A DISSERTATION FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

Molecular and genetic analyses of a major gene for seed protein content in soybean

By

Jang Young Eun

February, 2015

MAJOR IN CROP SCIENCE AND BIOTECHNOLOGY

DEPARTMENT OF PLANT SCIENCE

THE GRADUATE SCHOOL OF SEOUL NATIONAL UNIVERSITY

Molecular and genetic analyses of a major gene for seed protein content in soybean

JANG YOUNG EUN

DEPARTMENT OF PLANT SCIENCE
THE GRADUATE SCHOOL OF SEOUL NATIONAL UNIVERSITY

GENERAL ABSTRACT

Seeds store nutrients for conservation and propagation of genetic materials between two generations and propagation. Seed crops are cultivated to supply those nutrients. Soybean is important seed crop since it has high protein content to maintain body and generate energy. Continuous efforts of soybean breeders are directed to improve the protein quantity by identifying linked markers and quantitative trait loci (QTL). Danbaekkong has 48% of seed protein content that its major protein QTL was founded on Chr 20, surrounding region of Satt239 and Satt496 marker interval. We selected three residual heterozygous lines heterozygous in the Satt239-Satt496 region from two recombinant inbred line populations of Sinpaldalkong 2 x

Danbaekkong and Daewonkong x Danbaekkong soybean crosses to develop two sets of near isogenic lines (NILs), Prot_high$_{SD}$/Prot_low$_{SD}$ and Prot_high$_{DD}$/Prot_low$_{DD}$ with different seed protein content. The maternal parent cultivar Danbaekkong, carrying a high protein content allele, showed high seed protein content. In the genomic region harboring the introgressed Danbaekkong segment on Chr 20, we identified nucleotide differences between low- and high-protein NILs through whole genome sequencing. From the major protein QTL region, we found 66 non-synonymous single nucleotide polymorphisms and one frameshifts are overlapped between both low- and high-protein lines. After elimination of genes by expression during seed stages, we identified two genes, calcium-dependent protein kinase and exocyst subunit exo70 family protein, are more likely involved in see protein accumulation in soybean.

The gene expression changes of developing seed between Prot_low$_{DD}$ and Prot_high$_{DD}$ represent the pattern of up-regulation of many genes in Prot_high$_{DD}$ at 2 week after flowering (WAF), stage of early maturation. Up-regulation pattern of storage protein genes in Prot_high$_{DD}$ was increasing according to seed maturation. There is a block of gene expression pattern that raised up-regulation in Prot_low$_{DD}$, consist of genes involving protein degradation. Transcription factors were highly up-regulated in Prot_high$_{DD}$ at 2 WAF, including major transcription factor to regulate seed development. Gene expression of carbon precursors, protein, and oil

metabolic enzymes are compared between 1 and 2 WAF in Prot_low$_{DD}$ to identify initial gene expression at early maturation stage. *Glycine max* undergoes two rounds of whole-genome duplication; the number of genes involved in the three synthesis pathways is more than two times higher than that in *Arabidopsis*. Among these genes, five were conserved as single-copy genes and 44 were high copy gene families consisting of more than seven homolog members. We identified five differentially expressed genes in immature seeds aged between 1 and 2 WAF.

To find out protein expression change according to storage protein difference, additional proteomic research was hired for two NIL lines. Three enzymes, sucrose synthase, glyceraldehyde-3-phosphate dehydrogenase and ketol-acid reductoisomerase showed over two times differential expression between Prot_high lines and Prot_low lines. In this research, we developed two NIL pairs of high and low seed protein content derived from different genetic backgrounds. Genomic, trasncriptomic and proteomic analysis using these NILs indicates that development regulation from transcription factors and initial carbon metabolism is important to seed storage protein synthesis.

Key words: *Glycine max;* seed storage protein; near isogenic line; next generation sequencing; seed development

Student number: 2007-30866

CONTENTS

LIST OF FIGURES

The region of sequence comparison is shown on the rightmost vertical line. Green circles represent the locations of nine genes with non-synonymous SNPs of those indicated on the right side. Each SNP and indel site in the five genes is indicated by a black triangle, and the sequences of NIL lines are shown in blue (low-protein) and red (high-protein). If the SNP has a non-synonymous change, the amino acid change is shown below the SNP site. As Glyma20g17500 has 35 SNPs in its genic region, only three non-synonymous SNPs are shown in this figure.

Figure II-2 Predicted protein structures of calcium dependent protein kinase (CDPK) between low- and high-protein lines. The protein sequence of low- and high-protein lines were obtained from Williams 82 reference genome and Danbaekkong, respectively. Green arrows indicate the site of protein change caused by non-synonymous mutation. Black arrows point the protein structure change on serine/threonine kinase domain of CDPK.

Figure III-1 Soybean seed development during experimental period. a) Seeds at four stages of seed filling. Seeds sampling was started at 1 WAF and continued to 4 WAF with 7 days intervals. b) Seed size and mass of each seed filling stage. Length, thickness, width and weight values are average of 10 seeds and error bars are standard deviation

Figure III-2 Venn diagram of differentially expressed genes (DEGs) of differential expressed genes between Prot_high and Prot_low lines. Number

in each circles represents number of DEGs and the light blue, red, blue and yellow colors of circles indicate each 1, 2, 3 and 4 WAF, respectively.

Figure III-3 Heatmap and ontologies of differentially expressed genes related with seed storage product metabolic pathway. Log_2 fold changes of FPKM value between $Prot\_high_{DD}$ and $Prot\_low_{DD}$ were calculated. On the logarithmic color scale ranging from -3 to 3, dark blue represents at least 6-fold higher gene expression in $Prot\_high_{DD}$ comparison to $Prot\_low_{DD}$, and dark red represents 6-fold higher gene expression in $Prot\_low_{DD}$ comparison to $Prot\_high_{DD}$.

Figure III-4 Heatmap of transcription factors and TF gene expression in $Prot\_high_{DD}$ and $Prot\_low_{DD}$. a) Heatmap of log_2 fold change value between $Prot\_high_{DD}$ and $Prot\_low_{DD}$. b) Gene expression of major seed development regulator ABI3 and LEC1-like genes. c) Gene expression of C2C2-Zn-YABBY and b-ZIP.

Figure III-5 Circos showing the distribution of expressed genes involved in seed protein and oil storage metabolism and their duplication patterns. The exterior circle indicates *Glycine max* chromosomes. The three inner circles are gene distributions involved in synthesis of carbon precursor (red), protein (green), and oil (blue). Genes that are expressed only at one stage are depicted in red (1 WAF) and blue (2 WAF).

Figure III-6 Ks distributions of *G. max* synteny block.

Figure III-7 Fragments per kilobase of exon per million fragments mapped (FPKM) expression patterns of a paralogous gene pair in different synteny blocks at different times during seed development. Colored dots indicate the FPKM expression value of duplicated genes involved in each of the three metabolic pathways (carbon precursor, protein, and oil synthesis); trend lines were drawn using linear regressions.

Figure IV-1 Number of up-regulated protein spots in each Prot_high and Prot_low lines. Blue, red and green bars represent up-regulated protein spot in each 2, 3 and 4 WAF, respectively.

Figure IV-2 Expression change of four proteins, sucrose synthase, glyceradehyde-3-phosphate dehydrogenase and ketol-acid reductoisomerase and glycinin during seed development. Each NIL pairs are distinguished by different color. Strait lines are represent Prot_high lines and dashed lines represent Prot_low lines.

LIST OF TABLES

LIST OF ABBREVIATIONS

QTL	Quantitative trait loci
RIL	Recombinant inbred line
NIL	Near isogenic line
RHL	Residual heterozygous line
SSR	Simple sequence repeat
NGS	Nest generation sequencing
PCR	Polymerase chain reaction
SNP	Single nucleotide polymorphism
SSR	Simple sequence repeat
WGS	Whole genome sequencing
DAF	Day after flowering
WAF	Week after flowering
DEG	Differentially expressed gene
FPKM	Number of fragments per kilobase of exon per million fragments mapped
CDPK	Calcium dependent protein kinase
SUS	Sucrose synthase
INV	Invertase
LEC1	Leafy cotyledon 1
ABI3	Abscisic acid insentitive 3
FUS3	Fusca 3

LITERATURAL REVIEWS

Residual heterozygous line

Near isogenic lines (NIL) is one of the immortal breeding populations for research quantitative traits. NIL is efficient at investigation and characterization of (I) the effect of a single gene, which was identified by QTL analysis from RILs (Glover et al. 2004, Kazi et al. 2010) and (II) the cumulative effect of each locus by taking advantage of gene pyramiding (Moon et al. 2009). In rice, introgression of wild tiller gene *PROG1* from *Oryza rufipogon* to NILs using *O. sativa* as a recurrent parent was effective for gene cloning and characterization of expression pattern (Jin et al. 2008). Thus, NILs are a powerful genetic material for map-based cloning.

Donor chromosomal region of NILs can be taken from any resources like advanced backcrosses (BCs), recombinant inbred line (RILs), double haploid (DHs), heterozygous inbred families (HIFs), or other mapping populations (F_2/F_3) (Kooke et al. 2012). Among those resources, HIF, also known as residual heterozygous lines (RHLs) is the unique one (Tuinstra et al. 1997). RHLs are selected from populations through more than five generation of inbreeding. Most of the genetic background in each RHL is

homozygous except for the targeted heterozygous QTL region. To select balanced NILs for more precise mapping and homogeneous genetic background using RHL methods, original RIL populations have advantages to be larger size, and more advanced inbred generations (Haley et al. 1994, Tuinstra et al. 1997). For instance, some genetic researches identified genes from NILs developed from RILs. Two sets of NILs were developed from RIL to identify two different resistance genes to sudden death syndrome in soybean (Triwitayakorn et al. 2005). Soybean flowering time and maturity genes were also investigated using the same RHL-derived NILs (Thakare et al. 2010, Watanabe et al. 2009, Yamanaka et al. 2005).

Seed storage protein

Plants undergoes seed stage between two vegetative generations for propagation genetic materials. Seeds are sources of dietary protein that have various protein contents and the major seed storage proteins are classified to albumins, globulins and prolamins (Shewry et al. 1995). Storage proteins are synthesized as the precursor forms at the cytoplasmic side of endoplasmic reticulum (ER) and they are matured and deposited to protein storage vacuoles (PSVs) via vesicle transport system (Herman and Larkins 1999, Müntz 1998). Like this, seed protein storage is complex

metabolic process that involving transport, gene expression and signal transduction (Baud et al. 2008, Gao et al. 2012, Verdier and Thomson 2008, Weber et al. 2005). The essential elements for protein synthesis, nitrogen (N) and carbon (C) are transported from vegetative tissues to developing embryo (Angeles Núñez and Tiessen 2011, Masclaux-Daubresse et al. 2010). Sink capacity of N and C and N/C ratio in the embryo is major factors of amino acid synthesis for seed protein accumulation (Takahashi et al. 2003, Schmidt et al. 2011). N is converted from amino acid and C is dissolved from sucrose in developing embryo. Sucrose is a primary molecule of glycolysis which is involved in not only synthesis of starch and lipid but also interconversion of amino acids for storage protein synthesis (Pandurangan et al. 2012). The intermediate molecules including 3-phosphoglycerate, pyruvate, α-ketoglutarate and oxaloacetate are used as carbon backbones for amino acids synthesis.

The storage protein synthesis is a part of genetically determined process for seed development and maturation (Gao et al. 2012). In *Arabidopsis*, LEAFY COTYLEDON (LEC) 1 and 2, ABSCISIC ACID-INSENTITIVE 3 (ABI3) and FUSCA 3 (FUS3), encoding transcription factors (TFs), are known as ‘master regulators’ for seed development. These master regulators and other TFs regulate initiation of seed storage protein gene expression as intrinsic seed condition (Verdier and Thompson 2008). The phytohormone ABA and GA are other key factors regulating seed

maturation processes including the initiation of the maturation phase, filling of seed reserves and entrance into dormancy through signal transduction to master regulators (Finkelstein et al. 2002). In addition, sugars act as signal molecules that regulate gene expression of gene involved in seed filling and maturation notably triggered by sucrose/hexose ratio in the embryo (Weber et al. 2005). Therefore, seed filling regulation factors including TFs, hormones and sugars crosstalk in complex regulation network system (Abid et al. 2010).

Next generation sequencing

In recent years, sequencing methods using different methods with traditional Sanger method were hired to whole genome and transcriptome, so called as next-generation sequencing (NGS) technologies (Shendure and Ji 2008, Grada and Weinbrecht 2013). They have allowed to expand more sequence database in a shorter time than the Sanger method (Burger et al. 2008, Schmutz et al. 2010). Due to the rapid growth rate of sequencing technology in all areas of science, whole genome sequencing can be performed efficiently by NGS technologies at a substantially reduced cost and higher accuracy. These advantages allow to construct high resolution density maps

and genetic diversity analysis on cultivars, landraces, and wild species (Van et al. 2013). First plant genome scale sequencing using NGS technology was hired for hybridization-based resequencing of *A. thaliana* (Clark et al. 2007). After that, *de novo* plant genome sequencing has been performed using NGS for plants such as woodland strawberry, barley, tomato, pigeonpea and chickpea (Shulaev et al. 2011, Consortium PGS 2011, Consortium IBGS 2012, Varshney et al. 2012, Varshney et al. 2013). Resequencing analysis of related plant genomes are increased which based on reference sequence data (Nowrousian, 2010). Through NGS technology, transcriptome atlas of various crops are also rapidly increased (Yamakawa and Hakata 2010, Sekhon et al. 2011, Fasoli et al. 2012).

In soybean, the whole-genome of *G. max* was *de novo* sequenced at the 1.1 Gbs genome by a whole-genome shotgun approach (Schmutz et al. 2010). The genome of *G. soja* was resequenced to 43-fold average, resulted in a consensus sequence covering 97.65% of the *G. max* published genome sequence (Kim et al. 2010). Soybean transcriptome databases were established from analyses using NGS technology (Libault et al. 2010, Severin et al. 2010).

REFERENCES

Abid G, Jaquemin J-M, Sassi K, Muhovski Y, Toussaint A, Baudoin J-P (2010) Gene expression and genetic analysis during higher plants embryogenesis. Biotechnol Agron Soc Environ14: 667-680

Angeles-Núñez JG, Tiessen A (2011) Mutation of the transcription factor LEAFY COTYLEDON 2 alters the chemical composition of *Arabidopsis* seeds, decreasing oil and protein content, while maintaining high levels of starch and sucrose in mature seeds. J. Plant Physiol 168:1891-1900

Baud S, Dubreucq B, Miquel M, Rochat C, Lepiniec L (2008) Storage reserve accumulation in *Arabidopsis*: metabolic and developmental control of seed filling. The *Arabidopsis* book/American Society of Plant Biologists 6 doi: 10.1199/tab.0113

Burger, J.C., M.A. Chapman, and J.M. Burke (2008) Molecular insights into the evolution of crop plants. Am J Bot 95(2): 113–122.

Clark RM, Schweikert G, Toomajian C, Ossowski S, Zeller G, Shinn P, Warthmann N, Hu TT, Fu G, Hinds DA (2007) Common sequence polymorphisms shaping genetic diversity in *Arabidopsis thaliana*.

Science 317:338-342

Consortium IBGS (2012) A physical, genetic and functional sequence assembly of the barley genome. Nature 491:711-716

Consortium PGS (2011) Genome sequence and analysis of the tuber crop potato. Nature 475:189-195

Fasoli M, Dal Santo S, Zenoni S, Tornielli GB, Farina L, Zamboni A, Porceddu A, Venturini L, Bicego M, Murino V (2012) The grapevine expression atlas reveals a deep transcriptome shift driving the entire plant into a maturation program. Plant Cell 24:3489-3505

Finkelstein RR, Gampala SS, Rock CD (2002) Abscisic acid signaling in seeds and seedlings. Plant Cell 14:S15-S45

Gao Q, Yue G, Li W, Wang J, Xu J, Yin Y (2012) Recent Progress Using High-throughput Sequencing Technologies in Plant Molecular BreedingF. J. Integr. Plant Biol 54:215-227

Glover KD, Wang D, Arelli PR, Carlson SR, Cianzio SR et al. (2004) Near isogenic lines confirm a soybean cyst nematode resistance gene from PI 88788 on linkage group J. Crop Sci 44: 936-941

Grada A and Weinbrecht K (2013) Next-generation sequencing: methodology and application. J Invest Dermatol 133: e11

Haley SD, Afanador LK, Miklas PN, Stavely JR, Kelly JD. (1994) Heterogeneous inbred populations are useful as sources of near-isogenic lines for RAPD marker localization. Theor Appl Genet 88: 337-342

Herman EM, Larkins BA (1999) Protein storage bodies and vacuoles. Plant Cell 11:601-613

Jin J, Huang W, Gao JP, Yang J, Shi M et al. 2008. Genetic control of rice plant architecture under domestication. Nat Genet 40: 1365-1369

Kazi S, Shultz J, Afzal J, Hashmi R, Jasim M et al. (2010) Iso-lines and inbred-lines confirmed loci that underlie resistance from cultivar 'Hartwig' to three soybean cyst nematode populations. Theor Appl Genet 120: 633-644

Kooke R, Wijnker E, Keurentjes JJ (2012) Backcross populations and near isogenic lines. In Rifkin SA (ed) Quantitative Trait Loci (QTL). Springer, New York, pp 3-16

Kim MY, Lee S-H, Van K, Kim T, Jeong A, Choi I (2010) Whole-genome sequencing and intensive analysis of the undomesticated soybean

(*Glycine soja* Sieb. and Zucc.) genome. Proc Natl Acad Sci USA 107(51):22032-22037.

Libault M, Farmer A, Joshi T, Takahashi K, Langley RJ, Franklin LD, He J, Xu D, May G, Stacey G (2010) An integrated transcriptome atlas of the crop model *Glycine max*, and its use in comparative analyses in plants. Plant J 63:86-99

Müntz K (1998) Deposition of storage proteins. Plant Mol Biol 38:77-99

Masclaux-Daubresse C, Daniel-Vedele F, Dechorgnat J, Chardon F, Gaufichon L, Suzuki A (2010) Nitrogen uptake, assimilation and remobilization in plants: challenges for sustainable and productive agriculture. Annals of Botany doi:10.1093/aob/mcq028

Moon JK, Jeong SC, Van K, Maroof MAS, Lee S-H. (2009) Marker-assisted identification of resistance genes to soybean mosaic virus in soybean lines. Euphytica 169: 375-385

Nowrousian M (2010) Next-generation sequencing techniques for eukaryotic microorganisms: sequencing-based solutions to biological problems. Eukaryot. Cell 9(9): 1300–10.

Pandurangan S, Pajak A, Molnar SJ, Cober ER, Dhaubhadel S, Hernández-Sebastià C, Kaiser WM, Nelson RL, Huber SC, Marsolais F (2012)

Relationship between asparagine metabolism and protein concentration in soybean seed. J Exp Bot doi: 10.1093/jxb/ers039

Schmidt MA, Barbazuk WB, Sandford M, May G, Song Z, Zhou W, Nikolau BJ, Herman EM (2011) Silencing of soybean seed storage proteins results in a rebalanced protein composition preserving seed protein content without major collateral changes in the metabolome and transcriptome. Plant Physiol 156:330-345

Schmutz J, Cannon SB, Schlueter J, Mitros T (2010) Genome sequence of the palaeopolyploid soybean. Nature 463(7278): 178-183

Sekhon RS, Lin H, Childs KL, Hansey CN, Buell CR, de Leon N, Kaeppler SM (2011) Genome-wide atlas of transcription during maize development. Plant J 66:553-563

Shendure J, Ji H (2008) Next-generation DNA sequencing. Nat Biotechnol 26(10):1135-1145

Severin AJ, Woody JL, Bolon Y-T, Joseph B, Diers BW, Farmer AD, Muehlbauer GJ, Nelson RT, Grant D, Specht JE (2010) RNA-Seq Atlas of *Glycine max*: a guide to the soybean transcriptome. BMC Plant Biol 10:160

Shewry PR, Napier JA, Tatham AS (1995) Seed storage proteins: structure

and biosynthesis. Plant Cell 7:945-956

Shulaev V, Sargent DJ, Crowhurst RN, Mockler TC, Folkerts O, Delcher AL, Jaiswal P, Mockaitis K, Liston A, Mane SP (2011) The genome of woodland strawberry (Fragaria vesca). Nat Genet 43:109-116

Takahashi M, Uematsu Y, Kashiwaba K, Yagasaki K, Hajika M, Matsunaga R, Komatsu K, Ishimoto M (2003) Accumulation of high levels of free amino acids in soybean seeds through integration of mutations conferring seed protein deficiency. Planta 217:577-586

Thakare D, Kumudini S, Dinkins RD (2010) Expression of flowering-time genes in soybean E1 near-isogenic lines under short and long day conditions. Planta 231: 951-963

Triwitayakorn K, Njiti VN, Iqbal MJ, Yaegashi S, Town C, Lightfoot DA (2005) Genomic analysis of a region encompassing QRfs1 and QRfs2: genes that underlie soybean resistance to sudden death syndrome. Genome 48: 125-138

Tuinstra MR, Ejeta G, Goldsbrough PB (1997) Heterogeneous inbred family (HIF) analysis: a method for developing near-isogenic lines that differ at quantitative trait loci. Theor Appl Genet 95: 1005-1011

Van K, Kang YJ, Han K-S, Lee Y-H, Gwag J-G, Moon J-K, Lee S-H (2013) Genome-wide SNP discovery in mungbean by Illumina HiSeq. Theor Appl Genet 126:2017-2027

Varshney RK, Chen W, Li Y, Bharti AK, Saxena RK, Schlueter JA, Donoghue MT, Azam S, Fan G, Whaley AM (2012) Draft genome sequence of pigeonpea (Cajanus cajan), an orphan legume crop of resource-poor farmers. Nat Biotechnol 30:83-89

Varshney RK, Song C, Saxena RK, Azam S, Yu S, Sharpe AG, Cannon S, Baek J, Rosen BD, Tar'an B (2013) Draft genome sequence of chickpea (Cicer arietinum) provides a resource for trait improvement. Nat Biotechnol 31:240-246

Verdier J, Thompson RD (2008) Transcriptional regulation of storage protein synthesis during dicotyledon seed filling. Plant Cell Physiol 49:1263-1271

Watanabe S, Hideshima R, Xia ZJ, Tsubokura Y, Sato S et al. 2009. Map-based cloning of the gene Associated With the Soybean maturity locus E3. Genetics 182: 1251-1262

Weber H, Borisjuk L, Wobus U (2005) Molecular physiology of legume seed development. Annu Rev Plant Biol 56:253-279

Yamakawa H, Hakata M (2010) Atlas of rice grain filling-related metabolism under high temperature: joint analysis of metabolome and transcriptome demonstrated inhibition of starch accumulation and induction of amino acid accumulation. Plant Cell Physiol 51:795-809

Yamanaka N, Watanabe S, Toda K, Hayashi M, Fuchigami H, Takahashi R, Harada K. 2005. Fine mapping of the FT1 locus for soybean flowering time using a residual heterozygous line derived from a recombinant inbred line. Theor Appl Genet 110: 634-639

CHAPTER I

Identification of high seed protein allele effect on chromosome 20 in soybean using near isogenic lines derived from two recombinant inbred lines

ABSTRACT

Soybean produces the highest protein yield per unit area and provides protein through various food types for humans and animals. Soybean seed protein content is shown to be regulated by a small number of genes. Several research groups have identified soybean seed protein quantitative trait loci on linkage group I, now Chromosome 20. Using a recombinant inbred line population, seed protein QTL from high-protein cultivar Dangaekkong was mapped near marker Satt239 and Satt 496 on Chr 20. In this study, three different sets of near isogenic lines were selected from two different RIL populations, Sinpaldalkong 2 x Dangaekkong and Daewonkong

x Dangaekkong using marker-assisted selection (MAS) with Satt239 and Satt496. From background selection, genotype identities of markers between high- and low protein NILs on non-carrier chromosomes were more than 97% in three selected NILs. From phenotypic data, the average seed protein difference was 34.3 g kg^{-1} and the protein difference between $Prot\_high_{SD}$ and $Prot\_low_{SD}$ was more than 40 g kg^{-1}. For fine mapping of this seed protein QTL, four undetected microsatellite markers were newly developed between Satt239 and Satt496. These NILs will be effective materials for further characterization of this major QTL and for analyzing possible genetic interactions. Further studies using transcriptome and protein profiling of these sets of NILs should provide a better understanding about protein changes during seed development.

INTRODUCTION

Proteins are the essential nutrient for the human and animals that used to build body tissue and generate energy. Soybean [*Glycine max* (L.) Merr] is one of the major sources of protein which occupying about 70% of world protein meal (http://www.soystats.com/2009/Default-frames.htm). Soybean contains high protein and oil contents, approximately 40% and 20%, respectively (Clemente and Cahoon, 2009). Therefore, soybean breeders are interested in development of high seed protein cultivars and identification of genes that controlling seed storage protein quantity

Research groups have investigated soybean seed protein quantitative trait loci (QTLs). Over one hundred QTLs associated with soybean seed protein have been discovered (www.soybase.org). Using $F_{2:3}$ progenies from a cross between high-protein *G. soja* (PI 468916) and *G. max* (A81-356022), restriction fragment length polymorphism (RFLP) analysis indicated that QTLs for high-protein content were distributed on linkage groups (LGs) E and I (chromosomes (Chrs) 15 and 20) (Diers et al. 1992). However, these QTLs were identified only on Chr 15 with the recombinant inbred line (RIL) population of PI 97100 x Coker 237. The high-protein genotype PI 97100, whose seed protein content was lower than PI 468916, was used in this

population (Lee et al. 1996). Recently, introgression of a high-protein allele on Chr 20 from PI 468916 into A81-356022 increased seed protein about 12 to 30 g kg^{-1} and the QTL region was positioned by fine mapping (Nichols et al. 2006, Sebolt et al. 2000).

Precise mapping for protein QTL region on Chr 20 was performed with other populations. Using the high-seed protein accession *G. max* PI 437088A (>480 g kg^{-1}) as parental line of a recombinant inbred line (RIL) population, a major QTL for protein and oil content flanked by SSR markers Satt239 and Satt496 was identified on Chr 20. Five dominant random amplified polymorphic DNA (RAPD) markers were mapped in the interval between Satt239 and Satt496 and these RAPD markers increased the Satt239-Satt496 interval from 0.85 to 6.63 cM. The QTLs for protein and oil content and yield were mapped near marker OPAW13a, 0.84 cM away from Satt239. This QTL is now designated as *Prot 15-1* in Soybase (http://soybase.org) (Chung et al. 2003). In previous research, we investigated the major QTL that associated with a flanking marker Satt239 on Chr 20 for high-protein content in the Korean cultivar Dangaekkong (480 g kg^{-1} in seed protein) using a RIL population from Benning x Dangaekkong (Park 2002).

Neal isogenic lines (NILs) is immortal population that can be used to investigate the effect of QTL. NIL is generally constructed by successive

backcrosses and marker selections requires a considerable amount of time and labor. In other cases, NILs could be selected from inbred lines that are not completely homozygous also referred to as residual heterozygous line (RHL) or heterozygous inbred family (HIF). For instance, some genetic researches identified genes from NILs developed from RILs. Two sets of NILs were developed from RIL to identify two different resistance genes to sudden death syndrome in soybean (Triwitayakorn et al. 2005). Soybean flowering time and maturity genes were also investigated using the same RHL-derived NILs (Thakare et al. 2010, Watanabe et al. 2009, Yamanaka et al. 2005).

The objective of this research was to obtain NILs for the seed protein QTL flanked by SSR markers Satt239 and Satt496 in two RIL populations, Sinpaldalkong 2 x Dangaekkong and Daewonkong x Dangaekkong. Using these populations, we examined the effect of high seed protein allele on Chr 20 of *G. max* in terms of seed protein content depending on different genetic backgrounds and interactions between genotypes and environments. Additionally, candidate genes for seed protein content within a 5 Mb region of Chr 20, flanked by Satt239 and Satt496 were predicted.

MATERIAL AND METHODS

Plant materials and NIL development

Five RIL populations for RHL selection were assembled by Sinpaldalkong 2 x Dangaekkong, Daewonkong x Dangaekkong, Jinpumkong 2 x Danbaekkong, Pungsanamulkong x Danbaekkong and Ilpumgeomjeongkong x Danbaekkong crosses. At the F_7 generation, four F_7 seeds from RILs were planted in a greenhouse at the University Farm of Seoul National University (SNU) in Suwon for foreground selection of two SSR markers, Satt239 and Satt496 locus. One of the 137 individuals in the Sinpaldalkong 2 x Dangaekkong population (SD) and two from the 195 individuals from the Daewonkong x Dangaekkong populations (DD) were identified as heterozygous around Satt239 and Satt 496 locus at F_7 and segregated at the following F_8 generation.

