting of the BAC FPC contigs (i.e. physical map) would be especially useful for map integration. Because the Williams 82 genome sequence is available, the BES itself can help place some of BAC contigs on the Williams 82 genome sequence. Because only the BESs of MTP were available in the Forrest physical map, the power of anchoring BAC clones by sequence alignment was not as great as Williams 82 BACs in which all BAC ends were sequenced. To obtain high-quality integrated physical maps of Forrest and Williams 82, all BAC ends should be sequenced, so that would provide multiple anchors from any FPC contigs and provide more sequences for SSR, Indel or SNP marker identification.

In summary, we selected a core set of a Forrest  $\times$  Williams 82 population that was used for building a high-resolution genetic map. A framework map was constructed to demonstrate the integration of Forrest physical map with the Williams 82 maps. The results showed that the Forrest BACs indeed could fill gaps or help anchor the unanchored sequence scaffolds to improve the quality and coverage of Williams 82 genome sequence assembly.

**Acknowledgments** This research was supported by the National Center for Soybean Biotechnology. We thank Kerry M. Clark for the

RIL population development used in this study. We also thank Sean Blake, Cuilan Liu, Vinh Thi Nguyen, and P. Rajesh for assistance with DNA isolation and PCR preparation.

**Open Access** This article is distributed under the terms of the Creative Commons Attribution Noncommercial License which permits any noncommercial use, distribution, and reproduction in any medium, provided the original author(s) and source are credited.

### References

- Choi IY, Hyten DL, Matukumalli LK, Song Q, Chaky JM, Quigley CV, Chase K, Lark KG, Reiter RS, Yoon MS, Hwang EY, Yi SI, Young ND, Shoemaker RC, van Tassell CP, Specht JE, Cregan PB (2007) A soybean transcript map: gene distribution, haplotype and single-nucleotide polymorphism analysis. *Genetics* 176:685–696
- Cregan PB, Jarvik T, Bush AL, Shoemaker RC, Lark KG, Kahler AL, Kaya N, VanToai TT, Lohnes DG, Chung J, Specht JE (1999) An integrated genetic linkage map of the soybean genome. *Crop Sci* 39:1464–1490
- Fan J, Gunderson K, Bibikova M, Yeakley J, Chen J, Garcia EW, Lebruska L, Laurent M, Shen R, Barker D (2006) Illumina universal bead arrays. *Methods Enzymol* 410:57–73
- Frelichowski JE, Palmer MB, Main D, Tomkins JP, Cantrell RG, Stelly DM, Yu J, Kohel RJ, Ulloa M (2006) Cotton genome mapping with new microsatellites from Acala ‘Maxxa’ BAC-ends. *Mol Gen Genomics* 275:479–491
- Gai J, Wang Y, Wu X, Chen S (2007) A comparative study on segregation analysis and QTL mapping of quantitative traits in plants—with a case in soybean. *Front Agric China* 1:1–7
- Han Y, Chagne D, Gasic K, Rikkerink EHA, Beever JE, Gardiner SE, Korban SS (2009) BAC-end sequence-based SNPs and Bin mapping for rapid integration of physical and genetic maps in apple. *Genomics* 93:282–288
- Hisano H, Sato S, Isobe S, Sasamoto S, Wada T, Matsuno A, Fujishiro T, Yamada M, Nakayama S, Nakamura Y, Watanabe S, Harada K, Tabata S (2008) Characterization of the soybean genome using EST-derived microsatellite markers. *DNA Res* 14:271–281
- Howad W, Yamamoto T, Dirlwanger E, Testolin R, Cosson P, Cipriani G, Monforte AJ, Georgi L, Abbott AG, Arus P (2005) Mapping with a few plants: using selective mapping for microsatellite saturation of the *Prunus* reference map. *Genetics* 171:1305–1309
- Hwang TY, Sayama T, Takahashi M, Takada Y, Nakamoto Y, Funatsuki H, Hisano H, Sasamoto S, Sato S, Tabata S, Kono I, Hoshi M, Hanawa M, Yano C, Xia Z, Harada K, Kitamura K, Ishimoto M (2009) High-density integrated linkage map based on SSR markers in soybean. *DNA Res* 16:213–225
- Hyten DL, Choi IY, Song Q, Specht JE, Carter TE, Shoemaker RC, Hwang EY, Matukumalli LK, Cregan PB (2010) A high density integrated genetic linkage map of soybean and the development of a 1,536 universal soy linkage panel for QTL mapping. *Crop Sci* 50:960–968
- Jackson SA, Rokhsar D, Stacey G, Shoemaker RC, Schmutz J, Grimwood J (2006) Toward a reference sequence of the soybean genome: a multiagency effort. *Crop Sci* 46:S55–S61
- Kassem MA, Shultz J, Meksem K, Cho Y, 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



