RESEARCH ARTICLE
J. Crop Sci. Biotech. 10 (4) : 237 ~ 242

Gene Duplications Revealed during the Process of SNP Discovery in Soybean [*Glycine max* (L.) Merr.]

Chun Mei Cai[1,2], Kyujung Van[1], Suk-Ha Lee[1,3,*]

[1]*Department of Plant Science, Seoul National University, Seoul, 151-921, Korea*
[2]*Genetic Resource Division, National Institute of Crop Science, RDA, Suwon 441-707, Korea*
[3]*Research Institute for Agriculture and Life Sciences, Seoul National University, Seoul, 151-921, Korea*

Abstract

Genome duplication (i.e. polyploidy) is a common phenomenon in the evolution of plants. The objective of this study was to achieve a comprehensive understanding of genome duplication for SNP discovery by Thymine/Adenine (TA) cloning for confirmation. Primer pairs were designed from 793 EST contigs expressed in the roots of a supernodulating soybean mutant and screened between 'Pureunkong' and 'Jinpumkong 2' by direct sequencing. Almost 27% of the primer sets were failed to obtain sequence data due to multiple bands on agarose gel or poor quality sequence data from a single band. TA cloning was able to identify duplicate genes and the paralogous sequences were coincident with the nonspecific peaks in direct sequencing. Our study confirmed that heterogeneous products by the co-amplification of a gene family member were the main cause of obtaining multiple bands or poor quality sequence data in direct sequencing. Counts of amplified bands on agarose gel and peaks of sequencing trace suggested that almost 27% of nonrepetitive soybean sequences were present in as many as four copies with an average of 2.33 duplications per segment. Copy numbers would be underestimated because of the presence of long intron between primer binding sites or mutation on priming site. Also, the copy numbers were not accurately estimated due to deletion or tandem duplication in the entire soybean genome.

Key words: direct sequencing, genome duplication, single nucleotide polymorphism (SNP), soybean

Introduction

Single DNA base differences or small insertions and deletions (indels) at a specific base position, collectively referred to as single nucleotide polymorphisms (SNPs) were revealed by comparison of homologous DNA sequences (Brookes 1999). SNPs are valuable DNA markers because of abundance and high stability and often contribute directly to a phenotypic trait (Kim et al. 2004). Various methods have been developed to discover SNPs, but the most direct and simple approach is direct sequencing of PCR products from several individuals (Gupta et al. 2001). PCR primers are frequently derived from genes of interest or expressed sequence tag (ESTs). However, a high percentage of PCR primers failed to obtain sequence data due to the presence of multiple bands from PCR amplification or poor sequence quality from which appeared a single band on an agarose gel in soybean (Van et al. 2004; Zhu et al. 2003). This may be caused by the amplification of duplicate genes. However, no further work was done to validate duplication of genes in the soybean genome.

Gene duplication, arising from region-specific duplication or genome-wide duplication, is an important feature of genome evolution. Gene and genome duplication are recognized as driving forces in the evolution of eukaryotic genomes (Ohno 1970). Upwards of 80% of all angiosperms likely have polyploidy origins (Masterson 1994). The long-term evolution of polyploidy genomes is often followed by a process of "diploidization" that restores diploid-like pairing and disomic genetics (Wolfe 2001). The complete sequenc-

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

ing of *Arabidopsis thaliana* demonstrated that it has experienced three rounds of apparent whole genome duplication, with 82% of all genes and 80% of sequences residing in duplicated segments (Simillion et al. 2002).

Soybean [*Glycine max* (L.) Merr] has been hypothesized to be an ancient diploidized tetraploid based on evolutionary studies and haploid genome analysis (Hadley and Hymowitz 1973). The soybean genome is not fully sequenced at this time due to the intermediate genome size (1,115 Mbp, Arumuganathan and Earle 1991) and relatively complex genome organization. Nevertheless, large-scale EST data and physical mapping can be used to improve our knowledge of genome structure and genome evolution. Hybridization of RFLP probes indicated that more than 90% of probes are duplicated an average of 2.6 times in the soybean genome (Shoemaker et al. 1996). Marek et al. (2001) also found that each *G. max* RFLP probe identifies an average of 2.9 homoeologous regions against two bacterial artificial chromosome (BAC) soybean libraries. Recently, computational analyses of clustered 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). Thus, soybean can also be referred to as a 'paleopolyploid' genome.

