ORIGINAL PAPER

Dynamic genetic features of chromosomes revealed by comparison of soybean genetic and sequence-based physical maps

Woo Kyu Lee · Namshin Kim · Jiwoong Kim · Jung-Kyung Moon · Namhee Jeong · Ik-Young Choi · Sang Cheol Kim · Won-Hyong Chung · Hong Sig Kim · Suk-Ha Lee · Soon-Chun Jeong

Received: 12 May 2012 / Accepted: 21 December 2012 / Published online: 10 January 2013

Abstract Despite the intensive soybean [*Glycine max* (L.) Merrill] genome studies, the high chromosome number (20) of the soybean plant relative to many other major crops has hindered the development of a high-resolution genomewide genetic map derived from a single population. Here, we report such a map, which was constructed in an F_{15} population derived from a cross between *G. max* and *G. soja* lines using indel polymorphisms detected via a *G. soja* genome resequencing. By targeting novel indel markers to marker-poor regions, all marker intervals were reduced to under 6 cM on a genome scale. Comparison of the Williams 82 soybean reference genome sequence and our genetic map indicated that marker orders of 26 regions were discrepant with each other. In addition, our comparison showed seven misplaced and two absent markers in the current Williams 82 assembly and six markers placed on the scaffolds that were not incorporated into the pseudomolecules. Then, we showed that, by determining the missing sequences located at the presumed beginning points of the five major discordant segments, these observed discordant regions are mostly errors in the Williams 82 assembly. Distributions of the recombination rates along the chromosomes were similar to those of other organisms. Genotyping of indel markers and genome resequencing of the two parental lines suggested that some marker-poor chromosomal regions may represent introgression regions, which appear to be prevalent in soybean. Given the even and dense distribution of markers, our genetic map can serve as a bridge between genomics research and breeding programs.

Introduction

Soybean [*Glycine max* (L.) Merrill] has historically been a major crop in Eastern Asia and is now one of the most important crop plants used for seed protein and oil content,

Communicated by R. Snowdon.

W. K. Lee and N. Kim made equal contributions.

Electronic supplementary material The online version of this article (doi:10.1007/s00122-012-2039-8) contains supplementary material, which is available to authorized users.

W. K. Lee · N. Jeong · S.-C. Jeong (✉)
Bio-Evaluation Center, Korea Research Institute of Bioscience and Biotechnology, Cheongwon, Chungbuk 363-883, Republic of Korea
e-mail: scjeong@kribb.re.kr

N. Kim · J. Kim · S. C. Kim · W.-H. Chung
Korean Bioinformation Center, Korea Research Institute of Bioscience and Biotechnology, Daejeon 305-806, Republic of Korea

J.-K. Moon
National Institute of Crop Science, Rural Development Administration, Suwon 441-857, Republic of Korea

I.-Y. Choi
National Instrumentation Center for Environmental Management, Seoul National University, Seoul 151-921, Republic of Korea

H. S. Kim
Department of Crop Science, Chungbuk National University, Cheongju 361-763, Republic of Korea

S.-H. Lee
Department of Plant Science, Research Institute for Agriculture and Life Sciences, Plant Genomics and Breeding Institute, Seoul National University, Seoul 151-921, Republic of Korea

in terms of both cultivated area and production worldwide, and for its capacity to fix atmospheric nitrogen by intimate symbioses with microorganisms. Despite the complex nature of the soybean genome including paleoallopolyploidy, the high haploid chromosome number ($n = 20$), and large genome size (1,115 Mb) relative to model plants including *Arabidopsis*, rice, and *Medicago truncatula*, the soybean genome has been intensively studied, due to the importance of soybean as a major crop. Thus, there has been significant progress in the development of genomic resources and experimental tools needed to expedite soybean research (Stacey et al. 2004; Boerma et al. 2011). A recent culmination was the release of a genome sequence draft (Schmutz et al. 2010) comprised of 937.3 megabases (Mb) of 20 pseudomolecules corresponding to 20 chromosomes, representing about 84 % of the predicted 1,115-Mb genome (Arumuganathan and Earle 1991). Whole-genome resequencing studies of cultivated soybean and its wild progenitor, *G. soja*, using the draft genome sequence as a reference were subsequently reported (Kim et al. 2010; Lam et al. 2010).

Among the genomic resources, the high-resolution genomewide genetic linkage maps are particularly important because the genetic maps provide bridging information between breeding and genome sequence efforts (Lewin et al. 2009). In other words, genetic maps are necessary for a structural and functional understanding of the genetic make-up of a genome and for the localization of genes of interest through a genetic linkage analysis with mapped markers (Bernatzky and Tanksley 1986). However, a high-resolution genomewide genetic linkage map derived from a single population that covers the soybean genome evenly has yet to be established. Most of the soybean genomewide genetic maps consisting of 20 homologous linkage groups corresponding to the 20 soybean chromosomal pairs and containing over 1,000 markers have been constructed from the alignment of 20+ linkage groups derived from multiple populations (Cregan et al. 1999; Song et al. 2004; Choi et al. 2007; Hwang et al. 2009). The difficulty in constructing a genomewide genetic map from a single population is that, despite the thousands of soybean molecular markers developed, the available polymorphic markers are not evenly distributed in the soybean genome. The development of a high-resolution genomewide soybean genetic map from a single population may have been hindered also by the high haploid chromosome number of the soybean plant. The soybean contains 20 haploid chromosomes, almost double that of other intensively studied major crops such as rice ($n = 12$), maize (10), barley (7), and tomato (12), as well as in research plants including *Arabidopsis thaliana* (5), *Lotus japonicus* (6), and *M. truncatula* (8). Thus, twice the effort is likely needed to establish the genetic linkage relationship between telomeres and centromeres. Recently, Yang et al. (2008) reported a single population-based genomewide genetic linkage map, which coalesced into 20 linkage groups representing 20 soybean chromosomes without unlinked markers by a stringent cutoff criterion. However, the map contained only 421 markers with many regions of low marker density.

The integration of distinct maps in a consensus map is usually achieved by aligning common markers with the assumption that recombination distances are similar between the maps (Stam 1993; Hu et al. 2004). Recombination distances in a composite map are average values of the merged maps. However, the recombination distances among markers vary among different populations within a species (e.g. Tulsieram et al. 1992; Williams et al. 1995; Sanchez-Moran et al. 2002; Rockman and Kruglyak 2009). Thus, although a composite map may provide the best fitting order of markers and average recombination distances of the merged maps, it has been regarded as inappropriate for structural and functional genomics research, including positional cloning, segregation distortion, and segment rearrangements, variability of recombination events along chromosomes (Esch et al. 2007; Rockman and Kruglyak 2009).

Here, we described the development of a high-resolution genomewide soybean genetic map generated from a single population derived from a cross between “Hwangkeum” (*G. max*) and “IT182932” (*G. soja*). A unique opportunity to compare the single population-based genetic map with the soybean draft genome sequence allowed us to detect many large-scale misassembled regions in the soybean draft genome sequence. Several discordant regions were corrected by obtaining experimental evidence. Then, we compared the distribution of recombination events to the genome features. Furthermore, we used resequencing data of the genomes of Hwangkeum and IT182932 to substantiate putative *G. soja* introgression regions in Hwangkeum, which were implicated during the genetic mapping. As such, the soybean genetic map can be used as a bridge between genomics research and breeding programs.

Materials and methods

Plant materials

A population of 113 F_{15} recombinant inbred lines (RILs) obtained by single-seed descent from the F_2 generation of an interspecific cross between *Glycine max* “Hwangkeum” and *Glycine soja* Siebold & Zucc. “IT182932,” a wild annual progenitor of the soybean (referred to as the HI population), was used for linkage analysis. A set of 12 soybean variants including Pureun, Ilpumgeomjeong, Peking, PI96983, T245, Williams 82, Hwangkeum, IT182932, V94-5152, T181, Sowon, and Seoritae was used in the marker diversity test. Fresh leaves of field- or greenhouse-grown individuals of F_{15} HI population and soybean parents were harvested, and genomic DNA was extracted in accordance with the methods described by Saghai-Maroof et al. (1984).

Development and analysis of markers

To locate markers in the marker-poor linkage regions in the preexisting HI genetic map before the release of the draft soybean genome sequence, novel microsatellite markers were mainly generated from scaffolds of the preliminary soybean whole genome shotgun sequence assembly (version "Glyma0") released by the USDOE-Joint Genome Institute Community Sequencing Program (http://www.phytozome.net/soybean) in 2008. Microsatellite repeat sites (mainly, AT or AAT motifs) were detected in a scaffold sequence using more than 10 motif-repeated sequences as a query. Then, the microsatellite-containing sequence was used to design a pair of forward and reverse primers to PCR-amplify 100- to 250-bp lengths of the DNA fragment containing the microsatellite site. In addition, publicly available markers (Saghai Maroof et al. 2009; Hwang et al. 2009; Yang et al. 2010; Suh et al. 2011) that were presumed to map to the large marker-sparse regions were incorporated into the HI genetic map.

Soon after the release of the draft soybean genome sequence (Schmutz et al. 2010), 196,356 indels (−35 to +14 bp) were detected by the comparison between genome resequencing data of a wild soybean accession IT182932 and the soybean draft genome sequence (Kim et al. 2010). A total of 71,751 indels containing longer-than-two-base pair indels were used as a source of marker development. The remaining 124,605 single-base pair indels were excluded because they are below the resolution limit of the QIAxcel Biocalculator System (QIAGEN, Hilden, Germany). The DNA sequences of the 71,751 indels were classified by presence or absence of microsatellite repeat variations. Indels that contained mononucleotide repeats with a repeat number of ten or greater and multinucleotide repeats of 2–8 bp repeat units with a repeat number of five or greater in either Williams 82 or IT182932 genome sequences were identified as microsatellite sites. Forward and reverse primers were designed for amplification of 100- to 250-bp lengths of the DNA fragment containing the indel site with the web-based Primer3 platform (Rozen and Skaletsky 2000). Whenever possible, those microsatellite-repeats-based indels were not used to generate novel markers to minimize the occurrence of allelic dropout frequently observed in the PCR amplification of microsatellite loci (Gagneux et al. 1997).

