DNA Research **15**, 93–102, (2008) doi:10.1093/dnares/dsn001

Sequence Level Analysis of Recently Duplicated Regions in Soybean [*Glycine max* (L.) Merr.] Genome

Kyujung Van[1], Dong Hyun Kim[1], Chun Mei Cai[2], Moon Young Kim[1,3], Jin Hee Shin[1], Michelle A. Graham[4], Randy C. Shoemaker[4], Beom-Soon Choi[5], Tae-Jin Yang[1], and Suk-Ha Lee[1,3,*]

Department of Plant Science, Seoul National University, San 56-1, Sillim-dong, Gwanak-gu, Seoul 151-921, South Korea[1]; National Institute of Crop Science, Suwon 441-857, South Korea[2]; Research Institute for Agriculture and Life Sciences, Seoul National University, Seoul 151-921, South Korea[3]; Corn Insect and Crop Genetics Research Unit, USDA-ARS, Iowa State University, Ames, IA 50011, USA[4] and National Instrumentation Center for Environmental Management, Seoul National University, Seoul 151-921, South Korea[5]

(Received 15 November 2007; accepted on 16 January 2008; published online 11 March 2008)

Abstract

A single recessive gene, *rxp*, on linkage group (LG) D2 controls bacterial leaf-pustule resistance in soybean. We identified two homoeologous contigs (GmA and GmA′) composed of five bacterial artificial chromosomes (BACs) during the selection of BAC clones around *Rxp* region. With the recombinant inbred line population from the cross of Pureunkong and Jinpumkong 2, single-nucleotide polymorphism and simple sequence repeat marker genotyping were able to locate GmA′ on LG A1. On the basis of information in the Soybean Breeders Toolbox and our results, parts of LG A1 and LG D2 share duplicated regions. Alignment and annotation revealed that many homoeologous regions contained kinases and proteins related to signal transduction pathway. Interestingly, inserted sequences from GmA and GmA′ had homology with transposase and integrase. Estimation of evolutionary events revealed that speciation of soybean from *Medicago* and the recent divergence of two soybean homoeologous regions occurred at 60 and 12 million years ago, respectively. Distribution of synonymous substitution patterns, K_s, yielded a first secondary peak (mode $K_s = 0.10-0.15$) followed by two smaller bulges were displayed between soybean homologous regions. Thus, diploidized paleopolyploidy of soybean genome was again supported by our study.

Key words: BAC; divergence time; duplication; K_s; *Rxp*; soybean

1. Introduction

Legumes have begun to draw much attention through recent genomic and phylogenetic studies.[1] The crop legumes, such as *Lotus*, *Medicago*, *Pisum*, *Glycine*, *Phaselous*, and *Vigna*, also receive attention from researchers because they are economically important.[2] Asian countries have a long history of making different food products, such as soymilk, tofu, soybean sprouts, etc., with soybean seeds because its seed obtains high protein and oil content. Thus, soybean is considered a very valuable crop among legumes.[3] As approximately 20 000 species belong in the legume family, a wide range of genetic and morphological diversities can be observed.[2] Unlike many other plants, legumes have a symbiotic relationship with the soil-borne bacteria, *Rhizobia*. *Medicago truncatula* and *Lotus japonicus* were selected as model legume plants because these legumes had not only plant–*Rhizobium* interactions

Edited by Satoshi Tabata

* To whom correspondence should be addressed. Tel. +82 2-880-4545. Fax. +82 2-873-2056. E-mail: sukhalee@snu.ac.kr

for nitrogen fixation but also small genomes suitable for full genome sequencing.[4,5]

Most Papilionoids are diploids except *Glycine*. An ancient genome duplication occurred in *Glycine*, leading to $2n = 38$, 40 or 78–80 depending on annual/perennials or geographic locations.[6,7] Polyploidy has had an evolutionary impact on the structure of the soybean genome.[8–10] Using restriction fragment-length polymorphism (RFLP) analysis with nine populations (*Glycine max* × *G. soja* and *G. max* × *G. max*) of the *Glycine* subgenus *soja*, it was shown that the soybean genome presents about 2.55 copies per digest. This suggests that an additional round of genome duplication might have occurred in at least one of the original genomes.[8]

Other studies have supported those observations. RFLP and simple sequence repeat (SSR) analyses showed that parts of linkage groups (LGs) B1/S, H, and F of soybean genome shared homoeologous regions.[9] Other genetic mapping analysis suggested that extensive rearrangements and additional duplications were present in soybean genome.[10] Also, high similarity in physical organization between soybean duplicated regions and a high percentage of microsynteny were shown by characterizing bacterial artificial chromosome (BAC) clones of soybean and other model plants.[11,12] In addition, BACs containing FAD2 genes also contained a number of syntenic genes and were positioned on LG I and O, again indicating duplication of soybean genome.[13] Fluorescence *in situ* hybridization of BACs visualized segmental duplications within the soybean genome.[14] *M. truncatula* genome also presents segmental duplications identified by high-throughput genome sequencing.[15]

The processes of genome evolution and patterns of divergence can be studied by duplicate gene analysis.[16] Because the full genome sequence of many plants is not yet available, ESTs provide resources for studying evolutionary events such as ancient bursts of gene duplications. Because the accumulation of synonymous substitutions occurs stochastically over time, the level of divergence (age of duplication) is estimated by nucleotide substitution in coding sequences[2,17] Putative genome duplications events were identified with large EST collections from eight plant species using synonymous substitution measurements (K_s) of duplicated genes.[18] Soybean was estimated to have had two major genome duplications events at 15 million years ago (MYA) and 44 MYA. A genome duplication event also was observed in *M. truncatula* at ~58 MYA. With different calibration, duplications also were observed in both soybean and *M. truncatula*.[17] A mutigene approach combined with a phylogenetic approach suggested soybean and *Medicago* shared a round of gene duplications, along with about 7000 other legume plants.[19]

Xanthomonas axonopodis pv. *glycine* (*Xag*) causes bacterial leaf pustule (BLP) in soybean that occurs in Korea and the southern United States, where hot and humid weather conditions are prevalent.[20] Typically, small yellow-to-brown lesions with a raised lesion are formed in early development and develop into large necrotic lesions causing substantial losses in yield through premature defoliation.[21–24]

Twenty consensus LGs of soybean genome, representing the 20 soybean chromosomes, were reported[25] with a joined map from three different populations spanning 2400 cM in length. A total of 420 SSR markers were added to the integrated genetic linkage map and its length was expanded to 2523.6 cM of Kosambi map distance across 20 LGs.[26] And, 1141 single-nucleotide polymorphism (SNP) markers were later located on the soybean genetic map.[27] Among 20 LGs, we are interested in LG D2 because the recessive gene conditioning resistance to BLP, *rxp*,[28] was mapped to LG D2 only 3.9 cM away from Satt372.[29] Also, the *Rxp* locus linked to the malate dehydrogenase (*Mdh*) locus with an estimated recombination frequency of $15.2 \pm 3.8\%$.[30] In the process of BAC clone selection for 'chromosome walking' around *Rxp* region, we were able to create two contigs, which represent homoeologous regions of the soybean genome. Here, we describe the consequences of the duplication events around the *Rxp* region. Annotation, gene arrangement, and evolution events estimated by K_s (the number of synonymous substitution per synonymous site) will also be presented.

