Sequence Divergence of Recently Duplicate Genes in Soybean
(Glycine max L. Merr.)

Chun Mei Cai¹,², Kyujung Van¹, Moon Young Kim¹,³ and Suk-Ha Lee¹,³*

¹Department of Plant Science, Seoul National University, Seoul 151-921, Korea
²National Institute of Crop Science, Rural Development Administration, Suwon 441-857, Korea
³Research Institute for Agriculture and Life Sciences, Seoul National University, Seoul 151-921, Korea

Received April 1, 2008; accepted April 23, 2008

ABSTRACT

Sequence divergence of duplicate genes was investigated in soybean, a diploidized paleopolyploid. To examine gene duplication, a total of seven primers designed from expressed sequence tags (ESTs) were selected for this study because they produced (1) a single band but showed poor sequencing quality and (2) double bands on an agarose gel. After PCR amplification with genomic DNA, duplicate genes were identified by cloning and subsequent sequencing of twenty randomly-selected clones. Based on their alignment, two or three classes were identified in each amplicon. Linkage analysis confirmed that two duplicate loci were located on different chromosomes. The exon/intron structure was conserved between duplicate gene pairs. A detailed examination of the variation between duplicate pairs showed that the coding regions are highly conserved while many substitutions and insertion/deletions (indels) were identified in non-coding regions. In particular, non-homologs fragments were numerous and, on average, 82 bp in length within the non-coding region. Quantitative trait loci (QTL) for some major agronomic characters, such as seed protein and oil, seed yield, plant height, and corn earworm resistance, were positioned across duplicate loci of TC225246 and TC224550, suggesting that the gene families appear to have retained similar functions throughout genome duplication and evolution events.

Key words: divergence analysis, duplicate gene, Glycine max, quantitative trait loci (QTL).

INTRODUCTION

Gene duplication plays a major role in determining genome size and gene complement and in creating diverse/novel gene functions in eukaryotic genomes (Schlueter et al., 2006). Although duplicate genes can arise in a number of different ways, the most spectacular process for gene duplication may be whole genome duplication via polyploidization. Polyploidy has long been accepted as an important and recurring feature of genome evolution (Wendel, 2000). The long-term evolution of polyploid genomes is often followed by a diploidization process occurring through extensive genome rearrangement at both gene and chromosome levels (Wolfe, 2001).

Upwards of 80% of all angiosperms likely have a polyploid or paleopolyploid history (Wendel, 2000).
Comparative mapping studies as well as genome sequencing efforts revealed that both Arabidopsis and rice were paleopolyploids and that about 60% of their genome was duplicated (Vision et al., 2000; Goff et al., 2002; Simillion et al., 2002). In soybean, analyses of clustered expressed sequence tags (ESTs) revealed two rounds of genome duplication with disparate time estimates of ~15 and 44 MYA (Schlueter et al., 2004) or ~4 and 16 MYA (Blanc and Wolfe, 2004) depending on the parameter for calculating divergence time. Thus, soybean can be considered to have a "paleopolyploid" genome.

The polyploidy events resulted in a large set of homeologous chromosomal segments in the current soybean genome. For example, more than 90% of G. max RFLP probes detected approximately 2.6 hybridizing loci, leading to homeologous chromosomal segments estimated to range from 1.5 to 106.4 cM in length (Shoemaker et al., 1996). BAC hybridization, BAC-end sequencing, and fingerprinting studies have shown fairly extensive sequence conservation between homeologous blocks in soybean (Marek et al., 2001; Foster-Hartnett et al., 2002; Yan et al., 2003). More recently, it was shown that sequenced soybean homeologous BACs containing transcribed duplicate genes showed a range of gene conservation in both order and orientation (Schlueter et al., 2006; Schlueter et al., 2007; Van et al., 2008). However, as an initial step to understanding duplicate gene evolution and functional divergence, a detailed examination of sequence divergence is required.

Single nucleotide polymorphisms (SNPs) are becoming the most widely used genetic markers. The simplest approach to discover SNPs is direct sequencing based on primers designed to tentative consensus sequences (TCs) at The Institute for Genome Research (TIGR) (http://www.tigr.org), almost 26% of the primer sets produced multiple products on an agarose gel or poor-quality sequence data from what appeared to resemble a single band on the agarose gel (data not shown). In this study, seven primer sets were selected from these failures, of which four showed a single band with poor-quality sequence data and three showed double bands on an agarose gel (Table 1).

Genetic mapping of duplicate genes

The objectives of this study were to recover putative duplicate genes from sequencing failures by TA cloning and to investigate the sequence divergence of these duplicate genes. The duplicate loci were located on different chromosomes by linkage analysis and quantitative trait loci (QTL) for some major agronomic characters were surveyed across the duplicate loci. Additionally, we estimated the divergence times between the duplicate pairs of the soybean genome.

MATERIALS AND METHODS

Plant materials

An F2-derived soybean population of 90 recombinant inbred lines (RILs) that were developed by single-seed descent from the cross of 'Pureunkong' and 'Jinpumkong 2' was used for this study. Previously, a genetic map from this population was constructed using SSR markers (Kim et al., 2004).

Recovery of duplicate genes by TA cloning

During SNP discovery using direct sequencing based on primers designed to tentative consensus sequences (TCs) at The Institute for Genome Research (TIGR) (http://www.tigr.org), almost 26% of the primer sets produced multiple products on an agarose gel or poor-quality sequence data from what appeared to resemble a single band on the agarose gel (data not shown). In this study, seven primer sets were selected from these failures, of which four showed a single band with poor-quality sequence data and three showed double bands on an agarose gel (Table 1).

Each PCR primer set was used to amplify genomic DNA of 'Pureunkong' and 'Jinpumkong 2' as described by Van et al. (2004). PCR products showing a single band were purified using the AccuPrep™ PCR purification kit (Bioneer, Daejeon, Korea), while PCR products showing double bands were gel-extracted individually using S.N.A.P.™ purification columns (Invitrogen, Carlsbad, CA, USA). These purified PCR products were cloned into TOPO TA vectors (Invitrogen), and transformed into TOP10 E. coli cells (Invitrogen). Twenty independent clones were chosen randomly and sequenced with M13 forward and reverse primers on an ABI3730 with BigDye3.1 (Applied Biosystems, Foster City, CA, USA)(Van et al., 2004).

Genetic mapping of duplicate genes

After SNPs were surveyed between 'Pureunkong' and 'Jinpumkong 2' in individual member of duplicate gene pairs, member-specific PCR primers were