frontiers in
PLANT SCIENCE

REVIEW ARTICLE
published: 05 June 2013
doi: 10.3389/fpls.2013.00176

Gene divergence of homeologous regions associated with a major seed protein content QTL in soybean

*Puji Lestari[1,2†], Kyujung Van[1†], Jayern Lee[1], Yang Jae Kang[1] and Suk-Ha Lee[1,3]**

[1] Department of Plant Science, Research Institute for Agriculture and Life Sciences, Seoul National University, Seoul, Korea
[2] Indonesian Center for Agricultural Biotechnology and Genetic Resources Research and Development, Bogor, Indonesia
[3] Plant Genomics and Breeding Institute, Seoul National University, Seoul, Korea

Edited by:
Rajeev K. Varshney, International Crops Research Institute for the Semi-Arid Tropics, India

Reviewed by:
Damon Lisch, University of California at Berkeley, USA
Paula Casati, Centro de Estudios Fotosinteticos-CONICET, Argentina

***Correspondence:**
Suk-Ha Lee, Department of Plant Science, Seoul National University, San 56-1, Sillim-dong, Gwanak-gu, Seoul 151-921, Korea
e-mail: sukhalee@snu.ac.kr
† These authors have contributed equally to this work

Understanding several modes of duplication contributing on the present genome structure is getting an attention because it could be related to numerous agronomically important traits. Since soybean serves as a rich protein source for animal feeds and human consumption, breeding efforts in soybean have been directed toward enhancing seed protein content. The publicly available soybean sequences and its genomically featured elements facilitate comprehending of quantitative trait loci (QTL) for seed protein content in concordance with homeologous regions in soybean genome. Although parts of chromosome (Chr) 20 and Chr 10 showed synteny, QTLs for seed protein content present only on Chr 20. Using comparative analysis of gene contents in recently duplicated genomic regions harboring QTL for protein/oil content on Chrs 20 and 10, a total of 27 genes are present in duplicated regions of both Chrs. Notably, 4 tandem duplicates of the putative homeobox protein 22 (HB22) are present only on Chr 20 and this *Medicago truncatula* homolog expressed in endosperm at seed filling stage. These tandem duplicates could contribute on the protein/oil QTL of Chr 20. Our study suggests that non-shared gene contents within the duplicated genomic regions might lead to absence/presence of QTL related to protein/oil content.

Keywords: genome duplication, QTL, seed protein content, soybean, sequence divergence

INTRODUCTION

Since soybean [*Glycine max* (L.) Merrill] seed is a good source of protein and oil, it is grown widely throughout the world for its numerous uses, such as various edible products, animal feed and potential industrial applications (Vuong et al., 2007; Van et al., 2008; Kim et al., 2012). Although the wild soybean (*G. soja* Sieb. and Zucc.), an undomesticated form of the current soybean, is distributed in East Asia, including China, Taiwan, Russian Far East, the Korean Peninsula, and Japan, the origin or domestication site of soybean is still in controversy (Boerma and Specht, 2004; Van et al., in press). After soybean is introduced into Central and South America in the mid-1900's via North America in 1765, soybean becomes one of the major economically valuable crops in terms of the world's total production (Vuong et al., 2007; Stupar and Specht, 2013).

Although high seed protein content directs soybean products having greater nutritional value, the complexity of soybean genome made difficulty for rapid development of strategies in soybean breeding programs. Before the genomic era, SoyBase (http://soybase.org) is the main resource for quantitative trait loci (QTLs) for various traits and linkage map with 20 soybean chromosomes (Chrs). Also, classical, allozyme and other genetic markers such as restriction fragment length polymorphism (RFLP), amplified fragment length polymorphism (AFLP), simple sequence repeats (SSRs) and single nucleotide polymorphism (SNP), are publically available. Starting with the genome sequences of the *G. max* cultivar (Williams 82, Schmutz et al., 2010), the tremendous amount of sequence information generated by resequencing of *G. max* accessions and *G. soja* against the reference genome (Kim et al., 2010; Lam et al., 2010) would be more feasible for soybean improvement.

This review aims to introduce soybean genome complexity in terms of genome duplication and the recent researches of the major QTLs for seed protein content and to suggest gene divergence in homeologous regions related to this QTL with respect to genome duplication between two soybean Chrs 20 and 10.

SOYBEAN GENOME STRUCTURE

Genome duplication is a key process in the evolution of many lineages in flowering plants (Zhu et al., 2005; Flagel and Wendel, 2009). Following whole genome duplication, small-scale duplications are arisen from unequal crossing over and chromosomal anomalies (Freeling, 2009). After crossing over, several kinds of mechanisms including translocation, inversion, deletion and duplication play a considerable role during small duplications (Pagel et al., 2004). If whole genome duplications tend to increase the dosage gene simultaneously, small-scale duplications (tandem and segmental duplications) result in genes out of balance to maintain proper balance (Edger and Pires, 2009).