2. Methods

2.1. Primary BAC library screening

The constructed 'Iowa State' BAC library of soybean 'Williams 82' (gmw1)[31] was used for selection of BAC clones around *Rxp* locus. For the first round of PCR-based library screening, Satt 372 (forward, 5′-CAG AAA AGG AAT AAT AAC AAC ATC AC-3′; reverse, 5′-GCG AAA ACA TAA TTC ACA CAA AAG ACA G-3′), Satt486 (forward, 5′-GCG CAT GCA TTA CCA TAG GCT ATA ATA -3′; reverse, 5′-GGG GTC ATG CAT AAT AGA GAT AGA ACC A-3′), and Satt498 (forward, 5′-CAA CCC CGA AAT ACA ACT AAT GTT-3′; reverse, 5′-TGG TGA GGC TCA TTT TCA TAA GA-3′) were used as PCR primers, and pooled DNAs made by copies of the library by combining the overnight cell culture[31] were used as templates.

Basic PCR protocols were followed as described with minor alternations,[31] using a PTC-110 Peltier Thermal Cycler (MJ Research, Inc., Watertown, MA, USA). The components of the reaction mixture in

20 μL of total volume were 0.5 U of *Taq* polymerase (Invitrogen, Carlsbad, CA, USA) and the rests of components were the same.[31] Cycling conditions started with initial denaturation at 94°C for 3 min, followed by 35 cycles of 94°C for 30 s, 48°C for 30 s, and 72°C for 30 s and the final step was at 72°C for 2 min. The amplified PCR products were analyzed in 1.5% ethidium bromide-stained agarose.

Several rounds of the BAC library were screened systematically with in order, full-plate super pool DNA, individual full-plate pools, row and column super pools, and row and column pools. All PCRs were performed as describe earlier and DNA of 'Williams 82' was used as positive control for all screening processes.

2.2. Shotgun plasmid library and DNA sequencing

After BAC DNAs were prepared by a Plasmid Midi Kit (Qiagen, Hilden, Germany) and the insert size of each selected BAC clone was estimated,[31] the random plasmid library for shotgun sequencing was constructed with 10–15 μg of the extracted BAC clone DNA and pUC118 vector using Takara BKL Kit (Takara Bio, Inc., Otsu, Japan). The rest of methods were performed as described previously.[32]

Full sequencing of each BAC DNA was performed with the BigDye Terminator (v. 3.1) cycle sequencing kit (Applied Biosystems, Foster City, CA, USA). Cycle conditions for sequencing and analysis of BAC sequences were described.[32] Also, the individual sequences were assembled with Phred/Phrap software and the remaining gaps of each clone were closed by direct sequencing, using plasmid DNA.[32] Image v. 3.0 and FPC v. 4.7.9 were used for confirmation of BAC contig assignment.[33,34]

2.3. Secondary BAC library screening and sequence analysis

After the sequences of each BAC clone were aligned,[35] BAC end sequences (BES) were selected for extending BAC contigs. Primer3 program[36] was used to design primers for secondary BAC library screening. With primers derived from BES, the BAC library screening was performed again as described earlier. Addition to 'Iowa State' BAC library, 'Missouri' soybean BAC library (gmw2) consisted of *Bsty*I partially digested 'Williams 82' DNA was also used for screening.[37]

After BAC contigs were confirmed, alignment between BAC contigs and its alignment results were inspected with GBrowse (http://www.gmod.org/?q=node/71) and SynBrowse (http://www.synbrowse.org/). Also, gene annotation was conducted with the web-based gene prediction programs FGENESH (http://sun1.softberry.com/berry.phtml) and GeneMark (http://exon.gatech.edu/GeneMark/) against Medicago (legume plant) database. Putative amino acid sequences from the predicted genes were used as queries for searching similar known proteins using BLASTP. With each predicted gene of GmA at first, EST information was searched against *G. max* EST database at *G. max* Genome Database (http://bionary.agry.purdue.edu/GmaxGDB/index.php). Nucleotide blast or tblastn (http://www.ncbi.nlm.nih.gov/blast/Blast.cgi) was also used for searching EST information with each predicted gene of GmA′ or MtA, if no ESTs corresponding to the predicted genes in GmA were identified.

The rate of non-synonymous nucleotide substitution (K_a) and the fraction of synonymous substitutions (K_s) were obtained with the CODEML program[38] of the PAML package.[39] K_s was used to estimate the divergence time between two sequences. So, coding sequences of the predicted gene from two contigs were used for analysis of K_s, as described.[35] Divergence times (T) were calculated using a synonymous mutation rate of 6.1×10^{-9} substitutions per synonymous site per year[18,40] as $T = K_s/(2 \times 6.1 \times 10^{-9})$.

2.4. SNP detection, SNP genotyping, and generation of linkage map

To locate each BAC contig in LGs, contig-specific regions longer than 4.5 kb were surveyed. Seven different primer sets (Supplementary Table S1) were designed from these contig-specific regions using Primer3 (http://primer3.sourceforge.net/). And, the detection of SNP in the contig-specific regions between two soybean genotypes, Pureunkong and Jinpumkong 2, was followed.[3]

SNP capture probe (5′-GTT TTT TCA TCA ATC TTC CTC TAA A-3′) was designed to be complementary to the 5′ region from the SNP site within an amplicon using SBEPrimer version 1.1[41] and single base extension reactions followed by fluorescence polarization (FP) measurements were performed on a Victor3 microplate reader (PerkinElmer Life Science, Boston, MA, USA).[42] SNP primers were tested using genomic DNA of each parent and a mixture of both parents as artificial heterozygotes. The SNP primer was accepted as an SNP marker, only if the results of genotyping by AcycloPrime FP analysis were confirmed by sequencing data and then used for genotyping in a segregating population, an F_2-derived soybean population of 90 recombinant inbred lines (RILs) from the cross of Pureunkong and Jinpumkong 2.[43]

Genotyping data were automatically transferred to Microsoft Excel and the genotypes of the segregation population were determined, if the clusters were separated at least 40 mp (thousandth of the polarization unit) apart, at least seven times higher than standard deviation of the negatives (>99% at significant

level).[44] The construction of the linkage map with SNP marker genotyping data and integration of these markers on LGs were followed.[43] SNP genotyping data from the heterozygous line were considered as missing data. The five SSRs located on GmA′ sequences were additionally used for accurate mapping of GmA′, after the SSRs were identified by Sputnik: DNA microsatellite repeat search utility (http://cbi.labri.fr/outils/Pise/sputnik.html). LGs were designated according to the USDA genetic map,[26] and MapChart v. 2.1 was used for generation of linkage map.[45]

2.5. *Accession numbers*

Accession nos gmw1-20O10, EU028328; gmw1-24M16, EU028329; gmw1-29F06, EU028330; gmw1-89M01, EU028331; gmw2-77P21, EU028332. Sequence data from this article can be found in the GenBank/EMBL data libraries.