The objective of this study was to confirm that the failures of PCR and sequencing during SNP discovery were caused by genome duplication. With the PCR and sequencing failures, sequences of the duplicate genes were obtained by Thymine/Adenine (TA) cloning. And, the copy number of duplicate genes was estimated based on the failures of PCR and sequencing data.

Materials and Methods

Plant materials

Two genotypes of soybean, 'Pureunkong' and 'Jinpumkong 2' were used in this study. Pureunkong is characterized by small seed size, green seed coat, and green seed embryo, which are considered desirable seed traits for producing soybean sprouts (Kim et al. 1996). Jinpumkong 2 has the null alleles *Lx1, Lx2*, and *Lx3*, which leads to the lack of a beany taste associated with common cultivars (Kim et al. 1997). Genomic DNA was extracted from fully expanded leaves according to the procedure described by Rogers and Bendich (1994).

PCR amplification and direct sequencing

A total of 793 tentative contigs (TCs) were randomly selected from soybean cDNA library (Gm-c1078) derived from roots of seven-day-old 'Bragg' supernodulating mutant (http://www.tigr.org). These TC sequences were used as templates for primers designed by Primer3 (http://frodo.wi.mit.edu/) to produce amplicons of approximately 400-600 bp in length. PCR was performed in a 50$\mu\ell$ reaction volume with 2 U *Taq* DNA polymerase (Applied Biosystems, Foster city, CA, USA), following the manufacturer's recommended protocols and cycling conditions. After PCR products were confirmed by electrophoresis on 1.0% ethidium bromide-stained agarose gels, only PCR amplicons shown a single fragment were used as templates for further sequencing reactions (Kim et al. 2004).

Table 1. Number of PCR primers designed and results of PCR and sequence analysis

Primers designed	793	Percent of total (%)
Primers produced a single band	685	86
(Single band with good sequence quality)	(520)	(66)
(Single band with poor sequence quality)	(165)	(21)
Primers produced multiple bands	44	6
Primers produced no band	64	8

Analysis of PCR and sequencing failures by TA cloning

The primers used to study poor quality sequence from a single band and multiple bands were designed from TC218232 and TC204444, respectively. PCR products showing a single band with poor sequencing data 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 vector (Invitrogen) and transformed into TOP10 *E. coli* cells (Invitrogen). Ten independent clones from each PCR product were chosen randomly and sequenced with M13 forward and reverse primers on an ABI3730 with BigDye3.1 (Applied Biosystems) (Van et al. 2004).

Base callings were performed with SeqScape version 2.0 (Applied Biosystems). After vector sequences were removed, sequence alignment between paralogous sequences was performed with the software alignment program bl2seq (http://www.ncbi.nlm.nih.gov).

Table 2. Copy number of soybean sequences

Copy numbers	No. of primers			Percentage of total (%)
	PCR-based (determined by the number of bands on agarose gel)	Sequence-based (determined by the number of peaks on chromatogram)	Total	
1	-*	520	520	65.6
2	34	132	166	20.9
3	6	32	38	4.8
4	4	1	5	0.6

*PCR amplified band was divided into high quality sequence and poor quality sequence.

Results

Failure of obtaining sequence data from TCs

With PCR primer pairs designed from soybean TCs, four different patterns were shown based on PCR performance and sequencing analysis: no band, single band with high quality sequence (single peak), single band with poor quality sequence (noisy peaks), and multiple bands (Fig. 1). Out of the 793 primer sets, 685 (86%) amplified a single PCR product (Table 1). Multiple products were produced by 44 (6%) primer sets and 64 primers (8%) produced no PCR product. High quality sequence data from two soybean genotypes were obtained from 520 (66%) of the 793 primer sets. Good quality of sequence data could not be obtained from 165 cases (21%), although a single discrete PCR product was observed on an agarose gel.

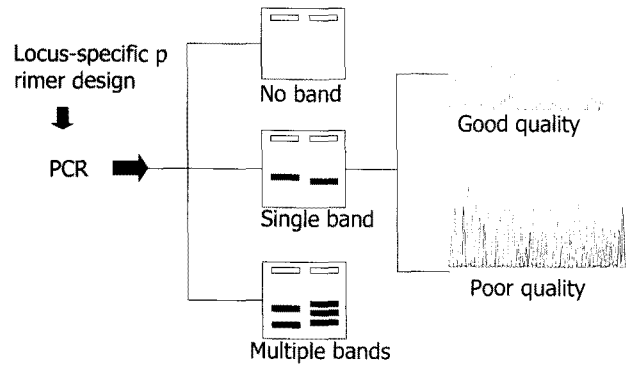

Fig. 1. Amplification and sequencing patterns of the PCR primer pairs derived from TCs for SNP discovery in soybean.