The moderately large soybean genome (1.1 Gb) with ancient and recent duplications demonstrates that soybean genome is complex (*Glycine max* v1.0 at http://www.phytozome.net/soybean.php). The second round of soybean whole genome duplication occurred approximately 13 million years ago and

this polyploidy event contributes to the soybean genome structure ranging from near-identical, rather divergent to latter more divergent, leading to dynamic and massive genome rearrangement (Wendel and Doyle, 2005; Schmutz et al., 2010). The predicted number of coding genes in soybean is higher than that of Arabidopsis and grape, possibly due to the genome duplication events in soybean's history (Sterck et al., 2007; Cannon and Shoemaker, 2012). Based on the homeologous relationships determined by genome assembly of integrated data from recently duplicated genomic segments (http://www.phytozome.net; http://www.soybase.org), homeologous blocks of duplicated segments were found in all 20 Chrs (**Figure 1**, gray ribbons). Multiple blocks on more than two Chrs indicate homoeologous retention and chromosomal rearrangements (Schmutz et al., 2010).

Various gene duplications should be useful as subject to evolutionary divergence because the mode of duplication can influence evolutionary outcomes and plant specific traits are affected by functional gene duplication (Kaessmann, 2010; Cannon and Shoemaker, 2012; Yang and Bharti, 2012). A large impact of segmental duplications has been reported on the evolution of genes involved in phenotypic traits such as disease resistance and developmental process. QTLs associated with corn earworm resistance, Sclerotinia stem rot, soybean cyst nematode, seed-related traits (size, weight, and yield) and contents of protein, oil and sucrose were conserved across homeologous genomic regions after genome duplication (Shin et al., 2008; Kim et al., 2009). Therefore, the integration of soybean genomics with relative phenotypic trait resources should facilitate the identification of homeologous chromosomal rearrangements and new duplicate

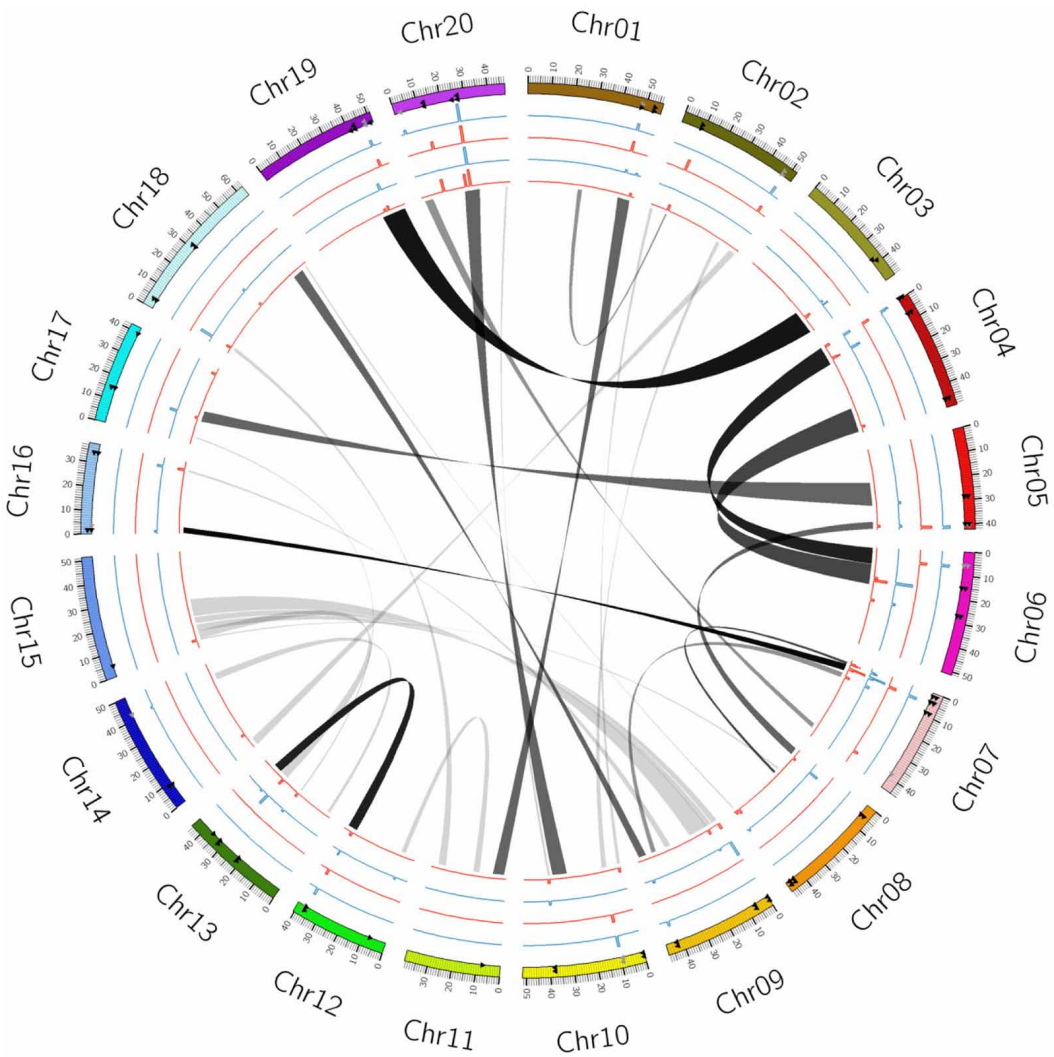

FIGURE 1 | Circular map showing homeologous relationships among 20 soybean chromosomes based on QTLs associated with protein and oil content in recently duplicated segments. Chr01–Chr20 denote chromosome number of the soybean genome. The scale of 0–50 kb represents physical locations. Gray ribbons represent as duplicated positions of the *G. max* genome after similar duplicated regions are grouped as bundles. Circular lines from the inner to the outer: 1, LOD of protein QTL (red); 2, R^2 of protein QTL (blue); 3, LOD of oil QTL (red); 4, R^2 of oil QTL (blue).

gene copies and help to identify informative QTLs related to desirable traits in soybean.