3. Results

3.1. *Identification of soybean duplicated regions by BAC selection*

To obtain BAC clones around the *Rxp* locus, we screened the gmw1 BAC library with three SSR markers, Satt372, Satt486, and Satt498. Among a total of six BAC clones identified by these SSR markers, we selected gmw1-29F06 and gmw1-24M16 for determining DNA sequences because they represented an SSR marker with long length (Fig. 1).

DNA sequences of gmw1-29F06 and gmw1-24M16 were aligned and created the GmA contig, comprised of 264 239 bp, including overlapped DNA sequences of ~73 kb (Fig. 1). Primers were then designed from BES of GmA for extending the contig. Clone gmw1-20O10 (120 kb) was selected from BES of gmw1-29F06 and primers designed from BES from gmw1-24M16 selected gmw1-89M01 to create an apparent extension of 175 kb. We were able to extend the contig longer with full sequenced gmw1-20O10 and gmw1-89M01 clones.

The full DNA sequences of gmw1-20O10 and gmw1-89M01 were compared with GmA, but DNA sequences of the expected overlapped regions showed only an approximate 90% match. To close the gap between gmw1-20O10 and gmw1-89M01, we were able to select gmw2-77P21 (94 kb) from another soybean BAC library, gmw2. After the DNA sequences of gmw2-77P21 were aligned with gmw1-20O10 and gmw1-89M01, the GmA′ contig (gmw1-20O10_ gmw2-77P21_ gmw1-89M01, 292 895 bp) was formed with 100% match (Fig. 1).

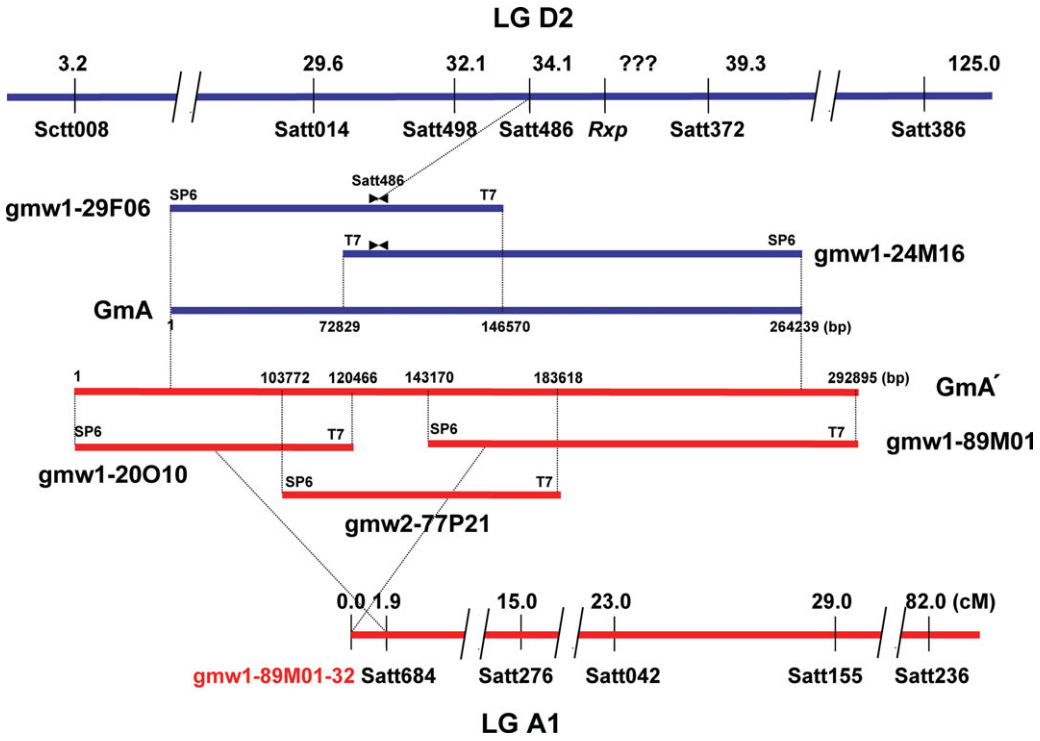


Figure 1. Schematic relationships between homologous regions (GmA and GmA′) containing the *Rxp* locus from LGs A1 (red) and D2 (blue). The positions (in cM) and their corresponding SSR markers are located on the upper and lower sides of the black solid and dashed bars for LG D2 and A1, respectively. Linkage map of LG D2 was taken from the Soybean Breeders Toolbox (http://soybase.org) and the mapped position of GmA′ is shown on the detailed genetic linkage map of LG A1 in the RIL population of Pureunkong × Jinpumkong 2. GmA was composed of gmw1-29F06 and gmw1-24M16. gmw1-20O10, gmw1-89M01 and gmw2-77P21 were made GmA′.

3.2. Mapping of soybean duplicated regions

To locate GmA′ on the soybean genetic linkage map, SNP genotyping was performed. First, unique regions longer than 4.5 kb in GmA′ were surveyed. Seven different unique regions were identified, and seven different primer sets were randomly designed from these seven contig-specific regions (Supplementary Table S1). One SNP locus between Pureunkong (deletion) and Jinpumkong 2 (A) was identified by primers (forward, 5′-TTC GTG CTA AGT GGA ACT TCT G-3′; reverse, 5′-TAC AAC AAC GAT GTT CAT GAC G-3′) designed between 159 723 and 160 465 bp of GmA′.

SNP genotyping of GmA′ was conducted with the RIL population from the cross of Pureunkong and Jinpumkong 2. The SNP marker locus was incorporated into the frame map,[43] placing GmA′ to the top of LG A1, 1.9 cM away from Satt684 in LG A1 (Fig. 1). Five SSRs identified by Sputnik on GmA′ were additionally analyzed. Only one designed SSR showed polymorphism between Pureunkong and Jinpumkong 2 (data not shown), and this was turned out to be Satt684, which was positioned between 64 495 and 64 682 bp of GmA′. On the basis of all genotyping and mapping data, we are able to determine that the duplicated regions are located on LG A1 for GmA′ and LG D2 for GmA (Fig. 1).