Cross	Size
Sinpaldalkong 2 x Dangaekkong	55
Daewonkong x Dangaekkong	59
Jinpumkong 2 x Danbaekkong	22
Pungsanamulkong x Danbaekkong	43
Ilpumgeomjeongkong x Danbaekkong	32

Table I-1 Cross of RIL populations for RHL selection and the size of each

lines.

After selection in the F_7 generation, F_8 plants from each F_7 line were classified into three groups as paternal, maternal, and heterozygous according to the genotypes of Satt 239 and Satt 496. Mature F_9 seeds from each parental genotype groups were harvested to evaluate seed protein content. Candidates of NILs from F_9 were selected with consideration of protein contents of seeds and genetic background of plants. NILs were constructed with two plants from each F_9 line that had the significant relationship between genotype of SSR markers and seed protein content based on different homozygous.

Phenotype evaluation

Crude seed protein from F_9-derived progenies were measured in both 2008 and 2009 at the Legume and Oil Crop Research Div., National Institute of Crop Science in Miryang-si. Each line was analyzed four times and the amount of crude protein was determined by the Dumas combustion method according to the Association of Analytical Communities (AOAC International. 1992) using Rapid N Cube (Elementar Analysensysteme GmbH, Hanau). For each sample, 0.1 g was quantified with an analytical balance, Sartorius BS224S (Sartorius AG, Goettingen). After perfect combustion, reduction, purification, and detection, nitrogen content was measured using Rapid N

Software v 3.4.0 (Elementar Analysensysteme). The conversion factor from nitrogen to protein was 6.25. Protein content, experimental error, and interaction between genotype and environment during two years were statistically analyzed according to Gomez and Gomez (1984).

Genomic DNA extraction and SSR marker analysis

Fresh leaves from RILs and selected progenies were sampled at 25 days after emerging (DAE) and used to extract genomic DNAs according to the urea extraction method (Shure at al. 1998). All plant samples were screened by two SSR markers, Satt 239 and Satt 496, based on foreground selection. For background screening of F_9 plants, SSR markers were selected from SoyBase (http://soybase.org) to cover all 20 soybean chromosomes and tested to the three parents cultivars as template using 72 SSR markers. The selected marker sets were applied for screening the genetic background of each population.

The component of PCR reaction mixture in 20 μl of total volume was 0.4 units of *Taq* polymerase (VIVAGEN Co., Sungnam, Korea), 1 X reaction buffer (750 mM Tris-HCl pH 8.5, 200 mM $(NH_4)_2SO_4$, 0.1% Tween 20, 25 mM $MgCl_2$, 0.5% enzyme stabilizer), 0.16 mM of each dNTP, 2 μM of template DNA, and 0.5 μM of each primer. PCR was performed in PTC-100

MJ Thermo cycler (MJ Research, Watertown, MA, USA). The amplified products were loaded on 3.5% acrylamide gel with Triple-Wide Mini-Vertical System CE (TWC 202-33, CBS Scientific Company Inc., Del Mar, CA, USA). Additional confirmation was performed by ABI 3730 automated DNA sequencer (Applied Biosystems, Foster City, CA, USA) and GeneMapper® software v3.7 (Applied Biosystems, Foster City, CA, USA).

New SSR marker development and *in silico* transcript screening

After the genomic sequence was taken from the DOE-JGI *G. max* sequence (http://www.phytozome.net/soybean), primers were designed by 'SSR screening and primer design method' in BatchPrimer3 (http://probes.pw.usda.gov/cgi-bin/batchprimer3/batchprimer3.cgi). Physical positions on Chr 20 with selected SSR markers were confirmed by BLAST search against *G. max* genome sequence in Phytozome. Feasible transcript selection was performed based on transcript sequences and compare annotation of protein sequence using blast P of predicted sequences from soybean transcript database in Phytozome.

RESULTS

Development of NILs and marker analysis

In the process of NIL selection for the major protein QTL on Chr 20, two SSR markers, Satt239 and Satt496, were used for foreground screening of flanking region in F_7, segregating F_8, F_9, and NIL progenies. One of 137 plants from SD (SD-44) and two of 195 from (DD-01 and DD-24) DD RILs were heterozygous at Satt 239 and Satt 496 (Fig. I-1). Subsequently, F_8 individuals were developed from these two RHLs and they showed segregation for Satt239 and Satt496 markers (Fig. I-1). Based on SSR marker analysis of Satt239 and Satt496, two F_8 plants, SD-44-10, DD-01-05 and DD-24-07 were predicted to carry the homozygous high protein allele from Danbaekkong. SD-44-03, DD-01-02 and DD-24-03 had the identical alleles as Sinpaldalkong 2 and Daewonkong at the two SSR loci, respectively. The remaining 13 F_8 plants were heterozygous for Satt239 and Satt496. These two NIL pairs of SD-44-03 / SD-44-10, DD-01-02/DD-01-05 and DD-24-03 / DD-24-07, which were developed from the SD and DD RHLs, were designated $Prot\_low_{SD}$ / $Prot\_high_{SD}$ (NIL_{SD}), $Prot\_low_{DD}1$/ $Prot\_high_{DD}1$ (NIL_{DD1}) and $Prot\_low_{DD}2$/ $Prot\_high_{DD}2$ (NIL_{DD2}) respectively.

No recombination occurred in the flanking region of these two markers

during the subsequent generation. In the screening for selection of NILs from F_8 plants, similarity between high- and low protein NIL in each NILs was usually over 97%. Among selected markers, 72 markers were identical between a high protein lines and a low protein lines.

The genetic backgrounds of one NIL pair from Sinpaldalkong 2 x Danbaekkong (NIL_{SD} pair) and 2 NIL pairs from Daewonkong x Danbaekkong (NIL_{DD1} and NIL_{DD2} pairs) were differed according to the F_7 from which they derived. Marker profiles showed that NIL_{SD} pair had nine pairs of chromosomes with no recombination, eight pairs with one recombination, and three pairs with double recombination. NIL_{DD1} pair comprised nine pairs of chromosome without recombination, six pairs with one recombination, four pairs with double recombination, and one pair with triple recombination. NIL_{DD2} pair consisted of eight chromosome pairs without recombination, six pairs with one recombination, four pairs with double recombination, and two pair with triple recombination.

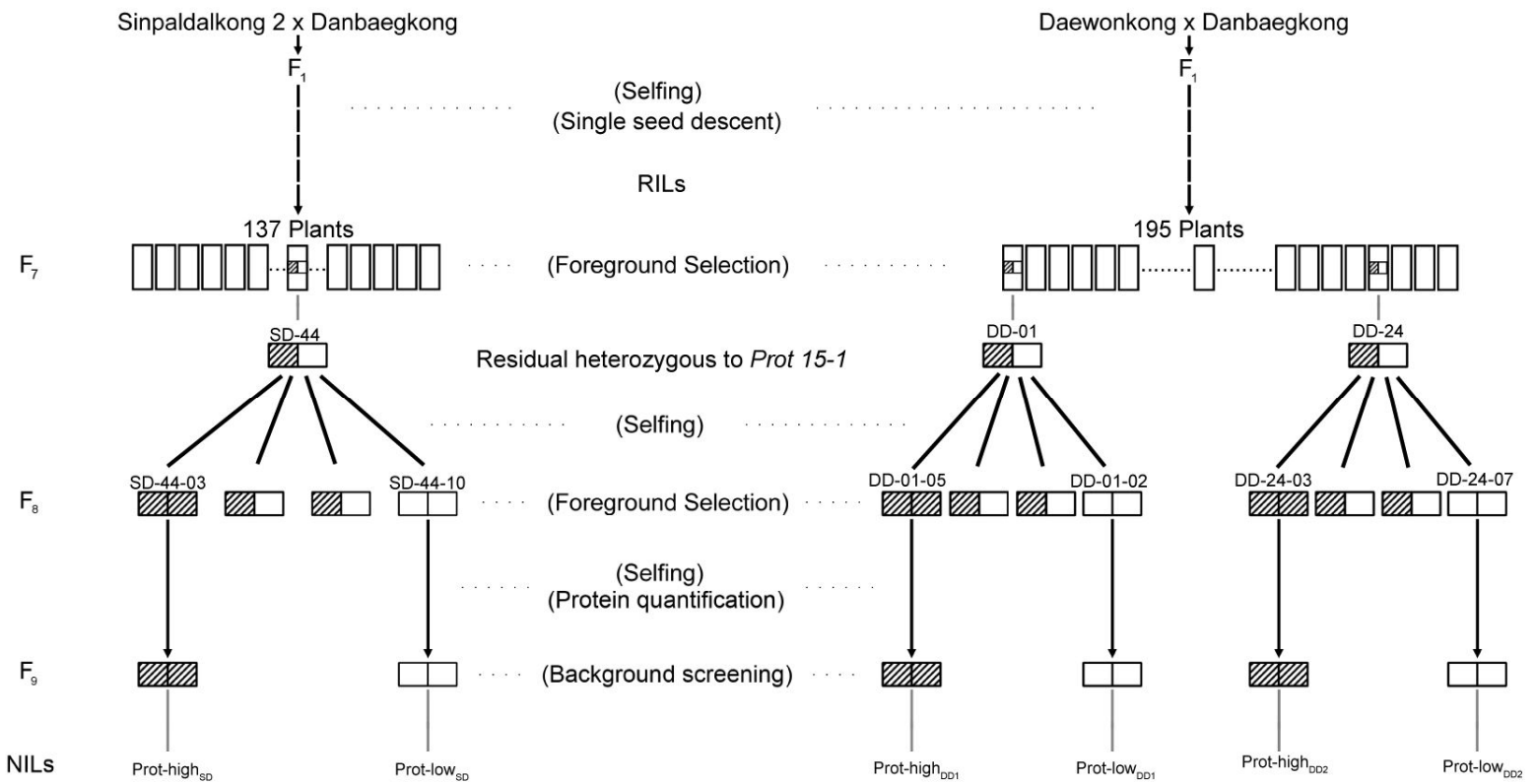

Sinpaldalkong 2 x Danbaegkong
Daewonkong x Danbaegkong
F1
F1
(Selfing)
(Single seed descent)
RILs
137 Plants
195 Plants
F7
(Foreground Selection)
SD-44
DD-01
DD-24
Residual heterozygous to Prot 15-1
(Selfing)
SD-44-03
SD-44-10
DD-01-05
DD-01-02
DD-24-03
DD-24-07
F8
(Foreground Selection)
(Selfing)
(Protein quantification)
F9
(Background screening)
NILs
Prot-high SD
Prot-low SD
Prot-high DD1
Prot-low DD1
Prot-high DD2
Prot-low DD2

Figure I-1 Schematic illustration of NIL construction to introgress high protein alleles of Dangaekkong using foreground and background selection. Foreground screening was performed to select residual heterozygous lines (RHLs) at protein content locus, *Prot 15-1* in two F_7 RIL populations of Sinpaldalkong 2 x Dangaekkong and Daewonkong x Dangaekkong using two flanking SSR markers, Satt239 and Satt496. Only three RHLs that showed genotypic segregation at the *Prot 15-1* locus in the advanced generation are shown in this figure. NILs were named by adding each first character of parental cultivar names: SD, Sinpaldalkong 2 x Dangaekkong; DD, Daewonkong x Dangaekkong. A striped box represents high protein alleles from Dangaekkong and a blank box represents low protein alleles from Sinpaldalkong 2 or Daewonkong.

Phenotype evaluation for protein content

Protein content in three parents, Sinpaldalkong 2, Daewonkong, and Dangaekkong was 407.2, 408.5, and 484.3 g kg^{-1} in 2009, respectively. Seed protein content according to Satt239 and Satt496 genotype of NILs was maintained for two years (Table I-2). In the first phenotype quantification of $F_{7:9}$ seeds, seed protein content was defined as associating with segregating SSR markers, Satt239 and Satt496. The highest protein content was shown in $Prot\_high_{SD}$. Differences according to genotypes between high- and low protein NILs were 42.2 g kg^{-1} at NIL_{SD} pair, 20.7 g kg^{-1} in NIL_{DD1} pair, and 29.8 g kg^{-1} at NIL_{DD2} pair. In the second year, seed protein differences increased to 44.8 g kg^{-1} at NIL_{SD} pair, 41.3 g kg^{-1} NIL_{DD1} pair, but decreased to 26.9 g kg^{-1} at NIL_{DD2} pair. Therefore, the effect of genotype of Satt239 and Satt496 to seed protein content was significant at the 5% level in three NILs.

Seed protein contents in three Prot_low NILs were decreased from 2008 to 2009 in common. Protein contents were decreased in the 2009 by 4.1 g kg^{-1} at $Prot\_high_{SD}$ and 9.5 g kg^{-1} at $Prot\_high_{DD2}$. However, the seed protein content of $Prot\_high_{DD1}$ was increased by 20.3 g kg^{-1} and consequently showed a significant interaction between genotype and environment. From this environmental effect, NIL_{DD2} pairs are identified only

NIL pair from Daewonkong x Danbaekkong that named as Prot_high$_{DD}$ and Prot_low$_{DD}$ (Table I-2).

Table I-2 Statistical analysis of seed protein contents of NILs about the genotype factor and interaction between genotype and environmental change by year.

Cross	Line	Seed protein content (g kg^{-1})		Everage (g kg^{-1}) [a]
		2008	2009	
Sinpaldalkong 2 x Danbaekkong (SD)	Prot_high$_{SD}$	486.0	481.9	483.9
	Prot_low$_{SD}$	443.8	437.1	440.5
Daewonkong x Danbaekkong (DD)	Prot_high$_{DD1}$	455.3	475.6	455.3[b]
	Prot_low$_{DD1}$	434.6	434.3	434.6
	Prot_high$_{DD2}$	470.5	461.0	465.8
	Prot_low$_{DD2}$	440.7	434.1	437.4

[a] significant between genotype and seed protein at 5% level

[b] significant genotype and environment at 1% level

Genetic composition of NILs on chromosome 20

We were able to identify genetic composition using SSR markers on Chr 20. In order to dissect the genotype of interval region between Satt239 and Satt496, we developed SSR markers between two markers (Fig. I-3). To develop new SSR markers, repeated sequences between Satt239 and Satt496 were identified and screened. Screening of microsatellite repeats was performed using Phytozome soybean genome sequence of Williams 82. After the design of primers for 29 putative SSRs between Satt239 and Satt496 from 2.4 Mb by BatchPrimer3, 16 primer sets were able to be amplified when DNA from Sinpaldalkong, Daewonkong, and Dangaekkong was used as template. Average marker interval was approximately 126 Kb and four out of the sixteen SSR markers showed polymorphism among three parents of NILs. Positions of these markers were 24.6, 24.7, 25.0 and 26.1 Mb on Chr 20, and repeat motif of three out of the four markers was $(AT)_n$ (Table I-3). Among four markers, AT4 and AT6 were detected at the same sites with BARCSOYSSR_20_0554 and BARCSOYSSR_20_0559 (Song et al. 2010) and the other two markers, ATT1 and AT13, were our SSR markers for fine mapping.

The segregation analysis of the SSR markers revealed that the Danbaekkong genomic region that was introgressed into the high-protein

NILs spanned from Satt239 (24.13 Mb) to Satt496 (26.50 Mb) and Satt354 (33.43 Mb) in Prot_high$_{DD}$ (previously NIL$_{DD2}$) and Prot_high$_{SD}$, respectively (Figure I-2). No segregation for the SSR markers Satt587 (3.73 Mb) in NIL$_{SD}$ and Satt354 (33.43 Mb) in NIL$_{DD}$ surrounding the interval of Satt239-Satt496 was observed between the low- and high- protein lines of NIL pairs (Figure I-2), indicating that the genetic background of Sinpalkong 2 and Daewonkong was fixed at these regions on Chr 20. However, Satt687 was 20.4 Mb away from Satt239 and the distance between Satt496 and Satt354 was 6.9 Mb.

Table I-3 List of SSR markers developed from the genome sequence between Satt239 and Satt496 on Chr 20.

Primer No.	Position	Expected PCR length	Forward Primer (5'-3')	Reverse Primer (5'-3')	Motif	poly morp hism	Remark[a]
AT1	24266894-24266675	220	ACAAACAACGCACCATTACT	CTCCTCCATAGCTAGTTCCATA	(AT)27	No	
AT2	24295926-24296114	189	ATCTTCCTCGAGTATGTTCCTA	TACCTGCTATGTTTTCAGACCT	(AT)30	No	BARCSOYSSR_20_0544
AT3	24347254-24347454	201	CGAGCCTCTAATTACAAACAG T	CTTCCGTATGAAGGTATTCG	(AT)28	No	BARCSOYSSR_20_0546
AT4	24617375-24617601	227	GCTTGAAATGTGTCTCTCTGTA	AGCTCATTTCCATGTGCTAT	(AT)34	Yes	BARCSOYSSR_20_0554
ATT1	24656864-24657118	225	GAGGACCATTTATTTTGCAAG CAC	CCAGTTTGCATTGATCATCATCT C	(ATT)15aat(ATT)9	Yes	
AT5	24803287-24803612	328	GCACTTGATTTGCTCAATGTTA CAG	GAAAATGGACAAAGTGGTGC	(AT)33	No	BARCSOYSSR_20_0558
AT6	24994083-24994281	199	CCACGAAAGATCAACAAGAT	ACACAAGACATGAGCAAGTTTC	(AT)30	Yes	BARCSOYSSR_20_0559
AT7	25090880-25091095	216	CCCATGTAGATCAAAGAATGA C	ACATTAGATGTTTGGGCATC	(AT)26	No	
AT8	25189083-25189256	174	GTCATTTCCATATACCTCTC	CTACCTTAGAGTTCCATGTCC	(AT)34	No	BARCSOYSSR_20_0560
AT9	25226423-25226793	371	GACACATTCTACTTTGCACTAC TC	CCTTGTTCTAATTTTCCCATGAT G	(AT)74	No	
AC1	25737546-25737781	236	GGTGCAAATATATGCCCTAGT	CCACACTTTTCTTCACAACA	(AC)25	No	BARCSOYSSR_20_0574
AT10	25772809-25772982	174	AATGCCTTAAATGACCCTTC	TGTGCAAACATGACTTCAAC	(AT)22	No	
AT11	25827590-25827791	202	CGCTACGATATCACCACTCTA	GTATTCGGTTAAGCATTGGT	(AT)29	No	
AT12	25894897-25895211	315	CTCTCTTCAAGATGGATAACCT CC	CCTAGGCGAATAGTTGTGTG	(AT)77	No	
AT13	26087651-26087867	220	CGGTTATGGGTTACTCTTCG	CCTGAAGATTGAGTGGAATGAAC C	(AT)82	Y	
AT14	26296626-26296985	443	GAGAGAAAAAAACACACACCC GAC	GCGAACACCCCTAATTTACAC	(AT)66	N	

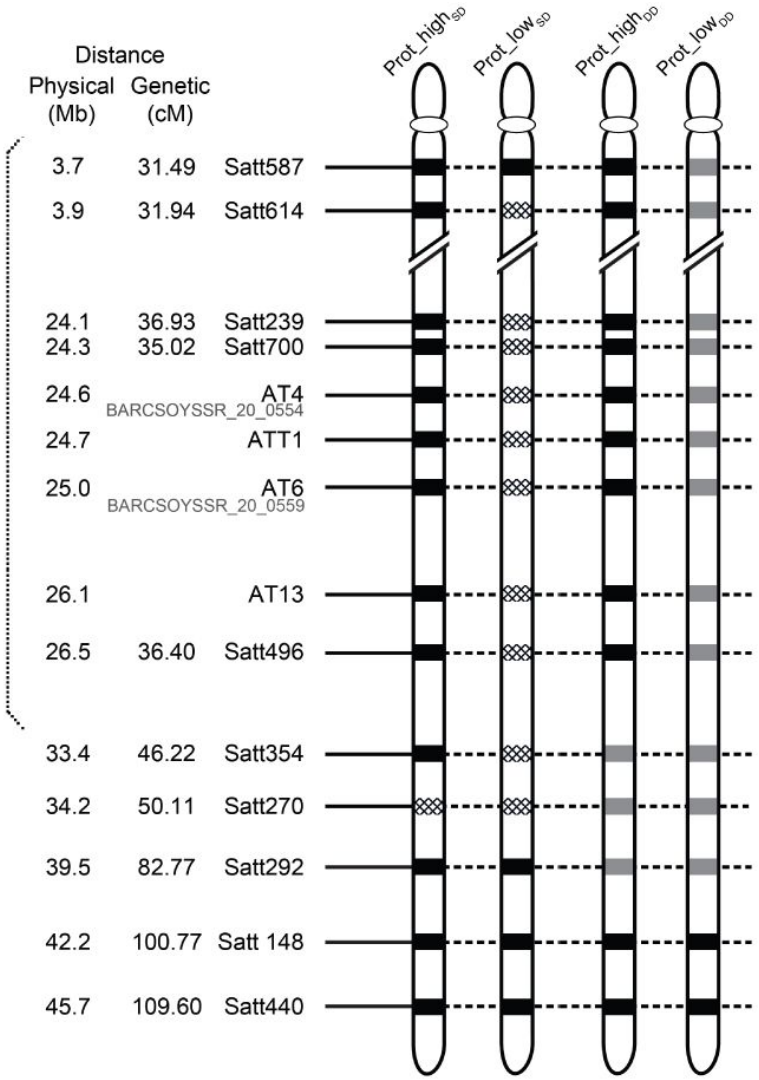


Figure I-2 Genetic composition on Chr 20 of three NIL pairs. Genotypes of Sinpaldalkong 2, Daewonkong and Danbaekkong in each locus were illustrated as crossed lines, gray and black boxes. Pericentromeric region is indicated by a dotted broken line at the leftmost side.

DISCUSSION

Dangaekkong has higher seed protein content than other existing Korean cultivars and its high protein trait was known as originated from *G.max*. Dangaekkong was developed from the cross between Tousan 69 and the high protein introduction line D76-8070 (Kim et al. 1996). D76-8070 breeding line has 495 g kg^{-1}seed protein content and is derived from a multiple crossing program (Hartwig 1990). The analysis of the pedigree of D76-8070 suggests that high protein QTL in Danbeagkong might have been introgressed from two high protein *G. max* accessions, Sioux (522 g kg^{-1}) and PI 96983 (510 g kg^{-1}). Two SSR markers, Satt239 and Satt496 were identified as flanking markers of high protein QTL from Dangaekkong in our previous research. The same flanking SSR markers were already mapped in other *G. max* accession and the QTL was named as *Prot 15-1* (Chung et al. 2003).

RHL selection about major protein QTL on Chr 20 was used to dissect genetic region and measure genetic effect of major seed protein content gene in our research. RHL selection ratios from our breeding processes were similar with the selection ratios from other previous research (Watanabe et al. 2011). Then our research is able to support that RIL populations from F_6 to F_8 generation according to population size could be suitable to detect one or two RHLs from 137 and 195 plants for a target

locus. Therefore, we successfully confirmed the utility of two generations of RHLs from two F_7 RIL populations which saved time and labor and allowed us to select NIL pairs with high genetic background identity.

The prior condition of original RIL population is important to selection result of RHL to NIL for fine mapping of target locus. Distributions of gene, genetic markers, recombination frequency of soybean genome are concentrated on distal regions (Ott 2011). This phenomenon was a cause of limitation of separation protein content gene from other locus to do fine mapping in our research. The position of major protein QTL region surrounding Satt239 and Satt496 located on pericentromeric region on Chr 20 and genes were seldom existed. Then the recombination event between Satt239 and Satt496 was not occurred during RIL breeding and NIL selection (Fig. I-2). Nevertheless of this limitation, selected NILs proved that the introgression on Chr 20 from Danbaekkong is highly associated with phenotypic differences in seed protein regardless of genetic backgrounds. The protein difference between $Prot\_high_{SD}$ and $Prot\_low_{SD}$ was more than 40 g $kg^{-1}$and the protein content of $Prot\_high_{SD}$ is similar with its high protein parent, Danbaekkong. Therefore, RHL selection about soybean seed protein content QTL has advantages to breeding several NILs that have various genetic backgrounds at one time.

$Prot\_high_{DD1}$ seed protein content increased 20 g kg-1 in 2009, unlike the other NILs decreased. Therefore, the response to environmental change

in 2008 and 2009 was analyzed only in NIL_{DD1} pairs. Seed protein content is known to be affected by changes in several environmental factors such as temperature and water stress and soil minerals (Carrera et al. 2009, Kumar et al. 2006, Rotundo and Westgate 2009, Yaklich et al. 2002, Laszlo 1994). Minerals in soil during two years were not significantly different, we assume that climate changes were major cause of seed filling difference. According to previous studies, responses of protein, oil and other residues to water stress bring a result of significant increase of seed protein concentration. In monthly weather reports in 2008 and 2009, a few localized torrential downpours and following high temperature changes came in Suwon on July and August in 2009. Thus differential protein filling of $Prot\_high_{DD1}$ inferred that the genetic background of $Prot\_high_{DD1}$ is sensitive to environment changes caused by water stress and temperature change during seed filling stages.

We have verified the effectiveness of RHL strategy by using specific genes or markers from existing F_7 generation RILs. This result in building three NIL pairs having various genetic backgrounds and high genetic identity between high and low protein genotypes. Three NIL pairs confirmed the effect of the seed protein QTL region that originated from Danbaekkong using selection with flanking markers Satt239 and Satt496. These NILs will be effective materials not only for identification of the major QTL and the candidate gene but also for investigating epistasis using their different

genetic backgrounds. Further studies using these NILs should provide a better understanding about transcription and translation changes during seed development through genome resequencing, transcriptome analysis, proteome analysis.

REFERENCES

Bolon YT, Joseph B, Cannon SB, Graham MA, Diers BW, Farmer AD, May GD, Muehlbauer GJ, Specht JE, Tu ZJ, Weeks N, Xu WW, Shoemaker RC, Vance CP (2010) Complementary genetic and genomic approaches help characterize the linkage group I seed protein QTL in soybean. BMC Plant Biol 10(1): 41

Brummer EC, Graef GL, Orf J, Wilcox JR, Shoemaker RC (1997) Mapping QTL for seed protein and oil content in eight soybean populations. Crop Sci 37(2): 370-378

Carrera C, Martinez MJ, Dardanelli J, Balzarini M (2009) Water deficit effect on the relationship between temperature during the seed fill period and soybean seed oil and protein concentrations. Crop Sci 49: 990-998

Chang C-I, Lee J-K, Ku K-H, Kim W-J (1990) Comparisom of soybean varieties for yield, chemical and sensory properties of soybean curds. Korean J Food Sci Technol 22: 439-444

Chung J, Babka HL, Graef GL, Staswick PE, Lee DJ, Cregan PB, Shoemaker RC, Specht JE (2003) The seed protein, oil, and yield QTL on soybean linkage group I. Crop Sci 43:1053-1067

Diers BW, Keim P, Fehr WR, Shoemaker RC (1992) RFLP analysis of

soybean seed protein and oil content. Theor Appl Genet 83: 608-612

Glover KD, Wang D, Arelli PR, Carlson SR, Cianzio SR, Diers BW (2004) Near isogenic lines confirm a soybean cyst nematode resistance gene from PI 88788 on linkage group J. Crop Sci 44:936-941

Gomez KA, Gomez AA (1984) Analysis of data from a series of experiments, In Statistical Procedures for Agriculture Research, Ed 2, John Willey & Sons, Hoboken, pp 316-356

Grigg D (1995) The pattern of world protein consumption. Geoforum 26: 1-17

Haley SD, Afanador LK, Miklas PN, Stavely JR, Kelly JD (1994) Heterogeneous inbred populations are useful as sources of near-isogenic lines for RAPD marker localization. Theor Appl Genet 88: 337-342

Hunts KE (1973) Economics of protein production. In JGW Jones (ed) The Biological Efficiency of Protein Production, Cambridge University Press, London, pp 45-68

Jin J, Huang W, Gao JP, Yang J, Shi M et al (2008) Genetic control of rice plant architecture under domestication. Nat Genet 40: 1365-1369

Kazi S, Shultz J, Afzal J, Hashmi R, Jasim M, Bond J, Arelli PR, Lightfoot DA (2010) Iso-lines and inbred-lines confirmed loci that underlie resistance

from cultivar 'Hartwig' to three soybean cyst nematode populations. Theor Appl Genet 120:633-644

Keurentjes JJB, Bentsink L, Alonso-Blanco C, Hanhart CJ, Vries HBD, Effgen S, Vreugdenhil D, Koornneef M (2007) Development of a near-isogenic line population of *Arabidopsis thaliana* and comparison of mapping power with a recombinant inbred line population. Genetics 175:891-905

Kumar V, Rani A, Solanki S, Hussain SM (2006) Influence of growing environment on the biochemical composition and physical characteristics of soybean seed. J Food Comp Anal 19: 188-195

Laszlo JA (1994) Changes in soybean fruit Ca^{2+} (Sr^{2+}) and K^{+} (Rb^{+}) transport ability during development. Plant Physiol 104: 937-944

Lee S-H, Bailey MA, Mian MAR, Carter TE, Shipe ER, Ashley DA, Parrott WA, Hussey RS, Boerma HR (1996) RFLP loci associated with soybean seed protein and oil content across populations and locations. Theor Appl Genet 93:649-657

Moon JK, Jeong SC, Van K, Maroof MAS, Lee S-H (2009) Marker-assisted identification of resistance genes to soybean mosaic virus in soybean lines. Euphytica 169: 375-385

Nichols DM, Glover KD, Carlson SR, Specht JE, Diers BW (2006) Fine

mapping of a seed protein QTL on soybean linkage group I and its correlated effects on agronomic traits. Crop Sci 46: 834-839

Ott A, Trautshold B, Sandhu D (2011) Using microsatellites to understand the physical distribution of recombination on soybean chromosome. PLoS ONE 6(7): e22306

Park Y-H (2002) SSR mapping for soybean seed protein and oil concentration across populations. Master. thesis. Seoul National Univ., Seoul

Rotundo JL, Westgate ME (2009) Meta-analysis of environmental effects on soybean seed composition. Field Crops Res 110: 147-156

Sebolt AM, Shoemaker RC, Diers BW (2000) Analysis of a quantitative trait locus allele from wild soybean that increases seed protein concentration in soybean. Crop Sci 40: 1438-1444

Singh P, Kumar R, Sabapathy SN, Bawa AS (2008) Functional and edible uses of soy protein products. Comp Rev Food Sci Food Safety 7: 14-28

Song Q, Jia G, Zhu Y, Grant D, Nelson RT, Hwang E-Y, Hyten DL, Cregan PB (2010) Abundance of SSR motifs and development of candidate polymorphic SSR markers (BARCSOYSSR_1. 0) in soybean. Crop Sci 50:1950-1960

Specht JE, Chase K, Macrander M, Graef GL, Chung J, Markwell JP,

Germann M, Orf JH, Lark KG (2001) Soybean response to water. Crop Sci 41(2): 493-509.

