Genomic analysis of a major seed protein/oil QTL region on soybean linkage group I

Bindu Joseph
Iowa State University

Follow this and additional works at: http://lib.dr.iastate.edu/etd
Part of the Agronomy and Crop Sciences Commons

Recommended Citation
Joseph, Bindu, "Genomic analysis of a major seed protein/oil QTL region on soybean linkage group I" (2009). Graduate Theses and Dissertations. 10650.
http://lib.dr.iastate.edu/etd/10650

This Dissertation is brought to you for free and open access by the Graduate College at Iowa State University Digital Repository. It has been accepted for inclusion in Graduate Theses and Dissertations by an authorized administrator of Iowa State University Digital Repository. For more information, please contact digirep@iastate.edu.
Genomic analysis of a major seed protein/oil QTL region on soybean linkage group I

by

Bindu Joseph

A dissertation submitted to the graduate faculty
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Plant Breeding

Program of Study Committee:
Randy C. Shoemaker, Co-Major Professor
Reid G. Palmer, Co-Major Professor
Madan K. Bhattacharyya
Max F. Rothschild
Dan Nettleton

Iowa State University
Ames, Iowa
2009

Copyright © Bindu Joseph, 2009. All rights reserved.
Dedicated to:

My husband Antony, my daughter Maria, and my parents Thankamma and Joseph. For their constant support, encouragement and inspiration to do my best.
TABLE OF CONTENTS

CHAPTER ONE: GENERAL INTRODUCTION 1
Introduction 1
Dissertation Organization 3
Literature Review 5
References 14

CHAPTER TWO: MOLECULAR GENETIC ANALYSIS OF A SEED PROTEIN/OIL QTL ON LINKAGE GROUP I (LG I) IN NEAR ISOGENIC LINES IN SOYBEAN 22
Abstract 22
Introduction 23
Results 25
Discussion 35
Materials and Methods 40
Acknowledgements 47
References 48
APPENDIX: SUPPLEMENTARY TABLE 62

CHAPTER THREE: STRUCTURAL AND TRANSCRIPTIONAL DIVERGENCE OF A MAJOR SEED PROTEIN/OIL QTL REGION AND ITS HOMOEEOLOGOUS REGION IN SOYBEAN 64
Abstract 64
Introduction 65
CHAPTER FOUR: RETROTRANSPOSONS WITHIN SYNTENIC REGIONS BETWEEN SOYBEAN AND MEDICAGO TRUNCATULA AND THEIR CONTRIBUTION TO LOCAL GENOME EVOLUTION

Abstract

Introduction

Materials and Methods

Results

Discussion

Acknowledgements

References

APPENDIX: SUPPLEMENTARY TABLES

CHAPTER FIVE: GENERAL CONCLUSIONS

Conclusions

References
ABSTRACT

The economic and nutritional value of soybeans is determined by its seed protein and oil contents. The genetic control of seed protein/oil is quantitative and many genomic regions associated with the regulation of these traits have been identified in soybean. The protein/oil quantitative trait locus (QTL) mapped to LG I (chromosome 20) is ubiquitous among high protein soybean lines. Identification of the candidate gene(s) underlying this major QTL would help us understand the genetic and molecular mechanisms regulating seed protein/oil accumulation in soybean. Map based cloning to identify the candidate genes for the LG I QTL was undertaken. A BAC based physical map of the region was developed by chromosome walking. Utilizing the genome sequence of soybean and the fine genetic mapping, the QTL region was delineated to an approximately 8.4 Mbp pericentromeric region on chromosome 20. Thirteen genes exhibiting differential expression in developing seeds of near isogenic-lines (NILs) contrasting in protein/oil contents, were identified as potential candidates for the QTL. The homoeologous region corresponding to the QTL region was identified on chromosome 10, spanning a 7.3 Mbp pericentromeric region. Evidence suggests these regions were products of the recent soybean genome duplication 10-14 million years ago (MYA). Comparative analysis of the two regions indicated that both regions are hot spots for retrotransposon accumulation. However, the chromosome 20 region showed preferential accumulation of retrotransposons. The extent of synteny was poor in the retrotransposon rich (gene poor) regions while the adjacent retrotransposon poor (gene rich) regions showed high levels of synteny and colinearity. The genes conserved across the regions showed evidence of sub-funtionalization. The conserved genes were evolving under purifying selection and were enriched for genes involved in stress responses and transcription
regulation. The abundance of retrotransposons in these regions points to the likely role of retrotransposons in the evolution of these regions. A detailed study of the retrotransposon insertions in BAC-sized regions around the QTL indicated intergenic and intragenic insertions. The majority of retrotransposon accumulations in the regions studied, happened in the past two million years. Retrotransposon insertions and selection against the insertions could be the predominant forces driving the evolution of this major protein/oil QTL region in soybean.
CHAPTER ONE
GENERAL INTRODUCTION

INTRODUCTION
The United States is the leading producer of soybeans in the world. Soybeans represent 25.7 Mha of crop production in the US, with a value of over $26 billion (Soy Stats, 2008). Soybean seeds contain approximately 40% protein and 20% oil. The protein meal derived from the seed is the primary source of protein in livestock feed.

Protein and oil contents of soybeans grown in various regions of the USA can deviate significantly. Seed protein is typically lower in the northwestern states than in southeastern soybean-growing states. Fifty-one years of data from Northern and Southern Uniform Tests showed that mean protein concentration was higher in the southern region compared to the northern region (Yacklich et al., 2002). In 2001, this regional difference spanned three percentage points, the largest ever observed in 17 years of survey data (Hurburgh, 2001). If the seed protein content is low, processors may not be able to derive the valuable 48% soybean meal. The regional variation in protein levels leads to a price discrimination to producers in the North central USA. The variation in soybean seed protein levels could be in part due to the genetic factors regulating seed protein.

Soybean seed protein/oil is a quantitative trait. Many QTL associated with seed protein/oil have been identified (Csanádi et al., 2001; Fasoula et al., 2004; Lee et al., 1996; Mansur et al., 1993; Panthee et al., 2005). A major QTL for seed protein/oil has been mapped to soybean linkage group I (LG I) (Diers et al., 1992; Seboldt et al., 2000). Inheritance of the high protein allele at LG I resulted in protein increases of 18-24 g/kg and was associated with
lower oil content (Diers et al., 1992). Identification of the candidate genes for this major QTL would help us understand the molecular mechanisms regulating seed protein/oil content in soybean. This knowledge could contribute to the development of plant breeding and genetic strategies to elevate the seed protein levels in commercial soybean cultivars.