3.3. Alignments, annotations, and K_s estimation

After BAC contigs were confirmed and inspected with GBrowse and SynBrowse, genes were annotated with FGENESH or GeneMark against the Medicago database. Fig. 2 shows a schematic representation of approximate gene lengths, gene locations, and homologous regions (linked by shaded lines). The 54 and 58 genes were predicated in GmA and GmA′, respectively (Fig. 2). Gene density along these two sequenced BAC contigs was approximately one gene per 5.0 kb. A similar gene density (45 predicted genes along 219 028 bp) was also detected in *Medicago* along Contig 962B (MtA, as of January 2007 at http://www.medicago.org). The *Medicago* contig showed homology with the two soybean contigs. Gene order was conserved among syntenic blocks, except for one case (GmA_18 versus GmA′_10 and GmA_27 versus GmA′_10), and the same orientation between the predicted genes was observed. Gene order was maintained in *Medicago*, although linearity was fragmented (Fig. 2).

Supplementary Table S2 lists all pairwise comparisons of the predicted genes among homoeologous contigs. Segments showing no homology with known genes were excluded for this table. Also, the EST information corresponding to the predicted genes was included in Supplementary Table S2 after three

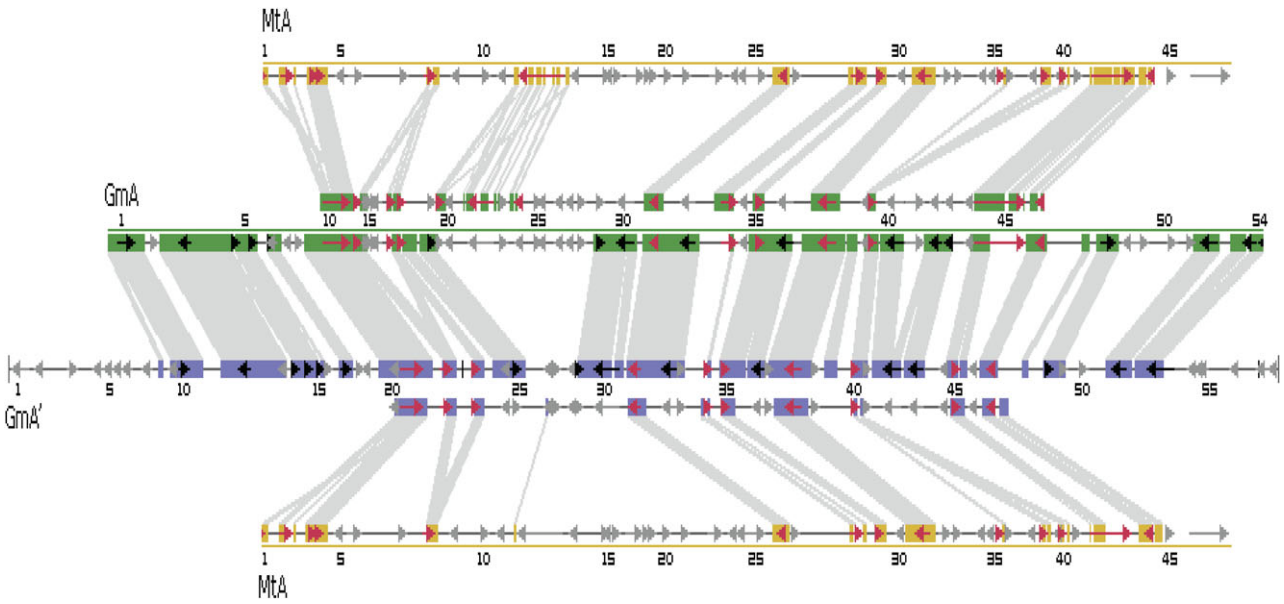
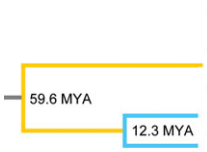


Figure 2. Comparative genome alignments among *M. truncatula* (MtA) and *G. max* (GmA and GmA′) based on discontinuous Megablast results. Colored arrows (red, common genes between *Medicago* and soybean contigs; black, common genes between soybean contigs; gray, unique genes) indicate the positions and orientation of predicated genes and their length presents the length of predicated gene. And, their locations were linked among three contigs, if the predicted genes were founded on all contigs. Darker color is used in filled link as higher *e*-value between two contigs was shown (Supplementary Table S2). Numbers indicate the number of genes annotated by FGENESH. A detailed description of gene annotations is listed in Supplementary Table S2.

BLAST programs, *G. max* Genome Database, nucleotide blast, and tblastn, were run. Eight regions were unique in GmA, whereas GmA′ had 13 unique regions. But, 22 among 45 segments of MtA were not similar to any of segments from the two soybean homoeologous regions. Twenty seven pairwise comparisons between soybean homoeologous regions were also observed in Supplementary Table S2. After each pairwise comparison was performed against BLASTP, nine of the 27 comparisons between soybean homoeologous regions showed >90% identity with putative genes with pretty low *e*-value. However, wide range of the conservation level was detected in the putative promoter regions and the introns between GmA and GmA′, averaging 59.7% and 54.9% in similarity for the putative promoter regions and the introns, respectively (Supplementary Table S2). Many homoeologous regions contained kinases and proteins related to signal transduction pathway. Interestingly, some unique segments from GmA and GmA′ showed homology with transposase (GmA′_07) or integrase (GmA_13, 14 and 15) (Supplementary Table S2). Among twenty seven pairwise comparisons between soybean homoeologous regions, nine comparisons showed alignment with MtA (Supplementary Table S2).

Using the maximum-likelihood method in the CODEML program,[39] synonymous (K_s) and non-synonymous (K_a) distance were estimated. This method was based on the F3 x 4 model of codon substitution,[38] explaining both transition/transversion and codon usage biases. Supplementary Table S3 shows the results of analysis between homologous gene pairs from each contig, along with percent identity of amino acid and cDNA sequences. The median K_a value (0.0426) between two soybean contigs was about 3.5 times smaller than the median K_s value (0.1472). Only one case in the K_a/K_s ratio was higher than 1 (1.5479, GmA_39 versus GmA′_40) (Supplementary Table S3). This K_a/K_s ratio might be non-significant because of moderate length of exons (282, 51, and 159 bp). When the substitutions per synonymous site (K_s) were plotted against the fraction of duplication events, secondary peaks were observed in the distribution for two contigs (data not shown). The first secondary peak (mode $K_s = 0.10-0.15$) followed by two smaller bulges (mode $K_s = 0.20-0.25$ and $0.30-0.35$) was displayed, indicating a burst of gene duplications in soybean. For GmA versus MtA, the median K_a value was 0.2888, which is 2.7 times smaller than the median K_s value (0.7755). MtA_12 showed homology with both GmA_21 and GmA_23. The K_s value for GmA_23 versus MtA_12 was extremely high because they aligned along 107 amino acids (Supplementary Table S3). The median K_a value (0.2876) between GmA′ and MtA was about 2.8 times smaller than the median K_s value (0.8003).