Thakare D, Kumudini S, Dinkins RD (2010) Expression of flowering-time genes in soybean E1 near-isogenic lines under short and long day conditions. Planta 231: 951-963

Tajuddin T, Watanabe S, Yamanaka N, Harada K (2003) Analysis of quantitative trait loci for protein and lipid contents in soybean seeds using recombinant inbred lines. Breed Sci 53(2): 133-140

Triwitayakorn K, Njiti VN, Iqbal MJ, Yaegashi S, Town C, Lightfoot DA (2005) Genomic analysis of a region encompassing *QRfs1* and *QRfs2*: genes that underlie soybean resistance to sudden death syndrome. Genome 48: 125-138

Tuinstra MR, Ejeta G, Goldsbrough PB (1997) Heterogeneous inbred family (HIF) analysis: a method for developing near-isogenic lines that differ at quantitative trait loci. Theor Appl Genet 95: 1005-1011

Wang HL, Swain EW, Kwolek WF, Fehr WR (1983) Effect of soybean varieties on the yield and quality of tofu. Cereal Chem 60: 245-248

Watanabe S, Hideshima R, Xia ZJ, Tsubokura Y, Sato S et al. 2009. Map-based cloning of the gene associated With the soybean maturity locus *E3*. Genetics 182: 1251-1262

Watanabe S, Xia Z, Hideshima R, Tsubokura Y, Sato S, Yamanaka N, Takahashi R, Anai T, Tabata S, Kitamura K, Harada K (2011) A map-based cloning strategy employing a residual heterozygous line reveals that the *GIGANTEA* gene is involved in soybean maturity and flowering. Genetics 188: 395-407

Wilcox JR, Cavins JF (1995) Backcrossing high Seed protein to a soybean cultivar. Crop Sci 35: 1036-1041

Yaklich RW, Vinyard B, Camp M, Douglass S (2002) Analysis of seed protein and oil from soybean Northern and Southern Region uniform tests. Crop Sci 42: 1504-1515

Yamanaka N, Watanabe S, Toda K, Hayashi M, Fuchigami H, Takahashi R, Harada K (2005) Fine mapping of the FT1 locus for soybean flowering time using a residual heterozygous line derived from a recombinant inbred line. Theor Appl Genet 110: 634-639

CHAPTER II

Nucleotide variations in the QTL region affecting seed protein content in NILs carrying high and low seed protein alleles on chromosome 20 in soybean (*Glycine max* [L.] Merr.)

ABSTRACT

Soybean is a valuable crop, as its seeds are rich in protein. A major quantitative trait locus (QTL) for soybean seed protein content has previously been identified in the simple sequence repeat (SSR) marker interval of Satt239-Satt496 on chromosome (Chr) 20. In the previous chapter, we selected three residual heterozygous lines heterozygous in the Satt239-Satt496 region from two recombinant inbred line populations of Sinpaldalkong 2 x Danbaekkong and Daewonkong x Danbaekkong soybean crosses to develop two sets of near isogenic lines (NILs) with different seed

protein content. The maternal parent cultivar Danbaekkong, carrying a high protein content allele, showed high seed protein content. In the genomic region harboring the introgressed Danbaekkong segment on Chr 20, we identified nucleotide differences between low- and high- protein NILs through whole genome sequencing using Illumina HiSeq 2000. The 66 non-synonymous single nucleotide polymorphisms and one frameshift that overlapped at the same positions in both of the NIL pairs were predicted to affect 30 genes in the high protein QTL region. Among these genes, five genes were selected as candidates with possible involvement in the accumulation of seed protein, showing gene expression during seed development. Two genes, calcium-dependent protein kinase and exocyst subunit exo70 family protein, are more likely involved in see protein accumulation in soybean.

INTRODUCTION

Intensive quantitative trait locus (QTL) mapping has been used to identify more than 100 QTLs affecting seed protein content in soybean (http://soybase.org). However, the molecular mechanism controlling seed protein accumulation is not fully understood. On Chr 20, 14 QTLs for seed protein content has been reported near the genomic region from 30 to 40 cM (http://soybase.org), explaining the high phenotypic variation of seed protein content independent of environmental conditions (Brummer et al. 1997, Chung et al. 2003, Diers et al. 1992, Nichols et al. 2006, Sebolt et al. 2000, Tajuddin et al. 2003). Among these QTLs, the simple sequence repeat (SSR) marker Satt239 was identified to be linked with high protein content using a recombinant inbred line (RIL) population of Moshidou Gong 503 (high protein, 463 g kg^{-1}) x Misuzudaizu (low protein, 360 g kg^{-1}) (Tajuddin et al. 2003). Another RIL population produced from a cross between PI 437088A (high protein, 480 g kg^{-1}) and Asgrow A3733 (low protein, 420 g kg^{-1}) showed the seed protein QTL *Prot 15-1* in the interval between Satt239 (36.9 cM in the composite linkage map of soybase) and Satt496 (36.4 cM) with >10 of LOD (Chung et al. 2003). Furthermore, fine mapping using near isogenic lines (NILs) developed from backcrosses of high protein *Glycine soja* PI 468916 into *G. max* accession A81-356022 revealed that the QTL for seed protein content was located between Satt239 and Satt496, even

though Satt239 was not segregated in these high- and low protein NILs (Bolon et al. 2010, Diers et al. 1992, Nichols et al. 2006, Sebolt et al. 2000). Through comparing transcript profiles between high protein and low protein NILs using Affymetrix® Soy GeneChip microarray and whole transcriptome sequencing, 13 candidate genes located in the seed protein QTL region on Chr 20 were found to show differential expressions between the NILs, including a Mov34-1 regulatory protein, a heat shock protein, and an ATP synthase (Bolon et al. 2010). The functional roles of these candidate genes in seed protein accumulation are not yet identified.

Recently, next-generation sequencing (NGS) technology has been used to conduct whole genome sequencing (WGS) in many crop species (Boetzer et al. 2011, Gao et al. 2012). Due to the reliability of NGS data, along with the high depth and low cost of NGS sequencing, the use of NGS has been extended from WGS of a typical model crop species to the resequencing of various populations (Boetzer et al. 2011, Gao et al. 2012). Resequencing can be employed to study population genetics and comparative genomics at the genome level (Mayer et al. 2011). Resequencing is also a powerful technology for developing numerous molecular markers, such as single nucleotide polymorphisms (SNPs), to fractionate the QTL region of an important agronomic trait and to construct high-density genetic maps (Huang et al. 2009, Yu et al. 2011).

In the present study, we sequenced the whole genomes of two NIL pairs and three parental genotypes to compare the sequences in the seed protein QTL region from Satt239 and Satt496 on Chr 20. This sequence analysis identified functional SNPs in several genes located within the QTL region between the high- and low protein lines, including protein kinase and exocyst complex subunit genes which likely play a role in seed protein accumulation.

MATERIAL AND METHODS

DNA extraction and SSR marker analysis

Genomic DNA was isolated from fresh soybean leaves according to the method of Shure et al. The quantity and quality of the DNA were determined using a NanoDrop ND-1000 Spectrophotometer (NanoDrop Technologies, Wilmington, DE, USA). Two SSR markers, Satt239 and Satt496, were first used for foreground selection relating to seed protein content. In addition, 72 SSR markers covering all 20 soybean chromosomes were employed to investigate the homogenous genetic background between the low protein and high protein NILs. To confirm the low protein and high protein NILs, additional SSR markers that were located in the Satt239 and Satt496 interval were developed. Repeated sequences between Satt239 and Satt496 on Chr 20 were obtained from Phytozome (http://www.phytozome.net/soybean.php), and specific primers were designed to amplify the repeated sequences using BatchPrimer3 (http://probes.pw.usda.gov/batchprimer3/index.html).

The PCR reaction mixture contained 0.4 units of *Taq* polymerase (Vivagen Co., Sungnam, Korea), 1 X reaction buffer (750 mM Tris-HCl pH 8.5, 200 mM $(NH_4)_2SO_4$, 0.1% Tween 20, 25 mM $MgCl_2$, 0.5% enzyme

stabilizer), 0.16 mM of each dNTP, 2 μM of template DNA, and 0.5 μM of each primer in a total volume of 20 μl. PCR was performed in a PTC-100 MJ Thermo cycler (MJ Research, Watertown, MA, USA). The amplified products were resolved using an ABI 3730 automated DNA sequencer (Applied Biosystems, Foster City, CA, USA) and GeneMapper® software v3.7 (Applied Biosystems). Additional repeated markers were analyzed on a 3.5% acrylamide gel in a Triple-Wide Mini-Vertical System CE (TWC 202-33, CBS Scientific Company Inc., Del Mar, CA, USA).

Analysis of sequence variations

The developed NILs ($Prot\_low_{SD}$ / $Prot\_high_{SD}$, and $Prot\_low_{DD}$ / $Prot\_high_{DD}$) and their parental soybean genotypes (Danbaekkong, Sinpaldalkong 2, and Daewonkong) were resequenced using Illumina HiSeq (Illumina, Co., San Diego, CA, USA). The adapter sequences of raw reads were removed, and the low-quality end sequences were trimmed out, with a quality value threshold of 20. To retrieve the genotype variants in the NIL set, the reads of NILs and cultivars were mapped against the reference genome Williams 82 (http://www.phytozome.net/soybean.php) using BWA software (http://bio-bwa.sourceforge.net) (Li and Durbin 2009). The genotypes, which were defined by the base positions of the reference genome, were

compared, and the positions that exhibited polymorphism among NIL pairs were retrieved, using SAMTools software (http://samtools.sourceforge.net/index.shtml) (Li et al. 2009). The genotypes of each base position were determined at a read depth ranging from 5 to 100 to resolve false genotype calling derived by sequencing errors and genome duplication. Chromosomal locations of the detected SNPs were assigned according to the reference soybean genome (http://www.phytozome.net/soybean.php).

RESULTS

Nucleotide variations in the high protein QTL region between low- and high-protein NILs

To perform fine mapping of the seed protein QTL, we investigated sequence differences between low- and high-protein NILs in the 20 to 31 Mb genomic region, including additional 4 Mb from each marker of Satt239 and Satt496 on Chr 20 through WGS using Illumina HiSeq 2000 (Table II-1). A total of 322 predicted gene IDs were detected in this genomic region, based on the Glyma1.1 gene set from Phytozome (http://www.phytozome.net/soybean.php). We produced an average of 179 million 100 bp paired reads from each of the seven soybean genotypes including four NILs and three parental cultivars, Danbaekkong, Sinpaldalkong 2, and Daewonkong (Table II-2). These reads were aligned to the *G. max* reference genome to identify the SNPs and small insertions/deletions (indels) in the 11 Mb genomic sequences on Chr 20 between low- and high-protein NILs. More than 92% of the Illumina HiSeq reads were properly mapped on the reference sequence, with a mapping depth of each genome (of seven soybean genotypes) ranging from 15- to 20-fold.

Table II-1 Summary of nucleotide differences between low- and high-protein lines of two NIL pairs from 20 Mb to 31 Mb on Chr 20

Varient type	No. of variants	Nongenic variants	1 kb upstream[b] (no. of genes)	Genic[a]							
						CDS[c] (no. of genes)					
				Total	5' UTR	Synonymous	Non-synonymous	Frameshift	Non-frameshift	3' UTR	Intron
SNP (1 bp)	13,403	12,627	225 (89)	551 (111)	29	30(24)	66(31)	NA	NA	16	410
Indel	985	890	34 (28)	61 (45)	5	NA	NA	1 (1)	1 (1)	1	53

[a]CDS, coding sequence; NA, not applicable; UTR, untranslated region.

[b]One kilobase upstream region of transcription start site was considered to be responsible for transcription regulation.

[c]All variants of high protein lines were determined using low protein lines as references.

The predicted SNPs and indels between the low- and high protein lines of two NIL pairs were subjected to additional filtering steps by comparing the alleles of the corresponding parents at the same positions. The NIL_{SD} and NIL_{DD} pairs had a total of 21,452 and 19,183 SNPs between low- and high protein lines, respectively. Among these, 13,403 SNPs were present at the same positions in two NIL pairs (Table II-1); the non-genic and genic regions harbored 12,627 and 551 SNPs, respectively, and 225 SNPs were located at 1 Kb upstream of the transcription start sites of 89 genes between low- and high protein NILs. The noncoding regions contained 455 SNPs, with 45 in 5' and 3' untranslated regions (UTRs) and 410 in introns. A total of 96 SNPs, including 30 synonymous and 66 non-synonymous SNPs, were located at the coding regions. The 20 to 31 Mb section surrounding the high protein QTL on Chr 20 contained 2,112 indels in the NIL_{SD} pair and 1,483 in the NIL_{DD} pairs. Of these indels, 985 were located at the same loci (Table II-1). A total of 61 indels were positioned within the genic boundaries, and only one indel in coding sequences appeared to be a putative frameshift mutation. A total of 31 genes were predicted to be functionally affected by 66 non-synonymous SNPs and one frameshift indel (Table II-1). Among these genes, twenty-three genes were involved in chromosome maintenance, nucleotide metabolism or disease resistance, such as helicase, RNase, endonuclease and NBS-LRR. Among the remaining seven genes, five exhibited transcript accumulation during seed development based on

soybean RNA-seq database SoyKB (Joshi et al. 2012) (Figure 1 and Table 3). These five genes included DCD (Development and Cell Death) domain protein (Glyma20g16100), calcium-dependent protein kinase 1 (Glyma20g17020), Exocyst subunit exo70 family protein (Glyma20g17500), double Clp-N motif-containing P-loop nucleoside triphosphate hydrolases superfamily protein (Glyma20g17560), MOS4-associated complex 3 (Glyma20g21330). The high protein lines (Prot_high$_{SD}$ and Pro-high$_{DD}$) had missense mutations in these five genes that caused amino acid substitutions compared with the low protein lines (Prot_low$_{SD}$ and Prot_low$_{DD}$; Table 3). The non-synonymous SNPs of CDPK and Exo70 between low- and high-protein lines were predicted to induce partial protein structure change (Figure II-2). CDPK showed the formation of beta-sheet on serine/threonine kinase domain (115-373) was partially blocked by amino acid change from 42Val to Ala. .

Table II-3 Non-synonymous SNPs and their amino acid changes between the low- and high protein NILs in the high protein QTL region on Chr 20.

Gene description by homology to *Arabidopsis thaliana*	Location	Codon change		Amino acid change (side chain polarity)	Gene expression in developing seeds[a]
		Prot_low	Prot_high		
bidirectional amino acid transporter	20,364,412..20,365,565	GCT	GTT	A74V (Both nonpolar)	no
DCD domain protein	22,282,845..22,286,101	ATT	CTT	I254L (Nonpolar)	yes
Calcium dependent protein kinase	23,983,441..23,991,,227	GTA	GCA	V42A (Nonpolar)	yes
S-adenosyl-L-methionine-dependent methyltransferases superfamily protein	24,439,521..24,440,457	AGT	AAT	T201I (Neutral to nonpolar	no
		CAG	CAT	L186M (Neutral to nonpolar)	
		CAT	TAT	H132Y (Basic to neutral)	
Exocyst subunit exo70 family protein	24,529,081..24,561,960	CGG	TGG	R146W (Basic to nonpolar)	yes
		ACA	GCA	T169A (Neutral to nonpolar)	
Double Clp-N motif-containing P-loop nucleoside triphosphate hydrolases superfamily protein	24,613,557..24,616,821	CAC	CGC	H454R (Basic)	yes
Nodulin MtN3 family protein	29,989,773..29,991,402	CAT	AAT	M1I (both nonpolar)	no
MOS4-associated complex 3	30,528,612..30,537,544	CGG	CAG	R16Q (Basic to neutral)	yes

[a] RNA-seq data were obtained from http://soybase.org/soyseq/

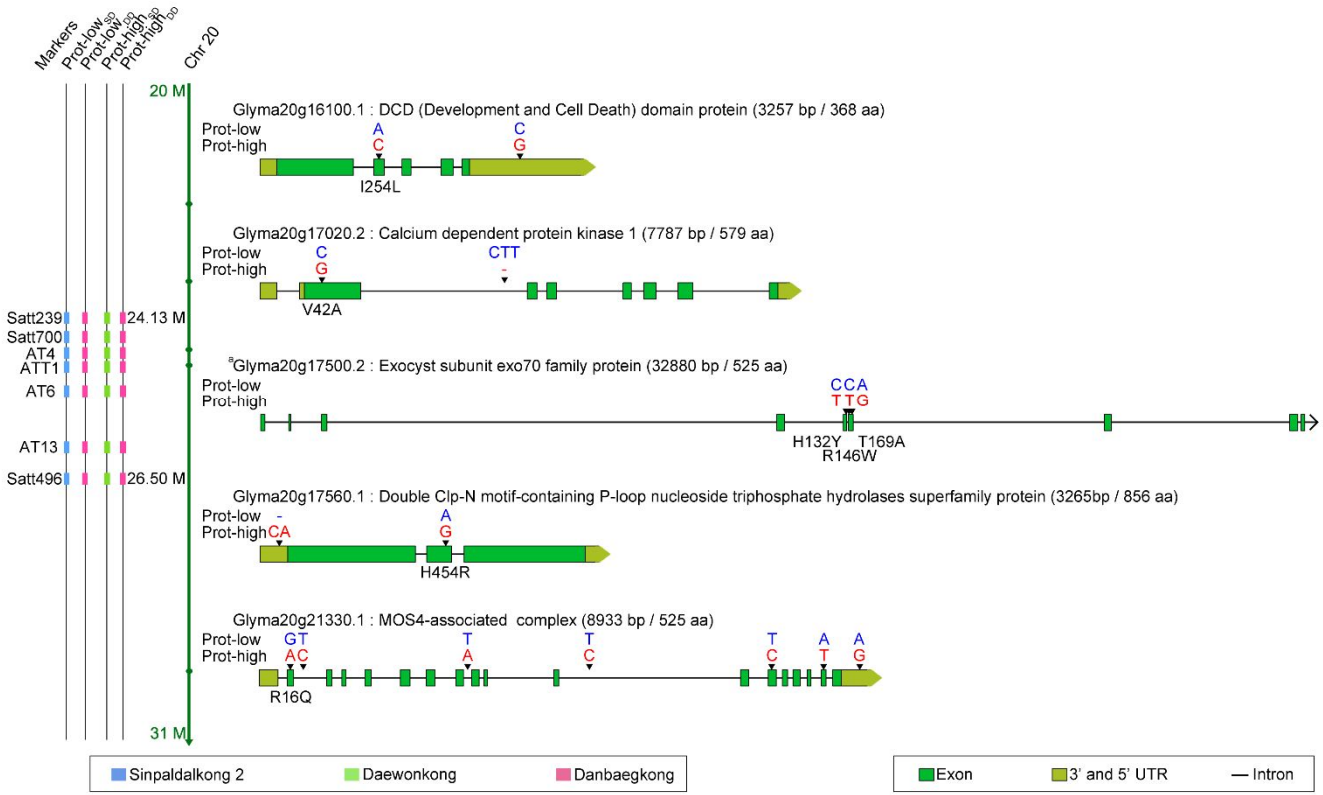

Markers
Prot-low SD
Prot-low DD
Prot-high SD
Prot-high DD
Chr 20
20 M
Glyma20g16100.1 : DCD (Development and Cell Death) domain protein (3257 bp / 368 aa)
Prot-low
Prot-high
A
C
C
G
I254L
Glyma20g17020.2 : Calcium dependent protein kinase 1 (7787 bp / 579 aa)
Prot-low
Prot-high
C
G
CTT
-
V42A
Satt239
Satt700
AT4
ATT1
AT6
24.13 M
AT13
Satt496
26.50 M
aGlyma20g17500.2 : Exocyst subunit exo70 family protein (32880 bp / 525 aa)
Prot-low
Prot-high
CCA
TTG
H132Y
T169A
R146W
Glyma20g17560.1 : Double Clp-N motif-containing P-loop nucleoside triphosphate hydrolases superfamily protein (3265bp / 856 aa)
Prot-low
Prot-high
-
CA
A
G
H454R
Glyma20g21330.1 : MOS4-associated complex (8933 bp / 525 aa)
Prot-low
Prot-high
GT
AC
T
A
T
C
T
C
A
T
A
G
R16Q
31 M
Sinpaldalkong 2
Daewonkong
Danbaegkong
Exon
3' and 5' UTR
Intron

Figure II-1 SNPs and indels of genes encoding proteins with amino acid changes between the low- and high protein NILs from protein QTL region on Chr 20. Vertical lines on the left side represent the chromosome and the region of sequence comparison. The leftmost line is Chr 20, and the next four lines show SSR marker haplotype of NILs. The region of sequence comparison is shown on the rightmost vertical line. Green circles represent the locations of nine genes with non-synonymous SNPs of those indicated on the right side. Each SNP and indel site in the five genes is indicated by a black triangle, and the sequences of NIL lines are shown in blue (-low) and red (-high). If the SNP has a non-synonymous change, the amino acid change is shown below the SNP site. As Glyma20g17500 has 35 SNPs in its genic region, only three non-synonymous SNPs are shown in this figure.

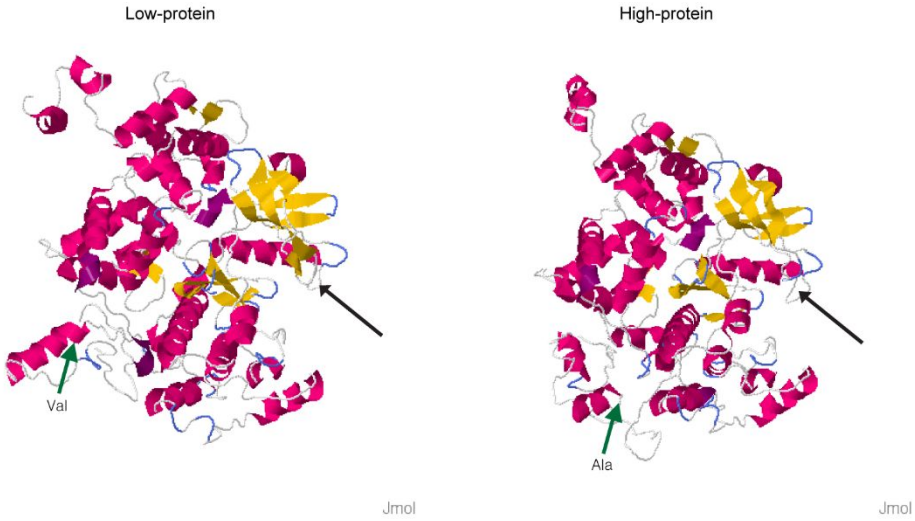


Figure II-2 Predicted protein structures of calcium dependent protein kinase (CDPK) between low- and high-protein lines. The protein sequence of low- and high-protein lines were obtained from Williams 82 reference genome and Danbaekkong, respectively. Green arrows indicate the site of protein change caused by non-synonymous mutation. Black arrows point the different beta-sheet structure on serine/threonine kinase domain of CDPK.

DISCUSSION

Mapping of introgression sites using SSR markers revealed that the introgressed regions included the major high seed protein QTL *Prot 15-1* between Satt239 (24.13 Mb) and Satt496 (26.50 Mb) on Chr 20 in the two pairs of NIL_{SD} and NIL_{DD} that we examined (Table II-1). In a previous study, we found that Danbaekkong has a favorable allele for increasing protein content at the Satt239 locus on Chr 20 (Park 2002). Another NIL population was previously developed by backcrossing the high protein *G. soja* (PI 468916) allele into a *G. max* background (A81-356022) (Bolon et al. 2010, Nichols et al. 2006). The introgressed region in this NIL is flanked by Sat_174 (24.54 Mb) and ssrqtl_38 (32.96 Mb) on Chr 20, corresponding to an approximately 8.4 Mb genomic sequence (Bolon et al. 2010).Though genotypic segregation was not observed at Satt239 between the Prot_low and Prot_high NILs, the protein QTL region of this NIL was located in the interval of Satt239 and Satt496 (Bolon et al. 2010). Therefore, we identified nucleotide variations between low- and high protein lines within the genomic region from 20 Mb to 31 Mb on Chr 20, which surrounds Satt239 to Satt496.

More recently, introgression sites of soybean NILs were efficiently mapped based on SNPs identified by NGS (Severin et al. 2010). *De novo* SNP discovery in NIL pairs with different protein contents was performed via RNA-seq analysis by NGS. Among the 387 SNPs identified between the

seed protein NILs, 142 are located in the well-known introgressed region on Chr 20. The positions of these coding DNA variations stretch from 12.12 Mb to 46.12 Mb (Severin et al. 2010). Comparative NGS analysis of DNA, as well as RNA, enables high mapping resolution, high marker density, and the prediction of non/missense or frameshift mutations using identified markers (Severin et al. 2010). In the present study, comparative analysis of NGS data between low- and high protein lines ($Prot\_low_{SD}$ vs. $Prot\_high_{SD}$ and $Prot\_low_{DD}$ vs. $Prot\_high_{DD}$) in two pairs of NILs revealed that 96 SNPs and two indels in coding sequences overlapped at the same positions in two NIL pairs (Table 2). The 66 non-synonymous SNPs and one frameshift were predicted to affect 30 genes in the seed protein QTL region. Among these genes, it is likely that five genes produce mRNA transcripts during seed development in soybean (Figure II-2 and Table II-3).

The accumulation of storage products consisting of starch, oil, and specialized protein during seed development is triggered by a complicated regulatory network in the embryo (Weber et al. 2005). This regulatory network is involved in genetic and biochemical reprogramming mediated by multiple pathways in response to sugar and phytohormones (Gibson 2004, Wobus and Weber 1999). Increased sugar concentration is thought to promote glycolysis and the tri-carbonic acid cycle pathway, leading to the production of a supplementary supply of carbon precursors for amino acid synthesis (Weigelt et al. 2009). Amino acid biosynthesis controls storage

protein synthesis (Weber et al. 2005). Additionally, protein phosphorylation of relevant enzymes is implicated in the metabolism of storage protein synthesis (Weber et al. 2005). In this study, one of the identified genes with missense mutations between low- and high protein lines was annotated as calcium (Ca^{2+})-dependent protein kinase (CDPK; Glyma20g17020). CDPK is required for storage product accumulation through phosphorylation of sucrose synthase in rice seeds (Asano et al. 2002). Arrested seed development occurs at an early stage in transgenic rice overexpressing *osCDPK2* (Morello et al. 2000). Diverse members of the rice *CDPK* gene family are preferentially upregulated during reproductive developmental stages in response to abiotic stress, and *OsCPK23* is specifically expressed in the early stage (S1) of seed development (Ray et al. 2007). Five *CDPK* genes, including *OsCPK23*, are upregulated at the endosperm stage in rice subjected to hormone treatment (Ye et al. 2009). In addition, CDPK proteins phosphorylate sucrose synthase in maize leaves under low-oxygen conditions, causing activation of the cleavage reaction (Subbaiah and Sachs 2001). CDPKs are closely related to sucrose nonfermenting-1-related protein kinase (SnRK1), which is a major component of the sugar signal transduction mechanism (Weber et al. 2005). These results provide evidence that phosphorylation is involved in metabolic regulation mediated by sugar, ABA, and/or certain stress conditions during seed maturation (Weber et al. 2005, Weber et al. 2010).