QTLs FOR SEED PROTEIN/OIL CONTENT

Seed protein content has been investigated extensively in many soybean breeding programs (Helms and Orf, 1998; Cober and Voldeng, 2000; Panthee et al., 2005). Since seed protein content is determined by the interaction of various genetic loci with environmental factors, traditional soybean breeding has been assisted by extensive linkage map analyses, which have been conducted to identify QTLs for protein and oil contents with a range of genetic backgrounds and in different environments (Diers et al., 1992; Csanadi et al., 2001; Jun et al., 2008). Various soybean lines such as wild and cultivated soybeans and genotypes from different countries have also been used to explore seed protein QTLs (Sebolt et al., 2000; Csanadi et al., 2001; Jun et al., 2008).

From a large number of studies performed to identify QTLs for seed protein content in soybean, approximately 108 and 124 QTLs with various phenotypic variations have been correlated with the seed protein and oil content, respectively, and these were located on all of the soybean Chrs (http://soybase.org). Over 61 QTLs are associated with the protein content in 17 different soybean populations (Vuong et al., 2007). The seed composition traits may be associated with seed sucrose content in soybean and a QTL for seed sucrose content on Chr 20 made a phenotypic contribution of greater than 10%, which may be a major QTL with a pleiotropic effect (Maughan et al., 2000). Combined with soybean genomic analysis, the QTLs for protein and their related traits could facilitate the rapid selection of significant protein QTLs and the identification of candidate genes regulating seed protein content.

A MAJOR QTL FOR SEED PROTEIN CONTENT IN RESPECT TO SOYBEAN GENOME DUPLICATION

Remarkable attention has been given to the major seed protein QTL mapped on Chr 20 [previously known as a linkage group (LG) I] because of the highest additive effect across many mapping populations and multiple environments (Brummer et al., 1997; Sebolt et al., 2000; Csanadi et al., 2001; Chung et al., 2003; Nichols et al., 2006). Accompanying with a reduced oil level, the application of marker-assisted selection to protein QTL on Chr 20 confirmed an increased production of protein in homozygous lines carrying alleles from a high protein parent (Diers et al., 1992; Yates et al., 2004) and the same correlation was also observed in different mapping populations using wild soybean as one of the parent (Brummer et al., 1997; Sebolt et al., 2000). The mapped QTL for protein and oil between Satt496 and Satt239 on Chr 20 had an additive effect of the PI 437088 alleles with increased protein level but reduced oil content (Chung et al., 2003). The near-isogenic line P-C609-45-2 was segregated at the smallest QTL interval on Chr 20, which corresponded to seed protein level (Nichols et al., 2006). Candidate genes identified by QTL analysis on Chr 20 have been associated with seed protein regulation and next-generation sequencing technology was also applied to an extensive investigation of the seed protein QTL on Chr 20 (Bolon et al., 2010; Severin et al., 2010). Although analyses of the linkage map and the major protein QTL on Chr 20 have been addressed using several approaches (Wang et al., 2006; Joseph, 2009; Qi et al., 2011), the regulation of seed protein content is not clear yet (Bolon et al., 2010). Furthermore, seed protein regulation may be related to soybean genome structure, such as gene duplication representing a primary source for gain of new gene function. It can be understood by whole genome and small-scale duplications facilitating an increase in biological complexity and evolutionary novelties (Van de Peer et al., 2009).

The recent genome duplication occurred frequently on many soybean Chrs, which is supported by the coincidence of several duplicate loci in the Chrs (Cannon and Shoemaker, 2012). Rearrangements of homeologous chromosomal regions are also observed in corresponding QTL regions related to both protein and oil traits. Based on Circos map, QTLs across duplicated regions were conserved, for example, Chr 4 vs. Chr 6 and Chr 3 vs. Chr 19 (**Figure 1**). Although Chr 20 shares high homology with the long arm of Chr 10, the major QTLs for seed protein content are only observed on Chr 20, not on its duplicated region of Chr 10 (**Figure 1**). It was reported that there is a close association between a QTL for seed composition in one member of a homeologous pair and a similar QTL on another duplicated pair (Shoemaker et al., 1996; Shin et al., 2008; Kim et al., 2009). However, protein and oil QTLs duplicated within interrelated homeologous regions showed rearrangement of the QTLs in homeologous pairs that occurred due to the recent duplication event (**Figure 1**; Shoemaker et al., 1996). The recent soybean genome structure shows that the major QTLs for soybean seed protein and oil are located mainly within not only homeologous regions (Chr 20) but also other homeologous regions (Chr 10) (**Figure 1**). The analysis of duplicated regions may suggest the rapid divergence of both regions at the chromosomal level (Chr 20 vs. Chr 10) (Picket and Meeks-Wagner, 1995).

