Sequence Comparisons of The Duplicated Regions in Soybean \( [Glycine \text{ max} \ (L.) \ Merr.] \) 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\(^2\), Beom-Soon Choi\(^3\), Tae-Jin Yang\(^1\), Suk-Ha Lee\(^{1,3}\)

\(^1\) Department of Plant Science, Seoul National University, Seoul, 151-921, Korea, Email kvan@snu.ac.kr
\(^2\) National Institute of Crop Science, Suwon, 441-857, Korea,
\(^3\) Research Institute for Agriculture and Life Sciences, Seoul National University, Seoul, 151-921, Korea,
\(^4\) Corn Insect and Crop Genetics Research Unit, USDA-ARS, Iowa State University, Ames, IA 50011, USA,
\(^5\) National Instrumentation Center for Environmental Management, Seoul National University, Seoul, 151-921, Korea

Abstract

A single recessive gene, \( \text{rxp} \), 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 \( \text{Rxp} \) region. With the RIL population from the cross of Pureunkong and Jinpumkong 2, SNP and SSR marker genotyping were able to locate GmA' on LG A1. Based on information in the Soybean Breeders Toolbox and our results, parts of LG A1 and LG D2 share duplicated regions. Estimation of evolutionary events revealed that speciation of soybean from \( \text{Medicago} \) and the recent divergence of two soybean homoeologous regions occurred at 60 and 12 million years ago, respectively. Thus, diploidized paleopolyploidy of soybean genome was again supported by our study.

Media summary

Recent Duplication in Soybean Genome was revealed again by sequence level analysis with BACs.

Key Words

BAC, divergence time, duplication, \( K_\alpha \), \( \text{Rxp} \), soybean

Introduction

Legumes have begun to draw much attention through recent genomic and phylogenetic studies and their economical value (Shoemaker et al. 2006). Ancient genome duplication occurred in \( \text{Glycine} \) and polyploidy has had an evolutionary impact on the structure of the soybean genome. Using restriction fragment length polymorphism (RFLP) analysis with nine populations of the \( \text{Glycine subgenus soja} \), the soybean genome presents about 2.55 copies per digest. RFLP and simple sequence repeat (SSR) analyses showed that parts of LGs B1/S, H, and F of soybean genome shared homoeologous region (Lee et al. 1999). Also, high similarity in physical organization between soybean duplicated regions and a high percentage of microsynteny were shown by characterizing BAC clones of soybean and other model plants. The processes of genome evolution and patterns of divergence can be studied by duplicate gene analysis. Putative genome duplications events were identified with large EST collections from eight plant species, using synonymous substitution measurements (\( K_\alpha \)) of duplicated genes (Schlueter et al. 2004). 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 \( \text{M. truncatula} \) at approximately 58 MYA. \( \text{Xanthomonas axonopodis pv. glycine} \ (\text{Xag}) \) causes bacterial leaf pustule (BLP) in soybean, causing substantial losses in yield through premature defoliation. Among 20 LGs, we are interested in LG D2 because the recessive gene conditioning resistance to BLP, \( \text{rxp} \), was mapped to LG D2 only 3.9 cM away from Satt372 (Van et al. 2004a). Also, the \( \text{Rxp} \) locus linked to the malate dehydrogenase (\( \text{Mdh} \)) locus. In the process of BAC clone selection for ‘chromosome walking’ around \( \text{Rxp} \) 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 \( \text{Rxp} \) region. Annotation, gene arrangement, and evolution events estimated by \( K_\alpha \) will be also presented.
Materials and Methods
gmw1 was used for selection of BAC clones around Rxp locus. For first round of PCR-based library screening, Satt372, Satt486, and Satt498 were used as PCR primers. Basic PCR protocols were followed as described with minor alternations (Marek and Shoemaker 1997). Several rounds of the BAC library were screened systematically and DNA of ‘Williams 82’ was used as positive control for all screening processes. Cycle conditions for shotgun sequencing of the random plasmid library and analysis of BAC sequences were described (Choi et al. 2006). Also, the individual sequences were assembled with Phred/Phrap software and remaining gaps of each clone were closed by direct sequencing. Image v. 3.0 and FPC v. 4.7.9 were used for confirmation of BAC contig assignment. After the sequences of each BAC clone were aligned, BAC end sequences (BES) were selected for extending BAC contigs. With primers derived from BES, the BAC library screening was performed again with gmw1 and gmw2 (Wu et al. 2005). Alignments between BAC contigs were inspected with GBrowse and SynBrowse. Also, gene annotation was conducted with FGENESH or GeneMark against Medicago database. The rate of non-synonymous nucleotide substitution ($K_a$) and the fraction of synonymous substitutions ($K_s$) were obtained with the CODEML program of the PAML package. $K_s$ was used to estimate the divergence time between two sequences. Divergence times ($T$) were calculated using a synonymous mutation rate of $6.1 \times 10^{-9}$ substitutions per synonymous site per year (Lynch and Conery 2000; Schlueter et al. 2004). To locate each BAC contig in LGs, seven different primer sets were designed from these contig-specific regions. And, the detection of SNP in the contig-specific regions between Pureunkong and Jinpumkong 2 was followed (Van et al. 2004b). SNP capture probe was designed and single base extension reactions followed by fluorescence polarization measurements were performed (Cai et al. 2005), with an F$_2$-derived soybean population of 90 recombinant inbred lines (RILs) from the cross of Pureunkong and Jinpumkong 2. The construction of the linkage map with SNP marker genotyping data and integration of these markers on LGs were followed.