Seed proteins of most plant species are synthesized within the endoplasmic reticulum (ER) and are then transported to protein storage vacuoles where they finally accumulate (Galili et al. 1993, Herman and Schmidt 2004, Herman 2008, Ibl and Stoger 2012). Large masses of seed storage proteins are concentrated in an aggregative form making them insoluble protein assemblies (Herman 2008). Protein bodies which are formed by the tubular ER sequester insoluble proteins and are released into cytoplasm. Subsequently, the proteins bodies are transferred to vacuoles by a direct ER-vacuole trafficking route that by passes the Golgi, different from the classical ER-Golgi-vacuole route for delivery of most of soluble secretary proteins (Galili 2004, Herman and Schmidt 2004, Michaeli et al. 2014). Extensive microscopic evidence suggested involvement of autophagy in internalization of protein bodies into the storage vacuole (Herman 2008, Reyes et al. 2011). In maize seed aleurone cells, storage proteins are transported from the ER to protein storage vacuoles using a special autophagic process (Reyes et al. 2011), similar to the ER-vacuole trafficking process also observed in wheat (Ibl and Stoger 2012). The direct ER to vacuole transport using autophagy appears to show specificity depending on the type of the storage proteins as well as the plant species and tissues (Reyes et al. 2011). Selection of cargo to be delivered by autophagosome is controlled by binding ofAtg8, embedded on the autophagosome membrane, to the cargo protein or cargo receptor protein (Li and Vierstra 2012). In

Arabidopsis, identification of two novel cargo receptor proteins, termed Autophagy Interacting proteins 1 and 2 (ATI1 and ATI2) also supported involvement of autophagy in the transport of functional proteins from the ER to the vacuole (Honig et al. 2012). These two genes were found to contain two consensus Atg8-binding motifs and also bind the *Arabidopsis* autophagy associated Atg8f protein through various approaches. In addition, a recent report revealed a subunit of the exocyst complex, Exo70B1 is co-transported with the autophagy-associated Atg8 protein to the vacuole, as a key player in autophagosome targeting into the vacuole (Kulich et al. 2013). The exocyst complex, consisting of eight subunits – Sec3, Sec5, Sec6, Sec8, Sec10, Sec15 Exo70 andExo84 – is originally involved in tethering of complex between vesicles and plasma membrane prior to exocytosis (Heider and Munson 2012). Existence of putative Atg8 interacting motifs within the Exo70B1peptide also indicates the formation of a tethering complex with Atg8 for autophagy-related Golgi-independent import into the vacuole in plants (Tzfadia and Galili 2013). In the present study, the gene (Glyma20g17500) showing the change of amino acids between low- and high-protein NILs was found to be a homolog of the *Arabidopsis* exocyst complex subunit Exo70. This exocyst complex subunit Exo70 gene possesses 35 SNPs in genic regions, of which three are non-synonymous (Figure II-2). Mutations in Exo70 subunits would cause target membrane-specific deviations, since Exo70 subunits are those to bind the target

membrane and to function as the landmark for vesicle targeting/tethering (Heider and Munson 2012, Kulich et al. 2013, Munson and Novick 2006, Žárský et al. 2009).

The three genes with missense mutations between the low- and high protein lines encode proteins with known functional domains (Table 3), including DCD domain protein (Glyma20g16100), double Clp-N motif-containing P-loop nucleoside triphosphate hydrolases superfamily protein (Glyma20g17560) and MOS4-associated complex (MAC) (Glyma20g21330). The DCD domain is plant specific and strongly induced during plant development and programmed cell death. The putative *A. thaliana* ortholog (At5g42050) of this gene is upregulated by external stresses including ozone stress, osmotic stress, and cold stress, as determined from transcript profiling data (Tenhaken et al. 2005_ENREF_33). While the *Arabidopsis* homolog of Clp-N motif-containing P-loop nucleoside triphosphate hydrolase superfamily protein appears to function in ATP binding, its biological role is unknown. The MOS4-associated complex (MAC) consist of MODIFIER OF snc1 4 (MOS4) associates, the Myb-transcription factor CELL DIVISION CYCLE 5 (AtCDC5), and the WD-40 repeat protein PLEIOTROPIC REGULATORY LOCUS 1 (PRL1) (Monaghan et al. 2009). The MAC 3A and 3B may function in the regulation of plant innate immunity.

In the current study, we developed two pairs of NILs with different seed protein contents, derived from two RHLs to the high protein QTL in the

interval of Satt239 and Satt496 on Chr 20 in the RIL populations. WGS analysis demonstrated that these NILs may be effective materials for identifying candidate genes likely involved in seed protein accumulation through survey of functional nucleotide changes in the QTL region. Of the identified candidate genes, two genes, *CDPK* and *Exo70* subunit, may be associated with synthesis and transport of seed storage protein, respectively. Further functional validation of the candidate genes, using transcriptome/proteome analysis, mutagenesis and association analysis, should shed light on the molecular regulation of seed protein accumulation during soybean seed development.

REFERENCES

Asano T, Kunieda N, Omura Y, Ibe H, Kawasaki T, Takano M, Sato M, Furuhashi H, Mujin T, Takaiwa F (2002) Rice SPK, a calmodulin-like domain protein kinase, is required for storage product accumulation during seed development phosphorylation of sucrose synthase is a possible factor. Plant Cell 14:619-628

Boetzer M, Henkel CV, Jansen HJ, Butler D, Pirovano W (2011) Scaffolding pre-assembled contigs using SSPACE. Bioinformatics 27:578-579

Bolon Y-T, Joseph B, Cannon SB, Graham MA, Diers BW, Farmer AD, May GD, Muehlbauer GJ, Specht JE, Tu ZJ (2010) Complementary genetic and genomic approaches help characterize the linkage group I seed protein QTL in soybean. BMC Plant Biol 10:41

Brummer E, Graef G, Orf J, Wilcox J, Shoemaker R (1997) Mapping QTL for seed protein and oil content in eight soybean populations. Crop Sci 37:370-378

Chung J, Babka H, Graef G, Staswick P, Lee D, Cregan P, Shoemaker R, Specht J (2003) The seed protein, oil, and yield QTL on soybean linkage group I. Crop Sci 43:1053-1067

Diers B, Keim P, Fehr W, Shoemaker R (1992) RFLP analysis of soybean seed protein and oil content. Theor Appl Genet 83:608-612

Galili G (2004) ER-derived compartments are formed by highly regulated

processes and have special functions in plants. Plant Physiol 136:3411-3413

Galili G, Altschuler Y, Levanony H (1993) Assembly and transport of seed storage proteins. Trends Cell Biol 3:437-442

Gao Q, Yue G, Li W, Wang J, Xu J, Yin Y (2012) Recent progress using high-throughput sequencing technologies in plant molecular breeding. J Integr Plant Biol 54:215-227

Gibson SI (2004) Sugar and phytohormone response pathways: navigating a signalling network. J Exp Bot 55:253-264

Hartwig EE (1990) Registration of soybean high-protein germplasm line'D76-8070'. Crop Sci 30:764-765

Heider MR, Munson M (2012) Exorcising the exocyst complex. Traffic 13:898-907

Herman E, Schmidt M (2004) Endoplasmic reticulum to vacuole trafficking of endoplasmic reticulum bodies provides an alternate pathway for protein transfer to the vacuole. Plant Physiol 136:3440-3446

Herman EM (2008) Endoplasmic reticulum bodies: solving the insoluble. Curr Opin Plant Biol 11:672-679

Honig A, Avin-Wittenberg T, Ufaz S, Galili G (2012) A new type of compartment, defined by plant-specific Atg8-interacting proteins, is induced upon exposure of *Arabidopsis* plants to carbon starvation. Plant Cell 24:288-303

Huang X, Feng Q, Qian Q, Zhao Q, Wang L, Wang A, Guan J, Fan D, Weng Q, Huang T (2009) High-throughput genotyping by whole-genome resequencing. Genome Res 19:1068-1076

Ibl V, Stoger E (2012) The formation, function and fate of protein storage compartments in seeds. Protoplasma 249:379-392

Joshi T, Patil K, Fitzpatrick MR, Franklin LD, Yao Q, Cook JR, Wang Z, Libault M, Brechenmacher L, Valliyodan B (2012) Soybean Knowledge Base (SoyKB): a web resource for soybean translational genomics. BMC genomics 13:S15

Kulich I, Pečenková T, Sekereš J, Smetana O, Fendrych M, Foissner I, Höftberger M, Žárský V (2013) *Arabidopsis* exocyst subcomplex containing subunit EXO70B1 is involved in autophagy-related transport to the vacuole. Traffic 14:1155-1165

Li F, Vierstra RD (2012) Autophagy: a multifaceted intracellular system for bulk and selective recycling. Trends Plant Sci 17:526-537

Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 25:1754-1760

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R (2009) The sequence alignment/map format and SAMtools. Bioinformatics 25:2078-2079

Mayer KF, Martis M, Hedley PE, Šimková H, Liu H, Morris JA, Steuernagel B, Taudien S, Roessner S, Gundlach H (2011) Unlocking the barley

genome by chromosomal and comparative genomics. Plant Cell 23:1249-1263

Michaeli S, Avin-Wittenberg T, Galili G (2014) Involvement of autophagy in the direct ER to vacuole protein trafficking route in plants. Front Plant Sci 5 doi: 10.3389/fpls.2014.00134

Monaghan J, Xu F, Gao M, Zhao Q, Palma K, Long C, Chen S, Zhang Y, Li X (2009) Two Prp19-like U-box proteins in the MOS4-associated complex play redundant roles in plant innate immunity. PLoS Pathog 5:e1000526

Morello L, Frattini M, Gianì S, Christou P, Breviario D (2000) Overexpression of the calcium-dependent protein kinase *OsCDPK2* in transgenic rice is repressed by light in leaves and disrupts seed development. Transgenic Res 9:453-462

Munson M, Novick P (2006) The exocyst defrocked, a framework of rods revealed. Nat Struct Mol Biol 13:577-581

Nichols D, Glover K, Carlson S, Specht J, Diers B (2006) Fine mapping of a seed protein QTL on soybean linkage group I and its correlated effects on agronomic traits. Crop Sci 46:834-839

Ray S, Agarwal P, Arora R, Kapoor S, Tyagi AK (2007) Expression analysis of calcium-dependent protein kinase gene family during reproductive development and abiotic stress conditions in rice (Oryza sativa L. ssp. indica). Mol Genet Genomics 278:493-505

Reyes FC, Chung T, Holding D, Jung R, Vierstra R, Otegui MS (2011) Delivery of prolamins to the protein storage vacuole in maize aleurone cells. Plant Cell 23:769-784

Sebolt A, Shoemaker R, Diers B (2000) Analysis of a quantitative trait locus allele from wild soybean that increases seed protein concentration in soybean. Crop Sci 40:1438-1444

Severin AJ, Woody JL, Bolon Y-T, Joseph B, Diers BW, Farmer AD, Muehlbauer GJ, Nelson RT, Grant D, Specht JE (2010) RNA-Seq Atlas of *Glycine max*: a guide to the soybean transcriptome. BMC Plant Biol10:160

Shure M, Wessler S, Fedoroff N (1983) Molecular identification and isolation of the *Waxy* locus in maize. Cell 35:225-233

Subbaiah CC, Sachs MM (2001) Altered patterns of sucrose synthase phosphorylation and localization precede callose induction and root tip death in anoxic maize seedlings. Plant Physiol 125:585-594

Tajuddin T, Watanabe S, Yamanaka N, Harada K (2003) Analysis of quantitative trait loci for protein and lipid contents in soybean seeds using recombinant inbred lines. Breed Sci 53:133-140

Tenhaken R, Doerks T, Bork P (2005) DCD–a novel plant specific domain in proteins involved in development and programmed cell death. BMC bioinformatics 6:169

Tzfadia O, Galili G (2013) The *Arabidopsis* exocyst subcomplex subunits

involved in a Golgi-independent transport into the vacuole possess consensus autophagy-associated atg8 interacting motifs. Plant Signal Behav 8:e26732

Weber H, Borisjuk L, Wobus U (2005) Molecular physiology of legume seed development. Annu Rev Plant Biol 56:253-279

Weber H, Sreenivasulu N, Weschke W (2010) Molecular physiology of seed maturation and seed storage protein biosynthesis. In Pua EC, Davey MR (ed) Plant Developmental Biology-Biotechnological Perspectives. Springer, Berlin, pp 83-104

Weigelt K, Küster H, Rutten T, Fait A, Fernie AR, Miersch O, Wasternack C, Emery RN, Desel C, Hosein F (2009) ADP-glucose pyrophosphorylase-deficient pea embryos reveal specific transcriptional and metabolic changes of carbon-nitrogen metabolism and stress responses. Plant Physiol 149:395-411

Wobus U, Weber H (1999) Seed maturation: genetic programmes and control signals. Curr Opin Plant Biol 2:33-38

Ye S, Wang L, Xie W, Wan B, Li X, Lin Y (2009) Expression profile of calcium-dependent protein kinase (CDPKs) genes during the whole lifespan and under phytohormone treatment conditions in rice (*Oryza sativa* L. ssp. indica). Plant Mol Biol 70:311-325

Yu H, Xie W, Wang J, Xing Y, Xu C, Li X, Xiao J, Zhang Q (2011) Gains in QTL detection using an ultra-high density SNP map based on

population sequencing relative to traditional RFLP/SSR markers. PloS one 6:e17595

Žárský V, Cvrčková F, Potocký M, Hála M (2009) Exocytosis and cell polarity in plants–exocyst and recycling domains. New Phytol 183:255-272

CHAPTER III

Gene expression profiling for seed protein and oil synthesis during seed development in soybean

ABSTRACT

Soybean (*Glycine max* L.) is one of the most important crops because of its high seed contents of protein and oil. However, little is known about the molecular regulation of seed filling during accumulation of seed protein and oil storage products. We identified soybean differentially expressed genes between low- and high-protein NIL lines during seed protein stage at 1, 2, 3 and 4 WAFs. Gene annotation using Mapman and cluster heatmap represent that genes related with seed reserve accumulation started to represent differentially expression at 2 WAF and more genes were up-regulated in Prot_high$_{DD}$ than Prot_low$_{DD}$ at this period. Transcription factors

including major regulators for seed development was identified as differentially expressed.

To understand metabolic gene regulation during the start of early seed maturation, we investigated *G. max* homologs of Arabidopsis genes involved in metabolic pathways of carbon precursors, protein, and oil, and analyzed gene expression patterns in immature seeds at 1 and 2 weeks after flowering (WAF). *G. max* undergoes two rounds of whole-genome duplication, the number of genes involved in the three synthesis pathways is more than two times higher than that in *Arabidopsis*. Among these genes, five were conserved as single-copy genes and 44 were high copy gene families consisting of more than seven homolog members. We identified five differentially expressed genes in immature seeds aged between 1 and 2 WAF, including *CELL WALL INVERTASE*, *BRANCHED-CHAIN AMINO ACID TRANSAMINASE*, *AMINO ACID PERMEASE*, *ALDEHYDE REDUCTASE*, and *BIOTIN CARBOXYL CARRIER PROTEIN*. Expression analysis of the duplicated genes on the synteny block revealed that the duplicated genes involved in protein synthesis had stronger positive correlations between their expression patterns than those of oil synthesis genes. This study provides novel insights into the molecular details of genes associated with soybean seed protein and oil synthesis pathways. These genes can be used as tools to improve seed nutrient composition.

INTRODUCTION

Soybean [*Glycine max* (L.) Merr] belongs to the family Leguminosae, also known as Fabaceae, along with many essential food legume crops such as peas, chickpeas, alfalfa, lentils, fava beans, and peanuts. Soybean has important nutritional benefits and is one of the most important crops, its average annual global production reached 217.6 million tons in 2005–2007 and is expected to increase by 2.2% annually to 371.3 million tons by 2030 (Masuda and Goldsmith 2009). Soybean is a very unique legume crop because it contains high protein and oil contents, approximately 40% and 20%, respectively (Clemente and Cahoon 2009). Continuous efforts of soybean breeders are directed to improve the oil and protein quality by identifying linked markers and quantitative trait loci (QTLs) (Burton 1985, Wilcox 1985).

A previous study of the relationship between protein, oil, and sugar metabolism in soybean seed reported a negative correlation between protein and oil content and protein and sugar content, but a positive correlation between sugar and oil content (Hymowitz et al. 1972). Consistent with these results, a QTL analysis of selected molecular markers revealed a negative relationship between protein and oil contents (Lee et al. 1996). There are currently more than 100 QTLs associated with protein and oil

content on all 20 chromosomes (Chrs) (http://www.soybase.org). Many seed protein and oil QTLs have been identified on Chr 20 (LG I) (Brummer et al. 1997, Sebolt et al. 2000, Specht et al. 2001, Tajuddin et al. 2003, Chung et al. 2003, Bolon et al. 2010).

The plant reproductive stage links two generations by producing seeds as a repository of genetic materials along with storage products such as protein and oil, which are metabolized by the embryo during germination. The seed maturation stage known as seed filling involves the accumulation of seed protein and lipid, this stage is crucial for establishing seed vigor and germination competence (Dornbos and Mullen 1991). Seed filling is dependent on metabolic network regulation and assimilate transport from a source (Weber et al.1997, Weber et al. 2005, Weichert et al. 2010). Seed protein and oil accumulation starts at an early stage of seed maturation after carbon precursor synthesis, and sucrose is the major assimilated carbon (Baud et al 2008). Substrates for seed protein and oil biosynthetic pathways are derived from sucrose degradation, glycolysis, and the TCA cycle (Baud et al. 2008, Baud and Lepiniec 2010). The accumulation of protein and oil in seed requires the transport and metabolism of sucrose and carbon precursors (Vigeolas et al. 2007, Weichert et al. 2010).

In the present study, we profiled gene expression between low-and high-protein NIL during seed development. Differentially expressed genes

were increased after 2 WAF, early maturation stage. Then we identified soybean homologs of *Arabidopsis* genes involved in the metabolic pathways of carbon precursor, protein, and oil. We analyzed the expression patterns of these genes at early stages of seed maturation, at 1 and 2 weeks after flowering (WAF). The soybean gene paralogs involved in metabolic pathways of carbon precursor, protein, and oil are located in synteny blocks generated by whole-genome duplication (WGD). The expression levels of these paralogs were compared to determine functional correlations between duplicated genes during biosynthesis of seed storage products.

MATERIAL AND METHODS

Plant materials and RNA extraction

A recombinant F_9 line ($Prot\_low_{DD}$) developed by selecting single-seed descendants following a cross between a low-protein cultivar (Daewonkong) and a high-protein cultivar (Danbaekkong) was used in this study. $Prot\_low_{DD}$ is homozygous for low-protein alleles in the seed protein QTL located in the interval of Satt239–Satt496 on Chr 20, which was inherited from Daewonkong. This experiment was conducted in a greenhouse at the experimental farm of Seoul National University (Suwon, Korea) on July 13, 2013. Seeds were sown 3 cm deep directly into the soil covered with plastic mulch, with a spacing of 20 cm between plants along the row and 60 cm between rows. The natural photoperiod was 11.5–14.5 h per day. Each pod was tagged by the flowering date. Pods were harvested to collect immature seeds at 7–8 days after flowering (DAF), 14–15 DAF, 21-22 DAF and 28-29 DAF, which are equivalent to 1, 2, 3 and 4 WAF. To minimize the effects of biological variations among individual soybean plants, hundreds of pods at each stage were gathered from multiple plants (Jones and Vodkin 2013). One gram of immature seeds was ground, and RNA was extracted using a modified TRIzol® (Invitrogen, CA, USA) method. RNA concentration and quality was evaluated by gel electrophoresis and

NanoDropTM-1000 Spectrophotometry (Thermo Scientific, MA, USA).

RNA-seq analysis

Total RNA sequencing of immature soybean seed samples at 1 and 2 WAF was performed using Illumina Hiseq 2000 (Illumina, Co, CA, USA). Then, cDNA libraries were constructed using TruSeq® RNA Sample Prep Kit v2 (Illumina Inc., CA, USA). Sequencing data were generated through the standard Illumina pipeline. Raw reads were filtered using the IlluQC tool contained in the NGSQC Toolkit v.2.3.3. The resultant 32,283,577 and 31,725,878 cDNA reads from 1, 2, 3 and 4 WAF soybean seeds were aligned to the *G. max* reference genome (ver. 1.0) provided by Ensembl Plants (http://plants.ensembl.org/Glycine_max/Info/Index) using the Bowtie2 aligner (Langmead and Salzberg 2012) and Tophat v.2 (Kim et al. 2013), with consideration of the inner distance of paired-end sequences and in a reference-guided manner. Gene expression values were calculated based on the number of fragments per kilobase of exon per million fragments mapped (FPKM) (Mortazavi et al. 2008). Data given in FPKM were obtained using Cufflinks, and comparisons of transcript/gene FPKMs in 1, 2, 3 and 4 WAF immature soybean seed samples were processed by Cuffdiff (Roberts et al. 2011). Differntially expressed genes between Prot_high$_{DD}$ and

Pro_low$_{DD}$ was classified using Mapman hierarchial ontology system (Thimm et al. 2004).

Survey of *G. max* homologs of *Arabidopsis thaliana* genes involved in sucrose degradation and protein and oil synthesis

A. thaliana genes involved in synthesis of carbon precursor, protein, and oil were identified in two public databases, the Plant Metabolic Network (PMN; http://www.plantcyc.org) and The Arabidopsis Information Resource (TAIR; http://www.arabidopsis.org). *G. max* genes homologous to the listed *A. thaliana* genes were retrieved from the soybean genome sequence database at Phytozome (http://www.phytozome.net).

Identification of synteny blocks in the soybean genome

The *G. max* WGD was estimated by comparing co-linearity of synteny blocks. *G. max* protein sequences were self-compared using BLASTP searches with an E-value threshold of $1e^{-5}$ and by parsing out the top five hits. Consequently, the collinear synteny blocks were calculated using

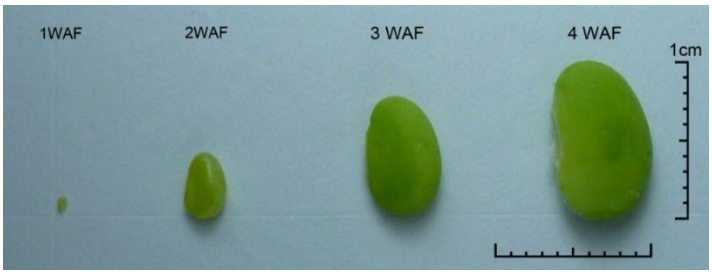
MCScanX software with default parameters (Wang et al. 2012). We calculated K_s values of the homologs within collinearity blocks using the perl script (add_ka_and_ks_to_collinearity.pl) from MCScanX. All K_s values were distributed; peak values were detected at 0.11 and 0.64, which corresponded to WGD. The synteny blocks at the peak had median values of 0.125; this K_s value was considered to be representative of the collinear blocks.

RESULTS

Characterization and developing soybean seeds and transcriptome analysis

We harvested immature seeds at 1, 2, 3 and 4 week after flowering (WAF) and each stage is corresponded to early cotyledon, early and late mature green stages, respectively (Figure III-1a). During the seed growth, average sizes like length, thickness and width were increased steadily until 3 WAF. While seed weight increased faster approximately 3-fold between 3 and 4 WAF, representing active seed storage product accumulation (Figure III-1b).

a)



b)

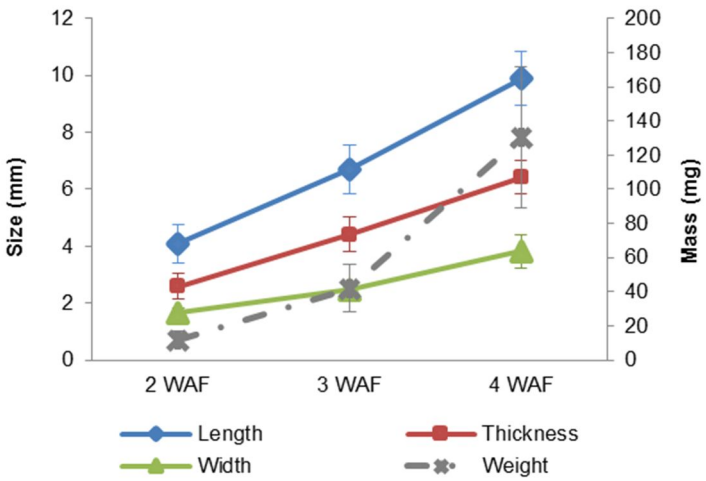


Figure III-1. Soybean seed development during experimental period. a) Seeds at 4 stages of seed filling. Seeds sampling was started at 1 WAF and continued to 4 WAF with 7 days intervals. b) Seed size and mass of each seed filling stage from 2 to 4 WAF. Length, thickness, width and weight values are average of 10 seeds and error bars are standard deviation.

Seeds from four WAFs from NIL pair (Prot_high$_{DD}$ and Prot_low$_{DD}$) consisted eight seed samples. High-throughput next generation transcriptome sequencing using RNA-seq was performed on eight samples of soybean seeds. We found the 54,140 transcriptionally active genes over four weeks of seed development. To identify differentially expressed genes (DEGs) between Prot_high$_{DD}$ and Prot_low$_{DD}$ lines, we compared FPKM value genes at each developing stage. There were 647 genes that represent differential expression in total (Figure III-2). Seed at globular stage (1 WAF), 3 genes were identified DEGs between Prot_high$_{DD}$ and Prot_low$_{DD}$. The number of DEGs between Prot_high$_{DD}$ and Prot_low$_{DD}$ were increased with seed maturation as 97, 279 and 364 genes at each 2, 3 and 4 WAF. Among the 97 DEGs at 2 WAF, number of up-regulated genes were 85 and 12 in Prot_high$_{DD}$ and Prot_low$_{DD}$, respectively. Meanwhile, the ratio of up-regulated genes in Prot_low$_{DD}$ increased to similar level of up-regulated genes in Prot_high$_{DD}$ at 3 and 4 WAFs. Over the three weeks period from 1 to 3 WAF, no DEGs maintained their expression regulation patterns, while during 3 and 4 WAFs, the 77 DEGs maintained their differential expression patterns. DEGs at 4 WAF were overlapped 17 genes with 2 WAF. (Figure III-2).

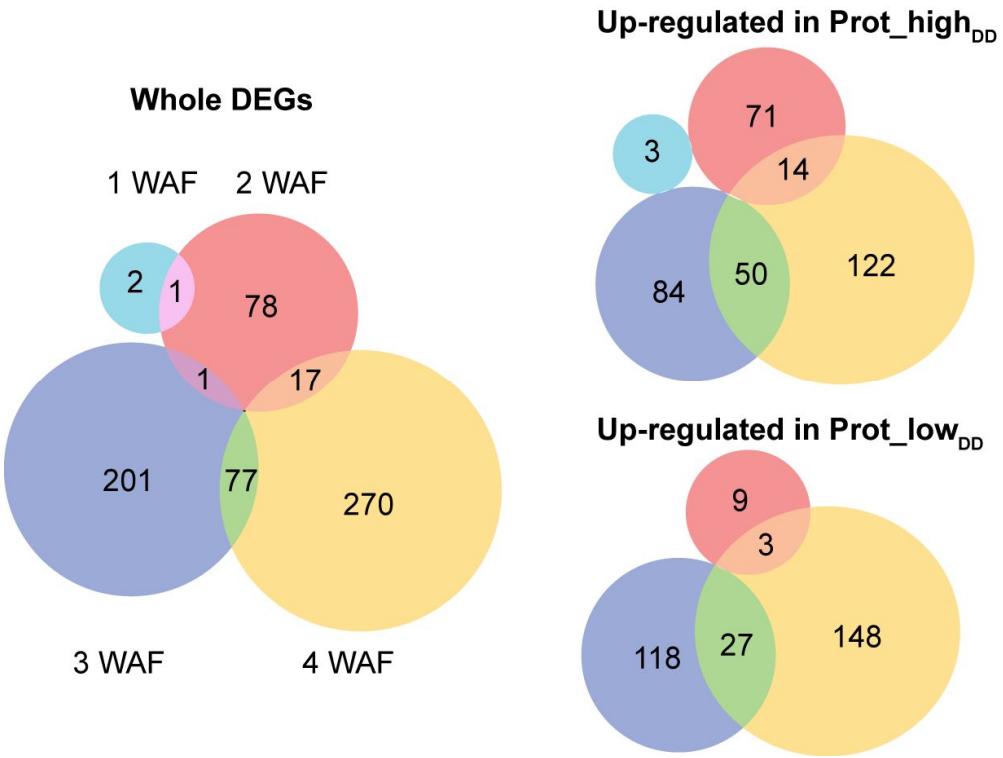


Figure III-2 Venn diagram of difnfferentially expressed genes (DEGs) of differential expressed genes between Prot-high and Prot-low lines. Number in each circles represents number of DEGs and the light blue, read, blue and yellow colors of circles indicate each 1, 2, 3 and 4 WAF, respectively.