Region-specific duplication or genome-wide polyploidization is a prominent feature in plant genome evolution, which continues to shape genomic diversity (Lynch, 2002). Among plants, close to 70% of flowering plants (angiosperms) or as high as 90% of pteridophytes have undergone at least one round of genome doubling during their evolution (Leitch and Bennett, 1997; Vision et al., 2000; Blanc and Wolfe, 2004). The high incidence of genome duplication in plants could be due to its potential impact on genetic diversity and adaptation (Lawton-Rauh, 2003).

The evolutionary history of soybean is no exception. Evolutionary studies and haploid genome analysis have suggested that soybean is a diplodized ancient tetraploid (Hadley and Hymowitz, 1973). Genetic mapping using RFLP probes in soybean detected markers duplicated across different linkage groups, a likely outcome of multiple rounds of genomic duplications (Shoemaker et al., 1996). Two of the large scale genome duplications in soybean were estimated to have happened 14 and 44 million years ago (MYA) (Schlueter et al., 2004). Following polyploidization, genomes undergo extensive restructuring, (Levin, 1983; Shaked et al., 2001; Song et al., 1995) which could affect the relative positions and linkage relationships among genes as well as the epistatic and pleiotropic expression patterns of duplicated genes (Lawton-Rauh, 2003).

Shoemaker et al. (1996) compared soybean linkage groups for homoeologous QTL for seed composition. They observed that chromosomal regions sharing markers also showed a
tendency to share seed composition QTL. Homoeologous regions retaining QTL would indicate conserved gene function for the QTL genes or functional redundancy of QTL genes (Shoemaker et al., 1996). It would be of interest to identify the homoeologus region corresponding to the LG I seed protein/oil QTL region, and to determine the divergence of the two regions to understand the evolution of these regions. The two specific objectives of my study were as follows.

1. To finely resolve and characterize the genomic region encompassing a major seed protein/oil QTL on LG I (chromosome 20).

2. To identify the homoeologous region corresponding to the LG I QTL region and determine the structural and transcriptional divergence of the two regions.

DISSEPTION ORGANIZATION

This dissertation is organized into five chapters. Chapter one provides general background information pertaining to the study and the objectives of the study, followed by a review of the current literature. The following three chapters represent original research, each in the form of a manuscript. The fifth chapter summarizes the conclusions from the study.

Chapter two, a journal paper prepared for submission to the journal *Genome Biology* entitled “Molecular genetic analysis of a seed protein/oil QTL on linkage group I (LG I) in near-isogenic lines in soybean”, is co-first authored by Bindu Joseph and Yung-Tsi Bolon. This chapter details the delineation of the QTL region onto chromosome 20 and the seed transcriptome analysis of near-isogenic lines (NILs) developed by the introgression of the LG I QTL region from the high protein accession PI468916 onto the low protein line A81-356022. The author contributions for this manuscript are; Bindu Joseph performed the
research and writing related to physical mapping, SSR genotyping and defining the QTL region onto soybean chromosome 20 under the supervision of Randy Shoemaker. Yung-Tsi Bolon performed the transcriptome analysis of NILs using soybean Affymetrix chip and wrote portions of the manuscript related to the transcriptome analysis under the guidance of Carroll Vance. Brian Diers and Jim Specht developed the NILs. Gregory May and Andrew Farmer generated the high throughput Illumina transcript data and carried out the preliminary analysis. Steven Cannon and Nathan Weeks performed the genome sequence annotations and aligned the transcripts to the soybean genome and provided the transcript counts. Michelle Graham performed the Gene Ontology (GO) annotations and the statistical analysis of the GO categories from the Affymetrix data. Zheng Jin Tu, Wayne Xu and Gary Muehlbauer performed the single feature polymorphism (SFP) analysis and other bioinformatic analysis.

Chapter three, a journal paper prepared for submission to the journal Genome, is entitled “Structural and transcriptional divergence of a major seed protein/oil QTL region and its homoeologous region in soybean”. This chapter details the comparative analysis of the QTL homoeologous regions. The pericentromeric location of the regions allowed us to determine the divergence of gene poor regions in comparison to its adjacent gene rich regions. The transcriptional divergence of the homoeologues were determined in terms of the differences in transcript abundance and tissue specificity in expression. The author contributions are; Bindu Joseph designed and performed the research and wrote the manuscript under the supervisory guidance of Randy Shoemaker. Michelle Graham performed the GO annotations. Yung-Tsi Bolon provided the RNA samples used for generation of Illumina transcript data. Brian Diers and Jim Specht developed the NILs. Gregory May and Andrew Farmer generated the high throughput Illumina transcript data and
carried out the preliminary analysis. Steven Cannon and Nathan Weeks performed the genome sequence annotations and aligned the transcripts to the soybean genome and provided the transcript counts.

Chapter four, a manuscript accepted pending revisions by the journal *The Plant Genome*, entitled “Retrotransposons within syntenic regions between soybean and *Medicago truncatula* and their contribution to local genome evolution”, is a closer look at two syntenic blocks among soybean and *Medicago truncatula* with emphasis on the contributions of retrotransposons to local genome evolution. These syntenic blocks represent gene poor and gene rich regions around the LG I QTL region. Bindu Joseph designed and carried out the research and wrote the manuscript under the guidance of Randy Shoemaker. Jessica Schuleter provided the soybean BAC sequence annotations for one of the two regions studied. Jianchang Du and Jianxin Ma annotated the retrotransposons. Michelle Graham assisted with the BAC annotation and retrotransposon identification by writing custom perl scripts.

**LITERATURE REVIEW**

**LG I seed protein QTL**

Soybean seed protein, like many other agronomic traits, is a quantitative trait. Many seed protein QTL have been identified in soybean (Csanádi et al., 2001; Fasoula et al., 2004; Lee et al., 1996; Mansur et al., 1993; Panthee et al., 2005). The QTL mapped to soybean LG I (chromosome 20) is considered a major QTL for seed protein/Oil. The LG I QTL was detected by Diers et al. (1992) using 60 F$_{2:3}$ lines derived from the mating of a high yielding *Glycine max* breeding line A81-356022 with a *Glycine soja* Siebold & Zucc. Accession
PI468916. Restriction fragment length polymorphism (RFLP) markers on LG I (A144, A407, A688, and K011) identified a *G. soja* segment that when homozygous, increased the seed protein content by 18-24g/Kg (Diers et al., 1992). Brummer et al. (1997) examined eight different populations of F₂-derived lines and identified a strong association of soybean protein and/or oil content with RFLPs A407 and A144 on LG I, in a population derived from the mating of a breeding line (M82-806) with a high protein line (HHP; 25% *G. soja* by pedigree). Sebolt et al. (2000) introgressed the high protein alleles from the *G. soja* donor parent PI468916 to the recurrent parent A81-356022 by back crossing (BC₃) and observed that the lines homozygous for the *G. soja* alleles at markers linked to the QTL increased the protein levels significantly compared to the lines homozygous for the recurrent parent alleles. However, *G. soja* alleles at the LG I QTL reduced seed yield, oil content and seed weight significantly and plants exhibited earlier maturity.