	MtA	GmA	GmA′
MtA		0.7654	0.6877
GmA	62.74		0.1498
GmA′	56.37	12.28	

Figure 3. K_s values and estimation of evolutionary events in three contigs. The number shown above the self-comparison diagonal represents estimation of median K_s values. Supplementary Table S3 showed statistics for K_s values between homologous regions of three contigs (GmA, GmA′ and MtA). And, divergence times in millions of years calculated as[18,40] $T = K_s/(2 \times 6.1 \times 10^{-9})$ are shown below the self-comparison diagonal. Colored boxes represent different evolutionary events: orange, *Medicago*–soybean speciation; sky blue, segmental duplication in soybean. Estimated dates of speciation and duplication events are given in the phylogenetic tree.

To determine the timing of the duplication event giving rise to the two contigs, the K_s value was used. Synonymous substitutions are thought to be evolutionarily neutral because the mutations cause no amino acid change[35] and therefore accumulate stochastically over time. K_s values less than 0.05 and greater than 1 were not included for searching for mixtures of normal distributions.[19] Divergence times (T) were estimated with K_s value and assumption of a mutation rate of 6.1×10^{-9} substitutions per synonymous site per year.[18,40] T ranged from 5.55 to 39.92 MYA between GmA and GmA′, and the median T was 12.3 MYA with low K_s value (0.1498) (Fig. 3, Supplementary Table S3). With MtA only included $0.05 < K_s < 1.0$, median K_s values were 0.7654 (0.5986 to 0.9342) and 0.6877 (0.4592 to 0.9128) for GmA and GmA′, respectively. Therefore, MtA and the soybean homoeologous contigs diverged at 56.4–62.7 MYA (Fig. 3, Supplementary Table S3). The two soybean homoeologous contigs were duplicated more recently, agreeing with the previous study.[17]

4. Discussion

4.1. *Paleopolypoidy of the soybean genome*

Diploidization or gene duplication is a process of switching from tetrasomic to disomic inheritance and a common process in plant genome evolution.[13,46] Many studies have suggested that paleopolyploidy is a common phenomenon in most plant species.[17,18] Previous studies suggested that the soybean genome has undergone two or more large-scale duplications and is probably an ancient polyploid.[8,11,18,47] In the present study, we identified and evaluated the duplication events in soybean genome with two contigs. In the process of full DNA sequencing of BAC clones, nucleotide sequences of gmw1-20O10 and gmw1-89M01 were not aligned

perfectly with GmA (gmw1-29F06_gmw1-24M16). Therefore, another round of BAC library screening was performed to close the gap between gmw1-20O10 and gmw1-89M01, and the GmA′ (gmw1-20O10_ gmw2-77P21_ gmw1-89M01) was made by selection of gmw2-77P21 (Fig. 1).

Although BAC-end sequences were used for BAC by BAC selection, the alignment of our two contigs was not perfect and gaps in alignment were observed. To locate these two homologous contigs, SNP genotyping was performed with the one SNP between Pureunkong (deletion) and Jinpumkong 2 (A). This SNP marker locus for GmA′ was located 1.9 cM away from Satt684 on LG A1, which SSR marker analysis was also able to be positioned between 64 495 and 64 682 bp of GmA′ (Fig. 1). Linkage maps from the Soybean Breeders Toolbox (http://soybase.org) were compared to locate the duplicated region (GmA versus GmA′). A comparison between the soybean composite maps for LG A1 and D2 indicated that homoeologous regions exist between them. In addition to the two homoeologous contigs, five RFLP markers were also common between the two LGs. Therefore, it suggested that GmA and GmA′ are indeed homoeologous.

4.2. *Genome dynamics among homoeologous regions*

Both soybean homoeologous contigs showed a gene density of 1 gene per 5 kb in this present study (Fig. 2). The gene density of soybean on LG G was estimated with 28 BAC ends[11] and subclone sequences from two contigs, to be approximately 1 gene per 14 kb. *Arabidopsis* and tomato showed similar gene density with an average of 1 gene per 5 kb, although more than a sevenfold difference in genome size is present between these two species.[11,48] Wheat, barley, and rice were compared near the Lrk locus and showed maximal density as 1 gene per 4–5 kb.[49]

The unusual relationship between physical distance and genetic distance for the homoeologous region on LG A1 was revealed by the comparison of the physical map and the genetic map in the distal region of LG A1 (Fig. 1). The physical distance between two markers (gmw1-89M01-32 and Satt684) on GmA′ was ~100 kb, whereas its genetic distance was observed to be 1.9 cM on the physical map of LG A1. Exhibition of 52.6 kb/cM in a physical-to-genetic distance ratio in this homoeologous region might represent high recombination region because not only duplication of genes and dispersed gene duplicates, which was single-gene duplications that were on different chromosome, occurred more often in high-recombination regions[50] but also the distal chromosomal regions showed high recombination rate.[50] And, rice genome showed high recombination frequency with 244 kb as the average physical distance per centimorgan, although a physical-to-genetic distance ratio was different depending on position along the chromosome.[51] High resolution mapping with various markers and genome sequencing would clarify the relationship between physical distance and genetic distance for this homoeologous region on LG A1.

Genic regions of the two soybean contigs and MtA retained gene structure in both order and orientation (Fig. 2). Similar conservations were also observed.[13,17,52] Sequence similarity was >60% in intergenic regions of most soybean homologous regions. This high level of similarity was also seen in previous studies,[11,13,52] suggestive of either a relatively recent duplication or concerted evolution.[53]

A homology search using BLASTP identified high homology with kinases and proteins related to signal transduction pathway within the homoeologous regions. Among them, the first aligned segment (GmA_01 versus GmA′_10) was very similar to putative receptor-like protein kinase INRPK1 (*Ipomoea nil* receptor-like protein kinase), which shows homology with Xa21, the rice pathogen recognition receptor.[54] INRPK1 and Xa21, as typical RPKs, were composed of extracellular leucine-rich repeat domain, transmembrane domain, and cytoplasmic kinase domain. Twenty one dominant loci responsible for resistance to bacterial blight disease indicated that a multiple gene family was involved in resistance to *Xanthomonas oryzae* pv. *oryzae*. And, sequence analysis between two classes of *Xa21* defined by the predicted amino acid sequence level suggested duplication was one of the roles in evolution of the *Xa21* gene family.[55] So, like *Xa21*, the soybean homoeologous regions contained important genes related to plant defense could be duplicated. However, it is difficult to say that *Rxp* encodes RPK for recognition of pathogen (*Xag*) by the extracellular domain and transduction of pathogen attack by an intracellular kinase because INRPK1 showed only 30% of homology with Xa21 at the amino acid sequence level.