Primers for marker analysis were custom made by Bioneer (Daejeon, Korea). The polymerase chain reaction (PCR) mixture contained 20 ng of total genomic DNA, 10 mM Tris–HCl (pH9.0), 30 mM KCl, 1.5 mM $MgCl_2$, 100 nM of each forward/reverse primer, 250 μM of each dNTP, and 1 unit of *Taq* polymerase for a total volume of 20 μl. Typically, PCR amplification was performed with an initial denaturation step at 94 °C for 3 min, followed by 34 cycles of 94 °C for 30 s, 50 °C for 30 s, and 72 °C for 30 s, with final extension steps of 72 °C for 5 min. The PCR products were resolved on the QIAxcel Biocalculator system (QIAGEN) using a 12 capillary QIAxcel DNA High Resolution Cartridge with the 0M700 method or the QIAxcel DNA Screening Cartridge with the AM420 method. The QIAxcel DNA High Resolution Cartridge was used to analyze PCR products expected to be shorter than 400 bp and for resolving PCR products that differ in lengths by only 2–9 nucleotides. The QIAxcel DNA Screening Cartridge was used to resolve PCR products that differ by more than 10 nucleotides. The PCR products were transferred from the thermocycler into the 96-well plate in an 8 × 12 strip format sample tray of the QIAxcel Biocalculator system. The PCR products were automatically injected into the capillary channel and subjected to electrophoresis according to the manufacturer's protocol. The QX alignment marker (QIAGEN), which consisted of 15- and 400-bp bands for the QIAxcel DNA High Resolution Cartridge method and 15 bp and 3 Kb bands for the QIAxcel DNA Screening Cartridge method, was automatically injected into the cartridge with each sample by the QIAxcel Biocalculator system and enabled the software to align the lanes. The QIAxcel Biocalculator system produced a digital gel image and an electropherogram for fragment analysis. The PCR products were sized using size markers provided by the manufacturer, and the size of the products was determined using the QIAxcel Biocalculator 3.0 software (QIAGEN). Diversity or polymorphism information content (PIC) values of a subset of indel markers were calculated by classifying allele sizes of each indel locus on a diverse set of 12 soybean variants listed above. The calculation of this marker diversity was described by Anderson et al. (1993).

Construction of linkage map

Linkage analysis between markers, estimation of recombination frequencies, and determination of the linear order of loci were performed using JoinMap 4.0 software program (Van Ooijen 2006) with the independence LOD between 2.0 and 10.0 for grouping and with the following settings for regression mapping: used linkages with recombination frequency smaller than 0.400 and LOD larger than 2.00, threshold for removal of loci with respect to jumps in goodness-of-fit 5.00, and Kosambi mapping function. Marker order within a group was verified using

MapMaker 3.0b (Lander et al. 1987). The marker order was determined using the Three point and Order commands and then was rechecked via repetitive use of the Compare command. In those cases, when a group contained more than 100 markers, more than two subsets of markers in the group were tested. The map presented in this study was finally drawn using MapChart (Voorrips 2002).

Assignment of marker positions on the soybean sequence-based physical map

The sequence and primers of each marker from the HI map were compared to 955 Mb of the soybean genome sequence including 20 chromosome-level pseudomolecules and 1,148 unanchored sequence scaffolds in the Phytozome database (http://www.phytozome.net/soybean) using BLASTN to determine the position of each marker in the genome assemblies. In those cases, when the source sequences of the markers from Hwang et al. (2009) were not available in public databases, the sequences were determined from the PCR products. When both primers and the sequence of a marker were placed at the same position, which was the case for most of the markers, the position was used as the location of the marker in the soybean genome sequence map. In cases where primer(s) of a marker were not positioned using BLASTN, the sequence hit by the source sequence of the marker was downloaded, to examine mismatches at primer site(s) or the absence of the primer sequence. Even if primers were not complete matches, the hit positions to which the marker sequences were matched 100 % except for potential differences possibly resulting from different soybean sources and sequencing errors of single-pass sequences were taken as the locations of the sequence-based markers in the soybean genome sequence map. In all those cases when primer or polymorphic sites were ambiguous, an approximate median position of PCR amplicon generated from each marker locus was taken as the marker's physical position for the assignment of marker positions on the soybean sequence-based physical map.

Long PCR

Long PCR amplification was performed to obtain missing DNAs corresponding to the break points of misassembled segments postulated on the basis of comparison between the genetic linkage map and the sequence map. Primers were designed from known sequences at the edges of the misassembled segments (Table S4). Long PCR was performed using a TaKaRa LA *Taq* kit (catalogue number RR002A, TaKaRa, Dalian, Japan) as recommended by the manufacturer. PCR cycles were initiated with a denaturation step of 1 min at 94 °C, followed by 30 two-step cycles of 10 s at 98 °C and 15 min at 52–58 °C depending on the primer annealing temperature, and terminated with a final extension step of 10 min at 72 °C. PCR products were directly sequenced or, in cases of multiple bands, were subcloned into a plasmid for sequencing.

Analysis of recombination rates

Average recombination rates were obtained by dividing the total linkage distance (cM) by the total physical length (Mb) for each linkage group. These estimates were not adjusted for differences in marker density. Recombination rate between two markers were determined by calculating their genetic distance (cM) divided by their physical distance (Mb). Recombination rates were calculated for each set of markers. Markers where their physical and genetic order did not match were not included in these calculations. Using the soybean genome annotation Glyma1.01 (Soybase, http://soybase.agron.iastate.edu/), the numbers of genes and transposable elements (TEs) of each marker interval were calculated and plotted against the recombination rate of each marker interval. Regression analysis was applied in logarithmic model.

Resequencing analysis of soybean genomes and identification of introgression segments

Seeds of Hwangkeum and Williams 82 were germinated at 25 °C in a pot in a dark chamber. After primary leaves of the germinated seedlings were opened, all shoots of the seedlings except their cotyledons were collected for genomic DNA extraction. Sequencing libraries were constructed according to the manufacturer's instructions (Illumina). Paired-end short reads were generated by applying the Illumina HiSeq2000 (Illumina). The single- and paired-end resequencing data of 48.8 Gb for IT182932 used in this study were previously reported by Kim et al. (2010).

Short reads were mapped against the Williams 82 reference genome (Glyma1) with Burrows Wheeler Aligner (Li and Durbin 2009). FixMateInformation and MarkDuplicates modules in Picard software package (http://picard.sourceforge.net/) were used for mate information correction and duplicate removal. After this step, short reads of 20.0, 16.5, and 36.9 Gb for Hwangkeum, Williams 82, and IT182932, respectively, were retained. Alignments around short indels were realigned with IndelRealigner, and base pair quality scores were recalibrated with CountCovariates and TableRecalibration modules in the Genome Analysis Toolkit (GATK) (DePristo et al. 2011). UnifiedGenotyper, VariantFiltration, and VariantAnnotator modules in GATK were used for SNP and short indel genotyping, filtering, and annotation. Soybean annotations were downloaded from http://www.soybase.org.

Introgression of genomic segments from wild to cultivated soybean was assessed using shared alleles between wild and cultivated soybeans as described by McNally et al. (2009) and Lam et al. (2010). As we used resequencing data from only two varieties, one each from wild and cultivated soybean, the assessment was simplified. SNPs with missing data and heterozygous genotypes in individual accessions were excluded. The genotypes of SNPs in a sliding 100-kb window were scored for IT182932 (wild soybean) and Hwangkeum (cultivated soybean) against the Williams 82 (cultivated soybean) draft genome sequence, and the percentage of shared SNPs of Hwangkeum per the number of SNPs for IT182932 was calculated in each window. Regions with a percentage higher than 50 % were defined as putative introgression regions (PIRs).

The sequencing data from this article have been deposited with the GenBank data library under Accession Nos. JQ898517–JQ898524, JS807314–JS807326, and JQ924190–JQ924195. The resequencing dataset of short reads for Hwangkeum and Williams 82K can be downloaded from the following ftp site: ftp://ftp.kobic.re.kr/soybean/hwangkeum_williams82.fastq/.

Results

Construction of a high-resolution genetic map before the release of the draft soybean genome sequence

Three different sources of markers were used to saturate a complete soybean genetic linkage map coalesced into 20 linkage groups representing 20 soybean chromosomes, which was constructed with 421 markers in the Hwangkeum (*G. max*) x IT182932 (*G. soja*) population (HI population) (Yang et al. 2008). First, 208 novel microsatellite markers were generated from scaffolds of the preliminary soybean whole genome shotgun sequence assembly (version “Glyma0”) released by the USDOE-Joint Genome Institute Community Sequencing Program (http://www.phytozome.net/soybean) in 2008. Of the Glyma0 scaffolds, we attempted to generate four microsatellite markers from Scaffold_1 to 19 each longer than 10 Mb by designing primers from the regions flanking the microsatellite repeats within the scaffolds. Sequences of the markers mapped in the HI genetic map were located on each of the scaffolds with BLAST values of $E \leq e^{-10}$ by BLAST searches against the soybean Glyma0 database. For the scaffolds between 1- and 10-Mb up to Scaffold_196, we attempted to generate one microsatellite marker from each scaffold when only one or two markers hit a scaffold in BLAST searches. When no marker hit a scaffold in the BLAST searches, we attempted to generate three microsatellite markers from the scaffold. In cases of scaffolds between 1-Mb and 44-kb in length, we attempted to generate one marker only from those scaffolds that no marker hit. We assumed that this strategy would provide novel markers in the marker-sparse regions of the HI map. Of the 430 attempted primer pairs, 208 primer pairs produced single polymorphic PCR products in each of the mapping parents, which were ultimately used to map the microsatellite sites in the HI population. Those markers were named with the scaffold number and the ascending order of the alphabet such as Sca_1c for third marker from Scaffold_1 (Table S1). The remaining 222 primers produced single monomorphic PCR products, multiple PCR products, or no product in a few cases. Thus, many marker-sparse regions still persisted after this endeavor.

The second source of markers used was the microsatellite markers reported by Hwang et al. (2009) during the course of this study. The reported 440 markers that were presumed to map to the large marker-sparse regions in the HI map and to the upper and lower ends of linkage groups were assessed for polymorphism between the two parental lines of the HI population. In total, 97 microsatellite markers, representing 22 % of the tested microsatellite markers, were mapped. Although the map reported by Hwang et al. (2009) had many large gaps in the middle of the linkage groups, the map contained many novel markers at the ends of several linkage groups. As a result, the markers from Hwang et al. (2009) extended the ends of several linkage groups including soybean chromosomes (Chr) 3, 6, 7, 9, and 11 or increased the marker density at the ends of the linkage groups. The third source of markers used in this study was the sequence-based markers, which were mostly developed for localization of genes of interest and were reported during this study (Saghai Maroof et al. 2009; Yang et al. 2010; Suh et al. 2011).