The map position of the LG I QTL was further delineated by Chung et al. (2003) using more markers than had been used in previous studies, in a population of 76 F₅-derived recombinant inbred lines (RILs) from the cross between a *G. max* germplasm accession PI 437088A (high protein) and a *G. max* cultivar Asgrow A3733 (low protein). The LG I QTL for protein, oil, and yield mapped to an interval flanked by the co-dominant SSR markers Satt496 and Satt239 with a genetic map distance of 6.6 cM in this population (Chung et al., 2003). PI437088 alleles at the markers nearest to the QTL increased the protein levels by 1.84% while decreased the oil content by 1.14%. All these reports provided strong evidence that the LG I QTL is ubiquitous among high protein soybean germplasm and further documented the negative phenotypic correlation between seed protein and oil content (Brim and Burton, 1979; Wilcox and Cavins, 1995).
Efforts to fine map the LG I QTL resulted in defining the QTL to a 3 cM interval between SSR marker Satt239 and amplified fragment length polymorphism (AFLP) marker ACG9b (Nichols et al., 2006). Fine mapping was carried out using two sets of backcross populations developed by additional backcrossing to the BC₃ populations (Seboldt et al., 2000) described above. The first set included three BC₄ populations, and the second set included four BC₅ populations. These populations segregated for different segments of the LG I QTL genomic region. LG I markers showed significant associations with protein and oil contents (P<0.001). Lines homozygous for the QTL allele from PI468916 had higher seed protein, lower seed oil, earlier maturity and lower yield (Nichols et al., 2006). Identification of the candidate gene(s) associated with the LG I QTL would help us to understand the genetic and molecular mechanisms regulating the seed protein/oil content, especially those underlying the negative correlation between protein and oil contents in soybean.

Genome duplication and the divergence of homoeologous regions

Evolutionary history of most eukaryotic genomes indicates the prevalence of whole genome duplications or polyploidy (Wolfe and Shields, 1997) which continues to contribute to genomic and genetic diversity (Lynch, 2002). Close to 70% of flowering plants (angiosperms) or as high as 90% of pteridophytes have undergone at least one round of genome doubling during their evolution (Leitch and Bennett, 1997; Vision et al., 2000; Blanc and Wolfe, 2004). Many plants that are considered diploids, are ancient polyploids (paleopolyploids) that became diplodized by extensive gene loss and genome rearrangements following polyploidy (Wolfe, 2001). The remnants of past polyploidy events exist as sets of duplicated chromosomal segments, and in Arabidopsis, these duplicated chromosomal
remnants from the most recent polyploidy span 70-90% of the genome (Blanc et al., 2003; Bowers et al., 2003). Genome restructuring and reshuffling following genome duplications often hamper the identification of more ancient duplications (Lawton-Rauh, 2003).

Functional evolution of duplicated genes

Following duplication, duplicated genes could experience varied functional fates such as functional redundancy, subfunctionalization, neofunctionalization, or pseudo-gene formation (Force et al., 1999; Lynch and Force, 2000; Prince and Picket, 2002). The classical model for duplicate gene evolution predicts that pseudo-gene formation or loss of one of the duplicated copies brought about by deleterious mutations is a more likely fate for duplicated gene pairs, though rarely, one of the duplicates can acquire new functions (Ohno, 1970). However, the classical model is inadequate to explain the empirical evidence of much higher preservation of duplicated genes than expected (Force et al., 1999). Force et al. (1999) proposed the Duplication-Degeneration-Complementation (DDC) model that predicts subfunctionalization, resulting from the acquisition of complementary degenerative mutations in regulatory regions of the duplicated genes that partition the function of the ancestral gene into its two descendents. The probability of duplicate pair retention through subfunctionalization versus neofunctionalization depends on population size and recombination between the pairs (Lynch et al., 2001). Because subfunctionalization does not depend on favorable mutations, it is more likely to explain the retention of duplicate pairs than neofunctionalization in organisms with small population sizes such as multicellular plants (Ganko et al., 2007).
Subfunctionalization in the expression of duplicated genes now referred to as the ‘expression subfunctionalization’ (reviewed in Doyle et al., 2008), is the partitioning of ancestral expression domains among duplicate genes. At the genomic level, this variation in expression relates to the genomic dominance when divergent genomes are combined into one nucleus as reported in synthetic allotetraploids of cotton (Adams et al., 2003) and Arabidopsis (Wang et al., 2006). The homoeologue bias in expression appears immediately after genome merger and is maintained following genome doubling (Flagel et al., 2008). It was concluded that hybridity and dosage imbalance during the polyploid formation could trigger expression subfunctionalization (Doyle et al., 2008). At the genic level, expression of homoeologues in polyploids suggested unequal contribution to the transcriptome (Adams et al., 2003; Hovav et al., 2008). Expression subfunctionalization observed in synthetic allopolyploids of cotton is with little sequence divergence. It acts far more quickly than classic subfunctionalization and could preserve a large number of homoeologous pairs from mutational decay by providing additional raw materials for subsequent evolution (Adams et al., 2003; Rapp and Wendel, 2005). However, in Arabidopsis, paleo-homoeologues from genome duplications show less expression subfunctionalization compared to other duplicates and may not be central to the preservation of homoeologues (Blanc and Wolfe, 2004; Casneuf, 2006; Ganko et al., 2007).

A study of patterns of gene expression and coding sequence divergence between duplicated genes in *Arabidopsis thaliana* indicated little relationship between expression divergence and synonymous substitutions (Ks) whereas a positive relationship was observed between expression divergence and non-synonymous substitutions (Ka). This positive
relationship is pronounced in the polyploidy-derived duplicates in comparison to the tandem or dispersed duplicates, suggesting that the strength of selection acting on protein sequence and expression pattern is correlated (Ganko et al., 2007). However, in cotton, expression subfunctionalization observed in synthetic allotetraploids is not associated with sequence divergence; therefore, epigenetic mechanisms also could explain the subfunctionalization. Except methylation (Wang et al., 2004, Madlung, 2002), epigenetic mechanisms for regulation of duplicate gene expression are largely unknown.