Results
To obtain BAC clones around the Rxp locus, we screened the gmw1 BAC library with Satt372, Satt486, and Satt498. We selected gmw1-29F06 and gmw1-24M16 for determining DNA sequences and created the GmA contig (Figure 1). Clone gmw1-20O10 was selected from BES of gmw1-29F06 and primers designed from BES from gmw1-24M16 selected gmw1-89M01. 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 from gmw2. After the DNA sequences of gmw2-77P21 were aligned with gmw1-20O10 and gmw1-89M01, the GmA’ contig was formed with 100% match. To locate GmA’ on the soybean genetic linkage map, SNP genotyping was performed. One SNP locus between Pureunkong (deletion) and Jinpumkong 2 (A) was identified seven contig-specific regions. SNP genotyping of GmA’ was conducted with the RIL population. The SNP marker locus was incorporated into the frame map placing GmA’ to the top of LG A1, 1.9 cm away from Satt684 in LG A1 (Figure 1). Based on 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.

After BAC contigs were confirmed and inspected with GBrowse and SynBrowse, genes were annotated with FGENESH or GeneMark against the Medicago database. The 54 and 58 genes were predicated in GmA and GmA’, respectively. Gene density along these two sequenced BAC contigs was approximately one gene per 5.0 kb. Gene order was conserved among syntenic blocks, except for one case and the same orientation between the predicted genes was observed. Gene order was maintained in Medicago, although linearity was fragmented (data not shown).

Using the maximum likelihood method in the CODEML program, $K_a$ and $K_s$ distances were estimated. 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. To determine the timing of the duplication event giving rise to the two contigs, the $K_s$ value between 0.05 and 1 was only used. $T$ 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) (Figure 2). With MtA, median $K_s$ values were 0.7654 and 0.6877 for GmA and GmA’, respectively. Therefore, MtA and the soybean
homoeologous contigs diverged at 56.4 to 62.7 MYA (Figure 2). The two soybean homoeologous contigs were duplicated more recently, agreeing with the previous study (Blanc and Wolfe 2004).

Figure 1. Schematic relationships between homologous regions (GmA and GmA') containing the Rxp locus from LGs A1 (red) and D2 (blue). GmA was composed of gmw1-29F06 and gmw1-24M16. gmw1-20O10, gmw1-89M01 and gmw2-77P21 were made GmA'.

Figure 2. $K_s$ values and estimation of evolutionary events in three contigs. The number shown above the self-comparison diagonal represents estimation of median $K_s$ values. Divergence times in millions of years calculated 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.

**Discussion**

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. Therefore, another round of BAC library screening was performed to close the gap between gmw1-20O10 and gmw1-89M01 and the GmA' was made by selection of gmw2-77P21 (Figure 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 (Figure 1). A comparison between the soybean composite maps for LG A1 and D2 indicated that homoeologous regions exist between them. Therefore, it suggested that GmA and GmA' are indeed homoeologous.

With comparisons of the sequences of the same gene from two species or gene family, counting of the number of non-synonymous and synonymous change is a good indicator of the degree of divergence between two sequences. Depending on $K_s$, as the background rate of evolution, the selection pressure in protein-coding
regions could be explained by deviations of the $K_a / K_s$ ratio. Short length of exons or highly divergent sequences could cause the $K_a / K_s$ ratio to be higher than 1. With the $K_s$ values to estimate divergence time for gene duplication, 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 (Figure 2), agreeing with the rapid diversification between 50-60 MYA in legumes (Shoemaker et al. 2006).

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.

References