Confirmation of duplicate genes by TA cloning

To examine the mechanism of poor quality sequence derived from a single band (Fig. 2A), primers (Forward: 5' - CCTCTCAAGGTGAGAAAGGAGA - 3'; Reverse: 5' - ATGAATGGCATATAGCATCGTG - 3') derived from TC218232 were studied in detail. After PCR amplification and cloning, 10 transformed clones were randomly selected and sequenced. Analysis of these 10 sequences showed that two groups of sequences with similar size were detected in the PCR products of TC218232 (Fig. 2B), even though they were amplified by the same primer set. The nonspecific peaks from sequences in PCR products amplified by the above primers were compared between direct sequencing and paralogous sequences obtained from TA cloning. At the beginning of the TC218232 sequence, the positions of the single base substitutions and indels between the paralogous sequences coincided with the positions of the double peaks seen in the direct sequencing results (Fig. 2). The occurrence of short indels following those first few substitutions caused a frameshift in the direct sequencing, resulting in the production of many false single nucleotide changes. The coincidence of paralogous sequences and direct sequencing confirmed that the poor sequence quality from a single band was due to the co-amplification of two or more members within gene family by the same primer set.

For double bands, PCR products of TC204444 amplified by primers (Forward: 5' - CCATGGTCATGCTAGGTAATTG - 3'; Reverse: 5' - CTCAGACCCTTCTCTTTCCAGA - 3') were gel-extracted individually for TA cloning. Ten clones were randomly chosen and sequenced for each of the double bands. Analysis of these 10 sequences showed that they were consistent in each of the double bands (data not shown). Therefore, two members of sequences with different sizes corresponding to the double bands on the agarose gel were detected. These results confirmed that the multiple bands from PCR amplification were likely due to the co-amplification of two or more paralogous genes with different sizes by the same primer set.

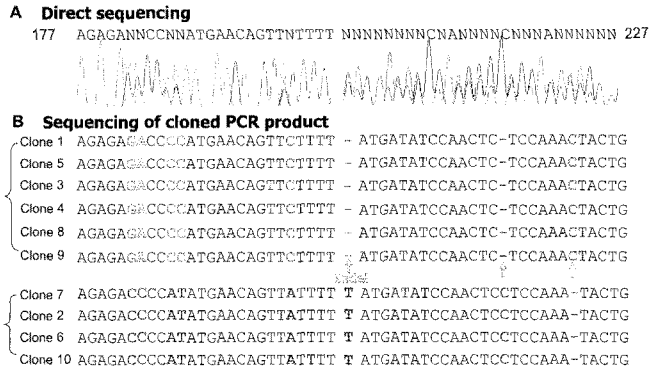

Fig. 2. Comparison of direct sequencing and sequencing data from ten clones generated by TA cloning with TC218232 PCR products of 'Pureunkong'.
(A) Results of partial chromatogram in direct sequencing results.
(B) The alignment of sequences from ten clones.

Estimates of copy number of gene families

Since poor quality sequence and multiple bands were caused by the co-amplification of duplicate genes, the failure data during SNP discovery could be used to estimate the degree of duplication, generally by counting the number of amplified bands on agarose gel or peaks of sequencing trace. Analysis of the average number of fragments amplified by each of 793 PCR primers showed that 26.3% detect, on average, 2.23 bands or peaks under high PCR annealing temperature (Table 2). These data suggest that more than 26% of the nonrepetitive sequences in soybean may on average be duplicated 2.23 times.

Discussion

The direct sequencing procedure is not always successful because either the amplification of multiple products with different sizes or the amplification of multiple paralogous regions with similar sizes appeared as a single-copy fragment on an agarose gel. Of the 793 primer sets tested, 44 (6%) and 165 (21%) were amplified multiple bands and produced poor quality sequence (Table 1), respectively, although much efforts were made to utilize a high annealing temperature to amplify unique sequences in the soybean genome. Other studies have also reported a similar incidence of these two unexpected outcomes. In order to develop SNPs, Zhu et al. (2003) directly sequenced fragments of 90 complete genes and 88 cDNAs. Of these, 3 and 20%, respectively, showed multiple amplification products and poor quality sequence in complete genes, while 14 and 24% showed multiple products and poor quality sequence in cDNAs. Subsequent work by Van et al. (2004) found that 12% of primers amplified multiple products and 10% showed poor sequence quality data out of 110 primer sets designed for ESTs.