The functional fate of genes via loss or retention of duplicate genes is not random. Specific functional categories preferentially retain or lose copies (Seoighe and Gehring, 2004; Blanc and Wolfe, 2004). Genes involved in signal transduction, and transcription are preferentially retained in duplicates while genes involved in DNA repair are mostly singletons (Blanc and Wolfe, 2004; Paterson et al., 2006). However, it should be noted that the patterns of divergence of duplicated genes may vary depending on the types of duplicate genes; e.g. tandem duplicates, dispersed duplicates and polyploidy derived duplicates. The distinction between single-gene duplicates versus duplicates from whole genome duplication (WGD) is important in evolutionary analysis (Reviewed in Conant and Wolfe, 2008). Duplicate gene decay is low for genes involved in kinase activity, transcription, protein binding and modification, and signal transduction pathways when created in large-scale gene duplication events, whereas gene decay is very high for such genes when created by individual, small-scale duplication events (Maere et al., 2005). The fact that transcription factors and kinases often form protein complexes and need to be present in stoichiometric quantities for their correct functioning, could be the reason for this bias (Papp et al., 2003;
Krylov et al., 2003). Duplicates retained after WGD diverge in expression quickly for a
given level of coding sequence divergence (Guan et al., 2007). Also WGD can lead to
retention of duplicates whose dosage balance is important such as ribosomal proteins, which
are rarely duplicated by single-gene duplication (Reviewed in Conant and Wolfe, 2008).

**Retrotransposons and genome evolution**

Among flowering plants, genome size ranges from less than 100 Mbp in members of
Rosaceae to more than 100,000 Mbp in members of Liliaceae (plant C-values database).
However, the variation in genome size is not correlated with the number of genes or the
biological complexity of the organism. This phenomenon is known as the C-value paradox
(Thomas, 1971). Genome size variation is associated with the variation in the amount of
repetitive DNA (Flavell et al., 1974) particularly long terminal repeat (LTR) retrotransposons
(SanMiguel et al., 1996). Retrotransposons account for ~9 % of the genome in Arabidopsis
(The Arabidopsis Genome Initiative, 2000) while in maize retrotransposons contribute to
greater than 70% of the genome size (Sanmiguel and Bennetzen, 1998). Amplification of
retrotransposons and polyploidization are considered major mechanisms of genome
expansion in plants. High correlation between genome size and retrotransposon copy number
was observed in a wide range of plant species with the same ploidy level (Vitte and
Bennetzen, 2006). Similarly, many-fold differences in genome size among closely related
species within the genera *Oryza*, *Vicia*, and *Gossypium* were attributed largely to species
specific amplification of retrotransposons (Piegu et al., 2006; Neumann et al., 2006; Hawkins
et al., 2006).
However, plant genomes are not destined to have a ‘one way ticket to genomic obesity’ (Bennetzen and Kellogg, 1997). Retrotransposon-mediated genome expansion is counterbalanced by the processes of unequal homologous recombination and illegitimate recombination that can remove retrotransposon DNA from the genome (Vitte and Panaud, 2003; Shirasu et al., 2000; Devos et al., 2002; Ma et al., 2004). Solo LTRs or intact elements without a target site duplication are evidence of retrotransposon removal by intra-strand recombination between LTRs of retroelements (Vitte and Panaud, 2003; Devos et al., 2002). It is now presumed that genome size is a function of both genome expansion and contraction forces (Devos et al., 2002; Bennetzen et al., 2005). The efficiency of these mechanisms could vary between genomic regions (Grover et al., 2007) resulting in certain genomic regions hot spots for retrotransposon accumulations. In Arabidopsis, the distribution of LTR retroelements was non-uniform and confined mainly to pericentromeric heterochromatin (Pereira, 2004). A negative correlation between gene density and transposable element (TE) accumulation was observed in Arabidopsis. Selection against gene disruption by TE insertion is thought to play a major role in keeping the Arabidopsis gene-rich regions free of retrotransposons (Wright et al., 2003).

Besides the impact on genome size variation, retrotransposon insertions into the coding or promoter regions of genes modulate gene function by regulation of gene expression or by formation of non-functional proteins (Hori et al., 2007; Xiao et al., 2008). Insertion of retroelements into the introns could be considered less deleterious, yet not inconsequential. Intronic insertions could cause splicing alterations and differential transcript accumulations (Marillonnet and Wessler, 1997; O’ Connor et al., 1999; Tighe et al., 2002). Comparative sequence analysis of orthologous genomic regions indicated that along with
gene deletions, insertions and duplications, retrotransposons could be agents of structural and functional divergence of genomic regions (Illic et al., 2003; Swigonova et al., 2005; Ma et al., 2005).

**Evolutionary history of the soybean genome**

Soybean is a member of the Papilionoideae subfamily that includes most economically important legumes. The tropical legumes soybean, *Phaseolus*, and *Vigna* diverged from their sister lineage comprising of temperate legumes *Medicago*, *Lotus*, and *Pisum* at around 50 MYA (Lavin et al., 2005). The genus *Glycine* is subdivided into two subgenera. Subgenus *Soja* contains the diploid annual species *Glycine max* and its progenitor, *Glycine soja*. The subgenus *Glycine* contains perennial species; both diploids and polyploids. While most of the papilionoids are diploids with x=8-10, members of *Glycine* show 2n=40 or greater (Goldblatt, 1981), suggesting ancient polyploidy (Shoemaker et al., 2006). Genetic mapping studies confirmed these findings. Shoemaker et al. (1996) compared the relative positions of RFLP probes across nine different mapping populations of soybean. More than 90% of the probes detected two or more hybridizing genomic fragments and ~ 60% detected three or more fragments. By comparing the markers duplicated across different linkage groups, it was determined that each chromosome segment is duplicated on an average 2.5 times. This was suggested to be the likely outcome of multiple rounds of genomic duplications (Shoemaker et al., 1996). Analyses of synonymous and non-synonymous distances between duplicate gene pairs identified from the soybean EST collection indicated that soybean had undergone major genome duplication events 14 and 44 MYA (Schlueter et al., 2004). Further it was shown that homoeologous regions in the soybean genome are a
mosaic of highly conserved and highly diverged regions and in general, duplicated genes involved in signaling or DNA binding were preferentially retained (Schlueter et al., 2006, Schlueter et al., 2007). Recently, the impact of genome duplication was studied by comparison of soybean homoeologous regions surrounding a disease resistance gene \( Rpg1-b \). This study revealed expansion of one of the homoeologues due to retrotransposon accumulation and homoeologue bias in the loss of genes following polyploidy (Innes et al., 2008). Identification of the homoeologous region corresponding to the soybean LG I protein/oil QTL region and comparative analysis of the two regions would help us understand the divergence and evolutionary processes of the homoeologus QTL regions in a paleopolyploid like soybean.