Construction of a high-resolution genetic map after the release of the draft soybean genome sequence

An improved version Glyma1 of the draft soybean genome sequence that contains 937.3 Mb of the 20 chromosome-level pseudomolecules and 17.7 Mb consisting of 1,148 unanchored sequence scaffolds was reported in 2010 (Schmutz et al. 2010). Subsequently, 71,751 of the putative insertion/deletion (indel) polymorphic sites where the insertions ranged from 2 to 14 bp and the deletions ranged from 2 to 35 bp were predicted by the comparison between genome resequencing data of a wild soybean accession IT182932 and the soybean draft genome sequence (Kim et al. 2010). Three major indel classes were identified by analyzing the DNA sequences of our indels with regard to the presence or absence of short tandem repeats (microsatellites): (1) 13,052 monomeric base pair expansions with the repeat number of ten or greater in either Williams 82 or IT182932 genome sequences, (2) 2,849 multi-base pair expansions of 2–5 bp repeat units with repeat numbers of five or greater [the criterion used by Song et al. (2010) to screen microsatellites in the Williams 82 genome sequence] in either Williams 82 or IT182932, (3) indels containing apparently random DNA sequences (Table 1). Of the multi-base pair expansions, di-, tri-, and tetranucleotide microsatellite sites that were found only in the Williams 82 genome sequences were 1,590. Thus, the classification results of our indels indicated that 70,161 (97.8 %) of the 71,751 putative indel polymorphic sites are novel indels generated from mutation events different from the 210,990 microsatellite loci identified by Song et al. (2010). As 33,065 (15.6 %) of the 210,990 microsatellite sites were proposed to be likely polymorphic between diverse soybean accessions, the overlapping rate (2.2 %) between our indels and the 210,990 microsatellite loci was too low. Whether this lower-than-expected overlapping rate was caused by the poor sequencing results at the short tandem repeat regions by the next-generation sequencing platforms that were used to generate the genome sequence of IT182932 needs to be examined in the future.

We constructed our high-resolution genetic map mainly using the markers generated from the putative indel polymorphic sites. At the beginning of the development of the large interval-targeted indel markers, we randomly chose 102 putative indel sites located on the lower telomeric end of Gm01 and on the upper telomeric ends of Gm02 and designed 102 primer pairs to amplify them to examine what portions of the putative indel sites could be used to generate markers. We obtained single polymorphic PCR bands in Hwangkeum and IT182932 from 72 primer pairs and single PCR bands only present in Hwangkeum from 2 primer pairs. The remaining 28 primer pairs produced single monomorphic PCR bands between Hwangkeum and IT182932 (13), more than two PCR products (14), or no PCR product (1). The results suggested that 72.5 % of the putative indel sites could be used to generate locus-specific markers in our HI population. In addition, we designed 18 primer pairs to amplify 18 microsatellite regions detected by BLAST searches against the Gm05 pseudomolecule to generate markers in the marker-sparse regions on Chr 5. Of those, nine primer pairs that produced single polymorphic bands between the parents were used for genetic mapping. The nine markers, which were designated GM001 thru GM017, were mapped to the expected chromosomal regions (Table S1).

These two initial trials suggested that the putative indels detected between Williams 82 and IT182932 genome sequences might be highly polymorphic between Hwangkeum and IT182932 and that the marker orders both on the HI genetic map and on the draft soybean genome sequence assembly might be collinear. Then, we designed primer pairs to generate indel markers targeted to both marker intervals of greater than 3 cM in the HI map at that time and the ends of the 20 genetic linkage groups, with an objective to reduce all marker intervals in the HI map to less than 2 cM. Of the 810 primer pairs, 462 (57 % of the

Table 1 Classification of indels

Indel class	
Repeat expansions	15,901
Monomeric	13,052
(T)n	5,913
(A)n	6,409
(C)n	366
(G)n	364
Dimeric	2,410
(AT)n	543
(TA)n	413
(AG)n	306
(TC)n	256
(GA)n	197
(CT)n	195
(TG)n	169
(AC)n	143
(GT)n	99
(CA)n	89
Trimeric	385
(AAG)n	28
(AAT)n	23
(GAA)n	21
(TTC)n	21
(TTA)n	20
(ATA)n	20
(AGA)n	17
(TTG)n	15
(TAT)n	15
(AAC)n	13
(ATT)n	13
(CTT)n	12
(ATG)n	12
(TAA)n	11
(TCT)n	11
Other (NNN)n	133
Tetrameric	47
Pentameric	7
Other	55,850
Total	71,751

attempted primer pairs) produced single polymorphic PCR products in the parents of the HI population and so they were used to map the indel sites in the HI population. The remaining 348 primer pairs produced single monomorphic, multiple PCR products, or no PCR product in a few cases. Mainly due to the lower than expected success rates of this initial targeted-marker generation, linkage analysis using mapped markers up to this stage indicated that ten marker intervals of greater than 6 cM still remained. Most of the generated markers were mapped to the expected chromosomal locations. However, several markers were mapped to chromosomes different from the chromosomes to which the markers were designed (Table 2), suggesting that a few regions of the pseudomolecules might be misassembled.

Table 2 List of markers or marker intervals discrepant between the genetic and sequence-based physical maps

Marker or marker interval	Position or interval (cM) on genetic map	Position or interval (bp) on the Williams 82 genome sequence assembly (Glyma1)	Description of discrepancy
Sca_193c	Chr 1:45.1	Scaffold_23:530760..530520	Located on unanchored Scaffold_23
GSINDEL9257–GSINDEL9007	Chr 1:95.5..101.4	Gm01:53767102..54582005	Order
Satt634–BI470504A/G	Chr 2:63.9..64.6	Gm02:11441753..11316722	Order
GMES1613	Chr 2:102.5	No hit	Absent in the Williams 82 assembly
GSINDEL20494–Sca-125c	Chr 3:24.2..26.8	Gm03:4127046..4301966	Order
Sca_341a	Chr 3:34.3	Scaffold_57:73125..73390	Located on unanchored Scaffold_57
AW277661–Sca-51b	Chr 4:63.6..64.3	Gm04:19660791..32803274	Order
Sca_16a	Chr 4:63.6..64.3	Gm18:26860865..26860718	Misplaced on Gm18
GMES6354	Chr 4:104.2	No hit	Absent in the Williams 82 assembly
GSINDEL37893–GSINDEL37855	Chr 5:8.3..9.3	Gm05:735303..848400	Order
Satt599–Sca_164d	Chr 5:92.5..110.0	Gm05:38121046..41691258	Order at the telomeric end
Satt_252–GSINDEL56531	Chr 6:94.9..103.4	Gm06:48582333..48415540	Order
A636.p1–A656.p1	Chr 9:42.1..43.2	Gm09:6455599..6477903	Order
GSINDEL121628	Chr 9:63.2	Gm13:35251705..35251881	Misplaced on Gm13
GSINDEL121634	Chr 9:64.0	Gm13:35274137..35274280	Misplaced on Gm13
GSINDEL99279–SM001	Chr 11:43.9..44.5	Gm11:7933139..7743622	Order
GSINDEL101692–Sca_115a	Chr 11:85.6..87.7	Gm11:18615873..30460752	Order
Sca-28-ta1	Chr 11:88.7	Scaffold_28:77699..77928	Located on unanchored Scaffold_28
Sca-28-ta2	Chr 11:88.7	Scaffold_28:206755..206945	Located on unanchored Scaffold_28
Sca-240a–Satt359	Chr 11:104.9..106.9	Gm11:36868404..36989799	Order
GSINDEL104502–CSSR18	Chr 11:117.1..133.0	Gm11:37952401..39168834	Order at the telomeric end
Sca-361a–Sca-358a	Chr 12:52.2..53.9	Gm12:17624967..22063214	Order
Sca-9c	Chr 12:53.9	Gm20:13282030..13281859	Misplaced on Gm20,
Sca-9d–Satt414.p2	Chr 12:53.9..54.8	Gm12:25716770..27558043	Order
Satt302–GSINDEL111315	Chr 12:69.6..70.8	Gm12:34979782..35108183	Order
SN314–GSINDEL113427	Chr 13:0.0..14.5	Approximately 14 Mbp	Order at the telomeric end where ribosomal DNA genes locate
GSINDEL121265–Sat_375	Chr 13:57.2..57.8	Gm13:33860360..34002148	Order
GMES1545	Chr 13:92.7	Gm17:3681499..3681724	Misplaced on Gm17
Satt577–GSINDEL124822	Chr 14:3.7..15.1	Gm14:623867..2156774	Order
GSINDEL134291–GSINDEL133985	Chr 14:107.1..113.4	Gm14:48919835..49708567	Order at the telomeric end
GSINDEL146736	Chr 15:11.3	Gm16:5869184..5869417	Misplaced on Gm16
A963.p1–GMES5448	Chr 15:54.6..63.2	Gm15:9360766..10403150	Order
SL301	Chr 16:47.6	Gm03:27658399..27658637	Misplaced on Gm03
GMES4288	Chr 17:94.8	Scaffold_41:166110..165873	Located on unanchored Scaffold_41
Sca-308a	Chr 17:95.7	Scaffold_41:55420..55585	Located on unanchored Scaffold_41
GMES4770–Sct_034.p2	Chr 17:101.0..101.6	Gm17:39398789..39521345	Order
Sca-4b–GSINDEL165018	Chr 18:51.1..51.6	Gm18:7753176..7519216	Order

Table 2 continued

Marker or marker interval	Position or interval (cM) on genetic map	Position or interval (bp) on the Williams 82 genome sequence assembly (Glyma1)	Description of discrepancy
GSINDEL176475–Sat_408	Chr 19:0.0..4.2	Gm19:21743..555440	Order at the telomeric end
Sat_174–Sca-1c	Chr 20:26.3..26.8	Gm20:6821145..28808576	Order
GSINDEL191623–GSINDEL191323	Chr 20:27.8..33.1	Gm20:31195240-32613841	Order
GSINDEL194650–GSINDEL194643	Chr 20:78.2..78.8	Gm20:41164749..41259620	Order

This observation prompted us to generate microsatellite markers from scaffold_21 to _30 of larger than 450 kb among the 1,148 unanchored sequence scaffolds in the soybean draft genome sequence (Schmutz et al. 2010). However, only scaffold_28 was mapped to Chr 11 by the two markers Sca-28-ta1 and Sca-28-ta2, probably because of highly repetitive and gene-poor nature of the scaffold sequences, as shown by Schmutz et al. (2010).