Supplementary Table S2 provides the information on insertion/deletion of sequences in terms of genome dynamics in *Medicago* and soybean. Several blocks of segments from each contig showing no homology with any other homologous regions were identified, indicating insertion or deletion of sequences within homologous contigs. Some segments draw attention because they were similar to transposase from *Medicago* (GmA′_07) and integrase from *Medicago* (GmA_13, 14 and 15). Potential transposable element (TE) activity is interesting because TE could be used for transposon mutagenesis in gene cloning and functional genomics in plants.[56,57] Many researchers have put efforts into identifying

and cloning of soybean TE via finding the mutable alleles because no active TE has yet been isolated from soybean.[57] The *w4*-mutable line contained an autonomous TE at the *W4* locus and the unstable *k2 Mdh1-n y20* chromosomal region caused by a non-autonomous TE were identified in soybean.[57,58] Interestingly, this unstable chromosomal region was mapped on LG H and this *Mdh* gene also located on LG D2,[30] near the *Rxp* locus studied in this research.

4.3. Genome evolution

With comparisons of the sequences of the same gene from two species or gene family, counting of the number of non-synonymous changes (amino acid sequences change) and synonymous change (no change of amino acids sequences) is a good indicator of the degree of divergence between two sequences.[59] A total of 23 homologous regions between two soybean contigs showed K_a values ranging from 0.0168 to 0.2807 and K_s values ranging from 0.0318 to 0.4797 (Supplementary Table S3). It suggested that the K_a/K_s ratio could test for assessing the protein-coding potentials of genomic regions.[60] Depending on K_s as the background rate of evolution, the selection pressure in protein-coding regions could be explained by deviations of this ratio.[59] In this study, the ratio was less than 1 except for one comparison (GmA_39 versus GmA′_40) because of very low K_s. Short length of exons or highly divergent sequences (<70% nucleotide identity) could cause the K_a/K_s ratio to be higher than 1.[60] Although their percentage of identity at nucleotide level was >95%, this homologous region (GmA_39 versus GmA′_40) showed very low K_s because of very short length of exon 2 (50 bp).

Paleopolyploid species like soybean indicate the presence of duplications by showing secondary peaks in the age distributions of paralogous pairs.[17] Three secondary peaks were observed in this study. Distributions were indicated with a major and first peak at mode $K_s = 0.10-0.15$. These data were consistent, and the first secondary peak at the same mode also identified after comparisons of pairs of paralogous genes in 14 model plant species including soybean.[17] In addition, two minor bulges were identified in our study, indicating additional duplication events in this homologous region between two contigs.

To estimate divergence time for gene duplication, the K_s values were used, assuming rates of synonymous substitution of 6.1×10^{-9} substitutions per synonymous site per year.[18,40,61] This soybean homologous region was mainly duplicated at 12.3 MYA in this study and the speciation event of soybean from *Medicago* at 60 MYA was also suggested (Fig. 3), agreeing with the rapid diversification between 50–60 MYA in legumes.[2,61] However, estimated ages of the secondary peaks could be different depending on the assumed substitution rates. It estimated a rate of silent-site substitution of 6.1 per silent site per billion years.[18,40,61] But, a synonymous rate of 1.5×10^{-8} substitutions per synonymous site per year for dicots was used to calculate the absolute date for duplication events.[17,62] With a synonymous rate of 1.5×10^{-8} substitutions, the average divergent time was 6.3 MYA in this study (median = 5.0 MYA), similar to estimation of the recent duplication (3.3–5.0 MYA) in soybean.[17] However, these estimates are only approximations because the rate of synonymous substitution is different among genes and species and generation time is also the factor for controlling mutational rate.[17]

Our study provides additional evidence of the paleopolyploidy of the soybean genome. We also showed that organization and sequence homology between duplicated segments were very similar. In this study, homoeologous regions were so similar that the contig on LG A1 was originally sequenced instead of that on LG D2, even though BAC-end sequences located near *Rxp* locus on LG D2 were used for BAC selection in genome sequencing. Thus, in future studies, to avoid walking in the wrong direction, BAC by BAC soybean genome sequencing should be performed in concert with whole-genome physical mapping because of high level of similarity between homologous contigs.

Acknowledgements: This research was supported by a grant for genome sequencing funded by Agricultural R&D Promotion Center, Technology Development Program for Agriculture and Forestry, the Ministry of Agriculture and Forestry, the Republic of Korea, and in part by a grant (code no. CG3121) for genetic mapping from the Crop Functional Genomics Center of the 21st Century Frontier Research Program funded by the Ministry of Science and Technology, the Republic of Korea. Dr K. Van and Mr Kim are the recipients of a fellowship from the BK21 program granted by the Ministry of Education and Human Resources Development (ME and HRD), the Republic of Korea. We also thank the National Instrumentation Center for Environmental Management at Seoul National University in Korea.

Supplementary Data: Supplementary data are available online at www.dnaresearch.oxfordjournals.org.