REFERENCES


Sanmiguel, P. and Bennetzen, J. L. 1998. Evidence that a recent increase in maize genome size was caused by the massive amplification of intergene retrotransposons. Ann. Bot. 82: 37-44.


CHAPTER TWO
MOLECULAR GENETIC ANALYSIS OF A SEED PROTEIN/OIL QTL ON LINKAGE GROUP I (LG I) IN NEAR ISOGENIC LINES IN SOYBEAN

A manuscript to be submitted to *Genome Biology*

Yung-Tsi Bolon¹*, Bindu Joseph²*, Steven Cannon³, Brian Diers⁴, Andrew Farmer⁵, Michelle Graham³, Gregory May⁵, Gary Muehlbauer⁶, James Specht⁷, Zheng Jin Tu⁸, Nathan Weeks³, Wayne Xu⁸, Randy Shoemaker³, Carroll Vance¹

ABSTRACT

The nutritional and economic value of soybean [*Glycine max* (L.) Merrill] is effectively a function of its seed protein and oil content. Insight into the genetic controls and regulatory molecular mechanisms involved in the deposition of these constituents in the developing seed is needed to guide future soybean improvement. A major quantitative trait locus (QTL) located on soybean LG I (chromosome 20) has the largest effect on seed protein/oil of any QTL detected to date. Identification of candidate genes for this QTL would enable a better understanding of how this QTL regulates protein/oil accumulation in the developing seed. A
pair of near-isogenic lines (NILs) contrasting seed protein/oil because of the introgression of a high protein QTL segment from the *G. soja* parent was used as a resource for the demarcation of the QTL region. We delineated the introgressed QTL region to less than 8.4Mbp of genomic sequence on chromosome 20 (=LG I). In parallel, gene expression profiles were obtained from multiple stages of seed fill of each NIL. Analysis of differential transcript accumulation uncovered four candidate genes located in the 8.4 Mbp segment using Affymetrix® Genechip and high-throughput Illumina® sequencing. To account for the simultaneous regulation of seed protein and oil by this QTL, we hypothesize two models, the negative regulation of protein or the positive regulation of oil, as potential modes of control. This study demonstrated the power of gene expression analysis to characterize contrasting NILs and also illustrated how two complementary gene profiling technologies may aid in the annotation of genes in the soybean genome

**INTRODUCTION**

Seed protein and oil traits are crucial to the value of the soybean crop [1]. Protein meal and oil are the major by-products of soybean processing, and high seed protein levels allow processors to derive meal with high nutritional value [2]. A better understanding of the genetic basis of seed protein and oil variation in soybean will be important for developing future breeding and targeting strategies to improve seed quality traits in commercial soybean cultivars.

Storage reserves account for the majority of the protein and oil in the seed [3, 4]. The period of seed development where these reserves accumulate is commonly referred to as the seed filling stage, a 4-5 week period of cell expansion once cell division is complete [5]. The
most prevalent seed storage proteins are beta-conglycinin and glycinin [5-8], and seed oil is commonly in the form of triacylglycerol [4, 8]. A number of diverse and interlinked processes, including photosynthesis, sucrose signaling and transport are associated with seed development and the regulation of complex traits [9, 10].

Genetic control of seed protein is inherited in a quantitative manner. Many quantitative trait loci (QTLs) associated with seed protein have been identified in soybean [11-15]. However, the seed protein QTL mapped to soybean linkage group I (LG I) is of particular interest due to its large additive effect which accounts for its consistent detection in many soybean mapping population [14, 16, 17] and across multiple growing environments [18]. LG I QTL was reported to account for upto 24-28% of the seed protein variation [19]. Inheritance of the high protein allele at LG I resulted in protein increases of 18-24 g/kg and was associated with lower oil concentration [16, 17, 19]. A negative phenotypic and genotypic correlation between soybean seed protein and oil content is well documented at the genotypic level [20-23], and this was also documented for the LG I QTL [19].

Using a recombinant inbred line population, Chung et al. (2003) defined the LG I protein QTL to a 6.6 cM interval. Nichols et al. (2006) then fine mapped the LG I protein QTL region to a 3 cM interval using BC$_2$F$_5$-derived NILs contrasting in seed protein and oil. While mapping is a valuable tool for localizing genetic regions of interest for a trait, the capabilities of mapping can be greatly enhanced by genomic approaches to identify genes that control traits of interest.

Studies of transcriptome changes, especially when used to contrast near-isogenic lines (NILs), have proven useful for the discovery of candidate genes of interest in soybean and
other oilseeds and crops [24-26]. Analysis of gene expression profiles have provided insight into the genes and pathways involved in the developing seed [27-31]. However, a genome-wide evaluation of transcript profile changes during seed development within soybean NILs contrasting in seed protein levels has not been reported and would be a valuable resource toward the dissection of seed protein control in soybean.

In the present study, we leveraged a combination of resources to characterize the seed protein/oil control at LG I by using: nearly identical soybean lines (NILs) that differed substantially in oil and protein [32], transcript profiling technologies by microarray chip and high-throughput sequencing and the newly available soybean genome sequence. The objectives of our study were 1) to define the borders of the genomic segment encompassing the LG I protein/oil QTL region, 2) to characterize gene expression in the developing seed of high vs. low protein NILs and, most importantly, 3) to identify candidate genes for this seed protein and oil QTL. The accomplishment of these objectives constitutes the first step toward understanding the genetic and molecular mechanisms underlying the regulation of seed protein and oil by this QTL.

RESULTS

Demarcation of the QTL region

Previous studies [19, 32] produced a fine map of the LG I protein QTL region. Populations of lines screened to create the fine genetic map of the LG I QTL [32] segregated for different parts of the LG I QTL from PI468916 (high protein donor parent) in the background of A81-356022 (low protein recurrent parent). The population P-C609-45-2, which was developed through five backcrosses (BC5), segregated for the smallest QTL
containing interval from PI468916. This population was confirmed to be segregating for the
LG I protein QTL based on field tests of the population.