COMPARISON OF DUPLICATED REGIONS ASSOCIATED WITH SEED PROTEIN CONTENT

Duplicated regions in plant genome that contain genes may cause gene retention/loss, where polyploidy commonly contributes an expansion of gene copy (Cannon et al., 2004). Since subsequent duplication leads mutated genes to alter their functions, soybean genome duplication may also act on gene regulation (Shoemaker et al., 1996; Schmutz et al., 2010). The concordance of homeologous regions with QTLs for seed protein content support common roles, which homeologous loci and genetic redundancy inherited quantitatively (Shoemaker et al., 1996). However, it is assumed that the absence of the QTL in Chr 10 is derived from the absence of gene contents which could be decayed or from insertion of genes into Chr 20 after recent duplication event. The major QTL for seed protein contents, Prot 15-1, is associated with markers Satt239 and Satt496 on Chr20: 24,867,385..28,878,629 and its duplicated region is located on Chr10: 30,286,648..34,294,718. Among 81 genes in both duplicated regions, a total of 27 genes commonly identified in both regions and 19 and 35 genes were present only on Chr 20 and Chr 10, respectively (**Table 1**). Since genome duplication also gives a large impact on gene content and retention rate for balancing (Edger and Pires, 2009), the QTL for soybean seed protein could be a good clue to trace duplicated genes associated with seed protein content.

Table 1 | Gene divergence of duplicated regions between Chr 20 (24,867,385..28,878,629) and Chr 10 (30,286,648..34,294,718).

Chromosome 20	Chromosome 10	Putative function
–*	Glyma10g23750	Core-2/I-branching beta-1,6-N-acetylglucosaminyltransferase family protein
–	Glyma10g23790	Uricase/urate oxidase/nodulin 35, putative
–	Glyma10g23800	Concanavalin A-like lectin protein kinase family protein
–	Glyma10g23810	Exocyst subunit exo70 family protein A1
–	Glyma10g23840	Double Clp-N motif-containing P-loop nucleoside triphosphate hydrolases superfamily protein
–	Glyma10g23910	Polynucleotidyl transferase, ribonuclease H-like superfamily protein
–	Glyma10g24030	Glycosyl hydrolase superfamily protein
Glyma20g17960	Glyma10g24060	GTP-binding family protein
Glyma20g17990	–	Urease accessory protein D
Glyma20g18010	–	Pentatricopeptide (PPR) repeat-containing protein
Glyma20g18280	–	Terpene synthase 03
–	Glyma10g24080	Expansin B2
–	Glyma10g24100	Double Clp-N motif-containing P-loop nucleoside triphosphate hydrolases superfamily protein
–	Glyma10g24120	Expansin B2
–	Glyma10g24190	Leucine-rich repeat transmembrane protein kinase family protein
–	Glyma10g24200	AT1G21280.1
–	Glyma10g24270	Gibberellin 2-oxidase
Glyma20g18290	Glyma10g24340	AMMECR1 family
Glyma20g18420	–	F-box and associated interaction domains-containing protein
Glyma20g18440	Glyma10g24350	U2 small nuclear ribonucleoprotein A
Glyma20g18450	–	Homeobox protein 22
Glyma20g18460	–	Homeobox protein 22
Glyma20g18520	–	Homeobox protein 22
Glyma20g18540	–	Homeobox protein 22
Glyma20g18550	Glyma10g24360	P-loop containing nucleoside triphosphate hydrolases superfamily protein
–	Glyma10g24400	P-loop containing nucleoside triphosphate hydrolases superfamily protein
Glyma20g18620	Glyma10g24420	P-loop containing nucleoside triphosphate hydrolases superfamily protein
Glyma20g18860	Glyma10g24420	P-loop containing nucleoside triphosphate hydrolases superfamily protein
–	Glyma10g24430	Vesicle-associated membrane protein 726
Glyma20g18870	Glyma10g24540	Protein kinase superfamily protein
Glyma20g18890	Glyma10g24550	Ankyrin repeat family protein
–	Glyma10g24570	Ankyrin repeat family protein
Glyma20g18900	Glyma10g24580	RING/U-box superfamily protein
Glyma20g18970	Glyma10g24580	RING/U-box superfamily protein
Glyma20g18980	Glyma10g24590	Peroxisomal 3-ketoacyl-CoA thiolase 3
Glyma20g18990	–	hAT transposon superfamily
–	Glyma10g24600	DZC (Disease resistance/zinc finger/chromosome condensation-like region) domain containing protein
Glyma20g19000	Glyma10g24620	Potassium channel beta subunit 1
Glyma20g19200	Glyma10g24630	Pectin lyase-like superfamily protein
Glyma20g19210	Glyma10g24650	Inosine triphosphate pyrophosphatase family protein
Glyma20g19250	–	pfkB-like carbohydrate kinase family protein
Glyma20g19470	–	Modifier of rudimentary [Mod(r)] protein
–	Glyma10g24670	P-loop containing nucleoside triphosphate hydrolases superfamily protein

(Continued)

(Continued)