References

1. Young, N. D. and Shoemaker, R. C. 2006, Genome studies and molecular genetics, Part 1: model legumes

exploring the structure, function and evolution of legume genomes, *Curr. Opin. Plant Biol.*, **9**, 95–98.
2. Shoemaker, R. C., Schlueter, J. and Doyle, J. J. 2006, Paleopolyploidy and gene duplication in soybean and other legumes, *Curr. Opin. Plant Biol.*, **9**, 104–109.
3. Van, K., Hwang, E.-Y., Kim, M. Y., Kim, Y.-H., Cho, Y.-I., Cregan, P. B. and Lee, S.-H. 2004, Discovery of single nucleotide polymorphisms in soybean using primers designed from ESTs, *Euphytica*, **139**, 147–157.
4. Sato, S. and Tabata, S. 2006, *Lotus japonicus* as a platform for legume research, *Curr. Opin. Plant Biol.*, **9**, 128–132.
5. Town, C. D. 2006, Annotating the genome of *Medicago truncatula*, *Curr. Opin. Plant Biol.*, **9**, 122–127.
6. Doyle, J. J., Doyle, J. L., Rauscher, J. T. and Brown, A. H. D. 2004, Diploid and polyploid reticulate evolution throughout the history of perennial soybeans (*Glycine* subgenus *Glycine*), *New Phytol.*, **161**, 121–132.
7. Zhu, H., Choi, H.-K., Cook, D. R. and Shoemaker, R. C. 2005, Bridging model and crop legumes through comparative genomics, *Plant Physiol.*, **137**, 1189–1196.
8. Shoemaker, R. C., Polzin, K., Labate, J., Spechet, J., Brummer, E. C., Olson, T., Young, N., Concibido, V., Wilcox, J., Tamulonis, J. P., Kochert, G. and Boerma, H. R. 1996, Genome duplication in soybean (*Glycine* subgenus *soja*), *Genetics*, **144**, 329–338.
9. Lee, J. M., Bush, A., Specht, J. E. and Shoemaker, R. C. 1999, Mapping duplicate genes in soybean, *Genome*, **42**, 829–836.
10. Lee, J. M., Grant, D., Vallejos, C. E. and Shoemaker, R. C. 2001, Genome organization in dicots. II. *Arabidopsis* as a bridging species to resolve genome duplication events among legumes, *Theor. Appl. Genet.*, **103**, 765–773.
11. Foster-Hartnett, D., Mudge, J., Larsen, D., Danesh, D., Yan, H., Denny, R., Peñuela, S. and Young, N. D. 2002, Comparative genomic analysis of sequence sampled from a small region on soybean (*Glycine max*) molecular linkage group G, *Genome*, **45**, 634–645.
12. Yan, H. H., Mudge, J., Kim, D.-J., Larsen, D., Shoemaker, R. C., Cook, D. R. and Young, N. D. 2003, Estimates of conserved microsynteny among the genomes of *Glycine max, Medicago truncatul*a and *Arabidopsis thaliana*, *Theor. Appl. Genet.*, **106**, 1256–1265.
13. Schlueter, J. A., Vasylenko-Sanders, I. F., Deshpande, S., Yi, J., Siegfried, M., Roe, B. A., Schlueter, S. D., Scheffler, B. E. and Shoemaker, R. C. 2007, The FAD2 gene family of soybean: insights into the structural and functional divergence of a paleopolyploid genome, *The Plant Genome (A supplement to Crop Sci.)*, **47**, S-14–S-26.
14. Pagel, J., Walling, J. G., Young, N. D., Shoemaker, R. C. and Jackson, S. A. 2004, Segmental duplication within the *Glycine max* genome revealed by fluorescence in situ hybridization of bacterial artificial chromosomes, *Genome*, **47**, 764–768.
15. Zhu, H., Kim, D.-J., Baek, J.-M., Choi, H.-K., Ellis, L. C., Küester, H., McCombie, W. R., Peng, H.-M. and Cook, D. R. 2003, Syntenic relationships between *Medicago truncatula* and *Arabidopsis* reveal extensive divergence of genome organization, *Plant Physiol.*, **131**, 1018–1026.
16. Gaut, B. S. and Doebley, J. F. 1997, DNA sequence evidence for the segmental allotetraploid origin of maize, *Proc. Natl. Acad. Sci. USA*, **94**, 6809–6814.
17. Blanc, G. and Wolfe, K. H. 2004, Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes, *Plant Cell*, **16**, 1667–1678.
18. Schlueter, J. A., Dixon, P., Granger, C., Grant, D., Clark, L., Doyle, J. J. and Shoemaker, R. C. 2004, Mining EST databases to resolve evolutionary events in major crop species, *Genome*, **47**, 868–876.
19. Pfeil, B. E., Schlueter, J. A., Shoemaker, R. C. and Doyle, J. J. 2005, Placing paleopolyploidy in relation to taxon divergence: a phylogenetic analysis in legumes using 39 gene families, *Syst. Biol.*, **54**, 441–454.
20. Narvel, J. M., Jakkula, L. R., Phillips, D. V., Wang, T., Lee, S. H. and Boerma, H. R. 2001, Molecular mapping of *Rxp* conditioning reaction to bacterial pustule in soybean, *J. Hered.*, **92**, 267–270.
21. Hartwig, E. E. and Johnson, H. W. 1953, Effect of the bacterial pustule disease on yield and chemical composition of soybeans, *Agron. J.*, **45**, 22–23.
22. Weber, C. R., Dunleavy, J. M. and Fehr, W. R. 1966, Effect of bacterial pustule on closely related soybean lines, *Agron. J.*, **58**, 544–545.
23. Kennedy, B. W. and Tachibana, H. 1973, Bacterial diseases, In: Caldwell, B. E.(ed.), *Soybeans: Improvement, Production, and Uses*, Madison, WI: American Society of Agronomy, 491–504.
24. Groth, D. E. and Braun, E. J. 1986, Growth kinetics and histopathology of *Xanthomonas campestris* pv. *glycines* in leaves of resistant and susceptible soybeans, *Phytopathology*, **76**, 959–965.
25. Cregan, P. B., Jarvik, T., Bush, A. L., Shoemaker, R. C., Lark, K. G., Kahler, A. L., Kaya, N., VanToai, T. T., Lohnes, D. G., Chung, J. and Specht, J. E. 1999, An integrated genetic linkage map of the soybean genome, *Crop Sci.*, **39**, 1464–1490.
26. Song, Q. J., Marek, L. F., Shoemaker, R. C., Lark, K. G., Concibido, V. C., Delannay, X., Specht, J. E. and Cregan, P. B. 2004, A new integrated genetic linkage map of the soybean, *Theor. Appl. Genet.*, **109**, 122–128.
27. Choi, I.-Y., Hyten, D. L., Matukumalli, L. K., Song, Q., Chaky, J. M., Quigley, C. V., Chase, K., Lark, K. G., Reiter, R. S., Yoon, M.-S., Hwang, E.-Y., Yi, S.-I., Young, N. D., Shoemaker, R. C., van Tassell, C. P., Specht, J. E. and Cregan, P. B. 2007, A soybean transcript map: gene distribution, haplotype and single-nucleotide polymorphism analysis, *Genetics*, **176**, 685–696.
28. Bernard, R. L. and Weiss, M. G. 1973, Qualitative genetics. In: Caldwell, B. E. (ed.), *Soybeans Improvement, Production, and Uses*, Madison, WI: American Society of Agronomy, pp.117–154.
29. Van, K., Ha, B.-K., Kim, M. Y., Moon, J. K., Paek, N.-C., Heu, S. and Lee, S.-H. 2004, SSR mapping of genes conditioning soybean resistance to six isolates of *Xanthomonas axonoppodis* pv. *glycines*, *Kor. J. Genetics*, **26**, 47–54.
30. Palmer, R. G., Lim, S. M. and Hedges, B. R. 1992, Testing for linkage between the *rxp* locus and nine isozyme loci in soybean, *Crop Sci.*, **32**, 681–683.