In parallel, a BAC-based physical map of this region was assembled that span QTL
test markers Satt239 and Satt496 [19] as well as three other SSR markers (Satt700,
Sat_219, and Sat_174) that mapped close to the QTL region (Fig. 1). BAC sequences were
generated that accounted for approximately 1.2 Mb of the QTL region. We used P-C609-45-2
population to link the physical map to the genetic map and identify recombination break
points in P-C609-45-2 to demarcate the protein QTL region. BAC sequence-derived SSR
markers that were polymorphic between A81-356022 and PI468916 were screened for
segregation in the P-C609-45-2 population. Since the QTL was segregating in P-C609-45-2
population, markers originating from within the QTL region were expected to segregate in
the population.

The soybean 8x whole genome sequence (www.phytozome.net, [33]) became
available by the end of 2008. Alignment of BAC sequences to the soybean whole genome 8X
assembly identified chromosome 20 (=LG I) as the best match to all the BACs in the
physical map. The order of BAC sequence alignment to chromosome 20 was in agreement
with the physical map (Fig. 1). Based on the finely mapped QTL position [32], the genomic
sequence on chromosome 20 between the markers Satt239 and Satt354 was considered the
putative QTL region. A total of 42 SSR markers (Suppl. Table 1) derived from the BAC
sequences and from the whole genome sequence spanning the QTL region plus five
previously mapped SSR markers [34] were screened for segregation in P-C609-45-2
population. Based on the positions of the segregating markers, the putative QTL region was
further delineated to an approximately 8.4 Mbp genomic sequence between Sat_174 and
ssrpqtl_38 on chromosome 20 (Fig. 1). Thirty-four of the 42 SSR markers segregated in the P-C609-45-2 population (Suppl. Table 1). The coordinates of the borders of the QTL region stretched from 24.54 Mbp to 32.92 Mbp on chromosome 20.

Phenotypic evaluation of seed protein and oil in NILs

To confirm the genotypes of the above NILs, markers for the segregating protein QTL region on LG I were verified in the parental lines and in two BC$_5$ NIL lines from the population P-C609-45-2-2, which was derived from a plant heterozygous for the LG I protein QTL region from the P-C609-45-2 population. One NIL (LoPro = LD0-15146) was homozygous for the region from the recurrent parent A81-356022, and the other (HiPro = LD0-15154) was homozygous for the high protein region from *G. soja* PI468916. The protein and oil phenotypes in the NILs were evaluated at four stages of seed fill (Fig. 2A). These four stages during seed fill were defined by seed size and could be harvested at the same time from the same plant for direct comparison. Stage one corresponded to 25-50 mg seed, stage two >50-100 mg, stage three >100-200 mg, stage four 200-300 mg, stage five mature seed. It is noteworthy that seed protein differences between genotypes, LoPro (low protein NIL, homozygous for A81-356022) and HiPro (high protein NIL, homozygous for PI468916), were apparent from the early stages of evaluation (Fig. 2B), and that difference remained consistent throughout the stages. Seed oil values, however, did not show as marked a contrast in the early stages (Fig. 2C). Both protein and oil phenotypes for the NILs were consistent with the previously reported values upon seed maturity (Fig. 2B, C; stage 5).
Gene expression changes during seed fill

To examine gene expression changes during seed fill, transcript profiles were evaluated in seeds of each genotype (LoPro or HiPro) from the four stages above. Differences in the transcriptomes of the NILs may reflect or affect the high and low protein and oil phenotypes seen in the lines. Using Student’s t-test ($P < 0.001$) to evaluate significance, Affymetrix® GeneChip probesets with at least 1.5-fold change between stages were identified at a false discovery rate (FDR) of less than five percent [35]. Gene expression changes across stages were evaluated with reference to the stage one profiles (stage 2 vs. stage 1, stage 3 vs. stage 1, stage 4 vs. 1). In both genotypes, no probesets from the stage 2 vs. stage 1 comparison qualified under the FDR <5% criterion, so this comparison was excluded from further analysis. The number of probesets representing differentially expressed genes in stage three compared to stage one was also greater in HiPro than in LoPro (716 vs. 616), and this difference was again apparent between stages four and one (2094 vs. 1294). Analysis of all probesets that represented gene expression changes over time in LoPro and HiPro revealed that they shared 18.2% of gene expression increases and 30.2% of decreases in total. In addition, an overall trend analysis was conducted to classify genes that increase or decrease across all four stages in both genotypes at $P < 0.01$ and FDR<0.05. Variations in transcriptome profiles reveal dissimilarities between the two genotypes and are presumably related to changes in protein and oil accumulation.
Temporal gene expression enrichment patterns in low vs. high protein lines during seed fill

During the progression through seed fill, NIL differentials in the temporal pattern of gene expression changes were detected (Table 1). Transcripts with accumulation changes were annotated and assigned to gene ontology categories in order to identify gene categories that were enriched under each condition within each genotype. Statistically overrepresented categories after Bonferroni correction [36] were compiled from the results of the Fisher Exact test [37]. Four different gene ontology (GO) slim categories [38] were statistically overrepresented during various stage progressions of seed development in LoPro, and ten different GO slim categories were statistically overrepresented in HiPro (Table 1).

Proteomic analysis of the seed filling stages in soybean provide evidence for the presence of proteins involved in sugar and polysaccharide metabolism, protein destination and storage, and metabolite transport [39]. Transcripts involved in the above processes were detected in this study; however, distinct differences were observed in the enrichment of these gene categories between genotypes. In the low protein/high oil LoPro, photosynthesis and sucrose or starch-related genes were overrepresented, with photosynthesis-related categories consistently down-regulated (Table 1). By comparison, significant transcript accumulation changes among photosynthesis-related genes were not overrepresented in the high protein and low oil line (Table 1). Interestingly, enrichment of photosynthesis-related transcripts was detected as a mild trend decrease in the high protein/low oil HiPro only at the last seed stage collected (data not shown). In contrast, significantly greater levels of photosynthesis-related transcript fold-change decreases were observed in the low protein/high oil LoPro line by stage three (Table 1). Previous studies have shown that an inverse relationship exists
between photosynthesis and lipid biosynthesis [40, 41], which is consistent with down-regulation of photosynthesis genes and increased oil accumulation in LoPro (Table 1, Fig. 2C).