Table 1 | Continued

Chromosome 20	Chromosome 10	Putative function
–	Glyma10g24740	Glyma10g24740.1
–	Glyma10g25070	Ferritin 4
–	Glyma10g25340	Plant U-Box 15
Glyma20g19550	Glyma10g25420	Protein of unknown function (DUF3741)
Glyma20g19580	–	Oligosaccharyltransferase complex/magnesium transporter family protein
Glyma20g19600	–	Glyma20g19600.1
Glyma20g19620	–	Glyma20g19620.1
–	Glyma10g25440	Leucine-rich repeat receptor-like protein kinase family protein
Glyma20g19640	Glyma10g25480	GATA transcription factor 1
–	Glyma10g25490	Uncharacterized protein family (UPF0016)
–	Glyma10g25500	Core-2/I-branching beta-1,6-N-acetylglucosaminyltransferase family protein
Glyma20g19670	Glyma10g25510	AT3G22520.1
Glyma20g19680	–	HSP20-like chaperones superfamily protein
Glyma20g19710	–	Hydroxyethylthiazole kinase family protein
–	Glyma10g25550	Hydroxyethylthiazole kinase family protein
Glyma20g19720	Glyma10g25560	Disease resistance-responsive (dirigent-like protein) family protein
Glyma20g19920	–	Disease resistance-responsive (dirigent-like protein) family protein
Glyma20g19930	Glyma10g25560	Disease resistance-responsive (dirigent-like protein) family protein
Glyma20g19940	–	Glyma20g19940.1
–	Glyma10g25570	Disease resistance-responsive (dirigent-like protein) family protein
Glyma20g19970	Glyma10g25620	RING/FYVE/PHD zinc finger superfamily protein
Glyma20g19980	Glyma10g25630	Heat shock protein 60
–	Glyma10g25640	AT5G41980.1
Glyma20g20010	Glyma10g25670	GRIM-19 protein
–	Glyma10g25680	HVA22 homologue E
Glyma20g20040	Glyma10g25690	BLISTER
Glyma20g20050	Glyma10g25700	BLISTER
Glyma20g20070	Glyma10g25710	Coenzyme F420 hydrogenase family/dehydrogenase, beta subunit family
–	Glyma10g25750	AT2G29880.1
Glyma20g20180	Glyma10g25760	YUCCA 3
Glyma20g20280	Glyma10g25760	YUCCA 3
Glyma20g20300	–	Protein kinase superfamily protein
–	Glyma10g25790	Beta-1,2-N-acetylglucosaminyltransferase II
–	Glyma10g25800	Disease resistance family protein / LRR family protein
–	Glyma10g26100	Glyma10g26100.1
–	Glyma10g26120	ROTUNDIFOLIA like 8
–	Glyma10g26150	Ubiquitin C-terminal hydrolase 3
–	Glyma10g26160	Disease resistance family protein/LRR family protein

(Continued)

Table 1 | Continued

**Indicates no G. max gene presents on the chromosome.*

A large inversion with synteny in the corresponding regions of Chr 20 and Chr 10 was detected by a dot plot comparison between these two Chrs (**Figure 2A**; Cannon et al., 2004). A positive linear synteny is also observed with a slight interruption (**Figure 2A**) and leads to survey the conserved blocks along with conserved genes (**Figures 2B**,**C**), showing a higher level of synteny with one another. Schmutz et al. (2010) suggested that most of the duplicated regions were conserved but interspersed with insertions/deletions and inversions. All of the syntenic blocks were conserved and some of the syntenic regions between Chr 20 and Chr 10 still obtained a few syntenic genes (**Figure 2C**), which may reflect the recent genome duplication event regarding gene content (Pagel et al., 2004; Cannon and Shoemaker, 2012).

Among the 19 genes present only on Chr 20, we identified the four tandem duplicates of homeobox protein 22 (HB22), which is reported as *Medicago truncatula* homolog expressed in