Gene annotation using Mapman allowed the assignment of 425 DEG classifications into 25 BINs (Table III-1). Other 222 DEGs were classified as not assigned. Soybean stores seed storage components such as protein, oil and starch. Genes included in those storage molecule accumulation, such as major and minor carbohydrate metabolism, lipid metabolism, amino acid metabolism, C1 metabolism and protein metabolism were 78. Number of DEGs identified as stress, redox and miscellaneous enzyme family (misc), RNA, signaling and development were 250 genes. At 2 WAF, DEGs including genes of lipid metabolism, hormone metabolism, protein, cell, development and transport categories started to represent up-regulation in Prot_high$_{DD}$ than up-regulation genes in Prot_low$_{DD}$ (Table III-1). In this classification, DEGs of soybean storage protein (glycinin, geta-conglycinin, legumin) were put into subBINS of ‘development’ (supplement).

Table III-1 Gene ontology of differentially expressed genes using Mapman.

BIN	Ontology	protein	No. of up-regulated genes			
			1 WAF	2 WAF	3 WAF	4 WAF
1	Photosynthesis	High				2
		low				1
2	Major carbohydrate metabolism	High			2	2
		low			1	2
3	Minor carbohydrate metabolism	High			1	1
		low				2
9	Mitochondrial electron transport / ATP synthesis	High				
		low			1	1
10	Cell wall	High		2	3	9
		low			9	7
11	Lipid metabolism	High		3	3	4
		low				6
13	Amino acid metabolism	High			1	1
		low				2
15	Metal handling	High			2	2
		low				
16	Secondary metabolism	High		2		3
		low			1	4
17	Hormone metabolism	High		4	5	5
		low		1	8	8
18	Co-factor and vitamine metabolism	High				
		low				
19	Tetrapyrrole synthesis	High				
		low				
20	Stress	High		2	5	8
		low			5	9
21	Redox	High		1		1
		low				1
23	Nucleotide metabolism	High		1		1
		low			1	1
24	Biodegradation of xenobiotics	High			1	
		low				
25	C1-metabolism	High			1	1
		low			1	
26	Miscellaneous enzyme families	High		15	11	16
		low		3	11	17
27	RNA	High		17	12	17
		low		1	19	28

(Continue on next page)

BIN	Ontology	protein	No. of up-regulated genes			
			1 WAF	2 WAF	3 WAF	4 WAF
28	DNA	High			1	1
		low			2	1
29	Protein	High		5	8	11
		low			13	6
30	Signaling	High		1	9	8
		low			13	20
31	Cell	High		1	1	3
		low			1	2
33	Development	High		2	8	14
		low			2	7
34	Transport	High		3	7	9
		low		1	3	5
35	not assigned	High	3	25	53	68
		low		6	54	46

Table III-1 (*continued)*

Expression patterns of differential expressed genes related with seed storage product metabolism during seed development

We investigated gene expression of 78 genes that involved in metabolism of carbohydrate, oil and protein, protein and sugar transporters, seed storage proteins during seed development. Clustering analysis showed gene expression relationship between Prot_high$_{DD}$ and Prot_low$_{DD}$ during seed development (Figure III-3). Genes related with seed reserve accumulation started to represent differentially expression at 2 WAF and more genes were up-regulated in Prot_high$_{DD}$ than Prot_low$_{DD}$ at this period. Genes of sucrose degradation enzymes like sucrose synthase (*SUS*) and cell wall invertase (*CWINV*) were differently regulated. *SUS* gene was up-regulated in Prot_high$_{DD}$ from 2 to 4 WAF and the differential expression was identified at 3 WAF, while the expression of *INV*s were up-regulated in Prot_low$_{DD}$ and differences between two lines were continuously increased to 4 WAF when represented differential expression. DEGs involved in protein metabolisms were occupied 39 genes and 22 genes among them had function of protein degradation. Expression patterns during seed development, genes encodes storage proteins were up-regulated at 2 WAF and increased to 4 WAF in Prot_high$_{DD}$ along with some genes involved in protein posttranslational modification, TAG synthesis of lipid metabolism

(Figure III-3). The clusters that showed week up-regulation in Prot_high$_{DD}$ at 2 WAF than changed to high up-regulation in Prot_low$_{DD}$ at 3 WAF contatined ubiquitin, subtilases and protease genes involved in protein degradation and sugar trasnproter. Meanwhile, transporters of peptides and oilgopeptides were up-reguated in Prot_high$_{DD}$ at 3 WAF. Genes involved in synthesis of carbohydrates such as starch, raffinose and sucrose were up-regulated in Prot_high$_{DD}$ (Figure III-3).

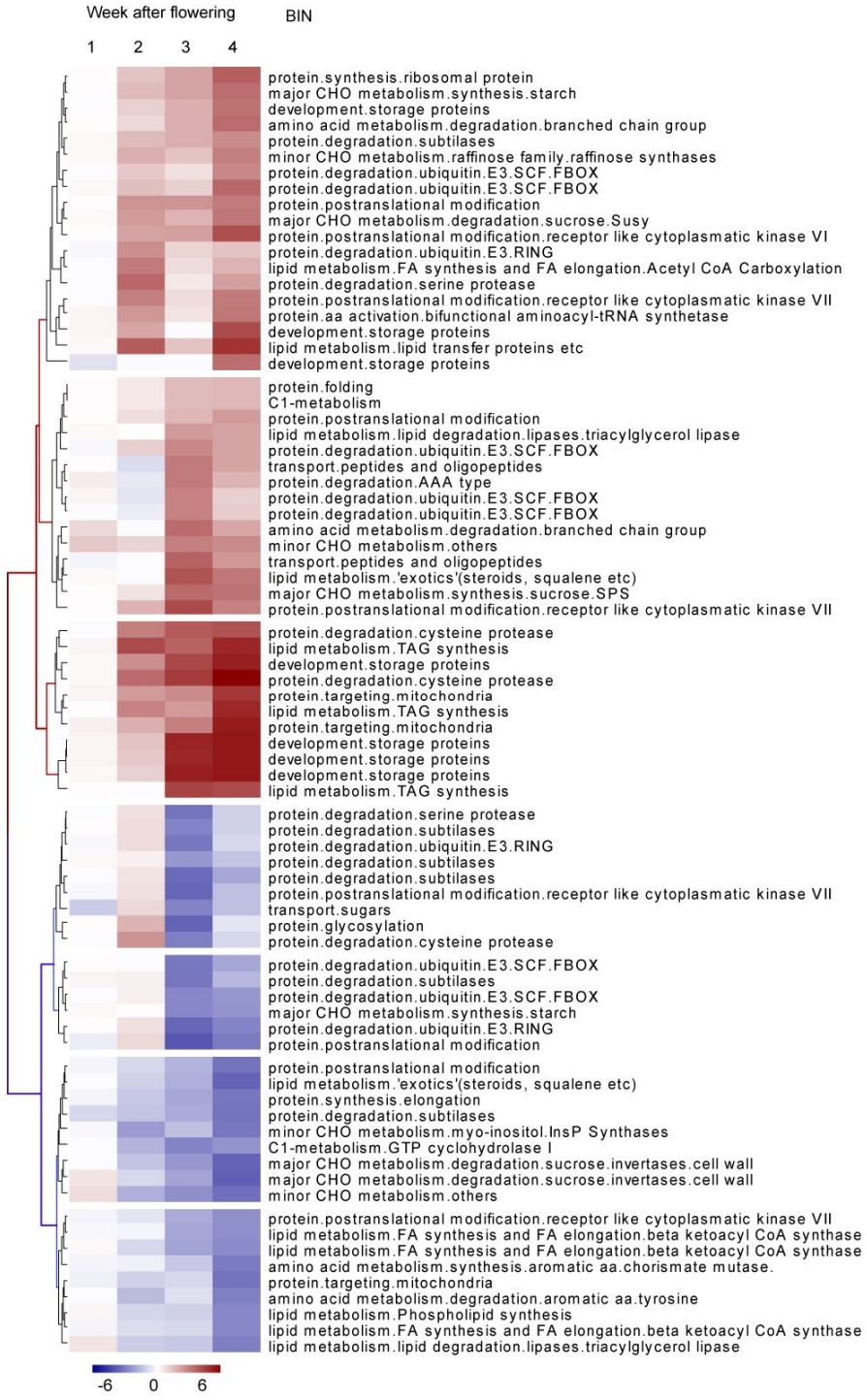

Week after flowering
BIN
1
2
3
4
protein.synthesis.ribosomal protein
major CHO metabolism.synthesis.starch
development.storage proteins
amino acid metabolism.degradation.branched chain group
protein.degradation.subtilases
minor CHO metabolism.raffinose family.raffinose synthases
protein.degradation.ubiquitin.E3.SCF.FBOX
protein.degradation.ubiquitin.E3.SCF.FBOX
protein.postranslational modification
major CHO metabolism.degradation.sucrose.Susy
protein.postranslational modification.receptor like cytoplasmatic kinase VI
protein.degradation.ubiquitin.E3.RING
lipid metabolism.FA synthesis and FA elongation.Acetyl CoA Carboxylation
protein.degradation.serine protease
protein.postranslational modification.receptor like cytoplasmatic kinase VII
protein.aa activation.bifunctional aminoacyl-tRNA synthetase
development.storage proteins
lipid metabolism.lipid transfer proteins etc
development.storage proteins
protein.folding
C1-metabolism
protein.postranslational modification
lipid metabolism.lipid degradation.lipases.triacylglycerol lipase
protein.degradation.ubiquitin.E3.SCF.FBOX
transport.peptides and oligopeptides
protein.degradation.AAA type
protein.degradation.ubiquitin.E3.SCF.FBOX
protein.degradation.ubiquitin.E3.SCF.FBOX
amino acid metabolism.degradation.branched chain group
minor CHO metabolism.others
transport.peptides and oligopeptides
lipid metabolism.'exotics'(steroids, squalene etc)
major CHO metabolism.synthesis.sucrose.SPS
protein.postranslational modification.receptor like cytoplasmatic kinase VII
protein.degradation.cysteine protease
lipid metabolism.TAG synthesis
development.storage proteins
protein.degradation.cysteine protease
protein.targeting.mitochondria
lipid metabolism.TAG synthesis
protein.targeting.mitochondria
development.storage proteins
development.storage proteins
development.storage proteins
lipid metabolism.TAG synthesis
protein.degradation.serine protease
protein.degradation.subtilases
protein.degradation.ubiquitin.E3.RING
protein.degradation.subtilases
protein.degradation.subtilases
protein.postranslational modification.receptor like cytoplasmatic kinase VII
transport.sugars
protein.glycosylation
protein.degradation.cysteine protease
protein.degradation.ubiquitin.E3.SCF.FBOX
protein.degradation.subtilases
protein.degradation.ubiquitin.E3.SCF.FBOX
major CHO metabolism.synthesis.starch
protein.degradation.ubiquitin.E3.RING
protein.postranslational modification
protein.postranslational modification
lipid metabolism.'exotics'(steroids, squalene etc)
protein.synthesis.elongation
protein.degradation.subtilases
minor CHO metabolism.myo-inositol.InsP Synthases
C1-metabolism.GTP cyclohydrolase I
major CHO metabolism.degradation.sucrose.invertases.cell wall
major CHO metabolism.degradation.sucrose.invertases.cell wall
minor CHO metabolism.others
protein.postranslational modification.receptor like cytoplasmatic kinase VII
lipid metabolism.FA synthesis and FA elongation.beta ketoacyl CoA synthase
lipid metabolism.FA synthesis and FA elongation.beta ketoacyl CoA synthase
amino acid metabolism.synthesis.aromatic aa.chorismate mutase.
protein.targeting.mitochondria
amino acid metabolism.degradation.aromatic aa.tyrosine
lipid metabolism.Phospholipid synthesis
lipid metabolism.FA synthesis and FA elongation.beta ketoacyl CoA synthase
lipid metabolism.lipid degradation.lipases.triacylglycerol lipase
-6
0
6

Figure III-3 Heatmap and ontologies of differentially expressed genes related with seed storage product metabolic pathway. Log2 fold changes FPKM value between Prot-high$_{DD}$ and Prot-low$_{DD}$ based on their gene expression data were calculated. On the logarithmic color scale ranging from -6 to 6, dark blue represents at least 100-fold higher gene expression in Prot-high$_{DD}$ comparison to Prot-low$_{DD}$, and dark red represents 100-fold higher gene expression in Prot-low$_{DD}$ comparison to Prot_high$_{DD}$.

Table III-2 Gene BINs and the number of genes included in each cluster group.

Group 1		
Starch synthesis	glucose-1-phosphate adenylyltransferase	1
Sucrose degradation	sucrose synthase	1
Minor chrbohydrate	galactinol-sucrose galactosyltransferase	1
Fatty acid synthesis	biotin carboxyl carrier protein	1
Lipid transfer	lipid transfer protein	1
Amino acid degradation	branched-chain alpha-keto acid dehydrogenase	1
Amino acid activation	tRNA synthetase class	1
Postranslational modification	receptor like cytoplasmatic kinase VI	1
	receptor like cytoplasmatic kinase VII	1
Protein degradation	subtilase	1
	C3HC4-type ring finger family ubiquitin	1
	f-box family ubiquitin	2
	serine carboxypeptidase	1
Storage protein	Legumin	1
	β-conglycinin	2
Group 2		
Sucrose synthesis	sucrose phosphate synthase	1
Minor chrbohydrate	aldose reductase	1
Lipid metabolism	hydroxyjasmonate sulfotransferase	2
Lipid degradation	triacylglycerol lipase	1
Amino acid degradation	branched-chain alpha-keto acid dehydrogenase	1
C1 metabolism	formyltetrahydrofolate synthetase	1
Postranslational modification	calcium-dependent protein kinase	1
	receptor like cytoplasmatic kinase VII	1
Protein degradation	f-box family ubiquitin	3
	AAA-type ATPase	1
Protein folding	embryo sac development arrest	1
Peptide transport	protein peptide transporter	2
Group 3		
Tag synthesis	oleosin	3
Protein targeting	mitochondrial import inner membrane translocase subunit	2
Protein degradation	cysteine protease	2
Storage protein	β -conglycinin	1
	glycinin	3

(Continue on the next page)

Table III-2 *(continued)*

Group 4		
Postranslational modification	receptor like cytoplasmatic kinase VII	1
Protein degradation	subtilase	3
	C3HC4-type ring finger family ubiquitin	1
	cysteine endopeptidase	1
	Serine carboxypeptidase	1
Protein glycosylation	galactosyltransferase	1
Sugar transport	mannitol transporter	1
Group 5		
Starch synthesis	plant glycogenin-like starch initiation protein	1
Postranslational modification	protein phosphatase 2C	1
Protein degradation	subtilase	1
	U-box domain-containing protein	1
	f-box family ubiquitin	2
Group 6		
Sucrose degradation	Cell wall invertase	2
Minor chrbohydrate	Myo-inositol-1-phosphate synthase	1
	aldo/keto reductase	1
Lipid metabolism	hydroxyjasmonate sulfotransferase	1
C1 metabolism	GTP cyclohydrolase	1
Protein elongation	protein elongation factor	1
Postranslational modification	protein phosphatase 2C	1
Protein degradation	subtilase	1
Group 7		
Fatty acid synthesis	3-ketoacyl-CoA synthas	3
Phospholipid synthesis	glycerol-3-phosphate acyltransferase	1
Lipid degradation	triacylglycerol lipase	1
Amino acid synthesis	chorismate mutase	1
Amino acid degradation	tyrosine aminotransferase	1
Protein targeting	mitochondrial processing peptidase alpha subunit	1
Postranslational modification	receptor like cytoplasmatic kinase VII	1

Expression patterns of differential expressed transcription factors

We identified 65 differentially expressed transcription factors (TFs) from 4342 genes of SoybeanTFDB hosted by the RIKEN plant science center. We found 65 TFs were involved in 22 families. The expression patterns of all differentially expressed TFs were analyzed during seed maturation (Table III-2, Figure III-4). Among these genes, 12, 6 and 11 TFs were found to be up-regulated in Prot_high$_{DD}$ at 2, 3 and 4 WAF, respectively. There were 12 and 30 TFs were up-regulated in Prot_low$_{DD}$ at 3 and 2 WAF (Table III-2). We analyzed expression patterns of differentially expressed TFs from log2 fold change of FPKM values (Figure III-4). ABI3/VP1, AP2/EREBP, bHLH, C2C2 (Zn)-YABBY, GRAS, R2R3_Myb and Trihelix started to represent up-regulated gene expression in Prot_high$_{DD}$ at 2 WAF. C2C2 (Zn)-YABBY were differentially expressed only at 2 WAF. Zf-HD was only DEG at 3 WAF that up-regulated in Prot_low$_{DD}$. C2C2 (Zn)-Dof and EIL were up-regulated DEGs in Prot_high$_{DD}$ only at 4 WAF, while b-ZIP, SRS and ZIM were up-regulated DEGs in Prot_low$_{DD}$. The expression pattern of TFs were categorized to eight clusters (Figure III-4). ABI3/VP1 (ABSCISIC ACID-INSENSITIVE 3 (ABI3)), CCAAT/HAP3 (HAP3, LEAFY COTYLEDON 1 (LEC1)-like), FUSCA 3 (FUS3) that known as major transcriptional

regulator during seed developlemt (Angeles-Núñez and Tiessen 2012), were commonly showed up-regulated pattern in Prot_high$_{DD}$ at 2 WAF. At 3 WAF, FPKM value in Prot_high$_{DD}$ of HAP3 gene was similar with Prot_low$_{DD}$ at 1 WAF and increased at 2 WAF, then HAP3 expression was decreased at 3 WAF in Prot_high$_{DD}$ while increased in Prot_low$_{DD}$. In the first cluster which represented highly up-regulated at 2 WAF included five C2C2 (Zn)-YABBY genes. Gene expression of C2C2 (Zn)-YABBY in Prot_low$_{DD}$ were decreased to low level at 2 and 4 WAFs. The other cluster represent highly up-regulated in Prot_low$_{DD}$ at 4 WAF included three b-ZIP genes which represented increased gene expression at 2 and 4 WAFs.

Table III-3 Classification of transcription factors that differentily expressed between Prot_high$_{DD}$ and Prot-low$_{DD}$during seed development

Transcription factor	Protein	No. of up-regulated genes in Prot_high		
		2 WAF	3 WAF	4 WAF
ABI3/VP1	High	1		
	Low			1
AP2/EREBP	High	2	1	
	Low		3	4
bHLH	High	1		2
	Low		1	3
bZIP	High			
	Low			3
C2C2(Zn) CO-like	High		1	
	Low			1
C2C2(Zn) Dof	High			1
	Low			
C2C2(Zn) GATA	High		1	1
	Low			
C2C2(Zn) YABBY	High	5		
	Low			
C2H2(Zn)	High			
	Low		1	4
CCAAT/HAP3	High			
	Low		1	
EIL	High			1
	Low			
GRAS	High	1	1	1
	Low			

(Continue on the next page)

Table III-3 *(continued)*

HB	High		1	2
	Low		1	
R2R3Myb	High	1		
	Low		2	4
Myb_related	High			
	Low			2
NAC	High		1	2
	Low			1
SBP	High			
	Low		1	4
SRS	High			
	Low			1
Trihelix	High	1		1
	Low			
WRKY(Zn)	High			
	Low		1	1
zf-HD	High			
	Low		1	
ZIM	High			
	Low			1

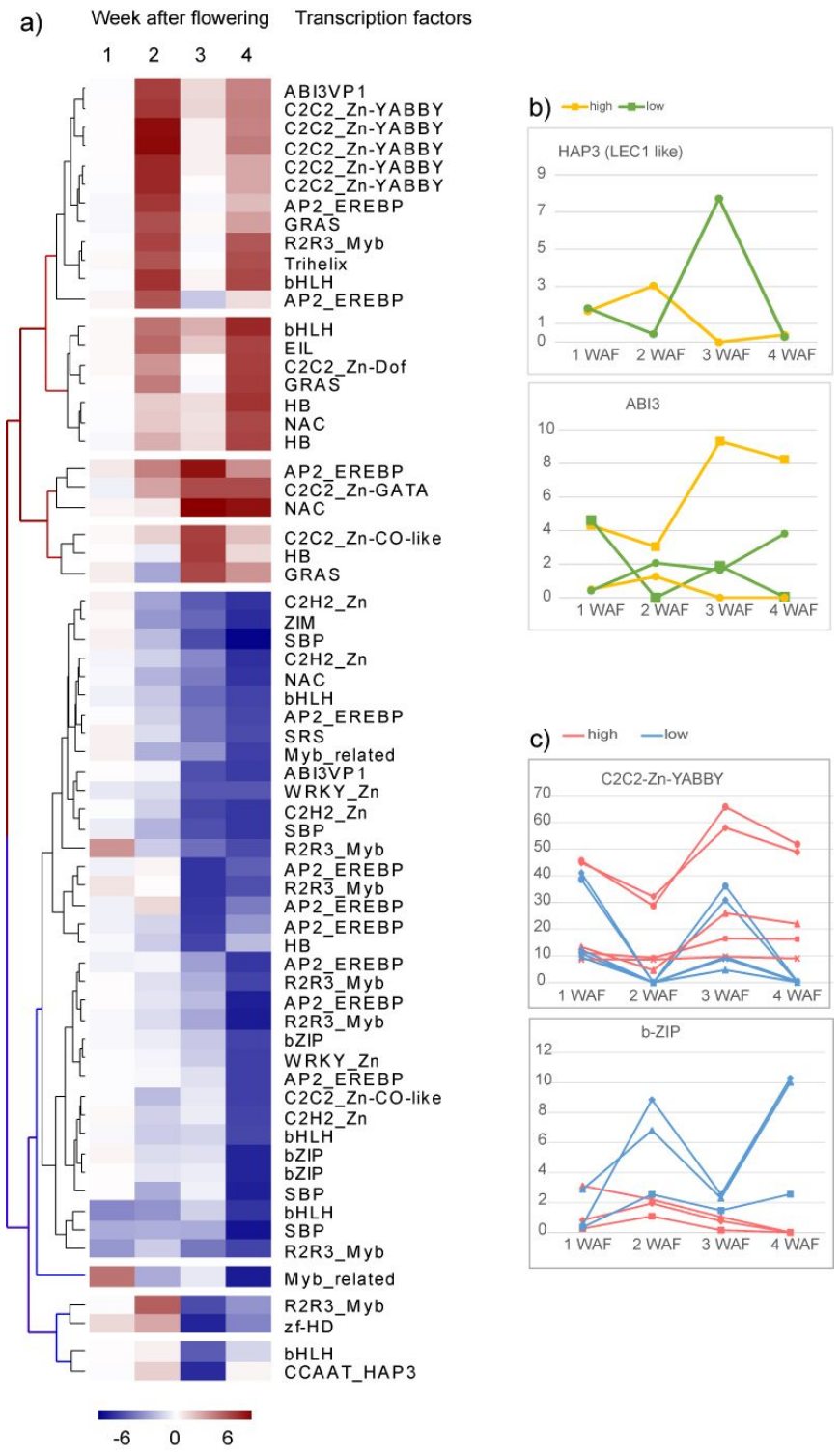

a)
Week after flowering
Transcription factors
1
2
3
4
ABI3VP1
C2C2_Zn-YABBY
C2C2_Zn-YABBY
C2C2_Zn-YABBY
C2C2_Zn-YABBY
C2C2_Zn-YABBY
AP2_EREBP
GRAS
R2R3_Myb
Trihelix
bHLH
AP2_EREBP
bHLH
EIL
C2C2_Zn-Dof
GRAS
HB
NAC
HB
AP2_EREBP
C2C2_Zn-GATA
NAC
C2C2_Zn-CO-like
HB
GRAS
C2H2_Zn
ZIM
SBP
C2H2_Zn
NAC
bHLH
AP2_EREBP
SRS
Myb_related
ABI3VP1
WRKY_Zn
C2H2_Zn
SBP
R2R3_Myb
AP2_EREBP
R2R3_Myb
AP2_EREBP
AP2_EREBP
HB
AP2_EREBP
R2R3_Myb
AP2_EREBP
R2R3_Myb
bZIP
WRKY_Zn
AP2_EREBP
C2C2_Zn-CO-like
C2H2_Zn
bHLH
bZIP
bZIP
SBP
bHLH
SBP
R2R3_Myb
Myb_related
R2R3_Myb
zf-HD
bHLH
CCAAT_HAP3
-6
0
6
b)
high
low
HAP3 (LEC1 like)
9
7
5
3
1
0
1 WAF
2 WAF
3 WAF
4 WAF
ABI3
10
8
6
4
2
0
1 WAF
2 WAF
3 WAF
4 WAF
c)
high
low
C2C2-Zn-YABBY
70
60
50
40
30
20
10
0
1 WAF
2 WAF
3 WAF
4 WAF
b-ZIP
12
10
8
6
4
2
0
1 WAF
2 WAF
3 WAF
4 WAF

Figure III-4 Heatmap of transcription factors and TF gene expression in Prot_high$_{DD}$ and Prot-low$_{DD}$. a) Heatmap of log_2 fold change value between Prot_high$_{DD}$ and Prot-low$_{DD}$. b) Gene expression of major seed development regulator ABI3 and LEC1-like genes. c) Gene expression of C2C2-Zn-YABBY and b-ZIP families that members showed similar fold change patterns.

G. max homologs of *A. thaliana* genes involved in synthesis of carbon precursor, protein, and oil

As described above, gene expression of storage protein is initiated at 14 days after flowering. We examined differently expressed genes in Prot_low$_{DD}$ between 1 and 2 WAF to understand initially expressed genes involve in seed storage reserve metabolism. Classification of the initial materials and end products of the metabolic pathways generated the following three pathways: carbon precursor, protein and oil synthesis. *A. thaliana* protein homologs involved in these three pathways were identified using the TAIR database (http://www.arabidopsis.org). Although one gene was involved in multiple pathways, we assigned each gene into a single pathway on the basis of its major function (Kim et al. 2013). We identified 445 members of 191 gene families in *A. thaliana.* The identified *A. thaliana* genes involved in carbon precursor synthesis, protein synthesis, and oil synthesis were 142, 210, and 93, respectively (Table III-2 and Table III-3). *G. max* had 1,012 orthologous counterparts of the *A. thaliana* genes, which belonged to 189 gene families. These *G. max* homologs were evenly distributed across all chromosomes; 319, 476, and 217 genes were involved in carbon precursor synthesis, protein synthesis, and oil synthesis, respectively (Figure III-5 and Table III-2).