31. Marek, L. F. and Shoemaker, R. C. 1997, BAC contig development by fingerprint analysis in soybean, *Genome*, **40**, 420–427.
32. Choi, S.-H., Kim, I.-C., Kim, D.-S., Kim, D.-W., Chae, S.-H., Choi, H.-H., Choi, I., Yeo, J.-S., Song, M.-N. and Park, H.-S. 2006, Comparative genomic organization of the human and bovine PRNP locus, *Genomics*, **87**, 598–607.
33. Soderlund, C., Longden, I. and Mott, R. 1997, FPC: a system for building contigs from restriction fingerprinted clones, *Bioinformatics*, **13**, 523–535
34. Sulston, J., Mallett, F., Staden, R., Rurbin, R., Horsnell, T. and Coulson, A. 1998, Software for genome mapping by fingerprinting techniques, *Comput. Appl. Biosci.*, **4**, 125–132.
35. Yang, T.-J., Kim, J. S., Kwon, S.-J., Lim, K.-B., Choi, B.-S., Kim, J.-A., Jin, M., Park, J. Y., Lim, M.-H., Kim, H.-I., Lim, Y. P., Kang, J. J., Hong, J.-H., Kim, C.-B., Bhak, J., Bancroft, I. and Park, B.-S. (2006). Sequence-level analysis of the diploidization process in the triplicated *FLOWERING LOCUS C* region in *Brassica rapa*, *Plant Cell*, **18**, 1339–1347.
36. Rozen, S. and Skaletsdy, H. 2000, Primer3 on the WWW for general users and for biologist programmers, *Methods Mol. Biol.*, **132**, 365–386.
37. Wu, X., Lee, G.-J., Blake, S., Pyatek, K., Huang, S., Wan, J., Stacey, G. and Nguyen, H. T. 2005, Six-dimensional BAC DNA pools—a new resource for soybean genome mapping, In: *Plant and Animal Genomes XIII Conference*, San Diego, CA, USA. Abstract 430.
38. Goldman, N. and Yang, Z. 1994, A codon-based model of nucleotide substitution for protein-coding DNA sequences, *Mol. Biol. Evol.*, **11**, 725–736.
39. Yang, Z. 1997, PAML: a program package for phylogenetic analysis by maximum likelihood, *Comput. Appl. Biosci.*, **13**, 555–556.
40. Lynch, M. and Conery, J. S. 2000, The evolutionary fate and consequences of duplicated genes, *Science*, **290**, 1151–1155.
41. Kaderali, L., Deshpande, A., Nolan, J. P. and White, P. S. 2003, Primer design for multiplexed genotyping, *Nucleic Acids Res.*, **31**, 1796–1802.
42. Cai, C. M., Van, K., Kim, M. Y. and Lee, S.-H. 2005, Optimization of SNP genotyping assay with fluorescence polarization detection, *Kor. J. Crop Sci.*, **50**, 361–367.
43. Kim, M. Y., Ha, B.-K., Jun, T.-H., Hwang, E.-Y., Van, K., Kuk, Y.-I. and Lee, S.-H. 2004, Single nucleotide polymorphism discovery and linkage mapping of lipoxygenase-2 gene ($L_{\chi 2}$) in soybean, *Euphytica*, **135**, 169–177.
44. Chen, X., Levine, L. and Kwok, P. Y. 1999, Fluorescence polarization in homogeneous nucleic acid analysis, *Genome Res.*, **9**, 492–498.
45. Voorrips, R. E. 2002, MapChart: software for the graphical presentation of linkage maps and QTLs, *J. Hered.*, **93**, 77–78.
46. Schlueter, J. A., Scheffler, B. E., Schlueter, S. D. and Shoemaker, R. C. 2006, Sequence conservation of homeologous bacterial artificial chromosomes and transcription of homeologous genes in soybean (*Glycine max* L. Merr.), *Genetics*, **174**, 1017–1028.
47. Cai, C. M., Van, K. and Lee, S.-H. 2007, Gene duplications revealed during the process of SNP discovery in soybean [*Glycine max* (L.) Merr.], *J. Crop Sci. Biotech.*, **10**, 237–242.
48. Ku, H. M., Vision, T., Liu, J. P. and Tanksley, S. D. 2000, Comparing sequenced segments of the tomato and *Arabidopsis* genomes: large-scale duplication followed by selective gene loss creates a network of synteny, *Proc. Natl. Acad. Sci. USA*, **97**, 9121–9126.
49. Feuillet, C. and Keller, B. 1999, High gene density is conserved at syntenic loci of small and large grass genomes, *Proc. Natl. Acad. Sci. USA*, **96**, 8265–8270.
50. Gaut, B. S., Wright, S. I., Rizzon, C., Dvorak, J. and Anderson, L. K. 2007, Recombination: an underappreciated factor in the evolution of plant genomes, *Nat. Rev. Genet.*, **8**, 77–84.
51. Chen, M., Presting, G., Barbazuk, W. B., et al. 2002, An integrated physical and genetic map of the rice genome, *Plant Cell*, **14**, 537–545.
52. Zhang, X.-C, Wu, X., Findley, S., Wan, J., Libault, M., Nguyen, H. T., Cannon, S. B. and Stacey, G. 2007, Molecular evolution of lysine motif-type receptor-like kinases in plants, *Plant Physiol.*, **144**, 623–636.
53. Wendel, J. F., Schnabel, A. and Seelanan, T. 1995, Bidirectional interlocus concerted evolution following allopolyploid speciation in cotton (*Gossypium*), *Proc. Natl. Acad. Sci. USA*, **92**, 280–284.
54. Lee, S.-W., Han, S.-W., Bartley, L. E. and Ronald, P. C. 2006, Unique characteristics of *Xanthomonas oryzae* pv. *oryzae* AvrXa21 and implications for plant innate immunity, *Proc. Natl. Acad. Sci. USA*, **103**, 18395–18400.
55. Song, W.-Y., Pi, L.-Y., Wang, G.-L., Gardner, J., Holsten, T. and Ronald, P. C. 1997, Evolution of the rice Xa21 disease resistance gene family, *Plant Cell*, **9**, 1279–1287.
56. Ramachandran, S. and Sundaresan, V. 2001, Transposons as tools for functional genomics, *Plant Physiol. Biochem.*, **39**, 243–252.
57. Xu, M. and Palmer, R. G. 2005, Genetic analysis and molecular mapping of a pale flower allele at the *W4* locus in soybean, *Genome*, **48**, 334–340.
58. Xu, M. and Palmer, R. G. 2005, Molecular mapping of *k2 Mdh1-n y20*, an unstable chromosomal region in soybean [*Glycine max* (L.) Merr.], *Theor. Appl. Genet.*, **111**, 1457–1465.
59. Hurst, L. D. 2002, The K_a/K_s ratio: diagnosing the form of sequence evolution, *Trends Genet.*, **18**, 486–487.
60. Nekrutenko, A., Makova, K. D. and Li, W.-H. 2002, The K_a/K_s ratio test for accessing the protein-coding potential of genomic regions: an empirical and simulation study, *Genome Res.*, **12**, 198–202.
61. Lavin, M., Herendeen, P. S. and Wojciechowski, M. F. 2005, Evolutionary rates analysis of Leguminosae implicates as rapid diversification of lineages during the tertiary, *Syst. Biol.*, **54**, 575–594.
62. Koch, M. A., Haubold, B. and Mitchell-Olds, T. 2000, Comparative evolutionary analysis of chalcone synthase and alcohol dehydrogenase loci in *Arabidopsis*, *Arabis*, and related genera (Brassicaceae), *Mol. Biol. Evol.*, **17**, 1483–1498.