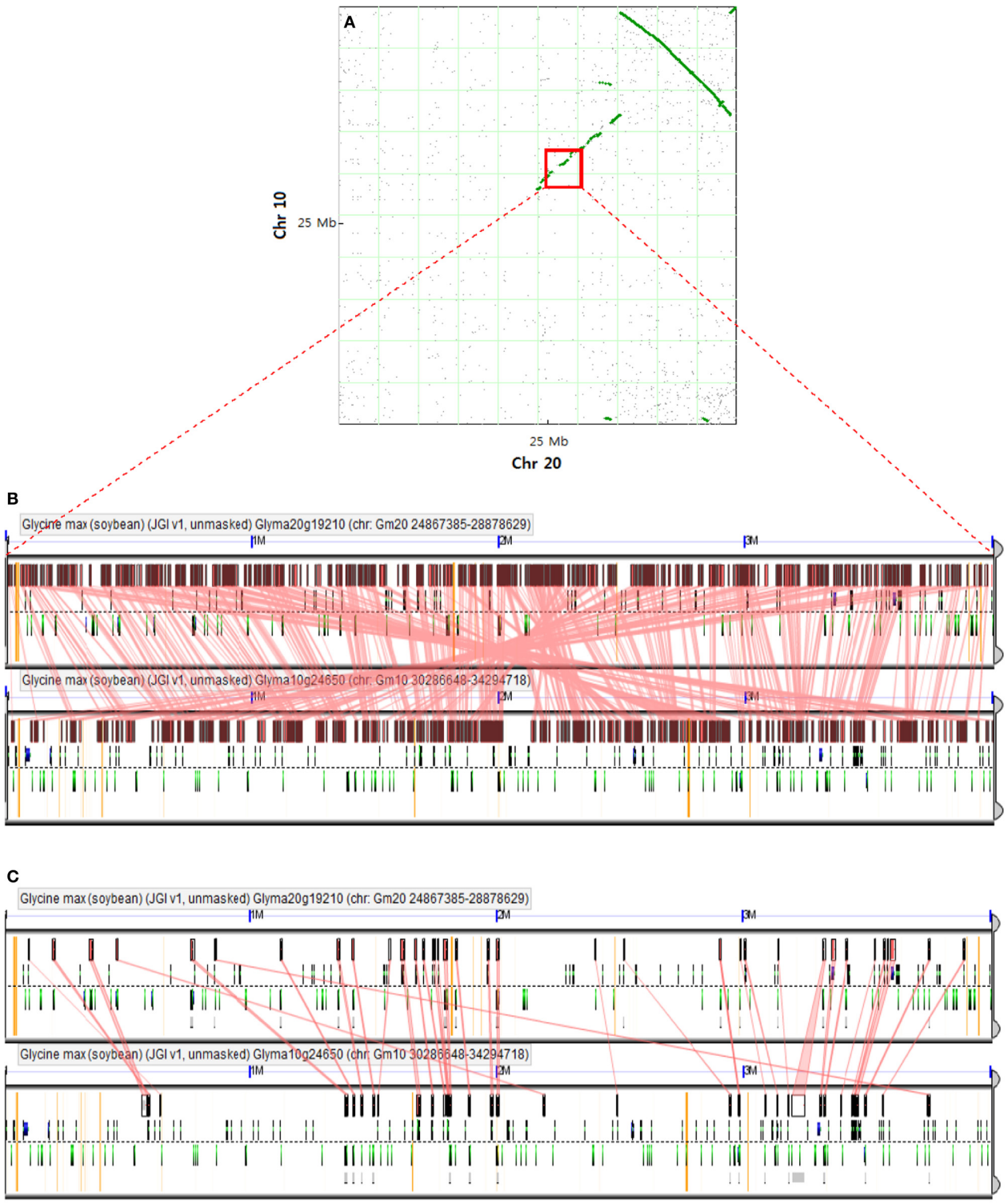

FIGURE 2 | Comparison of two soybean chromosomes. (A) Syntenic dot plot of soybean Chr 10 and Chr 20 [adapted from Cannon and Shoemaker (2012)]. Each dot represents homology of the predicted synteny between two chromosomes. A large inversion of synteny in the upper right quadrant is indicated by a line of homology dots that slope down to the right. A positive correlation based on synteny between two chromosomes is denoted by a discontinuous line with a slope moving up to the right. Red square denotes a determined region for alignment algorithm analysis. **(B)** The conserved blocks in the recently duplicated regions of Chr 10 and Chr 20 are based on the alignment of large genomic regions using LastZ algorithm, which is available at http://genomevolution.org/r/56zj. **(C)** Conserved genes in the recently duplicated regions of Chr 10 and Chr 20 based on predicted transcripts using Genome Threader algorithm (spliced alignment of peptides) at http://genomevolution.org/r/56zh.

endosperm at seed filling stage (Verdier et al., 2008). This previous report raises a possibility that these tandem duplicates could regulate the stage of seed filling in soybean and contribute the protein/oil QTL on Chr 20. In addition, several candidate genes identified by Soy GeneChip and transcriptome analyses are thought to be associated with protein content, which may help us understand soybean seed protein regulation, and ten genes were differentially expressed between NILs carrying high and low seed protein content alleles (Bolon et al., 2010).

SUMMARY

The accumulated genomic data can be used to identify functional genes of specific traits. Even this can provide a basis for predicted gene duplicates following modes of recent duplications. Here, in this review, we compared duplicated genomic regions, which are involved in seed protein content. Increased divergence after recent duplications resulted in the appearance or disappearance of QTLs related to protein and/or oil, suggesting gene retention/loss. Comparing gene and sequence divergence between recently duplicated genomic regions harboring a major QTL for seed protein content on Chr 20, 27 out of 81 genes were present in the homeologous regions of both Chr 20 and Chr 10. Several genes with over- and/or under- retained may be functional and contribute to seed protein content regulation. Therefore, the information of recently duplicated and diverged genes will provide insights into the identification of candidate genes of agronomically important trait.

ACKNOWLEDGMENTS

This research was supported by a grant from the Next-Generation BioGreen 21 Program (No. PJ008117) of the Rural Development Administration, Republic of Korea.

REFERENCES

Boerma, H. R., and Specht, J. E. (2004). *Soybeans: Improvement, Production and Uses.* Madison, WI: Am Soc of Agro.

Bolon, Y. T., Joseph, B., Cannon, S. B., Graham, M. A., Diers, B. W., Farmer, A. D., et al. (2010). Complementary genetic and genomic approaches help characterize the linkage group I seed protein QTL in soybean. *BMC Plant Biol.* 10:41. doi: 10.1186/1471-2229-10-41

Brummer, E. C., Graef, G. L., Orf, J., Wilcox, J. R., and Shoemaker, R. C. (1997). Mapping QTL for seed protein and oil content in eight soybean populations. *Crop Sci.* 37, 370–378. doi: 10.2135/cropsci1997.0011183X003700020011x

Cannon, S. B., Mitra, A., Baumgarten, A., Young, N. D., and May, G. (2004). The roles of segmental and tandem gene duplication in the evolution of large families in *Arabidopsis thaliana*. *BMC Plant Pathol.* 4:10. doi: 10.1186/1471-2229-4-10