Cloning is a good method to separate the heterogeneous sequences produced by the same primer sets. By TA cloning, our results confirmed that heterogeneous products generated by the co-amplification of members of a gene family were the main reason for production of multiple bands or poor quality sequence data in direct sequencing (Fig. 2). Since soybean is suspected to be an ancient tetraploid, its genome is considered to be highly duplicated (Shoemaker et al. 1996). Due to the nature of highly duplicated soybean genome, the efficiency of direct sequencing would decrease whereas the cost for SNP discovery would increase. Discovering the unique polymorphisms that distinguish the individual gene members within a gene family will be critical for overcoming this problem. Using this information, it will be possible to design locus-specific PCR primers that will amplify only a single member of a gene family. Thus, the problem of multiple products being sequenced simultaneously could be eliminated.

Previous methods involving DNA hybridization have provided some estimates of the potential degree of duplication within soybean genome, either by counting the number of hybridization signals in FISH mapping (Pagel et al. 2004) or the number of fragments that hybridize to a single RFLP probe (Marek et al. 2001; Shoemaker et al. 1996). PCR primers used for SNP discovery could amplify multiple gene family members and this is good evidence for co-amplification of multiple genes. In this study, almost 27% of nonrepetitive soybean sequences are probably present in 2.33 copies by counting of amplified bands on agarose gel and sequence peaks on chromatogram (Table 2). These data are lower than the observations of Shoemaker et al. (1996) which indicated that approximately 90% of the examined loci were present in 2.55 copies based on hybridization data. These underestimates may be due to the presence of long intron between primer binding sites. In addition, sequence divergence or mutation in the primer region was probably responsible for the lack of amplification in some duplicate genes.

The soybean genome was highly duplicated and probably a paleopolyploid (Foster-Hartnett et al. 2002; Schlueter et al. 2004; Shoemaker et al. 1996). The paleopolyploids are often followed by a diploidization process of switching from tetrasomic to disomic inheritance (Schlueter et al. 2006). In soybean, the process of diploidization has been quite slow and incomplete by deletion and tandem duplication, resulting in a mixture of diploid and tetraploid loci within the current genome (Schlueter et al. 2007). These duplicate loci may lead to complicate comparative genome research because duplications yield networks of synteny between genomes (Mudge et al. 2004). Even within a single genome, highly similar duplicated genomic regions may lead to the wrong direction in genome walking and map-based cloning in physical and genetic mapping, and genome sequencing studies in soybean (Paterson 2005). Also, duplicated genomic regions in a species could provide valuable insight into its genome structure and evolution, and transfer map information from well-mapped to poorly-mapped regions. Thus, understanding of genome duplication is quite essential in the successful molecular dissection of gene function in the future.

Acknowledgements

This work was supported by a grant from a grant (code no. CG3121) from the Crop Functional Genomic Center of the 21st Century Frontier Program of the Ministry of Science and Technology, Korea. Dr. K. Van is recipient of a fellowship from the BK21 program granted by the Ministry of Education & Human Resources Development (ME & HRD), the Republic of Korea. We are also thankful to the National Instrumentation Center for Environmental Management, Seoul National University.

References

Arumuganathan K, Earle ED. 1991. Estimation of nuclear DNA content of plants by flow cytometry. Plant Mol. Biol. Rep. 9: 229-241

Blanc G, Wolfe KH. 2004. Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. Plant Cell 16: 1667-1678

Brookes AJ. 1999. The essence of SNPs. Gene 234: 177-186

Foster-Hartnett D, Mudge J, Larsen D, Danesh D, Yan H, Denny R, Penuela S, Young ND. 2002. Comparative genomic analysis of sequences sampled from a small region on soybean (*Glycine max*) molecular linkage group G. Genome 45: 634-645

Gupta PK, Roy JK, Prasad M. 2001. Single nucleotide polymorphisms: A new paradigm for molecular marker technology and DNA polymorphism detection with emphasis on their use in plants. Current Sci. 80: 524-535