Diversity and distribution of indel variation in the marker-sparse regions

The low success rate of the initial targeted indel marker mapping suggested that many chromosomal regions might contain lower levels of polymorphism between Hwangkeum and IT182932 than between Williams 82 and IT182932. Otherwise, the low rate suggested that a large portion of the putative indels were false. Therefore, in the rest of our efforts to saturate the remaining marker-sparse regions, primer pairs designed from the putative indels were tested on a set of 12 diverse soybean accessions including Williams 82, Hwangkeum, and IT182932 listed in "Materials and methods" to obtain an estimated level of PCR product length polymorphism associated with the putative indels. Of the 405 tested primer pairs, 323 produced single polymorphic bands in the set of 12 soybean cultivars. The remaining 82 primer pairs produced single monomorphic ($n = 23$), multiple monomorphic (3), or multiple polymorphic PCR products (54) or, in 2 cases, no product. The overall accuracy of the indel detection including the primer pairs that produced single polymorphic bands and multiple polymorphic PCR products was 93 % (377/405). Of the 323 primer pairs that produced single polymorphic bands, 321 produced the same sized bands either from both Williams 82 and Hwangkeum or from both Hwangkeum and IT182932, while 2 (for GSINDEL89468 and GSINDEL70333) produced three different sized bands among Williams 82, Hwangkeum, and IT182932. Interestingly, only 227 of 323 (70 %) that produced single polymorphic bands between Williams 82 and IT182932 produced bands polymorphic between Hwangkeum and IT182932, and all the 227 primer pairs, suitable for development of single-locus markers, were used to map the indel sites in the HI population. Although our general strategy focused on the development of single-locus markers, of the 54 primer pairs that produced multiple polymorphic PCR products, 31 that produced different bands clearly separated from each other and showed polymorphic bands at the predicted size were used to generate markers. All marker loci genotyped using the 31 primer pairs were mapped to the targeted marker chromosomal regions and, consequently, were very useful for saturating large marker-sparse regions.

The mean polymorphism information content (PIC) of the indel loci detected by the primer pairs that produced single bands among the set of 12 soybean accessions was estimated to be 0.2734 (Table S2). This low diversity is because 138 of the 258 primer pairs produced unique polymorphic bands from IT182932, while producing the same sized bands in the remaining 11 soybean accessions. The analysis of the diversity of indels between wild and cultivated soybean populations has not been extensively studied. In the diversity analysis of single nucleotide polymorphisms (SNP) using resequencing data of 31 wild and cultivated soybeans, 35 % of SNPs were wild-specific, 5 % of SNPs were cultivated-specific, and 32 % of SNPs were present in both (Lam et al. 2010). With the assumption that indels would show the same pattern as SNPs, at least 35 % or as high as 67 % of indels detected between Williams 82 and IT182932 would be expected to show polymorphisms between Hwangkeum and IT182932. The success rate (56 %) of single-locus marker generation for the mapping in the HI population for this set of 405 primer pairs was lower than 72.5 % (for the 102 primer sets) and nearly equal to 57 % (for the 804 primer sets) of our two initial analyses described above. Consequently, many parts of the final high-resolution genomewide genetic map still have 3–6 cM marker intervals partly because of the lack of useful marker generation as a result of the lack of polymorphism of the available indels in many parts of the targeted chromosomal regions. In one case, two microsatellite markers, BARCSOYSSR_06_0382 and BARCSOY

SSR_06_0383, reported by Song et al. (2010) were mapped to the region of low indel polymorphism between GSINDEL48466 and GSINDEL48480 that remained as a marker-sparse region on Chr 6. Therefore, the chromosomal regions for which it was difficult to design polymorphic markers may be regarded as low polymorphic chromosomal regions in the soybean population or introgression regions in Hwangkeum from *G. soja*. To our surprise, our observations suggested that four chromosomal regions in the Hwangkeum genome may have been introgressed from *G. soja* (Table S3). In these regions, more than half of the indel markers polymorphic between Williams 82 (*G. max*) and IT182932 (*G. soja*) were shown to be monomorphic between Hwangkeum and IT182932. For example, across 13.4 cM between GMES1600 and GMES6736 on Chr 9, only 5 of 19 indels that were found to be polymorphic between Williams 82 and IT182932 were polymorphic between Hwangkeum and IT182932. This possibility was further tested by the alignment of the resequencing data of Hwangkeum and IT182932 genomes to the Williams 82 draft genome sequence, as described below.

Examination of discrepant regions between the genetic and sequence-based physical maps

We integrated a total of 1,154 sequence-based markers to the HI genetic map previously described by Yang et al. (2008). Data from the 1,575 markers were used for linkage analysis. The 20 linkage groups (LOD value of 10) corresponded to the 20 chromosomes of soybean (Table S1; Fig. 1). All marker intervals were below 6 cM. We integrated the genetic linkage map and the soybean genome sequence by assigning each marker to a unique position in the Williams 82 reference genome sequence (Glyma1) by BLAST searches. Since only unique sequences were used to construct the markers, we were able to unambiguously assign each marker to one location in the genome with the exception of a group of markers generated from the ribosomal DNA sequences. We noticed that, since the HI population consisted of only 113 recombinant inbred lines, one recombination corresponded to approximately 0.4 cM. The marker intervals of less than 0.4 cM were usually obtained by double crosses surrounding markers or by missing data points, indicating that our genetic mapping data were unable to determine the marker order below 0.4 cM. Therefore, where marker intervals were less than 0.4 cM due to a lack of recombination between markers, we placed the markers into the same order as they were found in the genomic sequence.

By comparing the positions and orders of the markers on the genetic and sequence-based physical maps, we were able to detect 41 discrepant segments between the two maps (Table 2; Fig. 1; Table S1), which corresponded to approximately 71 Mb (7.5 %) of the 950-Mb Williams 82 sequence assembly. First, the positions of 15 markers were different between the genetic and sequence-based physical maps: (a) 7 were located on different chromosomes from those identified by pseudomolecules of the soybean draft genome sequence; (b) 6 were located on the unanchored scaffolds (scaffold_23, _28, _41, and _57); (c) 2 were absent in the draft genome sequence. We predicted putative locations of the three anchored scaffolds by screening 1,000 N bps that were inserted for joining scaffolds at the time of the generation of the Glyma1 pseudomolecule version of the soybean draft genome sequence (Schmutz et al. 2010) and then attempted to retrieve sequences connecting the scaffolds to both sides of the 1,000 N bps by long PCR (Table S4). For scaffold_23 and _28, we failed to retrieve the sequences likely because the ends of the two scaffolds contained repetitive sequences. For scaffold_41, we were able to obtain 4.5 and 9.9 kb of the two connecting sequences that match with one of both ends of scaffold_41 and one of both sides flanking the 1,000 N bp, respectively (GenBank Accession No. JQ924193 and JQ924194). In other words, Scaffold_41 was inserted at a 1,000-N-bp position located at 39,027,038 position of pseudomolecule Gm17. Second, the orders of markers on the 26 regions greater than 0.4 cM in the genetic map were different from those detected by the BLAST searches in the draft genome sequence (Table 2). Notably, five discrepant regions were located at the ends of the chromosomes. With the assumption that these regions resulted from apparent misorientation of large scaffolds, we attempted to correct the errors. For the upper end of Chr 13, our BLAST searches of rDNA-derived markers against the draft genome sequence suggested that the disagreement was mainly caused by misplacement of rDNA repeat segment, which is usually found at the very end of the chromosomes in other plants. The rDNA sequences were inserted in a dispersed manner around the 14 Mb position of Gm13 and most rDNA sequences were likely filtered out during the assembly of the draft genome sequence. Thus, the discrepant segment, ranging from position 1 to 15,676,679 of Gm13, was inverted as indicated by a comparison of the genetic and sequence-based physical maps (Figs. 1, 2; Table S1). For the lower ends of Chr 5, Chr 11, and Chr 14 and upper end of Chr 19, our survey of the Genome Browser at Soybase (http://soybase.agron.iastate.edu/) indicated that the beginning points of the regions of disagreement likely corresponded to the break points between the two syntenic blocks from the same homeologous chromosome with the opposite orientation than that in the current soybean draft genome sequence (Fig. S1). By examining the end sequences of the soybean bacterial artificial chromosome (BAC) clones as well as Glycine

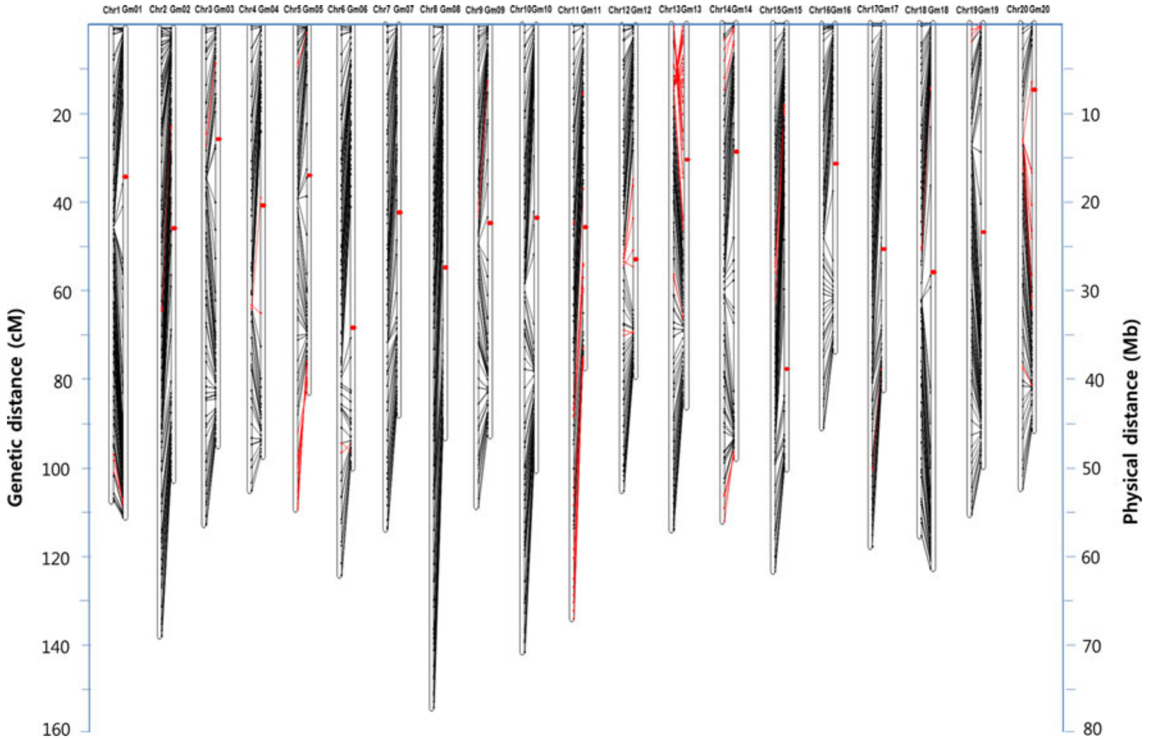

Fig. 1 Soybean linkage groups and their alignment to the soybean genome sequence. Linkage group (*left*)/pseudomolecule (*right*) pairs are presented for each chromosome. They are oriented such that the start of the bp and map distances (0) is at the top of each chromosome. The location of each marker is indicated by a *dot* on the genetic map and a *dot* on the genomic sequence. Corresponding location for each marker is indicated by a *black or red line*. The *red line* indicates marker-order disagreement regions between the genetic linkage map and pseudomolecule. Approximate centromere positions proposed by Schmutz et al. (2010) are indicated by the *thick red dot*s. The scale for the linkage groups (cM) is on the *left* and the scale for the genomic sequence (Mb) is on the *right*. Note that the recombination rate is not uniform across the chromosome with higher recombination rates at the distal ends and lower in the central centromeric region