Sucrose is well-known for its many roles during seed development [42-45]. Impaired storage metabolism has been linked with decreased sucrose levels [9], and sucrose may affect carbon flux at the transcriptional or post-transcriptional levels [46]. Up-regulation of sucrose and starch-related genes may be one mechanism that contributes to the final higher oil phenotype in LoPro vs. HiPro. Sucrose-related cis-acting motifs were discovered in the sequence analysis of the list of differentially expressed genes between the NILs (data not shown). In all, down-regulation of photosynthetic genes and up-regulation of sucrose-related genes may contribute to the high oil phenotype of LoPro.

In contrast, overrepresented categories of genes that were up- or down-regulated during seed fill in the high protein HiPro included such categories as nutrient reservoir and microtubule-related genes (Table 1). Up-regulation and accumulation of transcripts related to protein storage under the nutrient reservoir category is consistent with the high protein phenotype of HiPro (Fig. 2B). These genes include those that encode the expected beta-conglycinin and glycinin storage proteins (Table 1 and data not shown). High protein soybean lines were shown to accumulate higher levels of most beta-conglycinin and glycinin subunits [47, 48]. The presence of overrepresented microtubule-related gene categories in expression trend decreases (Table 1) suggests a role for fundamental transport mechanisms [49], and the slowing of cell expansion later in the development of the HiPro seed.
Overall transcript trends across stages were evaluated for enriched gene categories. Interestingly, while both the low protein LoPro and the high protein HiPro displayed an enrichment of increases in gene transcripts under the DNA binding molecular function only HiPro showed an enrichment of gene transcript increases under the regulation of transcription biological process. A significant proportion of gene increases devoted to transcriptional regulation may correspond to the high protein and low oil phenotype found in HiPro.

*Differentially expressed genes between NILs identified by microarray*

Direct analysis of transcript accumulation between the two genotypes showed few significant differences. Differentially expressed transcripts between the two genotypes were detected using one-way ANOVA and Student’s t-test ($P < .001$). At a false discovery rate of five percent or less, only 13 Affymetrix® probesets displayed at least 1.5-fold change between the two genotypes LoPro and HiPro. Strikingly, six probesets were detected at greater than four-fold change between the two genotypes (Fig. 3). Examination of the six probesets above (Fig. 3) revealed that they likely represent three genes according to EST and GenBank data, condensing the number of candidates from 13 to ten. All six of the probesets with the greatest fold change were detected as transcripts with greater abundance in LoPro than in HiPro at all four stages (Fig. 3). Probesets representing transcripts with greater abundance in HiPro than in LoPro, however, also existed (Table 2, Fig. 3).

An N-way ANOVA test was conducted to examine gene expression differences simultaneously across multiple factors, genotype and time (stage) within the genotype. At $P < 0.01$ and FDR<0.05 [35], a total of 62 probesets were detected with differential changes in
transcript accumulation using this method. Again, the six probesets with the most highly
differentially expressed values were represented (Table 2, Fig. 3).

Because the Affymetrix® Genechip analyses were performed using transcripts from
two different genotypes, the possibility of the presence of feature polymorphisms in the
transcripts that could alter probe to transcript affinity was high. Therefore, single feature
polymorphism (SFP) analysis was performed using the Affymetrix® Genechip data. This
method used an algorithm based on the Li-Wong model combined with a modified probe
level statistical method [50]. SFP analysis of the three genes above showed large affinity
differences to multiple probes on the Affymetrix® Genechip (Fig. 4A). Therefore, these
three genes were potentially polymorphic in one or more regions between the two genotypes
or completely absent in one genotype.

To further validate the microarray data, quantitative reverse transcriptase-polymerase
chain reaction (RT-PCR) was performed. Specific primers were designed for the three genes
and an actin control. Significant differences between LoPro and HiPro were observed for
Cand2 and Cand3 (Fig. 4B). However, no significant transcript level fold changes were
observed for Cand1 (Fig. 4B). Thus, only two of the three genes identified as upregulated in
LoPro in prior analyses were determined to display differentially accumulating transcripts
between the two genotypes by RT-PCR.

*Differentially expressed genes between NILs map to the LG I protein QTL*

The three most highly differentially accumulating transcripts identified by
Affymetrix® Genechip were aligned to the 8X soy genome sequence and found to reside
within the borders of the protein QTL region on chromosome 20 (=LG I) (Fig. 5A). Even though only two of the three were confirmed to accumulate differential levels of transcripts, allelic differences at the segregating QTL region are a potential source for polymorphisms between the two genotypes that could result in a candidate gene. Three additional candidates identified by Affymetrix® Genechip also mapped to the QTL region, one within 3 kb of Cand2 (Table 2, compare coordinates of #6 and #1-Cand2). Thus, six of the ten candidate transcripts identified by Affymetrix® Genechip (Table 2, #1 through #6) resided within the defined boundaries of the protein QTL region at LG I.

The genes identified by N-way ANOVA analysis were aligned to the genome sequence to show the range and distribution along the chromosomes according to the coordinates on the 8X soy genome sequence (Fig. 5B). The 8X soy genome sequence reveals a general bias toward gene-rich chromosome ends (author observation), a phenomenon that has been observed in other plant genomes [51]. However, a striking concentration of probes is observed on chromosome 20 at the protein QTL region (Fig. 5B). The presence of differentially accumulating transcripts in this region is consistent with the development of near-isogenic lines that display variation in seed protein phenotype and segregation of markers within the protein QTL region.

**Differentially expressed genes between NILs identified by high-throughput deep sequencing**

Since the Affymetrix® Soy Genechip does not represent the complete set of soybean genes, high-throughput Illumina® deep sequencing of the transcriptome was performed to confirm the microarray data and search for additional candidate genes. Using the same RNA samples prepared for microarray analysis as templates for high-throughput deep sequencing,
over 76 million reads were sequenced, each 36-46 nucleotides in length, using the Solexa GS 20 sequencer. Sequences were generated from each of the four stages in LoPro and in HiPro, producing at least 7 million reads per stage. Sequences were generated from random priming sites within transcript cDNA. Of these reads, 57.8% aligned to the 8X soybean genome sequence using GMAP [52] with up to 2 mismatches and allowing multiple mappings, and 37.3% aligned uniquely to the genome sequence.

Twelve differentially accumulated transcripts were detected with at least a four-fold change in expression at a \( P < 0.001 \) and were identified at the LG I protein QTL region using Illumina® high-throughput sequencing (Table 3). Putative genes were annotated using JGI gene calls and compared with plant EST and GenBank data sets (Table 3).