Hadley HH, Hymowitz T. 1973. Speciation and cytogenetics in soybeans: improvement, production, and uses, Ed. 1, Edited by B.E. Caldwell. American Society of Agronomy, Madison, WI. pp. 97-116

Kim MY, Ha B-K, Jun T-H, Hwang E-Y, Van K, Kuk YI, Lee S-K. 2004. Single nucleotide polymorphism discovery and linkage mapping of lipoxygenase-2 (*Lx2*) in soybean. Euphytica 135: 169-177

Kim SD, Hong EH, Seong YK, Kim YH, Lee SH, Kim HS, Ryu YH, Kim YS. 1996. A new soybean variety for sprouting "Pureunkong" with green seed coat and cotyledon, good seed quality. Rural Dev. Admin. J. Agric. Sci. (Upland and Industrial Crops) 38: 238-241

Kim SD, Kim YH, Park KY, Yun HT, Lee YH, Lee SH, Seong YK, Kim HS, Hong EH, Kim YS. 1997. A new beany taste-less soybean variety "Jinpumkong 2" with good seed quality. Rural Dev. Admin. J. Agric. Sci. (Upland and Industrial Crops) 39: 112-115

Marek LF, Mudge J, Darnielle L, Grant D, Hanson N, Paz M, Huihuang Y, Denny R, Larson K, Foster-Hartnett D, Cooper A, Danesh D, Larsen D, Schmidt T, Staggs R, Crow JA, Retzel E, Young ND, Shoemaker RC. 2001. Soybean genomic survey: BAC-end sequences near RFLP and SSR markers. Genome 44: 572-581

Masterson J. 1994. Stomatal size in fossil plants: evidence for polyploidy in majority of angiosperms. Science 264: 421-424

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

Ohno S. 1970. Evolution by gene duplication. Springer-Verlag, New York, N.Y.

Pagel J, Walling JG, Young ND, Shoemaker RC, Jackson SA. 2004. Segmental duplications within the *Glycine* max genome revealed by fluorescence in situ hybridization of bacterial artificial chromosomes. Genome 47: 764-768

Paterson AH. 2005. Polyploidy, evolutionary opportunity, and crop protection. Genetica 123: 191-196

Rogers SO, Bendich AJ. 1994. Extraction of total cellular DNA from plants, algae and fungi. In: Gelvin SB, Schilperoort RA (eds), Plant Molecular Biology Manual, 2nd ed. Kluwer Academic Publishers, Dordrecht, The Netherlands, pp D1:1-8

Schlueter JA, Dixon P, Granger C, Grant D, Clark L, Doyle JJ, Shoemaker RC. 2004. Mining EST databases to resole evolutionary events in major crop species. Genome 47: 868-876

Schlueter JA, Scheffler BE, Schlueter SD, Shoemaker RC. 2006. Sequence conservation of homeologous bacterial artificial chromosomes and transcription of homeologous genes in soybean (*Glycine max* L. Merr.). Genetics 174: 1017-1028

Schlueter JA, Vasylenko-Sanders IF, Deshpande S, Yi J, Siegfried M, Roe BA, Schlueter SD, Scheffler BE, Shoemaker RC. 2007. The FAD2 gene family of soybean: insights into the structural and functional divergence of a paleopolyploid genome. Plant Genome (A supplement to Crop Sci.) 47: S-14-S-26

Shoemaker RC, Polzin K, Labate J, Specht J, Brummer EC, Olson T, Young N, Concibido V, Wilcox J, Tamulonis JP, Kochert G, Boerma HR. 1996. Genome

duplication in soybean (*Glycine* subgenus *soja*). Genetics 144: 329-338

Simillion C, Vandepoele K, Van Montagu MCE, Zabeau M, Van de Peer Y. 2002. The hidden duplication past of *Arabidopsis thaliana*. Proc. Natl. Acad. Sci. U.S.A. 99: 13627-13632

Van K, Hwang EY, Kim MY, Kim YH, Cho YI, Cregan PB, Lee S-H. 2004. Discovery of single nucleotide polymorphisms in soybean using primers designed from ESTs. Euphytica 139: 147-157

Wolfe KH. 2001. Yesterday's polyploids and the mystery of diploidization. Nat. Rev. Genet. 2: 333-341

Zhu YL, Song QJ, Hyten SM, Fickus EW, Young ND, Cregan PB. 2003. Single-nucleotide polymorphism in soybean. Genetics 163: 1123-1134