max-WGS sequences retrieved from the GenBank databases (http://www.ncbi.nlm.nih.gov/) by BLAST searches, we were able to identify the putative beginning points of the regions of disagreement at the sequence level. Interestingly, those predicted positions corresponded to positions of the 1,000 N bps (Schmutz et al. 2010). For the Chr 5, Chr 14 and Chr 19, missing sequences at the beginning points were determined by sequencing long PCR products amplified from the genomic DNAs of IT182932, Hwangkeum and Williams 82 using primer pairs designed from the very ends of the agreement region for one and of the pseudomolecule for the other. Sequences of PCR fragments amplified from Williams 82 were fully determined (GenBank Accession No. JQ924190, JQ924192, and JQ924195). Single-pass sequences of PCR fragments amplified from IT182932 and Hwangkeum were compared with the Williams 82 sequences to check that we were sequencing the expected PCR products as well as to find polymorphic sites for generation of markers described below. Both ends of the sequences aligned perfectly to the respective end of the agreement region and of the pseudomolecule as predicted after removing the ambiguous sequence parts based on trace profiles of shotgun sequences, indicating that the sequences represented misassembled locations in the pseudomolecules. Therefore, the sequences were used as a bridge to connect the regions of agreement to the inverted regions of disagreement regions to obtain corrected assemblies of the pseudomolecules Gm05, Gm14, and Gm19. However, for the Chr 11, our PCR attempts using the scheme described above failed or produced numerous products. Then, the genomic DNA sequence corresponding to the soybean EST BE020413 on Chr 11 which is highly similar to the end of the agreement region was determined. A long PCR amplification using a pair of primers designed from the genomic DNA sequence containing BE020413 and the very end of the pseudomolecule produced several PCR products. One of them, approximately 10 kbp, was completely sequenced. One end of the sequence was aligned to the end of the pseudomolecule Gm11, but the other end of the sequence was not similar to the genomic sequence of BE020413, indicating that the sequenced PCR product was amplified from specific binding of the primer designed from the end of the pseudomolecule and from nonspecific binding of the primer designed from the genomic sequence of BE020413. Nevertheless, these two sequences, which were end-sequences

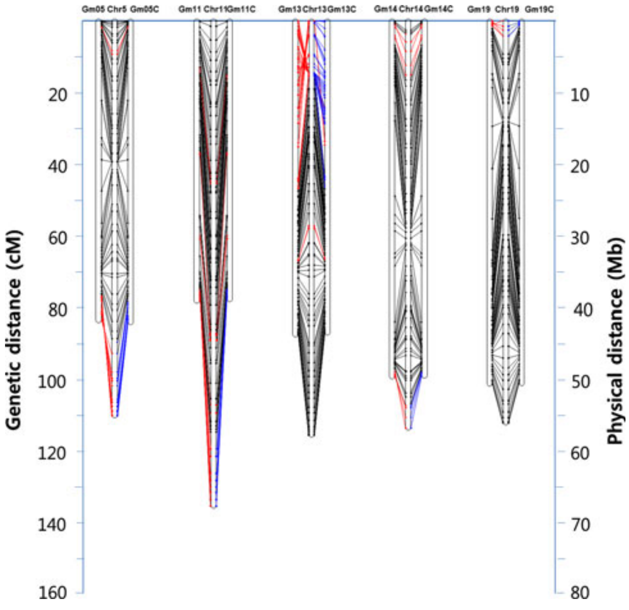

Fig. 2 Soybean linkage groups and their alignment to corrected pseudomolecules of Gm05, Gm11, Gm13, Gm14, and Gm19 in the soybean genome sequence. Pseudomolecule (*left*)/linkage group (*middle*)/corrected pseudomolecule (*right*) pairs are presented for each chromosome. They are oriented such that the start of the bp and map distances (0) is at the top of each chromosome. The location of each marker is indicated by a *dot* on the genetic map and a *dot* on the genomic sequence. The corresponding location for each marker is indicated by a *black*, *blue*, or *red line*. *The red line* indicates marker-order disagreement regions between genetic linkage map and pseudomolecule and *blue line* indicates corrected marker-order agreement regions between the genetic linkage map and pseudomolecule. The scale for the linkage groups (cM) is on the *left* and the scale for the genomic sequence (Mb) is on the *right*

of the potentially missing sequence in the pseudomolecule Gm11, were used to download the unplaced WGS sequences from the Glycine max-WGS trace achieves database (http://blast.ncbi.nlm.nih.gov/). As sequences of shotgun clones were determined by mate pair sequencing, we could assemble the sequences clone-by-clone considering the orientation and distance between the mate pair sequences of the clones. Finally, we obtained approximately 48.6 kbp of the potentially missing sequence with some ambiguous parts between the mate pair sequences (GenBank Accession No. JQ924191).

To substantiate that the sequences indeed correspond to the missing sequences that connect between the beginning points of the regions of disagreement and the ends of the current pseudomolecules, markers were generated from the polymorphic sites between their corresponding sequences obtained from Hwangkeum and IT182932. Indel polymorphisms were observed between Hwangkeum and IT182932 from the alignment of all sequences corresponding to the break points of the regions of disagreement of marker orders obtained from Gm05, Gm14, and Gm19. The indel polymorphisms were used to generate markers. The markers Chr5A1-L, Chr14-L, and Chr19-U were mapped to the predicted position in the genetic linkage map, respectively, at Chr 5, Chr 14, and Chr19. For Chr 11, three microsatellite repeat sites in the 48.7-kbp sequence were used to generate markers Chr11-La, Chr11-Lb, and Chr11-Lc. All the three markers were mapped to the predicted positions with a total genetic distance of 2.1 cM. Taken together, these results are convincing evidences that four regions of disagreement between genetic and sequence-based physical maps, which were located at the ends of Chrs 5, 11, 14, and 19, are misassembled regions in the pseudomolecules of the soybean draft genome sequence and indicated that although we cannot exclude unknown natural phenomena such as chromosomal rearrangements, most of the discrepant regions of marker orders between the genetic and sequence-based physical maps are likely misassembled regions in the pseudomolecules of the soybean draft genome sequence.

Integration of genetic linkage map and genome sequence

After adding the 6 markers generated from the misassembled regions at the telomeric ends of Gm05, Gm11, Gm14, and Gm19, the mapping data set containing a total of 1,581 markers were subject to a linkage analysis. Chr 1 contained the largest number of markers with 128 and Chr 4 contained the least number of markers with 59. The 20 linkage groups (LOD value of 10) corresponded to the 20 chromosomes of soybean (Table S1; Fig. 1). The map spanned a total length of 2,361.2 cM with an average of 1.5 cM between markers. The largest marker interval was 5.9 cM. The genetic length for each chromosome ranged from 91.9 cM (Chr 16) to 155.3 cM (Chr 8). Our total map distance was 45.9 cM longer than the 2,315.3 cM of map distance of the HI map reported in 2008 (Yang et al. 2008). One of the main reasons for this increase was addition of novel markers at the ends of linkage groups. For example, the addition of seven novel markers to the upper telomeric end of Chr 20, which was the shortest linkage group in the 2008 HI map, resulted in an increase of 14.3 cM. The addition of seven novel markers to the lower telomeric end of Chr 9 resulted in an increase of 11.8 cM. The marker density of each chromosome ranged from 403.6 to 853.4 kb/marker, with an average of 600.9 kb/marker (Table S5). There were 63 (6 %) marker intervals between 1 and 2 Mb and 63 (6 %) marker intervals larger than 2 Mb (Table S1). Relative to the centromeres localized by the centromeric repeat sequence positions in the soybean draft genome sequence, less than ten markers were distributed in the 10 Mb regions flanking the centromeres on each of the soybean chromosomes. Thus, most of markers were distributed in the paracentromeric or telomeric chromosomal regions.

Recombination frequency in a genomic context

The relative changes in recombination rates along each chromosome were examined by comparing the genetic and physical distances between the neighboring markers (Fig. 3; Fig. S2). The average recombination rate for each chromosome was similar to each other, ranging from 1.9 to 3.4 cM/Mb with a genomewide average of 2.5 cM/Mb (Table S5). However, recombination rates varied dramatically across the genome, from 0 to 107.0 cM/Mb. Recombination rates of approximately 170 marker intervals were higher than 20 cM/Mb, which is excess of eightfold as compared to the average of each chromosome (Fig. 3; Fig. S2). As has been demonstrated in numerous other systems (Nachman 2002; Morrell et al. 2012), we observed low recombination rates in centromeric and pericentromeric regions of each chromosome with low gene and high transposable element (TE) density and higher recombination rates in telomeric regions with high gene and low TE density (Fig. 3; Fig. S2).