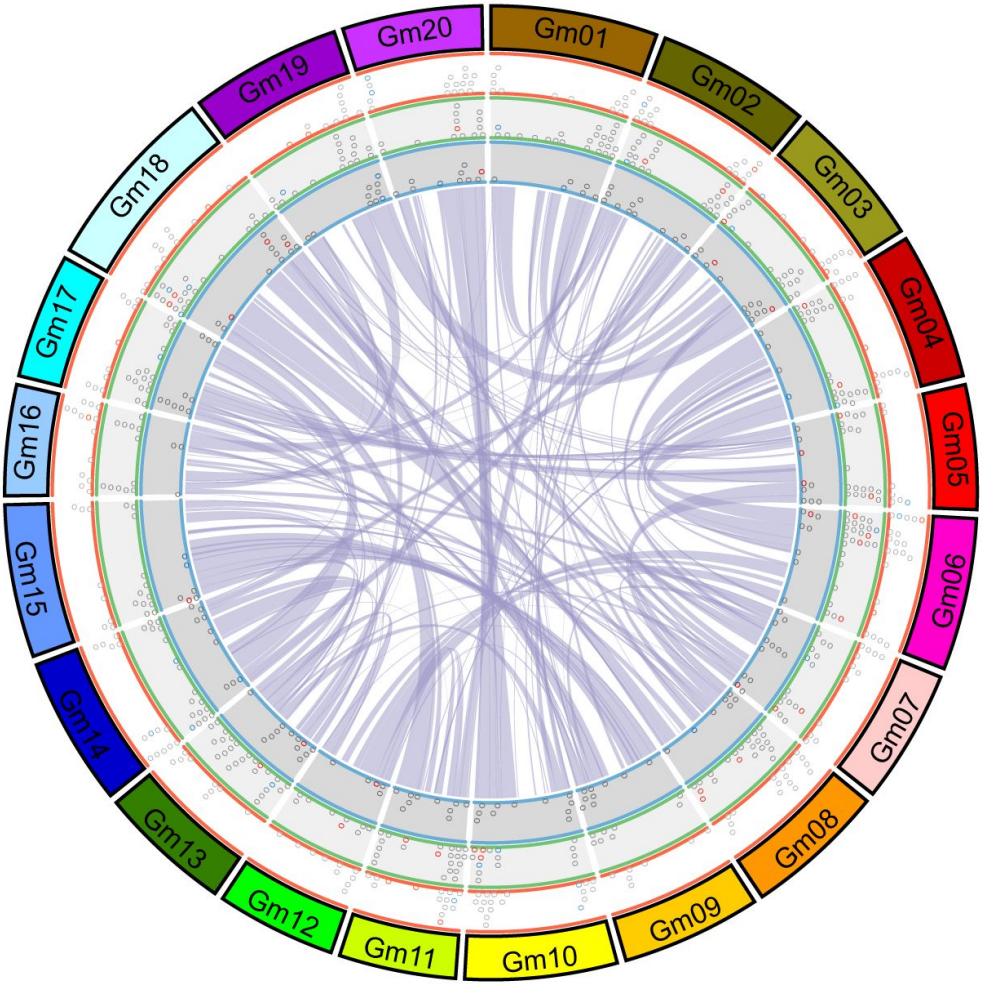


Figure III-5 Circos showing the distribution of expressed genes involved in seed protein and oil storage metabolism and their duplication patterns. The exterior circle indicates *Glycine max* chromosomes. The three inner circles are gene distributions involved in synthesis of carbon precursor (red), protein (green), and oil (blue). Genes that are expressed only at one stage are depicted in red (1 WAF) and blue (2 WAF).

Table III-4 Numbers of *Arabidopsis thaliana* genes and their *Glycine max* homologs involved in the accumulation of seed storage products, and the numbers of expressed genes during early seed-filling stages in *G. max.*

Pathway	No. of genes[a]		No. of expressed genes in *G. max*[b]		
	A. thaliana	***G. max***	**1 WAF**	**2 WAF**	**common**
Carbon precursor synthesis					
Sucrose transport	12	21	18	16	16
Sucrose degradation	41	113	76	77	72
Glycolysis	38	98	79	78	74
Acetyl-CoA biosynthesis	11	21	21	21	21
TCA cycle	40	66	53	51	50
Subtotal	142	319	247	243	233
Protein synthesis					
Ala biosynthesis	5	8	7	7	7
Arg, Gln, Glu, His, and Pro biosynthesis	53	145	116	108	107
Asn and Asp biosynthesis	15	34	30	28	28
Cys, Gly, and Ser biosynthesis	32	70	57	57	55
Ile, Leu, and Val biosynthesis	25	52	34	34	31
Lys, Thr, and Met biosynthesis	29	54	37	36	34
Phe, Trp, and Tyr biosynthesis	37	64	52	55	50
Transport of amino acid and protein	21	67	46	46	40
Subtotal	210	476	367	357	340
Oil synthesis					
Acyl-CoA hydrolysis	4	4	4	4	4
Fatty acid biosynthesis	27	57	45	41	39
Gamma-linolenate biosynthesis	2	9	8	8	8
Linoleate biosynthesis	21	63	46	48	46
Triacylglycerol biosynthesis	39	85	69	67	62
Subtotal	93	217	172	168	159
Total[2]	445	1012	786	768	732

Table III-5 The number of high copy member gene families in *Glycine max*

Gene	No. of genes in *G. max*	No. of genes in *A. thaliana*

and *Arabidopsis thaliana*.

Gene	No. of genes in *G. max*	No. of genes in *A. thaliana*
AMINO ACID PERMEASE	35	8
LONG-CHAIN ACYL-COA SYNTHETASE	25	8
GLYCEROL-3-PHOSPHATE ACYLTRANSFERASE	24	9
FATTY ACID DESATURASE	21	4
PHOSPHOENOLPYRUVATE CARBOXYLASE (PPC)	20	4
SERINE HYDROXYMETHYLTRANSFERASE (SHM)	19	7
PHOSPHOFRUCTOKINASE	19	7
GLYCERALDEHYDE 3-PHOSPHATE DEHYDROGENASE	19	7
SUCCINATE DEHYDROGENASE (SDH)	16	13
SUCROSE SYNTHASE	15	6
CELL WALL INVERTASE (CWIMV)	15	5
ASPARTATE AMINOTRANSFERASE (AAT)	15	6
GLUTAMINE SYNTHETASE (GLN)	14	6
ACONITASE (ACO)	14	3
PLASTIDIAL PYRUVATE KINASE	13	3
O-acyltransferase (WSD1-like) family protein	13	10
ENOLASE	13	3
VACULOLAR SORTING RECEPTOR	12	7
FRUCTOSE-BISPHOSPHATE ALDOLASE	12	7
SERINE ACETYLTRANSFERASE (SERAT)	11	5
METHIONINE SYNTHASE (MS)	11	3
HEXOKINASE	11	4
CYTOCHROME B5 ISOFORM (CB5)	11	3
BIDIRECTIONAL AMINO ACID TRANSPORTER	11	2
ALKALINE/NEUTRAL INVERTASE (A/N-Inv)	11	6
3-KETOACYL-COA SYNTHASE	11	5
2-ISOPROPYLMALATE SYNTHASE (IMS)	11	3
CHORISMATE MUTASE (CM)	10	3
PHOSPHOGLUCOMUTASE	9	4
LYSOPHOSPHATIDYL ACYLTRANSFERASE	9	4
BRANCHED-CHAIN AMINOTRANSFERASE (BCAT)	9	7

(Continue in next page)

arogenate dehydrogenase	9	2
aldehyde reductase (ADR)	9	2
ALANINE:GLYOXYLATE AMINOTRANSFERASE (AGT)	9	3
SUCROSE-PROTON SYMPORTER (SUC)	8	7
PEROXISOMAL NAD-MALATE DEHYDROGENASE (PMDH)	8	2
O-ACETYLSERINE (THIOL) LYASE (OAS)	8	4
GLUTAMATE DEHYDROGENASE (GDH)	8	3
CYSTEINE SYNTHASE (CS)	8	3
UDP-GLUCOSE PYROPHOSPHORYLASE	7	2
TRIOSEPHOSPHATE ISOMERASE (TIM)	7	2
DELTA 1-PYRROLINE-5-CARBOXYLATE SYNTHASE (P5CS2)	7	2
CARBAMOYL PHOSPHATE SYNTHETASE (CAR)	7	2
ANTHRANILATE SYNTHASE (ASA)	7	3

Table III-5 *(continued)*

Overall, the *G. max* genome contained approximately two times more homologous genes involved in the three biosynthesis pathways than *A. thaliana*, as expected from two rounds of WGD. Although WGD generated multiple copies of the genes in *G. max*, several genes were conserved as a single copy. We identified five single-copy genes in *G. max,* including *2,3-BIPHOSPHOGLYCERATE-INDEPENDENT PHOSPHOGLYCERATE MUTASE* (*PGAM*; Glyma18g45121) involved in glycolysis; *3-DEOXY-D-ARABINO-HEPTULOSONATE 7-PHOSPHATE SYNTHASE* (*DHS*; Glyma15g06020) and *TRYPTOPHAN BIOSYNTHESIS* (*TRP*; Glyma0435820) involved in Phe, Trp, and Tyr biosynthesis; and *ATP-PHOSPHORIBOSYL TRANSFERASE* (*ATP-PRT*; Glyma19G09080) and *HISTIDINOL-PHOSPHATE AMINOTRANSFERASE* (*HPA*; Glyma16g27220) involved in Arg, Gln, Glu, His, and Pro biosynthesis. There were no single-copy genes identified in oil synthesis. All identified single-copy genes were expressed at the early stages of seed maturation except *ATP-PRT* (Table III-4).

We examined the genes involved in seed storage product synthesis that have high copy numbers. Soybean genes were generally classified as high copy number genes when at least seven homologs were present, considering the two genome duplication events that occurred in *G. max* (Kim et al. 2013). A total of 182 soybean gene families were involved in the synthesis of carbon precursor, protein, and oil. Among these, 44 families

have more than seven members with a high copy number. *AMINO ACID PERMEASE*, which functions in amino acid transport, has the greatest number of copies at 35. The genes with the next-highest copy numbers were identified as *LONG-CHAIN ACYL-COA SYNTHETASE* (*LACS*) involved in linoleate biosynthesis, *GLYCEROL-3-PHOSPHATE ACYLTRANSFERASE* (*GPAT*) involved in triacylglycerol biosynthesis, *FATTY ACID DESATURASE* (*FAD*) involved in gamma-linolenate and linoleate biosynthesis, and *PHOSPHOENOLPYRUVATE CARBOXYLASE* (*PPC*) involved in Arg, Gln, Glu, His, and Pro biosynthesis (Table III-3 and Table III-4). These genes also were conserved as multiple copies in *A. thaliana* (Table III-4).

Among the 191 *Arabidopsis* gene families involved in the synthesis of carbon precursor, protein, and oil, nine gene families did not have orthologous counterparts in soybean. Most of these gene families were associated with protein synthesis (data not shown). These genes include *PHOSPHOGLYCERATE/BISPHOSPHOGLYCERATE MUTASE* (*PGM*) involved in glycolysis; *THREONINE ALDOLASE* (*THA*) family proteins involved in Cys, Gly, and Ser biosynthesis; *ISOPROPYLMALATE ISOMERASE* (*IPMI*) involved in Ile, Leu, and Val biosynthesis; aspartate semialdehyde dehydrogenase and pyridoxal-5-phosphate (PLP)–dependent genes involved in Lys, Thr, and Met biosynthesis; *DEFECTIVE* (*EMB*) *1144*, *PHOSPHORIBOSYLANTHRANILATE ISOMERASE* (*PAI*), and RNA 3'-

terminal phosphate cyclase/enolpyruvate transferase gene involved in Phe, Trp, and Tyr biosynthesis; and acyl-CoA thioesterase family protein gene involved in acyl-CoA hydrolysis (data not shown).

Differentially expressed genes in immature soybean seeds at 1 and 2 WAF

We analyzed the expression patterns of genes involved in the synthesis of carbon precursor, protein, and oil in early seed development stages at 1 and 2 WAF (Table III-3). In the 1 WAF immature soybean seed samples, 247, 367, and 172 genes involved in the synthesis of carbon precursor, protein, and oil were expressed in the Prot_low$_{DD}$ line, respectively (Figure III-5 and Table III-3). In 2 WAF samples, 243, 357, and 168 genes involved in the synthesis of carbon precursor, protein, and oil were identified. Most genes were expressed in both 1 and 2 WAF immature soybean seeds (Figure III-5 and Table III-3). The number of genes specifically expressed either at 1 or 2 WAF were 51 and 34, respectively, which is an FPKM value lower than 1.82 (Table III-5 and -6). Among the genes with an on-off expression pattern, five genes were identified as significantly differentially expressed genes (DEGs) in the 1 and 2 WAF samples (Table III-5 and -6). These five DEGs were homologs of *CELL*

WALL INVERTASE (*CWINV*; Glyma10g08670), *BRANCHED-CHAIN AMINOTRANSFERASE* (*BCAT*; Glyma07g30510), *AMINO ACID PERMEASE* (*AAP*; Glyma10g40130), *ALDEHYDE REDUCTASE (ADR*; Glyma20g37670), and *BIOTIN CARBOXYL CARRIER PROTEIN* (*BCCP*; Glyma13g06080).

We surveyed known QTLs controlling seed protein and oil traits that were co-localized within 3 Mb surrounding the five DEGs. Seven QTLs associated with seed protein and oil were detected near four genes of *CWINV* (Glyma10g08670), *AAP* (Glyma10g40130), *ADR* (Glyma20g37670), and *BCCP* (Glyma13g06080). *BCAT* (Glyma07g30510) did not overlap with any QTL for seed protein and oil traits (Table III-7). QTLs associated with seed oil, seed linoleic acid, and seed oleic acid contents were detected on Chrs 10, 13, and 20 (Table III-7).

Table III-6 Annotation of genes that were expressed only in 1 WAF soybean seeds.

Pathway	*A. thaliana* ID	Annotation	*G. max* ID	FPKM at 2 WAF
Sucrose degradation	At1g70730	*PHOSPHOGLUCOMUTASE*	Glyma03g05135	0.120376
Sucrose degradation	**At2g36190**	***CELL WALL INVERTASE***	**Glyma10g08670**	**9.60033**
Sucrose degradation	At3g13790	*CELL WALL INVERTASE*	Glyma20g03640	0.418039
Sucrose degradation	At4g34860	*ALKALINE/NEUTRAL INVERTASE*	Glyma06g00770	0.061067
Sucrose degradation	At5g11920	*6-&1-FRUCTAN EXOHYDROLASE*	Glyma20g03580	0.095802
Glycolysis	At1g32440	*PLASTIDIAL PYRUVATE KINASE*	Glyma11g04490	0.09715
Glycolysis	At2g36460	*FRUCTOSE-BISPHOSPHATE ALDOLASE*	Glyma10g07710	0.072073
Glycolysis	At4g04040	*MATERNAL EFFECT EMBRYO ARREST 51*	Glyma14g10971	0.164091
Glycolysis	At4g26270	*PHOSPHOFRUCTOKINASE*	Glyma05g36050	0.029784
TCA cycle	At5g50950	*FUMARASE*	Glyma02g01930	0.02739
Arg, Gln, Glu, His and Pro biosynthesis	At4g26900	*HIS HF*	Glyma18g12940	0.077586
Cys, Gly and Ser biosynthesis	At3g08860	*PYRIMIDINE 4*	Glyma03g04990	0.032411
Cys, Gly and Ser biosynthesis	At3g13110	*SERINE ACETYLTRANSFERASE*	Glyma14g01840	0.038101
Ile, Leu and Val biosynthesis	At1g70560	*TRYPTOPHAN AMINOTRANSFERASE OF ARABIDOPSIS 1*	Glyma01g03340	0.401044
Ile, Leu and Val biosynthesis	At1g70560	*TRYPTOPHAN AMINOTRANSFERASE OF ARABIDOPSIS 1*	Glyma02g04270	1.37185
Ile, Leu and Val biosynthesis	At1g74040	*2-ISOPROPYLMALATE SYNTHASE*	Glyma13g12484	0.031744
Lys, Thr and Met biosynthesis	At4g19710	*ASPARTATE KINASE-HOMOSERINE DEHYDROGENASE*	Glyma18g00600	0.029757
Lys, Thr and Met biosynthesis	At5g17920	*METHIONINE SYNTHASE*	Glyma17g23730	0.264116
Phe, Trp and Tyr biosynthesis	At3g06350	*MATERNAL EFFECT EMBRYO ARREST 32*	Glyma10g29970	0.127876
Phe, Trp and Tyr biosynthesis	At5g34930	arogenate dehydrogenase family	Glyma13g06340	0.041124
Phe, Trp and Tyr biosynthesis	At5g38530	*TRYPTOPHAN SYNTHASE BETA TYPE 2*	Glyma19g05630	0.077377
Transport of amino acid and protein	At1g58360	*AMINO ACID PERMEASE*	Glyma06g09280	0.26971
Transport of amino acid and protein	At1g77380	*AMINO ACID PERMEASE*	Glyma04g42520	0.039667
Transport of amino acid and protein	At1g77380	*AMINO ACID PERMEASE*	Glyma14g24370	0.053156
Transport of amino acid and protein	At3g52850	*VACULOLAR SORTING RECEPTOR*	Glyma07g14800	0.025593
Transport of amino acid and protein	At5g23810	*AMINO ACID PERMEASE*	Glyma18g07980	0.625543
Transport of amino acid and protein	**At5g49630**	***AMINO ACID PERMEASE***	**Glyma10g40130**	**4.15959**

(Continue in next page)

Pathway	*A. thaliana* ID	Annotation	*G. max* ID	FPKM at 2 WAF
Linoleate biosynthesis	At3g12120	*FATTY ACID DESATURASE*	Glyma15g23200	0.108045
Linoleate biosynthesis	At4g11030	AMP-dependent synthetase and ligase family protein family	Glyma04g32720	0.062235
Triacylglycerol biosynthesis	At1g02390	*GLYCEROL-3-PHOSPHATE ACYLTRANSFERASE*	Glyma19g40590	0.02798
Triacylglycerol biosynthesis	At2g38110	*GLYCEROL-3-PHOSPHATE ACYLTRANSFERASE*	Glyma01g27900	0.187259
Triacylglycerol biosynthesis	At2g38110	*GLYCEROL-3-PHOSPHATE ACYLTRANSFERASE*	Glyma20g16980	0.045681
Triacylglycerol biosynthesis	At3g02600	*LIPID PHOSPHATE PHOSPHATASE*	Glyma08g31706	0.686773
Triacylglycerol biosynthesis	At3g11430	*GLYCEROL-3-PHOSPHATE ACYLTRANSFERASE*	Glyma14g07290	0.043349
Fatty acid biosynthesis	At1g54870	aldehyde reductase family	Glyma19g42730	0.041634
Fatty acid biosynthesis	At1g74960	*FATTY ACID BIOSYNTHESIS*	Glyma15g20045	0.430114

Table III-6 *(continued)*

Table III-7 Annotation of genes that were expressed only in 2 WAF soybean seeds.

Pathway	*A. thaliana* ID	Annotation	*G. max* ID	FPKM at 2 WAF
Sucrose degradation	At1g70730	*PHOSPHOGLUCOMUTASE*	Glyma03g05135	0.120376
Sucrose degradation	**At2g36190**	***CELL WALL INVERTASE***	**Glyma10g08670**	**9.600330**
Sucrose degradation	At3g13790	*CELL WALL INVERTASE*	Glyma20g03640	0.418039
Sucrose degradation	At4g34860	*ALKALINE/NEUTRAL INVERTASE*	Glyma06g00770	0.061067
Sucrose degradation	At5g11920	*6-&1-FRUCTAN EXOHYDROLASE*	Glyma20g03580	0.095802
Glycolysis	At1g32440	*PLASTIDIAL PYRUVATE KINASE*	Glyma11g04490	0.09715
Glycolysis	At2g36460	*FRUCTOSE-BISPHOSPHATE ALDOLASE*	Glyma10g07710	0.072073
Glycolysis	At4g04040	*MATERNAL EFFECT EMBRYO ARREST 51*	Glyma14g10971	0.164091
Glycolysis	At4g26270	*PHOSPHOFRUCTOKINASE*	Glyma05g36050	0.029784
TCA cycle	At5g50950	*FUMARASE*	Glyma02g01930	0.02739
Arg, Gln, Glu, His and Pro biosynthesis	At4g26900	*HIS HF*	Glyma18g12940	0.077586
Cys, Gly and Ser biosynthesis	At3g08860	*PYRIMIDINE 4*	Glyma03g04990	0.032411
Cys, Gly and Ser biosynthesis	At3g13110	*SERINE ACETYLTRANSFERASE*	Glyma14g01840	0.038101
Ile, Leu and Val biosynthesis	At1g70560	*TRYPTOPHAN AMINOTRANSFERASE OF ARABIDOPSIS 1*	Glyma01g03340	0.401044
Ile, Leu and Val biosynthesis	At1g70560	*TRYPTOPHAN AMINOTRANSFERASE OF ARABIDOPSIS 1*	Glyma02g04270	1.37185
Ile, Leu and Val biosynthesis	At1g74040	*2-ISOPROPYLMALATE SYNTHASE*	Glyma13g12484	0.031744
Lys, Thr and Met biosynthesis	At4g19710	*ASPARTATE KINASE-HOMOSERINE DEHYDROGENASE*	Glyma18g00600	0.029757
Lys, Thr and Met biosynthesis	At5g17920	*METHIONINE SYNTHASE*	Glyma17g23730	0.264116
Phe, Trp and Tyr biosynthesis	At3g06350	*MATERNAL EFFECT EMBRYO ARREST 32*	Glyma10g29970	0.127876
Phe, Trp and Tyr biosynthesis	At5g34930	arogenate dehydrogenase family	Glyma13g06340	0.041124
Phe, Trp and Tyr biosynthesis	At5g38530	*TRYPTOPHAN SYNTHASE BETA TYPE 2*	Glyma19g05630	0.077377
Transport of amino acid and protein	At1g58360	*AMINO ACID PERMEASE*	Glyma06g09280	0.26971
Transport of amino acid and protein	At1g77380	*AMINO ACID PERMEASE*	Glyma04g42520	0.039667
Transport of amino acid and protein	At1g77380	*AMINO ACID PERMEASE*	Glyma14g24370	0.053156
Transport of amino acid and protein	At3g52850	*VACULOLAR SORTING RECEPTOR*	Glyma07g14800	0.025593
Transport of amino acid and protein	At5g23810	*AMINO ACID PERMEASE*	Glyma18g07980	0.625543
Transport of amino acid and protein	**At5g49630**	***AMINO ACID PERMEASE***	**Glyma10g40130**	**4.15959**

(Continue in next page)

Table III-7 *(continued)*

Pathway	*A. thaliana* ID	Annotation	*G. max* ID	FPKM at 2 WAF
Linoleate biosynthesis	At3g12120	*FATTY ACID DESATURASE*	Glyma15g23200	0.108045
Linoleate biosynthesis	At4g11030	AMP-dependent synthetase and ligase family protein family	Glyma04g32720	0.062235
Triacylglycerol biosynthesis	At1g02390	*GLYCEROL-3-PHOSPHATE ACYLTRANSFERASE*	Glyma19g40590	0.02798
Triacylglycerol biosynthesis	At2g38110	*GLYCEROL-3-PHOSPHATE ACYLTRANSFERASE*	Glyma01g27900	0.187259
Triacylglycerol biosynthesis	At2g38110	*GLYCEROL-3-PHOSPHATE ACYLTRANSFERASE*	Glyma20g16980	0.045681
Triacylglycerol biosynthesis	At3g02600	*LIPID PHOSPHATE PHOSPHATASE*	Glyma08g31706	0.686773
Triacylglycerol biosynthesis	At3g11430	*GLYCEROL-3-PHOSPHATE ACYLTRANSFERASE*	Glyma14g07290	0.043349
Fatty acid biosynthesis	At1g54870	aldehyde reductase family	Glyma19g42730	0.041634
Fatty acid biosynthesis	At1g74960	*FATTY ACID BIOSYNTHESIS*	Glyma15g20045	0.430114

Table III-8 Five *Glycine max* genes and their homologs that are differentially expressed at 1 and 2 WAF; QTLs linked to seed protein and oil synthesis were detected within a 3 Mb region surrounding the gene location.

Annotation	Gene ID	FPKM[1]		Reported QTL
		1 WAF[2]	2 WAF	
Cell wall invertase	Glyma10g08670	0	9.60	Seed oil 19-3 (Panthee et al. 2005) Seed Glu 1-7 (Panthee et al. 2006B)
Branched-chain amino acid transaminase	Glyma07g30510	4.30	0	-
Amino acid permease	Glyma10g40130	0	4.16	Seed oil 29-3, Seed linoleic 6-7 (Li et al. 2011) Seed oleic 6-10 (Bachlava et al. 2009)
Aldehyde reductase	Glyma20g37670	145.76	0	Seed oleic 6-4 and Seed linoleic 6-4 (Bachlava et al. 2009)
Biotin carboxyl carrier protein	Glyma13g06080	7.16	0	Seed oil 24-4 (Qi et al. 2011) Seed Ala 1-4, Seed Val 1-3, Seed Cys 1-2, and Seed Met 1-1 (Panthee et al. 2006B) Seed Met plus Cys 1-2 (Phantee et al., 2006A)

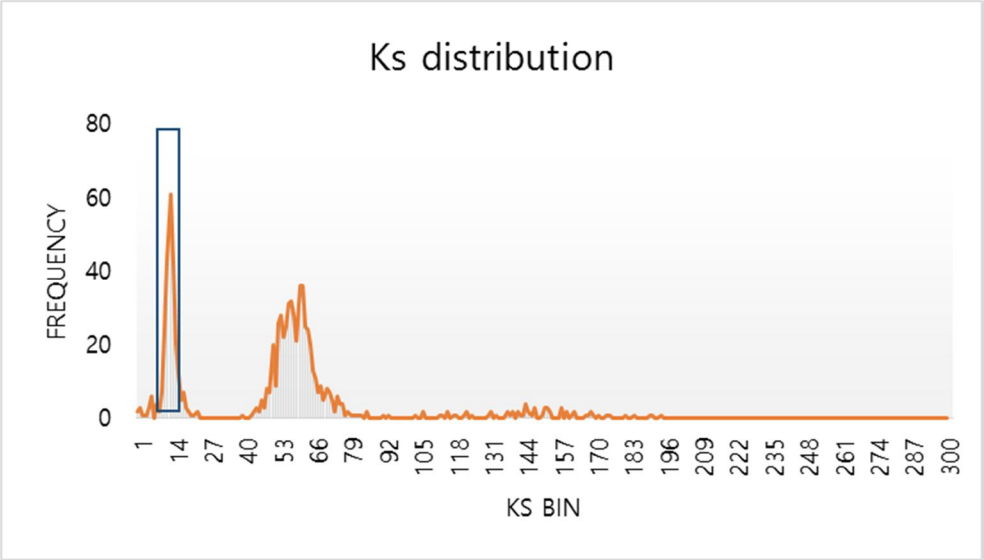
[1]FPKM, fragments per kilobase of exon per million fragments mapped.

[2]WAF, weeks after flowering

Expression patterns of paralogous soybean genes involved in the synthesis of carbon precursor, protein, and oil

In the *G. max* genome, 825 pairs of synteny blocks were identified. The distribution of K_s values of each pair of the duplicated genes exhibited two distinct peaks, which represented two rounds of WGD in soybean (Figure III-6). The medians of the first and second K_s peaks ranged from approximately 0.10–0.15 and 0.50–0.67, respectively, which identifies the recent and ancient WGD events. Four synteny blocks at the K_s peak of 0.10–0.15 contained at least three paralogous pairs of genes involved in synthesis of carbon precursor, protein, or oil, which were assumed to be derived from a recent WGD. These synteny blocks showed duplication co-linearity between chromosomes (Figure III-7). Based on the transcriptome data of early immature seeds in the low-protein line, we identified expression pattern differences between duplicated genes in pairs of synteny blocks generated by the recent WGD.