- Lightfoot DA (2008) Soybean genomics: developments through the use of cultivar “Forrest”. *Int J Plant Genomics* 2008:793158
- Lister C, Dean C (1993) Recombinant inbred lines for mapping RFLP and phenotypic markers in *Arabidopsis thaliana*. *Plant J* 4:745–750
- Nelson JC (1997) QGENE: software for marker-based genomic analysis and breeding. *Mol Breed* 3:239–245
- Sargent DJ, Cipriani G, Vilanova S, Gil-Ariza D, Arus P, Simpson DW, Tobutt KR, Monfort A (2008) The development of a bin mapping population and the selective mapping of 103 markers in the diploid *Fragaria* reference map. *Genome* 51:120–127
- Schmutz J, Cannon SB, Schlueter J, Ma J, Mitros T, Nelson W, Hyten DL, Song Q, Thelen JJ, Cheng J, Xu D, Hellsten U, May GD, Yu Y, Sakurai T, Umezawa T, Bhattacharyya MK, Sandhu D, Valliyodan B, Lindquist E, Peto M, Grant D, Shu S, Goodstein D, Barry K, Futrell-Griggs M, Abernathy B, Du J, Tian Z, Zhu L, Gill N, Joshi T, Libault M, Sethuraman A, Zhang XC, Shinozaki K, Nguyen HT, Wing RA, Cregan P, Specht J, Grimwood J, Rokhsar D, Stacey G, Shoemaker RC, Jackson SA (2010) Genome sequence of the palaeopolyploid soybean. *Nature* 463:178–183
- Schuelke M (2000) An economic method for the fluorescent labeling of PCR fragments. *Nat Biotechnol* 18:233–234
- Sharopova N, McMullen MD, Schultz L, Schroeder S, Sanchez-Villeda H, Gardiner J, Bergstrom D, Houchins K, Melia-Hancock S, Musket T, Duru N, Polacco M, Edwards K, Ruff T, Register JC, Brouwer C, Thompson R, Velasco R, Chin E, Lee M, Woodman-Clikeman W, Long MJ, Liscum E, Cone K, Davis G, Coe EH Jr (2002) Development and mapping of SSR markers for maize. *Plant Mol Biol* 48:463–481
- Shoemaker RC, Grant D, Olson T, Warren WC, Wing R, Yu Y, Kim H, Cregan P, Joseph B, Futrell-Griggs M, Nelson W, Davito J, Walker J, Wallis J, Kremitski C, Scheer D, Clifton SW, Graves T, Nguyen H, Wu X, Luo M, Dvorak J, Nelson R, Cannon S, Tomkins J, Schmutz J, Stacey G, Jackson S (2008) Microsatellite discovery from BAC end sequences and genetic mapping to anchor the soybean physical and genetic maps. *Genome* 51:294–302
- Shultz JL, Kurunam D, Shopinski K, Iqbal MJ, Kazi S, Zobrist K, Bashir R, Yaegashi S, Lavu N, Afzal AJ, 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
- Shultz JL, Kazi S, Bashir R, Afzal JA, Lightfoot DA (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 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
- 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
- Vision TJ, Brown DG, Shmoys DB, Durrett RT, Tanksley SD (2000) Selective mapping: a strategy for optimizing the construction of high-density linkage maps. *Genetics* 155:407–420
- Vuong TD, Sleper DA, Shannon JG, Nguyen HT (2010) Novel quantitative trait loci for broad-based resistance to soybean cyst nematode (*Heterodera glycines* Ichinohe) in soybean PI 567516C. *Theor Appl Genet* 121:1253–1266
- Wu XL, He CY, Wang YJ, Zhang ZY, Dongfang Y, Zhang JS, Chen SY, Gai JY (2001) Construction and analysis of a genetic linkage map of soybean. *J Genet Genomics* 28:1051–1061
- Wu XL, Larson SR, Hu ZM, Palazzo AJ, Jones TA, Wang RRC, Jensen KB, Chatterton NJ (2003) Molecular genetic linkage maps for allotetraploid *Leymus* (Triticeae). *Genome* 46:627–646
- Wu C, Sun S, 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* 14:319–326
- Wu X, Zhong G, Findley SD, Cregan P, Stacey G, Nguyen HT (2008) Genetic marker anchoring by six-dimensional pools for development of a soybean physical map. *BMC Genomics* 9:28
- Xia Z, Tsubokura Y, Hoshi M, Hanawa M, Yano C, Okamura K, Ahmed TA, Anai T, Watanabe S, Hayashi M, Kawai T, Hossain KG, Masaki H, Asai K, Yamanaka N, Kubo N, Kadowaki K, Nagamura Y, Yano M, Sasaki T, Harada K (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
- Yan JB, Tang JH, Meng YJ, Ma XQ, Teng WT, Chander S, Li L, Li JS (2006) Improving QTL mapping resolution based on genotypic sampling: a case using a RIL population. *Acta Genet Sin* 33:617–624
- Yang K, Moon JK, Jeong N, Back K, Kim HM, Jeong SC (2008) Genome structure in soybean revealed by a genome wide genetic map constructed from a single population. *Genomics* 92:52–59
- You FM, Huo N, Gu YQ, Luo MC, Ma Y, Hane D, Lazo GR, Dvorak J, Anderson OD (2008) BatchPrimer3: a high throughput web application for PCR and sequencing primer design. *BMC Bioinform* 9:253