The quantitative relationship between the recombination rate and the gene or the TE density was examined by plotting the gene and TE content of each map interval against the corresponding recombination rate (Fig. 4). For this analysis, we used the corrected assemblies for the pseudomolecules Gm05, Gm11, Gm13, Gm14, Gm17, and Gm19, as described above. The remaining discrepant regions between the genetic and sequence-based physical maps were excluded, as shown by discontinuities in Fig. 3. In addition, 0 recombination rate values generated by co-segregating markers were not included. Regression analysis showed a significant positive relationship between the recombination rate and the percentage of gene sequence in an interval (Fig. 4). There was a significant negative relationship between the recombination rate and the percentage of TE sequence, the largest non-gene component in the soybean genome, in an interval (Fig. 4) and, vise verse, a significant positive relationship between the recombination rate and the percentage of TE-excluded DNA sequence in an interval (Fig. 4). Relative to a similar analysis with *Brachypodium* genome reported by Huo et al. (2011), gradient of our logarithmic model curve between the recombination rate and the percentage of TE sequence is

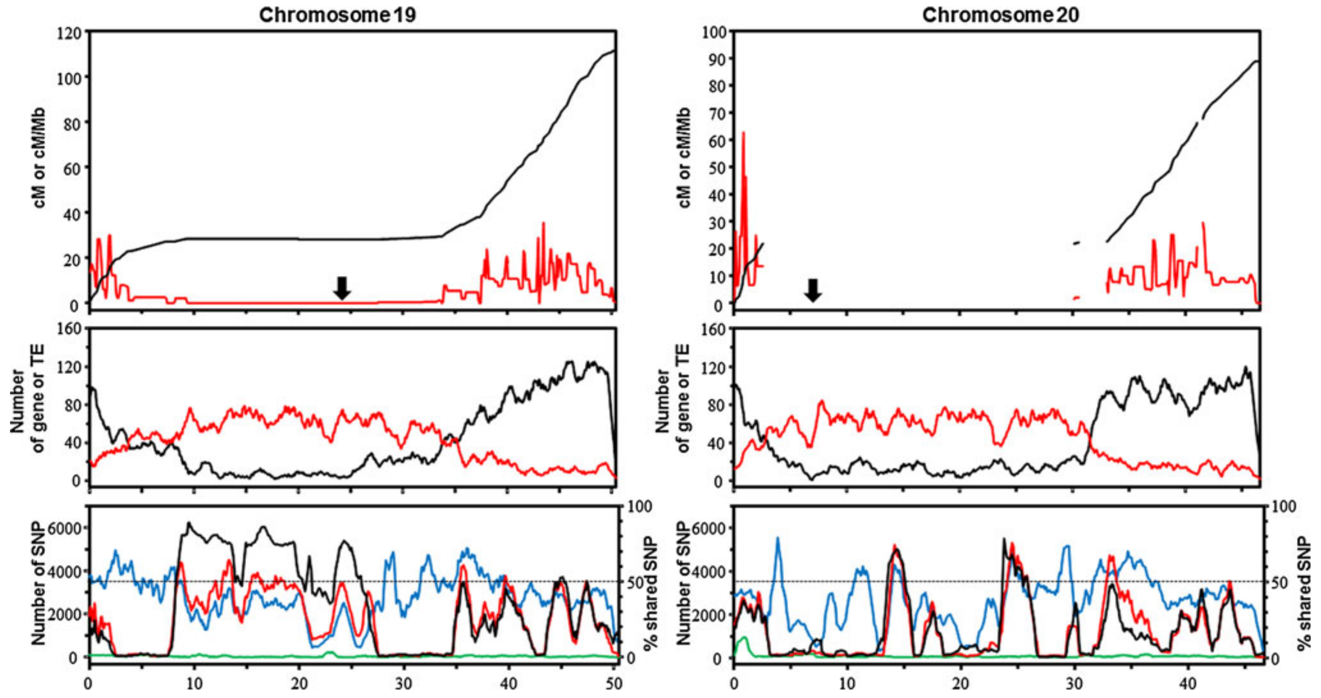

Fig. 3 Diversity along soybean chromosomes 19–20. The horizontal axes are in units of million base pairs along the Williams 82 reference genome and approximate centromere positions proposed by Schmutz et al. (2010) are indicated by the *thick arrows*. Top panel shows relationship between physical and genetic positions (cM, *black line*), and corresponding recombination rates (cM/Mb, *red line*) calculated from 100-kb sliding windows for the genomic regions between adjacent markers; the discrepant regions between the genetic and sequence-based physical maps are shown by discontinuities. *Middle panel* shows numbers of genes per 100 kb (*black line*) and numbers of transposable elements (TE) per 100 kb. *Bottom panel* shows numbers of single nucleotide polymorphic (SNP) sites per 100 kb (*left y axis*) and percentages of shared SNPs of Hwangkeum per the number of SNPs for IT182932; *green lines* are Williams 82K, *blue lines* IT182932, *red lines* Hwangkeum, and *black lines* % shared SNP

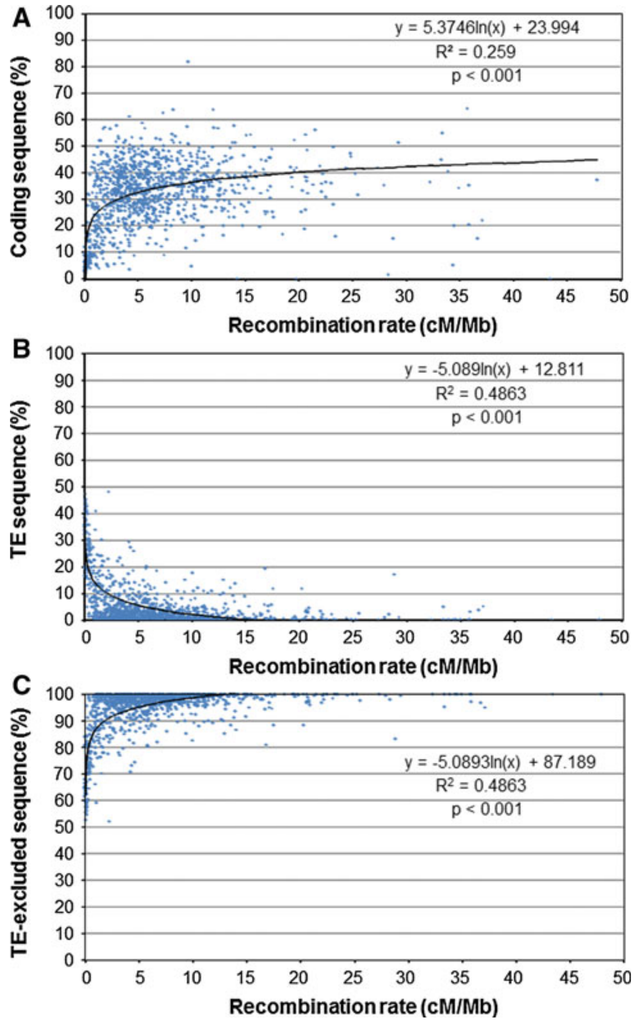

Fig. 4 Correlation of genome features with recombination rates. The percentages of coding sequence (**a**), transposable element (TE) sequence (**b**), and TE-excluded sequence (**c**) contained in each interval were calculated and plotted against the recombination rate (cM/Mb). A logarithmic model was applied to each dataset and the correlation coefficient calculated

much steeper in the low recombination rate intervals of 0–10 cM/Mb likely because many of those intervals contain <10 % TE sequences. This difference suggests that a majority of the approximately 100 Mb of unassembled repetitive sequences in the current Williams 82 genome assembly (Schmutz et al. 2010) should have been incorporated into those intervals.

The overall recombination rate for the soybean was lower than those for *Arabidopsis*, *Brachypodium*, and rice with small genomes below 500 Mb and higher than those for maize and sorghum with large genomes above 500 Mb (Table S6). In particular, genetic map lengths of both soybean and maize were approximately 2,350 cM, even though the maize genome is 2.45 times larger than the soybean genome. However, the question remained whether the estimated recombination rate of one species is unusually high or low relative to that of other well-studied species, as attempted by the genome studies of *Arabidopsis lyrata* (Kuittinen et al. 2004) and *Brachypodium distachyon* (Huo et al. 2011). The correlation between genome sizes and recombination rates has not been examined. Recent completion of numerous high-quality genome sequences driven by the next-generation sequencing platforms (Table S6) provided a unique opportunity to address this question. Interestingly, the adjusted recombination rate value of 5.9 cM/Mb of soybean, calculated by dividing the map length by the sequenced genome size excluding TE, was similar to those of the highly studied *A. thaliana*, rice, and maize. However, a test of the variability of the recombination rates estimated from the sequenced genome sizes is not significantly different from that of recombination rates estimated from the TE-excluded sequenced genome sizes (*F* test, $P = 0.619$). Therefore, our results indicated that TE contents are not the sole and major determinant of genomewide average recombination rates. Other chromosomal features such as the number and location of centromeres and the density of recombination hot spots may play an important role in determining genomewide average recombination rates. Our results suggested that the recombination rate of the soybean genome is not unusually high.

Identification of introgression regions in Hwangkeum

Four PIRs in Hwangkeum from *G. soja* were detected from the genotyping of indel markers developed in this study. Numerous introgression regions were recently predicted from the analysis of about 5× coverage resequencing data of 31 soybean accessions (Lam et al. 2010). In this study, we predicted the genomewide introgression region by aligning 21× coverage of genome resequencing data of Hwangkeum generated in this study and those of IT182932 reported by Kim et al. (2010) to the *G. max* reference sequence. A Williams 82 line (referred to as Williams 82K) that have been maintained in Korea Research Institute of Bioscience and Biotechnology (KRIBB) was resequenced to 17× coverage as a check. For each of the three accessions, the alignment indicated that more than 93 % of the reference genome was covered with at least five sequencing reads (Table S7). The mapping of the resequencing data of Williams 82K showed a degree of SNP call rate (approximately 1,000–2,000 SNP/100 kb) on small regions of Chrs 3, 7, 12, 14, 15, and 20 (Fig. 3; Fig. S2), consistent with the Kingwa introgression spots detected in the Illumina Infinium SNP genotyping (Haun et al. 2011) and showed extremely low or no SNP variation in the other chromosomes. This indicates that our mapping of the resequencing data is sound. The number of

SNP detected in IT182932 was approximately two times higher than that in Hwangkeum. Of the approximately 3.1 million SNPs, which were predicted from IT182932 and Hwangkeum against the Williams 82 sequence assembly, 473,886 (38.3 % of Hwangkeum and 19.7 % of IT182932) were shared between the two variants (Fig. S3).