Figure III-6 Ks distributions of *G. max* synteny block



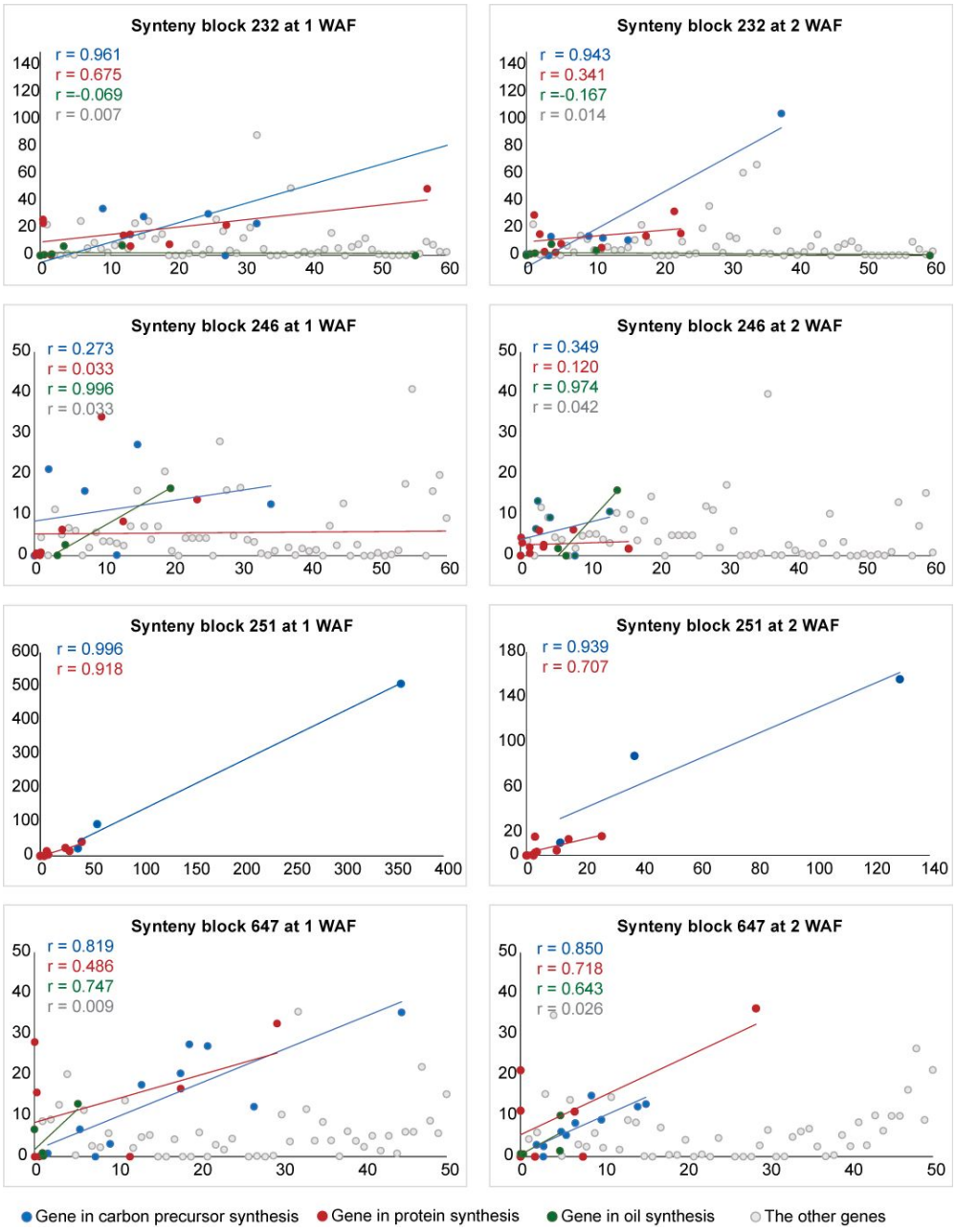

Synteny block 232 at 1 WAF
r = 0.961
r = 0.675
r =-0.069
r = 0.007
Synteny block 232 at 2 WAF
r = 0.943
r = 0.341
r =-0.167
r = 0.014
Synteny block 246 at 1 WAF
r = 0.273
r = 0.033
r = 0.996
r = 0.033
Synteny block 246 at 2 WAF
r = 0.349
r = 0.120
r = 0.974
r = 0.042
Synteny block 251 at 1 WAF
r = 0.996
r = 0.918
Synteny block 251 at 2 WAF
r = 0.939
r = 0.707
Synteny block 647 at 1 WAF
r = 0.819
r = 0.486
r = 0.747
r = 0.009
Synteny block 647 at 2 WAF
r = 0.850
r = 0.718
r = 0.643
r = 0.026
Gene in carbon precursor synthesis
Gene in protein synthesis
Gene in oil synthesis
The other genes

Figure III-7 Fragments per kilobase of exon per million fragments mapped (FPKM) expression patterns of a paralogous gene pair in different synteny blocks at different times during seed development. Colored dots indicate the FPKM expression value of duplicated genes involved in each of the three metabolic pathways (carbon precursor, protein, and oil synthesis); trend lines were drawn using linear regressions.

Within four pairs of the synteny blocks, there were 80 paralogous gene pairs, among which 28, 37, and 15 were involved in the synthesis of carbon precursor, protein, and oil, respectively. The Pearson correlation coefficient (*r*) was considered to be moderate when the values ranged over 0.3–0.6 and strong when $r > 0.6$. Within the four pairs of synteny blocks, paralogous gene pairs involved in the synthesis of seed storage products had stronger correlations than other gene pairs (Figure III-7). In synteny block 232, the gene pairs of carbon precursor and protein synthesis had strong and moderate positive correlations at both 1 and 2 WAF, respectively (Figure III-7). The paralogous genes participating in the oil synthesis pathway likely diverged functionally after the recent WGD because synteny block 232 had a negative correlation, but synteny blocks 246 and 647 had positive correlations. In immature seeds at 1 and 2 WAF, the duplicated genes involved in the synthesis of carbon precursor and protein showed strong or moderate correlations in mRNA expression in all synteny block pairs, indicating functional conservation after the recent WGD (Figure III-7).

DISCUSSION

In this study, we extended our investigation on soybean seed accumulation by transcriptome analysis. We classified DEGs into functional categorize and clustered expression patterns of genes involved in storage reserves metabolic pathway and transcription factors between $Prot\_high_{DD}$ and $Prot\_low_{DD}$ during seed maturation from 1 to 4 WAF. Although previous studies investigated whole-transcript profiling in developing soybean seeds (Jones and Vodkin 2013, Severin et al. 2010), very little research has focused on genes involved in the synthesis of carbon precursors, protein, and oil. The accumulation of seed storage products to vacuole complex is started with maturation stage (Le et al. 2007). The metabolic pathways for synthesize storage products are regulated by complex transcriptional network that specified to their developmental stage. (Weber et al. 2010). We found differentially expressed transcription factors that are major regulator of seed development and genes that involved in seed storage prodect metablic pathways. There were more up-regulated genes in $Prot\_high_{DD}$ than up-regulated genes at $Prot\_low_{DD}$ which indicates gene expression regulation in early seed maturation is important in soybean seed reserve accumulation (Figure III-2). From clustering analysis according to gene regulation change between $Prot\_high_{DD}$ and $Prot\_low_{DD}$, gene involved in

metabolic pathways had more hp-regulated genes at 2 WAF in Prot_high$_{DD}$ than transcription factors (Figure III-3 and -4).

The metabolic pathway of seed reserves such as protein, oil and starched are depend on their carbon precursor synthesis, and sucrose is the major assimilated carbon (Baud et al. 2008). We identified the genes of had opposite expression patterns that *SUS* was up-regulated in Prot_high$_{DD}$ and two *CWINV* were up-regulated in Prot_low$_{DD}$. SUS and INV degrade sucrose into hexoses. SUS produce fructose and UDP glucose from sucrose degradation. INV also produce fructose from sucrose, but INV produce glucose instead of UDP glucose. Cell wall invertase (CWINV) in insoluble invertase that expending sink tissue and assimilate unloading by the concentration gradient of sucrose (Weber et al. 2005). INV can produce twice hexose than SUS, INV does not regulate actively in condition of low oxigen, such as developing seed (Koch 2004). Transcription level of SUS is correlated with starch accumulation and protein storage regulation gene expression (Weber et al. 2010). In *Arabidopsis*, regulator transcription factors including LEC1 and 2, FUS3. ABI3 are involved in seed development (Vicente-Carbajosa and Carbonero 2005). LEC1 induces expression of FUS1 and ABI3 which proceed seed storage protein synthesis by interaction with seed storage protein gene (Kagaya et al. 2005). We found five C2C2(Zn)-YABBY genes were up-regulated in Prot_high$_{DD}$ and three b-ZIP were up-

regulated in Prot_low$_{DD}$. The domain of b-ZIP transcription factors can also interact to seed storage protein genes and induce expression (Vicente-Carbajosa and Carbonero 2005). C2C2(Zn)-YABBY has important role for cell fate and organ development and function of these gene family in *Arabidopsis* are highly conserve that all YABBY genes have similar roles (Golz and Hudson 1999).

We attempted to identify the orthologous counterparts of *A. thaliana* genes involved in the synthesis of seed storage proteins and oil in *G. max* using the currently available genome data. We investigated expression changes of these genes during early immature seed development at 1 and 2 WAF, and characterized the paralogous gene expression patterns on the pairs of synteny blocks. Although previous studies investigated whole-transcript profiling in developing soybean seeds (Severin et al. 2010, Jones and Vodkin 2013), very little research has focused on genes involved in the synthesis of carbon precursors, protein, and oil.

After the increase of gene copy number by large-scale WGD or small-scale duplication (SSD), most organisms control gene expression levels against the dosage imbalance (De Smet et al.2013). *G. max* is a diplodized paleopolyploid, in which many genes exist as multiple copies (Schmutz et al. 2010). However, five genes were found to be conserved as a single copy: *PGAM*, *DHS*, *TRP*, *ATP-PRT*, and *HPA*. *PGAM*

encodes phosphoglycerate mutase, which is a key enzyme catalyzing the reversible interconversion of 3-phosphoglycerate and 2-phosphoglycerate during sugar metabolism in the glycolysis pathway (Mazarei et al. 2003). *DHS* and *TRP* are active in the Phe, Trp, and Try synthesis pathway. *DHS* encodes 3-deoxy-D-arabino-heptulosonate 7-phosphate synthase, which is the first enzyme in the shikimate pathway. In higher plants, this pathway provides chorismate as the precursor of three aromatic amino acids, Phe, Tyr, and Trp (Zhang et al. 2011). The *TRP* protein, phosphoribosyl anthranilate, catalyzes the second committed step of Trp biosynthesis from chorismate (Rose et al. 1997). *ATP-PRT* and *HPA* are active in His metabolism. Phosphoribosyl transferase encoded by *ATP-PRT* catalyzes the first committed step of His biosynthesis. This enzyme is feedback-inhibited by the end product (L-His) of this pathway (Winkler 1987, Alifano et al. 1996, Ohta et al. 2000). *HPA* encodes histidinol-phosphate aminotransferase, which catalyzes the formation of L-histidinol phosphate from imidazole-acetol phosphate and glutamate in the histidine biosynthesis pathway (Zhang et al. 2014).

Single-copy genes might be sensitive to dosage balance, that is, it is important that they are maintained in the correct relative dose because an increase in copy number might unbalance their interactions with other proteins (De Smet et al. 2013). Genomic-scale studies comparing single-copy genes of various plants with *A. thaliana* indicate

that single-copy genes commonly have features of housekeeping functions, such as protein catabolism and synthesis, RNA processing, and DNA repair (Armisén et al 2008, De Smet et al 2013). Our results suggest that some duplicated genes with protein catabolic and synthetic activities are quickly fractionated after duplication because the genes are sensitive to dosage balance, which results in single-copy genes.

We detected 44 gene families with high copy number members, including genes involved in transport and oil synthesis. The highest copy number gene (*AAP*) encodes amino acid permease. This protein transports amino acids into the cell, and high-affinity with ligand is important to maintain the amino acids and to prevent leakage out of the cell (Fischer et al. 1998). *LACS*, *GPAT*, and *FAD* are involved in fatty acid synthesis. Acyl-CoA synthetase is essential for *de novo* fatty acid catabolism by producing acyl-CoA. *LACS* encodes acyl-CoA synthetase, which specifically activates fatty acid metabolism of chain lengths 12 to 20 carbon atoms (Soupene and Kuypers 2008). LACS activity is important for the synthesis of soybean major fatty acids such as linoleic acid (C-18:3) and oleic acid (C-18:1) (Scrimgeour 2005). *GPAT* encodes glycerol-3-phosphate 1-O-acyltransferase, which catalyzes the first committed step in the pathway for *de novo* synthesis of membrane and storage lipids. The enzyme transfers an acyl group from acyl-CoA or acyl-ACP to the *sn*-1 position of *sn*-glycerol-3-phosphate (Yang et al. 2012).

FAD encodes fatty acid desaturase, which converts a single bond between two carbon atoms to a double bond in a fatty acyl chain. The resultant unsaturated fatty acid is essential for membrane function and lipid storage (Los and Murata 1998).

We analyzed alterations in gene expression patterns of immature soybean seeds at 1 and 2 WAF because induction of seed protein gene expression was reported to occur at 2 WAF (Mienke et al. 1981). We identified five DEGs of *G. max*: Glyma07g30510, Glyma10g40130, Glyma10g08670, Glyma13g06080, and Glyma20g37670. Glyma10g08670 encodes cell wall invertase (CWINV), which binds to the cell wall and catalyzes the cleavage of transported sucrose (Weber et al. 1997). In maize (*Zea mays* L.), CWINV controls the levels of seed glucose and fructose, which likely influences the expression of genes involved in seed filling (Chourey et al. 2012). Glyma07g30510 is a homolog of *BCAT*, which encodes branched-chain amino acid transaminases; these enzymes catalyze the last step of synthesis and/or the initial step of degradation of branched-chain amino acids (BCAA) such as Leu, Ile, and Val (Diebold et al. 2002). A previous study using *BCAT*-null *A. thaliana* mutants reported that BCAT contributes to the natural variation of BCAA levels, glutamate recycling, and free amino acid homeostasis in seeds (Angelovici et al. 2013). Glyma10g40130 is a member of the *AMINO ACID PERMEASE* homologs described above.

Glyma20g37670 is a homolog of *ALDEHYDE REDUCTASE* (*ADR*). This enzyme catalyzes the *in vitro* reduction of carbonyl groups on saturated aldehydes and on alpha- and beta-unsaturated aldehydes with five or more carbons; its function *in planta* has not been extensively investigated (Yamauchi et al. 2011). Glyma13g06080 encodes biotin carboxyl carrier protein (BCCP), which catalyzes the irreversible carboxylation of acetyl-CoA to produce malonyl-CoA in fatty acid biosynthesis. In *A. thaliana*, overexpression of BCCP reduces the fatty acid content and elevates the protein content in seeds (Chen et al. 2009). We also determined that *CWINV*, *AAP*, *ADR*, and *BCCP* were co-localized with QTLs associated with seed protein and oil content traits including oil, linoleic acid, oleic acid, Ala, Cyc, Glu, Met and Val (Phanthee 2005, Phanthee 2006A, Phanthee 2006B, Bachlava et al 2009, Li et al 2011, Qi et al 2011). These results suggest that the identified DEGs likely affect seed protein and oil accumulation, and may be important in the control of soybean seed protein and oil contents.

Transcriptome analysis of immature soybean seeds at 1 and 2 WAF showed that paralogous genes involved in the synthesis of carbon precursor and protein, which are located on synteny block pairs, have stronger positive correlations between their expression patterns than those of oil synthesis genes. Duplicated genes have a short lifespan during which they eventually become non-functionalized, neo-

functionalized, and sub-functionalized (Lynch et al. 2000). In this study, the gene duplications involved in protein synthesis were expressed during seed maturation, which likely indicates the maintenance of gene function after WGD. Genes involved in oil synthesis showed a negative correlation at 1 and 2 WAF in synteny block 232. *GPAT*, *FAD*, *LACS*, and *ADR* are oil synthesis genes on this block; they are multicopy gene families, and some members of these families are not expressed in immature soybean seeds. Gene duplication also elevates gene functions and induces genetic redundancy, which may lead to phenotypic changes. Mutation of these genes rapidly occurred, and genes would not be expressed (non-functionalization) or expression would be reduced (sub-functionalization) (Lynch et al. 2000).

The basic pathways of seed storage product synthesis are likely to be relatively well conserved between *A. thaliana* and *G. max.* Thus, a candidate gene approach may be a powerful first step to verifying the molecular nature of genes involved in synthetic pathways of carbon precursor, protein, and oil. Transcript profiling of these homologous genes can identify additional key enzymes affecting soybean seed storage proteins and oils. Protein synthesis-related genes duplicated by WGD likely retain their function during soybean seed maturation. The results of this study provide a deeper understanding of seed storage pathways and the evolution of protein and oil metabolic genes in soybean.

REFERENCES

Alifano P, Fani R, Liò P, Lazcano A, Bazzicalupo M, Carlomagno MS, Bruni CB (1996) Histidine biosynthetic pathway and genes: structure, regulation, and evolution. Microbiol Rev 60(1): 44

Angeles-Núñez JG, Tiessen A (2012) Regulation of *AtSUS2* and *AtSUS3* by glucose and the transcription factor *LEC2* in different tissues and at different stages of *Arabidopsis* seed development. Plant Mol Biol 78:377-392

Angelovici R, Lipka AE, Deason N, Gonzalez-Jorge S, Lin H, Cepela J, Buell R, Gore MA, DellaPenna D (2013) Genome-wide analysis of branched-chain amino acid levels in *Arabidopsis* seeds. Plant Cell 25(12): 4827-4843

Armisén D, Lecharny A, Aubourg S (2008) Unique genes in plants: specificities and conserved features throughout evolution. BMC Evol Biol 8(1): 280

Bachlava E, Dewey RE, Burton JW, Cardinal AJ (2009) Mapping and comparison of quantitative trait loci for oleic acid seed content in two segregating soybean populations. Crop Sci 49(2): 433-442

Baud S, Dubreucq B, Miquel M, Rochat C, Lepiniec L (2008) Storage reserve accumulation in *Arabidopsis*: metabolic and developmental

control of seed filling. The *Arabidopsis* book/American Society of Plant Biologists 6 doi: 10.1199/tab.0113

Baud S, Lepiniec L (2010) Physiological and developmental regulation of seed oil production. Prog Lipid Res 49(3): 235-249

Bolon YT, Joseph B, Cannon SB, Graham MA, Diers BW, Farmer AD, May GD, Muehlbauer GJ, Specht JE, Tu ZJ, Weeks N, Xu WW, Shoemaker RC, Vance CP (2010) Complementary genetic and genomic approaches help characterize the linkage group I seed protein QTL in soybean. BMC Plant Biol 10(1): 41

Brummer EC, Graef GL, Orf J, Wilcox JR, Shoemaker RC (1997) Mapping QTL for seed protein and oil content in eight soybean populations. Crop Sci 37(2): 370-378

Burton JW (1985) Breeding soybeans for improved protein quantity and quality. In: Shibles R (ed) Proc 3rd World Soybean Res. Conf. Westview Press, Boulder/CO, pp 361- 367

Chen M, Mooney BP, Hajduch M, Joshi T, Zhou M, Xu D, Thelen JJ (2009) System analysis of an *Arabidopsis* mutant altered in de novo fatty acid synthesis reveals diverse changes in seed composition and metabolism. Plant Physiol 150(1): 27-41

Chourey PS, Li QB, Cevallos-Cevallos J (2012) Pleiotropy and its dissection

through a metabolic gene *Miniature1* (*Mn1*) that encodes a cell wall invertase in developing seeds of maize. Plant Sci 184: 45-53.

Chung J., Babka HL, Graef GL, Staswick PE, Lee DJ, Cregan PB, Shoemaker RC, Specht JE (2003) The seed protein, oil, and yield QTL on soybean linkage group I. Crop Sci 43(3): 1053-1067

Clemente TE, Cahoon EB (2009) Soybean oil: genetic approaches for modification of functionality and total content. Plant Physiol 151(3): 1030-1040

De Smet R, Adams KL, Vandepoele K, Van Montagu MC, Maere S, Van de Peer Y (2013) Convergent gene loss following gene and genome duplications creates single-copy families in flowering plants. Proc Natl Acad Sci U S A 110(8): 2898-2903

Diebold R, Schuster J, Däschner K, Binder S (2002) The branched-chain amino acid transaminase gene family in *Arabidopsis* encodes plastid and mitochondrial proteins. Plant Physiol 129(2): 540-550

Dornbos DL Jr, Mullen RE (1991) Influence of stress during soybean seed fill on seed weight, germination, and seedling growth rate. Can J Plant Sci 71(2): 373-383

Fischer WN, André B, Rentsch D, Krolkiewicz S, Tegeder M, Breitkreuz K, Frommer WB (1998) Amino acid transport in plants. Trends Plant Sci

3(5): 188-195.

Golz JF, Hudson A (1999) Plant development: YABBYs claw to the fore. Curr Biol 9:R861-R863

Hymowitz T, Collins FI, Panczner J, Walker WM (1972) Relationship between the content of oil, protein, and sugar in soybean seed. Agron J 64(5) 613-616

Jones SI, Vodkin LO (2013) Using RNA-Seq to profile soybean seed development from fertilization to maturity. PLoS one 8:e59270

Kagaya Y, Toyoshima R, Okuda R, Usui H, Yamamoto A, Hattori T (2005) LEAFY COTYLEDON1 controls seed storage protein genes through its regulation of FUSCA3 and ABSCISIC ACID INSENSITIVE3. Plant Cell Physiol 46:399-406

Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL (2013) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol 14(4): R36

Kim MY, Kang YJ, Lee T, Lee S-H (2013) Divergence of Flowering-Related Genes in Three Legume Species. Plant Genome 6(3) doi:10.3835/plantgenome2013.03.0008

Koch K (2004) Sucrose metabolism: regulatory mechanisms and pivotal roles in sugar sensing and plant development. Curr Opin Plant Biol

7:235-246

Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. Nat Methods 9(4): 357-359.

Le BH, Wagmaister JA, Kawashima T, Bui AQ, Harada JJ, Goldberg RB (2007) Using genomics to study legume seed development. Plant Physiol 144:562-574

Lee SH, Bailey MA, Mian MAR, Carter TE Jr, Shipe ER, Ashley DA, Parrott WA, Hussey RS, Boerma HR (1996) RFLP loci associated with soybean seed protein and oil content across populations and locations. Theor Appl Genet 93(5-6): 649-657

Li H, Zhao T, Wang Y, Yu D, Chen S, Zhou R, Gai J (2011) Genetic structure composed of additive QTL, epistatic QTL pairs and collective unmapped minor QTL conferring oil content and fatty acid components of soybeans. Euphytica 182(1): 117-132

Los DA, Murata N (1998) Structure and expression of fatty acid desaturases. Biochim Biophys Acta-Lipids and Lipid Metabolism 1394(1): 3-15

Lynch M, Force AG (2000) The origin of interspecific genomic incompatibility via gene duplication. Am Nat 156(6): 590-605

Masuda T, Goldsmith PD (2009) World soybean production: area harvested, yield, and long-term projections. International Food and Agribusiness

Management Review 12(4): 143-162

Mazarei M, Lennon KA, Puthoff DP, Rodermel SR, Baum TJ (2003) Expression of an *Arabidopsis* phosphoglycerate mutase homologue is localized to apical meristems, regulated by hormones, and induced by sedentary plant-parasitic nematodes. Plant Mol Biol 53(4): 513-530

Meinke D W, Chen J, Beachy RN (1981) Expression of storage-protein genes during soybean seed development. Planta 153(2): 130-139.

Mortazavi A., Williams BA, McCue K, Schaeffer L, Wold B (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods 5(7): 621-628

Ohta D, Fujimori K, Mizutani M, Nakayama Y, Kunpaisal-Hashimoto R, Münzer S, Kozaki A (2000) Molecular cloning and characterization of ATP-phosphoribosyl transferase from *Arabidopsis*, a key enzyme in the histidine biosynthetic pathway. Plant Physiol 122(3): 907-914

Panthee DR, Pantalone VR, Sams CE, Saxton AM, West DR, Orf JH, Killam AS (2006) Quantitative trait loci controlling sulfur containing amino acids, methionine and cysteine, in soybean seeds. Theor Appl Genet 112(3): 546-553

Panthee DR, Pantalone VR, Saxton AM., West DR, Sams CE (2006)

Genomic regions associated with amino acid composition in soybean. Mol Breed 17(1), 79-89.

Panthee DR, Pantalone VR, West DR, Saxton AM, Sams CE (2005) Quantitative trait loci for seed protein and oil concentration, and seed size in soybean. Crop Sci 45(5): 2015-2022

Qi ZM, Wu Q, Han X, Sun YN, Du XY, Liu CY, Jiang HW, Hu GH, Chen QS (2011) Soybean oil content QTL mapping and integrating with meta-analysis method for mining genes. Euphytica, 179(3), 499-514.

Roberts A, Pimentel H, Trapnell C, Pachter L (2011) Identification of novel transcripts in annotated genomes using RNA-Seq. Bioinformatics 27(17) 2325-2329

Rose AB, Li J, Last RL (1997) An allelic series of blue fluorescent trp1 mutants of *Arabidopsis thaliana*. Genetics 145(1): 197-205

Schmutz J, Cannon SB, Schlueter J, Ma J, Mitros T, Nelson W, Hyten DL, Song Q, Thelen JJ, Cheng J, Xu D, Hellsten U, May GD, Yu Y, Sakurai T, Umezawa T, Bhattacharyya MK, Sandhu D, Valliyodan B, Lindquist E, Peto M, Grant D, Shu S, Goodstein D, Barry K, Futrell-Griggs M, Abernathy B, Du J, Tian J, Zhu L, Gill N, Joshi T, Libault M, Sethuraman A, Zhang X-C, Shinozaki K, Nguyen HT, Wing RA, Cregan P, Specht J, Grimwood J, Rokhsar D, Stacey G, Shoemaker

RC, Jackson CA (2010) Genome sequence of the palaeopolyploid soybean. Nature 463: 178-183

Scrimgeour C (2005) Chemistry of fatty acids. In: Shahidi F (ed) Bailey's industrial oil and fat products, 6th edn. John Wiley & Sons, New Jersey, pp 1-43

Sebolt AM, Shoemaker RC, Diers BW (2000) Analysis of a quantitative trait locus allele from wild soybean that increases seed protein concentration in soybean. Crop Sci 40(5): 1438-1444

Severin AJ,Woody JL, Bolon YT, Joseph B, Diers BW, Farmer AD, Muehlbauer GJ, Nelson RT, Grant D, Specht JE, Graham MA, Cannon SB, May GD, Vance CP, Shoemaker RC (2010) RNA-Seq Atlas of *Glycine max*: a guide to the soybean transcriptome. BMC plant biol 10(1): 160

Soupene E, Kuypers F A (2008) Mammalian long-chain acyl-CoA synthetases. Exp Biol Med 233(5): 507-521

Specht JE, Chase K, Macrander M, Graef GL, Chung J, Markwell JP, Germann M, Orf JH, Lark KG (2001) Soybean response to water. Crop Sci 41(2): 493-509.

Tajuddin T, Watanabe S, Yamanaka N, Harada K (2003) Analysis of quantitative trait loci for protein and lipid contents in soybean seeds

using recombinant inbred lines. Breed Sci 53(2): 133-140

Vicente-Carbajosa J, Carbonero P (2005) Seed maturation: developing an intrusive phase to accomplish a quiescent state. Int J Dev Biol 49:645

Vigeolas H, Waldeck P, Zank T, Geigenberger P (2007) Increasing seed oil content in oil-seed rape (Brassica napus L.) by over-expression of a yeast glycerol-3-phosphate dehydrogenase under the control of a seed-specific promoter. Plant Biotechnology Journal 5(3): 431-441

Wang Y, Wang X, Paterson AH (2012) Genome and gene duplications and gene expression divergence: a view from plants. Annals of the New York Academy of Sciences 1256(1): 1-14

Weber H, Borisjuk L, Wobus U (1997) Sugar import and metabolism during seed development. Trends Plant Sci 22: 169–174

Weber H, Borisjuk L, Wobus U (2005) Molecular physiology of legume seed development. Annu Rev Plant Biol 56:253-279

Weber H, Sreenivasulu N, Weschke W (2010) Molecular physiology of seed maturation and seed storage protein biosynthesis. In Pua EC, Davey MR (ed) Plant Developmental Biology-Biotechnological Perspectives. Springer, Berlin, pp 83-104

Weichert N, Saalbach I, Weichert H, Kohl S, Erban A, Kopka J, Weber H (2010). Increasing sucrose uptake capacity of wheat grains

stimulates storage protein synthesis. Plant physiol 152(2): 698-710

Wilcox JR (1985) The uniform soybean tests, Northern states, Agronomy department, Purdue university

Winkler ME (1987) Biosynthesis of histidine. *Escherichia coli* and *Salmonella typhimurium*. in: Neidhardt FC, Ingraham JL, Low KB, Magasanik B, Schaechter M, Umbarger HE (ed) Cellular and Molecular Biology, American Society for Microbiology, Washington DC, pp 395-411

Yamauchi Y, Hasegawa A, Taninaka A, Mizutani M, Sugimoto Y (2011) NADPH-dependent reductases involved in the detoxification of reactive carbonyls in plants. J Biol Chem 286(9): 6999-7009

Yang W, Simpson JP, Li-Beisson Y, Beisson F, Pollard M, Ohlrogge JB (2012) A land-plant-specific glycerol-3-phosphate acyltransferase family in *Arabidopsis*: substrate specificity, sn-2 preference, and evolution. Plant Physiol 160(2): 638-652

Zhang Y, Yi L, Lin Y, Zhang L, Shao Z, Liu Z (2014) Characterization and site-directed mutagenesis of a novel class II 5-enopyruvylshikimate-3-phosphate (EPSP) synthase from the deep-sea bacterium *Alcanivorax sp*. L27. Enzyme Microb Technol 63: 64-70

Zhang ZZ, Li XX, Zhu BQ, Wen YQ, Duan CQ, Pan QH (2011) Molecular

characterization and expression analysis on two isogenes encoding 3-deoxy-D-arabino-heptulosonate 7-phosphate synthase in grapes. Mol Biol Rep 38(7): 4739-4747

CHAPTER IV

Comparative analysis of protein expression using 2-DE during seed protein accumulation

ABSTRACT

Seed storage protein accumuation is important prcess of soybean [*Glycine max* (L.) Merr.]. This accumulation process is complex including enzyme expression that regulated by the various factors. A proteomic approach was employed to determine the enzyme expression change during seed filling between low- and high- protein soybean NILs. Soybean seed was analyzed at 2, 3 and 4 week after flowering using two-dimensional gel electrophoresis and matrix-assisted laser desorption ionization time-of-flight mass spectrometry. Protein spots that represented different intensity between low- and high-protein lines and common expression patterns between two pairs

of NILs were selected from 2-DE analysis. Among thireen selected protein, three protein such as sucrose synthase, glyceraldehyde-3-phosphate dehydrogenase and ketol-acid reductoisomerase were identified as differentily expressed enzyme between low- and high-protein lines during seed development. The function of these three enzymes were involved in sugar and amino acid metabolisms

INTRODUCTION

Plants accumulate endogenous storage compounds during seed development to survive until germination in proper environment (Gao et al. 2012). Seed storage compounds consisting of protein, starch and lipids are valuable nutrition source for animals. Seeds of grain crops including wheat, rice and soybean are harvested for human food and animal feed. Soybean [*Glycine max* (L.) Merr.] is one of the major grain crops for vegetable protein because it contains high protein and oil contents, approximately 40% and 20%, respectively (Clemente and Cahoon, 2009). The content of seed protein is mainly influenced by their intrinsic genetic factors than environmental changes (Brummer et al. 1997, Chung et al. 2003, Lee et al.1996, Sebolt et al. 2000). Seed storage proteins including glycinin and β-conglycinin are synthesized during a 4- to 5- week period of seed filling. Accumulation of major storage proteins and their mRNAs initiates at shortly after the cessation of cell division in developing cotyledons (Meinke et al. 1981, Bolon et al. 2010).