Cannon, S. B., and Shoemaker, R. C. (2012). Evolutionary and comparative analysis of the soybean genome. *Breed Sci.* 61, 437–444. doi: 10.1270/jsbbs.61.437

Chung, J., Babka, H. L., Graef, G. L., Staswick, P. E., Lee, D. J., Cregan, P. B., et al. (2003). The seed protein, oil and yield QTL on soybean linkage group I. *Crop Sci.* 43, 1053–1067. doi: 10.2135/cropsci2003.1053

Cober, E. R., and Voldeng, H. D. (2000). Developing high-protein, high-yield soybean populations and lines. *Crop Sci.* 40, 39–42. doi: 10.2135/cropsci2000.40139x

Csanadi, G., Vollmann, J., Stift, G., and Lelley, T. (2001). Seed quality QTLs identified in a molecular map of early maturing soybean. *Theor. Appl. Genet.* 103, 912–919. doi: 10.1007/s001220100621

Diers, B. W., Keim, P., Fehr, W. R., and Shoemaker, R. C. (1992). RFLP analysis of soybean seed protein and oil content. *Theor. Appl. Genet.* 83, 608–612. doi: 10.1007/s10577-009-9055-9

Edger, P. P., and Pires, J. P. (2009). Gene and genome duplication: the impact of dosage-sensitivity on the fate of nuclear genes. *Chrom. Res.* 17, 699–717. doi: 10.1007/s10577-009-9055-9

Flagel, L. E., and Wendel, J. E. (2009). Gene duplication and evolutionary novelty in plants. *New Pathol.* 183, 557–564. doi: 10.1111/j.1469-8137.2009.02923.x

Freeling, M. (2009). Bias in plant gene content following different sorts of duplication: tandem, whole-genome, segmental, or by transposition. *Annu. Rev. Plant Biol.* 60, 433–453. doi: 10.1146/annurev.arplant.043008.092122

Helms, T. C., and Orf, J. H. (1998). Protein, oil, and yield of soybean lines selected for increased protein. *Crop Sci.* 38, 707–711. doi: 10.2135/cropsci1998.0011183X003800030015x

Joseph, B. (2009). *Genomic Analysis of a Major Seed Protein/oil QTL Region on Soybean Linkage Group I.* Graduate theses and dissertation. Paper 10650. Available online at: http://lib.dr.iastate.edu/etd/10650

Jun, T. H., Van, K., Kim, M. Y., Lee, S. H., and Walker, D. R. (2008). Association analysis using SSR markers to find QTL for seed protein content in soybean. *Euphytica* 162, 179–191. doi: 10.1007/s10681-007-9491-6

Kaessmann, H. (2010). Origins, evolution and phenotypic impact of new genes. *Genome Res.* 20, 1313–1326. doi: 10.1101/gr.101386.109

Kim, K. D., Shin, J. H., Van, K., Kim, D. H., and Lee, S.-H. (2009). Dynamic rearrangements determine genome organization and useful traits in soybean. *Plant Physiol.* 151, 1066–1076. doi: 10.1104/pp.109.141739

Kim, M. Y., Lee, S., Van, K., Kim, T. H., Jeong, S. C., Choi, I. Y., et al. (2010). Whole genome sequencing and intensive analysis of the undomesticated soybean (*Glycine soja* Sieb. And Zucc.) genome. *Prod. Natl. Acad. Sci. U.S.A.* 107, 22032–22037. doi: 10.1073/pnas.1009526107

Kim, M. Y., Van, K., Kang, Y. J., Kim, K. H., and Lee, S.-H. (2012). Tracing soybean domestication history: from nucleotide to genome. *Breed. Sci.* 61, 445–452. doi: 10.1270/jsbbs.61.445

Lam, H. M., Xu, X., Liu, X., Chen, W., Yang, G., Wong, F. L., et al. (2010). Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection. *Nat. Genet.* 42, 1053–1059. doi: 10.1038/ng.715

Maughan, P. J., Maroof, M. A. S., and Buss, G. R. (2000). Identification of quantitative trait loci controlling sucrose content in soybean (*Glycine max*). *Mol. Breed.* 6, 105–111. doi: 10.1023/A:1009628614988

Nichols, D. M., Glover, K. D., Carlson, S. R., Specht, J. E., and Diers, B. W. (2006). Fine mapping a seed protein QTL on soybean linkage group I and its correlated effects on agronomic traits. *Crop Sci.* 46, 834–839. doi: 10.2135/cropsci2005.05-0168

Pagel, J., Walling, J. G., Young, N. D., Shoemaker, R. C., and Jackson, S. A. (2004). Segmental duplications within the *Glycine max* genome revealed by fluorescence *in situ* hybridization of bacterial artificial chromosomes. *Genome* 47, 764–768. doi: 10.1139/g04-025

Panthee, D. R., Pantalone, V. R., West, D. R., Saxton, A. M., and Sams, C. E. (2005). Quantitative trait loci for seed protein and oil concentration and seed size in soybean. *Crop Sci.* 45, 2015–2022. doi: 10.2135/cropsci2004.0720

Picket, F., and Meeks-Wagner, R. (1995). Seeing double: appreciating genetic redundancy. *Plant Cell* 7, 1347–1356. doi: 10.1105/tpc.7.9.1347