To quantitatively examine the correlation of the distribution of SNPs along each chromosome in Hwangkeum versus IT182932 soybeans, the number of SNPs in a sliding 100-kb window was plotted along the sequence map of each chromosome. The average number of SNPs per 100-kb window was 130 in Hwangkeum and 252 in IT182932 (Fig. 3; Fig. S2). The genotypes of SNPs in a sliding 100-kb window were scored for each individual and the percentage of shared SNPs of Hwangkeum per the number of SNPs for IT182932 was calculated in each window. Among the regions where there were more SNPs in both Hwangkeum and IT182932 than the average, more than 50 chromosomal regions were identified as PIRs in Hwangkeum where the percentage of shared SNPs was higher than 50 %. Half of these PIRs were identified in the gene-rich chromosomal regions. Interestingly, they always corresponded to one of the approximately 40 chromosomal regions where Hwangkeum had higher number of SNPs than IT182932. The remaining PIRs were identified in the gene-poor pericentromeric regions. Interestingly, several of those PIRs, which were located on Chrs 3, 16, 18, and 19, contained much higher-than-average numbers of SNPs and showed higher than 70 % shared SNPs. It would be interesting to further explore how and when such PIRs in the pericentromeric regions where recombination rates are extremely low have originated. However, they may not be identified by conventional genetic mapping but by other methods including genome resequencing performed in this study. Two chromosomal regions, which were on Chr 2 and Chr 10, showed higher-than-average SNPs in Hwangkeum and much lower-than-average SNPs in IT182932. We presume that these two regions may be putative introgression regions in Williams 82 from *G. soja*. Among the four PIRs that were presumed to be introgression regions in Hwangkeum from *G. soja* on the basis of indel marker genotyping (Table S3), three that were predicted on Chr 3, Chr 9, and Chr 15 were supported by more than 50 % shared SNP and one that was predicted on Chr 6 was supported by more than 40 % shared SNPs. The potential introgression regions on Chr 6 contained a misoriented Williams 82 assembly region, thereby indicating that shared SNP percentage might be underestimated. Therefore, our PIR identification results that used resequencing data of two parental soybean genomes in our mapping population not only supported the PIRs predicted by the marker genotyping but also provided many additional PIRs.

Discussion

The development of a comprehensive genomewide genetic map for soybean provides a resource for understanding the dynamic genetic features of its physical chromosomes. Using publicly available markers as well as markers generated from indel polymorphisms detected by comparing a cultivated and a wild soybean genome sequence (Kim et al. 2010), we have developed a comprehensive soybean genetic linkage map of densely and evenly spaced markers in a single population. Relative to the previously published soybean maps that had several intervals larger than 10 cM in which no markers were present, the average intervals between markers in the current map was 1.5 cM and the largest interval was 5.9 cM. Comparison of the genetic linkage map with the soybean draft genome sequence (Schmutz et al. 2010) revealed that colinearity of the genetic and physical maps was mostly conserved with putative erroneously assembled or naturally occurring discordant regions in 7.5 % of the draft genome sequence. We then obtained convincing evidence to affirm the correct placement of five misoriented regions, estimated to be the largest portion of the misassembled regions. In addition, one unanchored scaffold was positioned by a long PCR approach. Seven marker sequences indicated that small portions of the pseudomolecules of the draft soybean genome sequence likely consisted of misplaced scaffolds. Two mapped markers were absent in the current Williams 82 assembly. These results suggested that most of the non-collinear regions between the genetic and sequence-based physical maps detected in this study likely represent errors in the Williams 82 assembly, although some of the regions may represent chromosomal rearrangements, as evidenced by the identification of a reciprocal translocation of chromosomal segments between chromosomes 11 and 13 in two wild soybean accessions (Findley et al. 2010). In other words, although the current Williams 82 sequence assembly did incorporate information of the 356,157 BAC end sequences, the soybean genome sequence essentially obtained from the whole genome shotgun sequence assembly (Schmutz et al. 2010) likely has some degree of assembly error as observed in the whole genome shotgun assemblies of rice and maize (International Rice Genome Sequencing Project 2005; Ganal et al. 2011). A comparison of the map-based sequence and two whole genome shotgun assemblies of the rice genome revealed that a substantial portion of the contigs from the two assemblies were non-homologous, misaligned, or provided duplicate coverage (International Rice Genome Sequencing Project 2005). The draft genome sequence of maize was obtained in quite a similar way to that of soybean. A comparison between the maize genome sequence and the two independent genetic maps revealed that 172 mapped markers were absent in the current maize assembly and, among many, five putative major misassembled regions (Ganal et al. 2011). Taken together, our results suggested that the soybean draft genome sequence contains some degree of inaccuracy, an inevitable downfall in the whole genome shotgun sequence assembly method. Therefore, our results can be used for future improvements of the Williams 82 reference sequence.

Genetic recombination is a vital component of crop genetics and breeding, as it generates new combinations of genes available for crop improvement. Genomewide distributions of recombination have been described for many animal and plant species including human, mouse, *Arabidopsis*, rice, maize, tomato, wheat, and *Brachypodium* (Gaut et al. 2007; Huo et al. 2011). In these cases, recombination rates in the repeat-rich centromeric and pericentromeric regions are much lower than rates in the gene-rich distal regions (Drouaud et al. 2006; Jensen-Seaman et al. 2004; Wright et al. 2003; Wu et al. 2003). Although variations of recombination rates along soybean chromosomes were previously examined using composite maps (Schmutz et al. 2010; Ott et al. 2011), the quantitative relationship between the recombination rate and the gene or the TE density in soybean was poorly addressed likely because of the limitation of the composite maps. We identified the physical locations of the marker sequences on the soybean chromosomes. Then, we studied the physical distribution of recombination on chromosomes by comparing the linkage map with the sequence-based physical map. The corrected pseudomolecules were analyzed to determine the recombination rate. The uneven distribution patterns of recombination on chromosomes were clear. High recombination was generally observed in the telomere regions as compared to the centromere regions. These observations are in line with several previous studies of other plant species, which showed a suppressed recombination in centromeric and pericentromeric regions and enhanced recombination in telomeric regions (Chen et al. 2002; Schnable et al. 2009; Saintenac et al. 2009; Yu et al. 2009). In agreement with studies in rice, maize, wheat, and *Brachypodium* (Anderson et al. 2006; Dvorak et al. 2004; Tian et al. 2009; Huo et al. 2011), our genomic composition study indicated that the recombination rate correlated positively with gene density and negatively correlated with TE density in soybean. The adjusted recombination rate value of soybean calculated by dividing the map length by the sequenced genome size excluding TE was similar to those of the highly studied rice and maize, suggesting that the overall recombination rate of soybean is not unusually high. Collectively, the overall pictures of recombination rate along chromosomes were not different from those of other organisms.

Previous genomewide genetic mapping efforts of soybean have frequently ended up with 20 plus linkage groups, which is more than the soybean haploid chromosome number, with many intervals longer than 20 cM. The marker-rich regions were interpreted to be closely associated with gene-rich regions in terms of a relationship of marker and gene distributions (Choi et al. 2007; Cregan et al. 1999; Marek et al. 2001; Mudge et al. 2004; Ott et al. 2011), as attempted in a number of higher plants (Sandhu and Gill 2002; Erayman et al. 2004; Carels et al. 1995; Barakat et al. 1997; Anderson et al. 2006). Genotyping of candidate indel markers in soybean accessions revealed that the four chromosomal regions that are gene-rich and marker-poor are the PIRs in Hwangkeum from *G. soja*. The mapping of the resequencing data of Hwangkeum and IT182932 against the draft genome sequence supported the four regions as PIRs, and far more such regions like these were identified. Given that Lam et al. (2010) predicted a total of 431 potential regions of introgression from the resequencing analysis of 31 wild and cultivated soybean genomes, approximately 50 PIRs predicted in this study may not be surprising. As to which portions of the PIRs identified by the simple comparison method of two *G. max* and *G. soja* data are true introgression regions needs to be rigorously tested by analyzing the resequencing data of multiple soybean accessions. Nevertheless, our results suggested that marker-poor chromosomal regions may not always correspond to gene-poor and low polymorphic regions, but some of those regions may represent introgression regions in *G. max* from *G. soja*, which are prevalent in soybean (Lam et al. 2010). Several important crop species, including tomato, barley, and wheat, have been improved by use of their wild relatives. However, wild soybean has been poorly used for soybean improvement with limited success (Stupar 2010). Consistent with a recent report that introgressions are prevalent in soybeans (Lam et al. 2010), our results suggest that the influence of wild introgressions on the soybean germplasm has been underestimated. Further studies of the predicted introgression regions should provide insight into the ancient breeding history as well as the utility of wild soybeans in improving cultivated soybeans.

Acknowledgements This work was supported mainly by a grant from the Next-Generation BioGreen 21 Program (No. PJ008124), Rural Development Administration, and partly by the Korea Research Institute of Bioscience and Biotechnology Research Initiative Program. We are grateful to Dr. Perry Cregan for his critical comments on an early version of this paper and Dr. Scott Jackson for discussions.

References

Anderson JA, Churchill GA, Autrique JE, Tanksley SD, Sorrells ME (1993) Optimizing parental selection for genetic linkage maps. Genome 36:181–186

Anderson LK, Lai A, Stack SM, Rizzon C, Gaut BS (2006) Uneven distribution of expressed sequence tag loci on maize pachytene chromosomes. Genome Res 16:115–122

Arumuganathan K, Earle ED (1991) Nuclear DNA content of some important plant species. Plant Mol Biol Rep 9:208–218

Barakat A, Carels N, Bernardi G (1997) The distribution of genes in the genomes of Gramineae. Proc Natl Acad Sci USA 94:6857–6861

Bernatzky R, Tanksley SD (1986) Toward a saturated linkage map in tomato based on isozymes and random cDNA sequences. Genetics 112:887–898

Boerma R, Wilson R, Ready E (2011) Soybean genomics research program strategic plan. Plant Genome 4:1–11

Carels N, Barakat A, Bernardi G (1995) The gene distribution of the maize genome. Proc Natl Acad Sci USA 92:11057–11060

Chen M, Presting G, Barbazuk WB, Goicoechea JL, Blackmon B, Fang G, Kim H, Frisch D, Yu Y, Sun S et al (2002) An integrated physical and genetic map of the rice genome. Plant Cell 14:537–545

Choi IY, Hyten DL, Matukumalli LK, Song Q, Chaky JM, Quigley CV, Chase K, Lark KG, Reiter RS, Yoon MS et al (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 et al (1999) An integrated genetic linkage map of the soybean. Crop Sci 39:1464–1490

DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M et al (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 43:491–498

Drouaud J, Camilleri C, Bourguignon PY, Canaguier A, Berard A, Vezon D, Giancola S, Brunel D, Colot V, Prum B et al (2006) Variation in crossing-over rates across chromosome 4 of *Arabidopsis thaliana* reveals the presence of meiotic recombination “hot spots”. Genome Res 16:106–114