Proteomic analysis offers a new approach to discovering the genes and pathways that are crucial for metabolism (Salekdeh and Komatsu 2007). Protein identification using 2-demensional electrophoresis (2-DE) and mass spectrometry is successfully introduced to large-scale protein profiling of developing soybean seed and identified 422 proteins including seed storage

protein synthesis (Hajduch et al. 2005). In recent years, soybean complete genome reference map and transcriptome profile allow identification of protein more easily and accurately by comparing (Schmutz J. 2010, Severin et al. 2010). Many proteomic researches were performed and the number of proteomic database is increased in various soybean genotypes at developmental stages and in response to environmental stress (Brechenmacher et al. 2012, Komatsu et al. 2011. Front Plant Sci et al.2012, Tavakolan et al. 2013).Therefore, functional identification and expression patterns of accumulated proteins from developing seed is accessible from published proteome databases (http://oilseedproteomics.missouri.edu, http://bioinformatics.towson.edu/Soybean_Seed_Proteins_2D_Gel_DB/Home.aspx).

We developed two NIL pairs of high and low seed protein content derived from different genetic backgrounds. Proteome changes in the NILs, Prot-low$_{SD}$, Prot-high$_{SD}$, Prot-low$_{DD}$ and Prot-high$_{DD}$ were analyzed during their seed development stages from 2 to 4 week after flowering (WAF). Accumulated protein spots among four NIL lines and three development stage were imaged and analyzed using 2-DE. Furthermore, significant differentially expressed proteins were identified using mass spectrometry.

MATERIAL AND METHODS

Plant materials

Two NIL pairs were developed by RIL selection from two RIL populations of Sinpaldalkong 2 x Danbaekkong (SD) and Daewonkong x Danbaekkong (DD) cross. Among 137 SD and 195 DD lines, SD-44 and DD-24 lines were identified as residual heterozygous in the interval Satt239-Satt496 on Chr 20 at F_7 generation. Two NIL pairs (Prot-high$_{SD}$, Prot-low$_{SD}$, Prot-high$_{DD}$ and Prot-low$_{DD}$) were derived from SD-44 and DD-24 for segregation in the interval Satt239-Satt496 at F_8. The corresponding relationship between seed protein content and segregated marker region was evaluated from field test during two years. To harvesting samples for RNA extraction, 24 plants from each NIL were grown in a greenhouse from 13th June in Suwon, Korea. Soybean flowers were tagged to each flower from R1 to R2 stage. Immature pods were harvested from over 10 plants at 14-15 DAF (DAF, 2 WAF), 21-22 DAF (3 WAF) and 28-29 DAF (4 WAF) and seeds were isolated from pods and collected by 500 mg per each stage. Seed samples were fresh frozen in liquid nitrogen before store at -70℃.

Protein two-dimensional gel electrophoresis

Total protein was isolated from developing seed according to a modified phenol extraction method (Herkman and Tanaka 1986). 500 mg of each immature seed was pulverized to a fine powder with liquid nitrogen, pestle and mortar. Grinded samples were homogenized in sample lysis solution composed with 7M urea , 2M thiourea containing 4% (w/v) 3-[(3-cholamidopropy) dimethyammonio]-1-propanesulfonate (CHAPS), 1% (w/v) dithiothreitol (DTT) and 2% (v/v) pharmalyte and 1mM benzamidine. Proteins were extracted during one hour at room temperature with vortexing and following centrifugation at 15,000xg for one hour at 15°C. Only soluble fraction was used for two-dimensional gel electrophoresis. Protein concentration was assayed by Bradford method (Bradford et al. 1976).

To separate proteins according to isoelectric point, IPG dry strips (4-10 NL IPG, 24cm, Genomine, Korea) were equilibrated for 12-16hours with 7M urea, 2M thiourea containing 2% CHAPS, 1% DTT, 1% pharmalyte. Each sample was respectively loaded with 200 μg. Isoelectric focusing (IEF) was performed at 20°C using a Multiphor II electrophoresis unit and EPS 3500 XL power supply (Amersham Biosciences, NJ, USA) following manufacturer's instruction. For IEF, the voltage was linearly increased from 150 to 3,500 V during 3 hours for sample entry followed by constant 3,500 V,

with focusing complete after 96 kVh. Prior to the second dimension, strips were incubated for 10 minutes in equilibration buffer (50mM Tris-Cl, pH6.8 containing 6M urea, 2% SDS and 30% glycerol), first with 1% DTT and second with 2.5% iodoacetamide. Equilibrated strips were inserted onto SDS-PAGE gels (20 x 24cm, 10-16%). SDS-PAGE was performed using Hoefer DALT 2D system (Amersham Biosciences) following manufacturer's instruction. 2D gels were run at 20°C for 1,700Vh. And then 2D gels were coomassie G250 stained as described by (Anderson et al. 1991). Quantitative analysis of digitized images was carried out using the PDQuest version 7.0 (BioRad, CA, USA) software according to the protocols provided by the manufacturer. Quantity of each spot was normalized by total valid spot intensity. After normalization, spots were selected by intensity between high and low protein NIL lines by fold change>=2 in both genetic background. Then spots that over 2 fold changes intra- same growth stage and protein QTL genotype were excluded from selection.

Protein identification using peptide mass fingerprinting (PMF)

To identify protein spots differentially accumulated among samples, a small gel piece of each protein spot was excised from coomassie G250 stained gels for PMF. Gels are digested with trypsin (Promega, WI, USA),

mixed with α-cyano-4-hydroxycinnamic acid in 50% acetonitrile and 0.1% TFA. Peptides resulting from trypsin digestion were subjected to matrix-assisted-laser-desorption-time of flight (MALDI-TOF) analysis using Microflex LRF 20 (Bruker Daltonics, German) (Fernandez et al.1998). Spectra were collected from 300 shots per spectrum over mass-to-charge (m/z) range 600-3000 and calibrated by two point internal calibration using Trypsin auto-digestion peaks (m/z 842.5099, 2211.1046). Peak list was generated using Flex Analysis 3.0. Threshold used for peak-picking was followed by 500 for minimum resolution of monoisotopic mass, 5 for S/N. Peptide mass were searched using MASCOT program (http://www.matrixscience.com) for protein identification by PMF. The parameters for the database search were trypsin as the cleaving enzyme, a maximum of one missed cleavage, iodoacetamide (Cys) as a complete modification, oxidation (Met)as a partial modification, monoisotopic masses and a mass tolerance of ±0.1 Da. PMF acceptance criteria is probability scoring. Additionally, proteins with less than 4% sequence coverage and more than 25% deviation between theoretical and experimental MW and pI values were discarded from positive proteins.

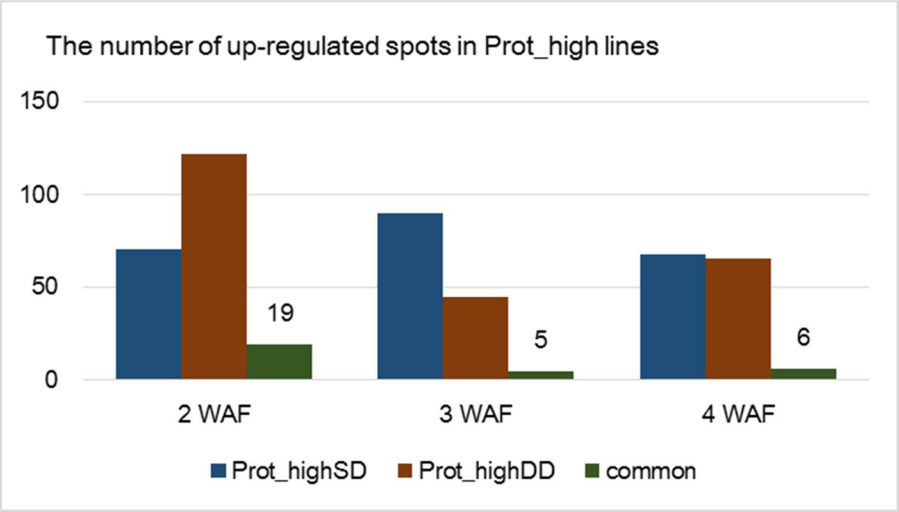
RESULTS AND DISCUSSION

Characterization of developing soybean seed and outline of protein expression

Since seed storage proteins are started to synthesize at mid-cotyledon stage, the experimental period of this study need to include before and after of mid-cotyledon stage. We harvested immature seeds at 2, 3 and 4 WAF and each stage is corresponded to early and late cotyledon and mature green stages, respectively (Figure III-1a). During the two weeks of seed growth, average sizes like length, thickness and width were increased steadily. While seed weight increased faster approximately 3-fold between 3 and 4 WAF, representing seed filling initiation (Figure III-1b).

Three development stages from 2 NIL pairs (Prot-high$_{SD}$, Prot-low$_{SD}$, Prot-high$_{DD}$ and Prot-low$_{DD}$) consisted 12 seed samples. Whole crude proteins were extracted from samples were resolved and detected using two-dimensional eletrophoresis (2-DE) followed by Coomassie Blue Staining. Expression of protein spots at the same developing stages were compared between Prot-high and Prot-low lines. The number of up-regulated spots was higher in Prot-high SD at 2WAF and contrary at 3 WAF. Therefore, the expression of proteins was similar between two Prot-high lines at 4 WAF (Figure IV-1).

a)



b)

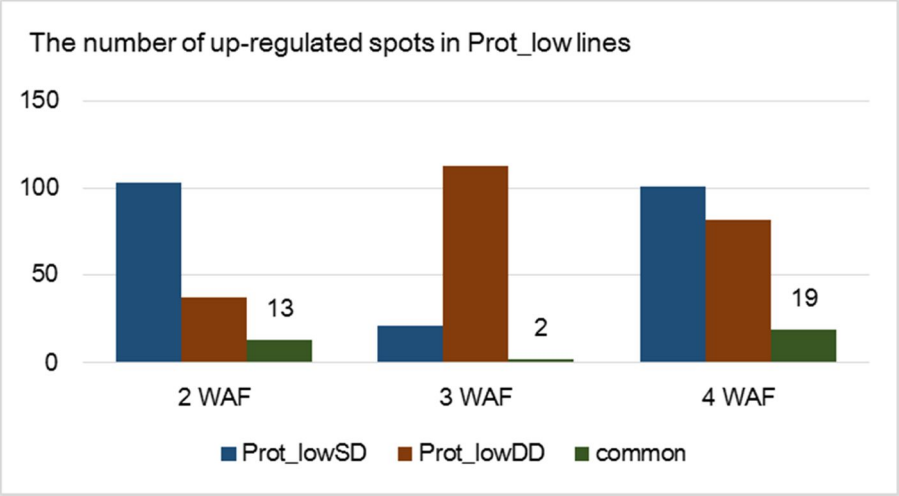


Figure IV-1 Number of up-regulated protein spots in each Prot_high and Prot-low lines. Blue, red and green bars represent up-regulated protein spot in each 2, 3 and 4 WAF, respectively.

We selected commonly expressed protein spots which show spot intensity differences more than 2 times at each WAF. These spot intensities were also compaired between each Prot_high$_{SD}$ and Prot_high$_{DD}$, and Prot_low$_{SD}$ and Prot_low$_{DD}$ for minimize the effect of genetic background. Selectied spots were identified using peptide mass fingerprinting PMF. From the PMF result, we identified three genes that involved in seed storage reserve metabolism differently expressed at 2 and 3 WAF, sucrose synthase (SUS), glyceraldehyde-3-phosphate dehydrogenase (GAPDH) and ketol-acid reductoisomerase (KARI) (Table IV-1). SUS and GAPDH are involved in carbon precursor metabolism of seed storage reserves. SUS is one of sucrose degradation enzyme that produce fructose and UDP-glucose. GAPDH is enzyme involved in conversion of glycerol 3-phosphate to D-glycerate 1,3 bisphosphate (Baud et al. 2008, Vigeolas et al. 2007). KARI involved in branched amino acid synthesis which catalyzes (R)-2,3-dihydroxy-3-methylbutanoate and $NADP^+$ to (S)-2-hydroxy-2-methyl-3-oxobutanoate and NADPH (Singh and Shaner 1995, Schomburg and Stephan 1995). Expression of SUS was up-regulated in Prot_high at 2 WAF, than in changed to show more expression in Prot_low at 4 WAF. GAPDH and KARI were up-regulated in Prot_high at 4 and 4 WAF, respectively.

Table IV-1 Identification of protein spots that had different spot intensity over fold change 2 between between Prot_high and Prot_low in both SD and DD NIL pairs at same developmental period.

Protein annotation in *G. max*	Up-regulation			Spot	Match/ %Cov	Theoretical MW/pI	Experimental MW/pI	Protein homologs
	2	3	4					
sucrose synthase	Prot_high		Prot_low	5817	24/28	92.65/5.94	98.02/6.90	Glyma02g40740, Glyma03g37441, Glyma09g08550, Glyma09g29710, Glyma11g33240, Glyma13g17421, Glyma14g39070, Glyma15g16171*, Glyma15g20180, Glyma16g34290, Glyma17g05067, Glyma19g40041
glyceraldehyde-3-phosphate dehydrogenase		Prot_high		8412	14/42	55.08/8.43	49.12/9.14	Glyma03g22790, Glyma04g36860, Glyma05g06420, Glyma06g18110, Glyma06g18120, Glyma11g37360, Glyma16g09020, Glyma18g01330, Glyma19g22780
ketol-acid reductoisomerase			Prot_high	5601	13/41	63.69/6.85	69.23/6.49	Glyma04g35256, Glyma06g43546, Glyma12g14420, Glyma12g33760, Glyma13g36730,

*Identified as differentially expressed gene from transcriptome analysis in developing seed

The expression patterns of proteins represent SUS, GAPDH and KARI expressions in Prot_high were highest at 3 WAF. In Prot_low these three genes which indicated that synthesis of seed storage products is most active in 3 WAF. In the previous research about protein expression profiling during seed filling, six SUS isoforms were identified by 2-D electrophoresis (Hajduch et al. 2007). Among this SUS isoforms, the relative abundance of three proteins were increased until 3 WAF then decreased after this and other one protein accumulated until 4 WAF.

In this research, we identified protein expression in developing seed and the regulatory change of sucrose degradation, glycolysis and branch amino acid are important to seed protein contents.

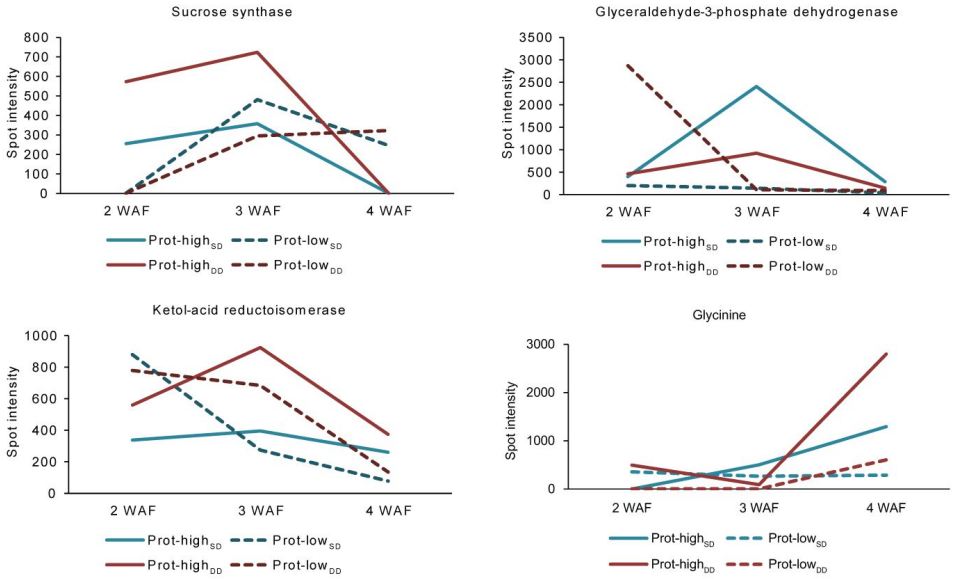


Figure IV-2 Expression change of three proteins, sucrose synthase, glyceradehyde-3-phosphate dehydrogenase and ketol-acid reductoisomerase during seed development. Each NIL pairs are distinguished by different color. Strait lines are represent Prot_high lines and dashed lines represnt Prot-low lines.

REFERENCES

Anderson NL, Esquer-Blasco R, Hofmann JP, Anderson NG (1991) A two-dimensional gel database of rat liver proteins useful in gene regulation and drug effects studies. Electrophoresis 12:907-913

Baud S, Dubreucq B, Miquel M, Rochat C, Lepiniec L (2008) Storage reserve accumulation in *Arabidopsis*: metabolic and developmental control of seed filling. The *Arabidopsis* book/American Society of Plant Biologists 6 doi: 10.1199/tab.0113

Bolon Y-T, Joseph B, Cannon SB, Graham MA, Diers BW, Farmer AD, May GD, Muehlbauer GJ, Specht JE, Tu ZJ (2010) Complementary genetic and genomic approaches help characterize the linkage group I seed protein QTL in soybean. BMC plant biol 10:41

Brechenmacher L, Nguyen THN, Hixson K, Libault M, Aldrich J, Pasa-Tolic L, Stacey G (2012) Identification of soybean proteins from a single cell type: the root hair. Proteomics 12:3365-3373

Brummer E, Graef G, Orf J, Wilcox J, Shoemaker R (1997) Mapping QTL for seed protein and oil content in eight soybean populations. Crop Sci 37:370-378

Chung J, Babka H, Graef G, Staswick P, Lee D, Cregan P, Shoemaker R, Specht J (2003) The seed protein, oil, and yield QTL on soybean linkage group I. Crop Sci 43:1053-1067

Clemente TE, Cahoon EB (2009) Soybean oil: genetic approaches for modification of functionality and total content. Plant Physiol 151(3): 1030-1040

Fernandez J, Gharahdaghi F, Mische SM (1998) Routine identification of proteins from sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE) gels or polyvinyl difluoride membranes using matrix assisted laser desorption/ionization-time of flight-mass spectrometry (MALDI-TOF-MS). Electrophoresis 19:1036-1045

Gao Q, Yue G, Li W, Wang J, Xu J, Yin Y (2012) Recent Progress Using High-throughput Sequencing Technologies in Plant Molecular Breeding. J Integr Plant Biol 54:215-227

Hajduch M, Ganapathy A, Stein JW, Thelen JJ (2005) A systematic proteomic study of seed filling in soybean. Establishment of high-resolution two-dimensional reference maps, expression profiles, and an interactive proteome database. Plant Physiol 137:1397-1419

Komatsu S, Yamamoto A, Nakamura T, Nouri M-Z, Nanjo Y, Nishizawa K, Furukawa K (2011) Comprehensive analysis of mitochondria in roots

and hypocotyls of soybean under flooding stress using proteomics and metabolomics techniques. J Proteome Res 10:3993-4004

Lee S-H, Bailey MA, Mian MAR, Carter TE, Shipe ER et al (1996) RFLP loci associated with soybean seed protein and oil content across populations and locations. Theor Appl Genet 93: 649-657

Meinke D, Chen J, Beachy R (1981) Expression of storage-protein genes during soybean seed development. Planta 153:130-139

Ohyanagi H, Sakata K, Komatsu S (2012) Soybean Proteome Database 2012: update on the comprehensive data repository for soybean proteomics. Front Plant Sci 3 doi: 10.3389/fpls.2012.00110

Salekdeh GH, Komatsu S (2007) Crop proteomics: aim at sustainable agriculture of tomorrow. Proteomics 7:2976-2996

Schmutz J, Cannon SB, Schlueter J, Ma J, Mitros T, Nelson W, Hyten DL, Song Q, Thelen JJ, Cheng J (2010) Genome sequence of the palaeopolyploid soybean. Nature 463:178-183

Schomburg D, Stephan D (1995) Ketol-acid reductoisomerase. In Schomburg D, Stephan D (ed) Enzyme Handbook 9. Springer, Berlin, pp 433-437

Sebolt A, Shoemaker R, Diers B (2000) Analysis of a quantitative trait locus

allele from wild soybean that increases seed protein concentration in soybean. Crop Sci 40:1438-1444

Severin AJ, Woody JL, Bolon Y-T, Joseph B, Diers BW, Farmer AD, Muehlbauer GJ, Nelson RT, Grant D, Specht JE (2010) RNA-Seq Atlas of *Glycine max*: a guide to the soybean transcriptome. BMC plant biol 10:160

Singh BK, Shaner DL (1995) Biosynthesis of branched chain amino acids: From test tube to field. Plant Cell 7:935-944

Tavakolan M, Alkharouf NW, Khan FH, Natarajan S (2013) SoyProDB: A database for the identification of soybean seed proteins. Bioinformation 9:165-167

Vigeolas H, Waldeck P, Zank T, Geigenberger P (2007) Increasing seed oil content in oil-seed rape (*Brassica napus* L.) by over-expression of a yeast glycerol-3-phosphate dehydrogenase under the control of a seed-specific promoter. Plant Biotechnology Journal 5:431-441

국문초록

단백질은 동물의 신체 구성요소와 에너지원으로 이용되는 필수영양소의 하나이므로 단백질의 섭취는 인간에게 매우 중요하며, 이러한 식품 단백질 공급원 중 가장 큰 비율을 차지하는 것이 콩에서 유래한 단백질이다. 콩은 높은 단백질과 지방의 함량을 동시에 가지고 있는 작물로서 콩의 주 육종 목표 중 하나는 단백질의 공급을 위해 고단백의 우수한 품종을 개발하는 것이다. 고단백 콩 품종의 육성을 위해 단백질 함량과 관련된 유전자를 찾으려는 많은 연구가 이루어졌으며, 이러한 연구들에서 콩의 단백질 함량과 관련한 주 양적형질유전자좌(QTL)인 *Prot 15-1* 이 염색체 20번의 두 SSR 마커, Satt239와 Satt496 사이에 위치한다고 알려진 바 있다.

단백콩은 현재 재배되는 품종 중에서도 가장 높은 수준인 48%의 높은 단백질을 가지고 있고 단백콩의 단백질 조절 주 QTL의 위치가 Satt239와 연관되어 있다. 본 연구에서는 단백콩을 부본으로 하는 교배에서 유래된 RIL에서 Satt239와 Satt496의 유전형이 이형접합인 RHL을 선발하고 이 RHL을 세대진전하여 *Prot 15-1* 의 위치가 다른 유전형을 가지는 두 NIL 쌍을 선발하였다. 이 두 NIL 쌍의 단백질 함량을 측정함으로써 단백콩에서 유래한 *Prot 15-1* 의 유전형이 콩의 단백질 함량 증가에 관련이 있음을 확인할 수 있었다.

Prot 15-1 위치에서 단백질 함량 증가에 영향을 미치는 단백콩의 유전자를 탐색하기 위해 고, 저단백 두 NIL쌍의 염기서열 다형성을 NGS를 이용해 조사하

였다. 두 쌍의 NIL의 고 저단백 사이 66개의 non-synonymous SNP 와 1개 frameshift indel 이 31개의 유전자의 공통된 위치에서 발견되었다. 이 유전자들의 기능에서 종자 단백질 합성과의 관련, 또한 종자에서의 유전자 발현 여부를 조사하여 calcium dependent protein kinase (CDPK) 와 exocyst complex subunit 인 Exo70을 콩 종자 단백질 함량 조절 기능을 하는 유전자일 것으로 예상하였다.

위와 같은 염기서열의 차이가 단백질 함량에 영향을 미치는 과정에서 일어나는 유전자와 단백질 발현 변화를 개화 후 1, 2, 3, 4주의 종자 성숙 단계별로 조사하였다. 선발된 두 쌍의 NIL 중 Prot_high$_{DD}$ 와 Prot_low$_{DD}$ 의 미성숙 종자에서 추출한 RNA를 RNA-seq 을 이용하여 전체 전사체의 발현을 분석하였다. 고, 저단백 계통에서 종자성숙과정 중 한 단계라도 발현 차이가 확실한 유전자(DEG)의 수는 647개였고 DEG의 수는 종자성숙과정이 진행될수록 증가하였다. DEG 중 종자저장산물 합성과 수송, 종자 발달과 관련된 유전자 87개를 선발하여 이들의 고, 저단백 계통간 유전자 조절 패턴을 파악한 결과 종자 저장단백질인 glycinin과 -conglycini뿐 아니라 자당 분해와 관련된 유전자인 sucrose synthase 등이 고단백 계통에서 높은 발현을 보이는 것을 확인하였다. 반면 또 다른 자당 분해 유전자 invertase는 저단백 계통에서 더 높은 유전자 발현을 보였다. 전사조절인자 (TF) 유전자 중 종자 저장 단백질 유전자의 발현 조절에 관여하는 인자에는 종자 성숙에 관여하는 주조절자에 해당하는 유전자들이 있는데, 이들 중 LEC1과 ABI3 유전자의 발현이 고단백 계통과 저단백 계통 사이에 발현 조절 차이를 보였다. 또한 종자 저장 단백질 유전자 조절부위에 결합하는 것으로 알려진 b-ZIP 유전자는 저단백 계통에서 더 높은 발현량을 나타내었다.

고단백과 저단백 계통 NIL 두 쌍 모두의 단백질을 2차원 전기영동으로 분리 후 정량분석 하였다. 각 스팟의 밀도가 고단백과 저단백 사이에서 두 배 이상이며 각 NIL 에서 시간에 따른 발현 변화가 유사한 13개의 스팟을 선발한 후 각 스팟의 단백질을 MALDI-TOF를 이용하여 질량분석, 동정하였다. 이 중 종자 단백질 함량 조절과 관련 있는 유전자로 sucrose synthase, glyceraldehyde-3-phosphate dehydrogenase와 ketol-acid reductoisomerase가 선발되었다. 이 유전자들은 공통적으로 고단백 계통에서 개화 후 3주차에 단백질 발현이 증가하는 변화 양상을 보였다.

본 실험의 결과에서, QTL *Prot 15-1* 위치의 염기서열 차이는 종자 단백질 함량에서 차이를 가져올 뿐 아니라 단백질 함량의 조절과 관련한 다른 유전자들의 발현 변화 역시 유도한다는 사실을 알 수 있었다. 따라서 염기서열 차이가 있는 후보 유전자가 실제 단백질 함량의 차이와 관련이 있는지 다른 계통을 이용하여 확인하는 연구와 함께 두 유전자들의 기능을 파악하는 분자생물학적 연구가 필요할 것으로 보인다.

주요어: 콩; 종자 저장 단백질; 근동질 유전자 계통; 차세대 유전자 분석; 종자 발달

학번: 2007-30866