Qi, Z. M., Wu, Q., Han, X., Sun, Y. N., Du, X. Y., Liu, C. Y., et al. (2011). Soybean oil content QTL mapping and integrating with meta-analysis method for mining genes. *Euphytica* 179, 499–514. doi: 10.1007/s10681-011-0386-1

Schmutz, J., Cannon, S. B., Schlueter, J., Ma, J., Mitros, T., Nelson, W., et al. (2010). Genome sequence of the palaeopolyploid soybean. *Nature* 463, 178–183. doi: 10.1038/nature08670

Sebolt, A. M., Shoemaker, R. C., and Diers, B. W. (2000). Analysis of a quantitative trait locus allele from wild soybean that increases seed protein concentration in soybean. *Crop Sci.* 40, 1438–1444. doi: 10.2135/cropsci2000.4051438x

Severin, A. J., Woody, J. L., Bolon, Y.-T., Joseph, B., Diers, B. W., Farmer, A. D., et al. (2010). RNA-Seq atlas of *Glycine max*: a guide to the soybean transcriptome. *BMC Plant Biol.*

10:160. doi: 10.1186/1471-2229-10-160

Shin, J. H., Van, K., Kim, D. H., Kim, K. D., Jang, Y. E., Choi, B.-S., et al. (2008). The lipoxygenase gene family: a genomic fossil of shared polyploidy between *Glycine max* and *Medicago truncatula*. *BMC Plant Biol.* 8:133. doi: 10.1186/1471-2229-8-133

Shoemaker, R. C., Polzin, K., Labate, J., Specht, J., Brummer, E. C., Olson, T., et al. (1996). Genome duplication in soybean (*Glycine* subgenus *soja*). *Genetics* 144, 329–338.

Sterck, L., Rombauts, S., Vandepoele, K., Rouze, P., and Van de Peer, Y. (2007). How many genes are there in plants (and way are they there)? *Curr. Opin. Plant Biol.* 10, 199–203. doi: 10.1016/j.pbi.2007.01.004

Stupar, R. M., and Specht, J. E. (2013). Insights from the soybean (*Glycine max* and *Glycine soja*) genome: past, present, and future. *Adv. Agron.* 118, 177–204.

Van, K., Kim, D., Cai, C. M., Kim, M. Y., Shin, J. H., Graham, M. A., et al. (2008). Sequence level analysis of recently duplicated regions in soybean [*Glycine max* (L.) Merr.] genome. *DNA Res.* 15, 93–102. doi: 10.1093/dnares/dsn001

Van, K., Kim, M. Y., Shin, J. H., Kim, K. D., Lee, Y.-H., and Lee, S.-H. (in press). "Molecular evidence for soybean domestication," in *Advances in Genomics of Plant Genetic Resources,* Vol 2*, Genomics of Plant Genetic Resources to Improve Crop Production, Food Security and Nutritional Quality*, ed R. Tuberosa (Springer).

Van de Peer, Y., Maere, S., and Meyer, A. (2009). The evolutionary significance of ancient genome duplications. *Nat. Rev. Genet.* 38, 615–643. doi: 10.1038/nrg2600

Verdier, J., Kakar, K., Gallardo, K., Le Signor, C., Aubert, G., Schlereth, A., et al. (2008). Gene expression profiling of *M. truncatula* transcription factors identifies putative regulators of grain legume seed filling. *Plant Mol. Biol.* 67, 567–580. doi: 10.1007/s11103-008-9320-x

Vuong, T. D., Wu, X., Pathan, M. D. S., Valliyodan, B., and Nguyen, H. T. (2007). "Genomics approaches to soybean improvement," in *Genomics Assisted Crop Improvement,* Vol. 2, eds R. K. Varshney and R. Tuberosa (Dordecht: Genomics Application in Crops), 243–279. doi: 10.1007/978-1-4020-6297-1_11

Wang, Y., Yao, J., Zhang, Z. F., and Zheng, Y. L. (2006). The comparative analysis based on maize integrated QTL map and meta-analysis on plant height QTLs. *Chin. Sci. Bull.* 51, 2219–2230. doi: 10.1007/s11434-006-2119-8

Wendel, J. F., and Doyle, J. J. (2005). "Polyploidy and evolution in plants," in *Plant Diversity and Evolution*, ed R. J. Henry (Wallingford: CABI Publishing), 97–117.

Yang, N. D., and Bharti, A. K. (2012). Genome-enabled insights into legume biology. *Annu. Rev. Plant Biol.* 63, 283–305. doi: 10.1146/annurev-arplant-042110-103754

Yates, J. L., Harris, D. K., and Boerma, H. R. (2004). "Marker-assisted selection around a major QTL on Chr 20 increases seed protein content in backcross-derived lines of soybean," in *Soy2004 the 10th Biennial Conference of the Cellular and Molecular Biology of the Soybean* (Columbia, MO), 67.

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. doi: 10.1104/pp.104.058891

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 06 December 2012; accepted: 17 May 2013; published online: 05 June 2013.

Citation: Lestari P, Van K, Lee J, Kang YJ and Lee S-H (2013) Gene divergence of homeologous regions associated with a major seed protein content QTL in soybean. Front. Plant Sci. ***4****:176. doi: 10.3389/fpls.2013.00176*

This article was submitted to Frontiers in Plant Genetics and Genomics, a specialty of Frontiers in Plant Science.