Dvorak J, Yang ZL, You FM, Luo MC (2004) Deletion polymorphism in wheat chromosome regions with contrasting recombination rates. Genetics 168:1665–1675

Erayman M, Sandhu D, Sidhu D, Dilbirligi M, Baenziger PS, Gill KS (2004) Demarcating the gene-rich regions of the wheat genome. Nucleic Acids Res 32:3546–3565

Esch E, Szymaniak JM, Yates H, Pawlowski WP, Buckler ES (2007) Using crossover breakpoints in recombinant inbred lines to identify quantitative trait loci controlling the global recombination frequency. Genetics 177:1851–1858

Findley SD, Cannon S, Varala K, Du J, Ma J, Hudson ME, Birchler JA, Stacey G (2010) A fluorescence in situ hybridization system for karyotyping soybean. Genetics 185:727–744

Gagneux P, Boesch C, Woodruff DS (1997) Microsatellite scoring errors associated with noninvasive genotyping based on nuclear DNA amplified from shed hair. Mol Ecol 6:861–868

Ganal MW, Durstewitz G, Polley A, Bérard A, Buckler ES, Charcosset A, Clarke JD, Graner EM, Hansen M, Joets J et al (2011) A large maize (*Zea mays* L.) SNP genotyping array: development and germplasm genotyping, and genetic mapping to compare with the B73 reference genome. PLoS ONE 6:e28334. doi:10.1371/journal.pone.0028334

Gaut BS, Wright SI, Rizzon C, Dvorak J, Anderson LK (2007) Recombination: an underappreciated factor in the evolution of plant genomes. Nat Rev Genet 8:77–84

Haun WJ, Hyten DL, Xu WW, Gerhardt DJ, Albert TJ, Richmond T, Jeddeloh JA, Jia G, Springer NM, Vance CP et al (2011) The composition and origins of genomic variation among individuals of the soybean reference cultivar Williams 82. Plant Physiol 155:645–655

Hu XS, Goodwillie C, Ritland KM (2004) Joining genetic linkage maps using a joint likelihood function. Theor Appl Genet 109:996–1004

Huo N, Garvin DF, You FM, McMahon S, Luo MC, Gu YQ, Lazo GR, Vogel JP (2011) Comparison of a high-density genetic linkage map to genome features in the model grass *Brachypodium distachyon*. Theor Appl Genet 123:455–464

Hwang TY, Sayama T, Takahashi M, Takada Y, Nakamoto Y, Funatsuki H, Hisano H, Sasamoto S, Sato S, Tabata S et al (2009) High-density integrated linkage map based on SSR markers in soybean. DNA Res 16:213–225

International Rice Genome Sequencing Project (2005) The map-based sequence of the rice genome. Nature 436:793–800

Jensen-Seaman MI, Furey TS, Payseur BA, Lu Y, Roskin KM, Chen CF, Thomas MA, Haussler D, Jacob HJ (2004) Comparative recombination rates in the rat, mouse, and human genomes. Genome Res 14:528–538

Kim MY, Lee S, Van K, Kim TH, Jeong SC, Choi IY, Kim DS, Lee YS, Park D, Ma J et al (2010) Whole-genome sequencing and intensive analysis of the undomesticated soybean (*Glycine soja* Sieb. and Zucc.) genome. Proc Natl Acad Sci USA 107:22032–22037

Kuittinen H, de Haan AA, Vogl C, Oikarinen S, Leppälä J, Koch M, Mitchell-Olds T, Langley CH, Savolainen O (2004) Comparing the linkage maps of the close relatives *Arabidopsis lyrata* and *A. thaliana*. Genetics 168:1575–1584

Lam HM, Xu X, Liu X, Chen W, Yang G, Wong FL, Li MW, He W, Qin N, Wang B et al (2010) Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection. Nat Genet 42:1053–1059

Lander ES, Green P, Abrahamson J, Barlow A, Daly MJ, Lincoln SE, Newberg LA (1987) MAPMAKER: an interactive computer package for constructing primary genetic linkage maps of experimental and natural populations. Genomics 1:174–181

Lewin HA, Larkin DM, Pontius J, O’Brien SJ (2009) Every genome sequence needs a good map. Genome Res 19:1925–1928

Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 25:1754–1760

Marek LF, Mudge J, Darnielle L, Grant D, Hanson N, Paz M, Huihuang Y, Denny R, Larson K, Foster-Hartnett D et al (2001) Soybean genomic survey: BAC-end sequences near RFLP and SSR markers. Genome 44:572–581

McNally KL, Childs KL, Bohnert R, Davidson RM, Zhao K, Ulat VJ, Zeller G, Clark RM, Hoen DR, Bureau TE et al (2009) Genomewide SNP variation reveals relationships among landraces and modern varieties of rice. Proc Natl Acad Sci USA 106:12273–12278

Morrell PL, Buckler ES, Ross-Ibarra J (2012) Crop genomics: advances and applications. Nat Rev Genet 13:85–96

Mudge J, Huihuang Y, Denny RL, Howe DK, Danesh D, Marek LF, Retzel E, Shoemaker RC, Young ND (2004) Soybean bacterial artificial chromosome contigs anchored with RFLPs: insights into genome duplication and gene clustering. Genome 47:361–372

Nachman MW (2002) Variation in recombination rate across the genome: evidence and implications. Curr Opin Genet Dev 12:657–663

Ott A, Trautschold B, Sandhu D (2011) Using microsatellites to understand the physical distribution of recombination on soybean chromosomes. PLoS ONE 6:e22306. doi:10.1371/journal.pone.0022306

Rockman MV, Kruglyak L (2009) Recombinational landscape and population genomics of *Caenorhabditis elegans*. PLoS Genet 5:e1000419. doi:10.1371/journal.pgen.1000419

Rozen S, Skaletsky HJ (2000) Primer3 on the WWW for general users and for biologist programmers. In: Krawetz S, Misener S (eds)

Bioinformatics methods and protocols: methods in molecular biology. Humana Press, Totowa, pp 365–386

Saghai Maroof MA, Glover NM, Biyashev RM, Buss GR, Grabau EA (2009) Genetic basis of the low-phytate trait in the soybean line CX1834. Crop Sci 49:69–76

Saghai-Maroof MA, Soliman KM, Jorgensen RA, Allard RW (1984) Ribosomal DNA spacer length polymorphism in barley: Mendelian inheritance, chromosomal location, and population dynamics. Proc Natl Acad Sci USA 81:8014–8018

Saintenac C, Falque M, Martin OC, Paux E, Feuillet C, Sourdille P (2009) Detailed recombination studies along chromosome 3B provide new insights on crossover distribution in wheat (*Triticum aestivum* L.). Genetics 181:393–403

Sanchez-Moran E, Armstrong SJ, Santos JL, Franklin FC, Jones GH (2002) Variation in chiasma frequency among eight accessions of *Arabidopsis thaliana*. Genetics 162:1415–1422

Sandhu D, Gill KS (2002) Gene-containing regions of wheat and the other grass genomes. Plant Physiol 128:803–811

Schmutz J, Cannon SB, Schlueter J, Ma J, Mitros T, Nelson W, Hyten DL, Song Q, Thelen JJ, Cheng J et al (2010) Genome sequence of the palaeopolyploid soybean. Nature 463:178–183

Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, Liang C, Zhang J, Fulton L, Graves TA et al (2009) The B73 maize genome: complexity, diversity, and dynamics. Science 326: 1112–1115

Song QJ, Marda 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

Song QJ, Jia GF, Zhu YL, Grant D, Nelson RT, Hwang EY, Hyten DL, Cregan PB (2010) Abundance of SSR motifs and development of candidate polymorphic SSR markers (BARCSOYSSR_1.0) in soybean. Crop Sci 50:1950–1960

Stacey G, Vodkin L, Parrott WA, Shoemaker RC (2004) National Science Foundation-sponsored workshop report. Draft plan for soybean genomics. Plant Physiol 135:59–70

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

Stupar RM (2010) Into the wild: the soybean genome meets its undomesticated relative. Proc Natl Acad Sci USA 107: 21947–21948

Suh SJ, Bowman BC, Jeong N, Yang K, Kast C, Tolin SA, Saghai Maroof MA, Jeong SC (2011) The *Rsv*3 locus conferring resistance to Soybean mosaic virus is associated with a cluster of coiled-coil nucleotide-binding leucine-rich repeat genes. Plant Genome 4:55–64

Tian Z, Rizzon C, Du J, Zhu L, Bennetzen JL, Jackson SA, Gaut BS, Ma J (2009) Do genetic recombination and gene density shape the pattern of DNA elimination in rice long terminal repeat retrotransposons? Genome Res 19:2221–2230

Tulsieram L, Compton WA, Morris R, Thomas-Compton M, Eskridge K (1992) Analysis of genetic recombination in maize populations using molecular markers. Theor Appl Genet 84:65–72

Van Ooijen JW (2006) JoinMap®4, software for the calculation of genetic linkage maps in experimental population. Kyazma B.V., Wageningen

Voorrips RE (2002) MapChart: software for the graphical presentation of linkage maps and QTLs. J Hered 93:77–78

Williams CG, Goodman MM, Stuber CW (1995) Comparative recombination distances among *Zea mays* L. inbreds, wide crosses and interspecific hybrids. Genetics 141:1573–1581

Wright SI, Agrawal N, Bureau TE (2003) Effects of recombination rate and gene density on transposable element distributions in *Arabidopsis thaliana*. Genome Res 13:1897–1903

Wu J, Mizuno H, Hayashi-Tsugane M, Ito Y, Chiden Y, Fujisawa M, Katagiri S, Saji S, Yoshiki S, Karasawa W et al (2003) Physical maps and recombination frequency of six rice chromosomes. Plant J 36:720–730

Yang K, Moon JK, Jeong N, Back K, Kim HM, Jeong SC (2008) Genome structure in soybean revealed by a genomewide genetic map constructed from a single population. Genomics 92:52–59

Yang K, Jeong N, Moon JK, Lee YH, Lee SH, Kim HM, Hwang CH, Back K, Palmer RG, Jeong SC (2010) Genetic analysis of genes controlling natural variation of seed-coat and flower colors in soybean. J Hered 101:757–768

Yu Q, Tong E, Skelton RL, Bowers JE, Jones MR, Murray JE, Hou S, Guan P, Acob RA, Luo MC et al (2009) A physical map of the papaya genome with integrated genetic map and genome sequence. BMC Genomics 10:371. doi:10.1186/1471-2164-10-371