Close comparison of the candidates identified by high-throughput sequencing (Table 3, #1 and #8) showed the presence of the two most highly differentially accumulated transcripts identified by Affymetrix® Genechip analysis (Table 2, #1-Cand2 and #3-Cand3). However, the positioning of the soybean target sequence from the Affymetrix® Genechip for Cand3 did not directly conform to the predicted gene model in the region from the 8X soybean genome release. Examination of the coordinates of the most highly differentially accumulated transcripts revealed a distance of 3.7 Mb between these two candidate genes (Fig. 5A).

Interestingly, two pairs of candidates from the Illumina® deep sequencing analysis (Table 3, #2 and #3, #5 and #6) appeared in the same region with overlapping chromosome coordinates but on opposite strands. Potential candidates with sequence homology to known proteins included an ethylene receptor and a glutamyl-tRNA synthetase that presented
differentially accumulated transcripts at only one stage, as well as a putative ammonium transporter (Table 3). Examination of the available Affymetrix® Soy Genechip equivalents, however, did not provide support for the ethylene receptor and ammonium transporter candidates (Table 3). In all, the union of Affymetrix® Soy Genechip and Illumina® deep sequencing transcriptome data yielded 13 candidate genes (Table 2, Table 3).

**DISCUSSION**

This study provided the rare opportunity to intersect structural mapping and molecular expression studies. Here, we report the first seed development study conducted to compare transcriptomes of soybean NILs with contrasting seed protein and oil content, providing candidate genes for involvement in the regulation of protein and oil accumulation in the soybean seed.

*Seed protein and oil relationships*

It has long been documented that seed protein and oil content are inversely correlated in the soybean seed [20-23]. Low oil alleles are consistently cotransmitted with high protein alleles in many instances [22] and attempts to separate these two traits through chromosomal recombination in the NILs used in this study have not been successful [32]. It has been hypothesized that this relationship may be due to either very tight linkage or pleiotropic effects [19]. Whether one phenotype directly or indirectly results in the other is unknown, and the timing of events regarding differential accumulation of contrasting protein and oil levels in the seed is uncertain. GmDof4 and GmDof11 transcription factors, however, have
been reported to activate genes involved in lipid biosynthesis and simultaneously suppress the expression of storage protein genes [57]. Transcriptional suppression of some aspect of seed protein accumulation could be envisioned for the low protein/high oil NIL homozygous for the *G. max* allele of the LG I QTL. However, transcriptional suppression of seed oil accumulation in the NIL homozygous for the *G. soja* allele (assuming a repulsion-based pleiotropy of the two alleles of the candidate gene underlying this QTL) would be envisioned to occur in a time frame late in seed fill. This assumption is due to the observation that the rate of seed oil accumulation in HiPro did not differ from that of LoPro until the last stage of seed fill (Fig. 2). Although the high protein line matures slightly earlier and generally yields less seed than the low protein line [19, 32] these differences do not fully account for the striking differences in NIL seed protein content observed at the early stages of seed fill. Whether additional differences in the morphology or composition of the seed exist between the near-isogenic lines remains to be seen. Further detailed investigation is in progress to investigate the temporal and spatial distribution and partitioning of candidate gene expression that may govern the relationship between protein and oil accumulation in the developing soybean seed.

**Candidates for regulation of seed protein and oil**

We identified fourteen candidate genes mapping to the protein QTL region at LG I that may play a role in the regulation of seed protein and oil. Of the thirteen candidates that displayed differentially accumulating transcripts, eleven were found at high levels in the low protein line with low or no detectable levels in the high protein line. Based on sequence homology searches to protein databases, these candidates include a potential regulatory
protein in the Mov34-1 family, a heat shock protein Hsp22.5 and an ATP synthase (Table 3). While the Mov34-1 candidate appeared to possess versatile domains for the potential regulation of multiple processes, transcripts isolated from this candidate region contained numerous stop codons, raising the possibility of non-coding genes. The same was true for a number of the other candidates and may account for the high percentage of genes with no matches to the Uniprot protein database [54]. Evidence for the expression of heat shock proteins during the stress-independent development of the seed has been observed [55, 56]. Interestingly, heat shock protein genes were found to be expressed at higher levels in the low protein line of a near-isogenic line pair in barley [26], a phenomenon also observed in the low protein line of this study, especially in the third stage of seed fill. Previous studies have detailed an indirect relationship among the accumulation of storage proteins, lipid biosynthesis, and photosynthesis in the seed, correlating to the availability and distribution of ATP [40, 41, 57, 58]. Further investigation into the modulation of ATP synthase levels on energy status and storage product accumulation in the soybean seed will shed light on the potential role for ATP synthase as a candidate gene. Additional candidate genes are being identified from earlier stages of seed development and from an atlas of other tissues through differential analysis of transcriptome profiles of the near-isogenic line pair.

Potential modes of regulation for seed protein and oil

The low protein line was converted into a higher protein line upon inheritance of a G. soja allele at the LG I protein QTL region. However, the low protein line is the high oil line, and a number of scenarios may explain how gene expression differences may relate to variation in protein and oil phenotypes in the seed.
Protein content may be positively regulated by the expression of a gene that increases protein production in the high protein line. Alternatively, it may be negatively regulated by expression of a gene in a low protein line that inhibits or lowers protein accumulation and allows for increased oil accumulation. Significant protein differences could then be observed at an earlier stage than oil differences, as seen in Fig. 2. Inhibition of protein accumulation could take place at many levels, including transcriptional and post-transcriptional control and regulation of protein synthesis, transport, and turnover. The presence of candidate genes with non-coding segments raises the possibility of regulation at the transcriptional level that may affect the transcription of genes outside of the list of candidates shown in this study.

Differences in transcriptome profiles presumably correlate with the differences in protein and oil accumulation between the NILs. The low protein line, for example, did not display enrichment in transcript trend increases of genes involved in the regulation of transcription seen in the high protein HiPro (Table 1). These regulators of transcription included numerous bZIP and MYB factors (data not shown), and extensive analysis of cis-regulatory elements of seed storage proteins has demonstrated interaction of these elements with bZIP or MYB factors [59, 60]. It is possible that transcription of a candidate gene in LoPro results in negative regulation of transcriptional regulators or key factors involved in high protein accumulation. The presence of sequence polymorphisms in gene sequences or promoter regions within the segregating region of the protein QTL may account for the low or absent levels of candidate gene transcripts accumulating in HiPro versus LoPro (Table 3).

In another scenario, oil content is positively regulated. Gene expression or transcript accumulation leading to a higher oil phenotype occurs and may act in concert with other
factors to directly or indirectly correspond to lowered accumulation of protein. A transcription factor could initiate this effect. In support of this model, batch analysis of the promoter regions of the genes with the greatest differentially accumulated transcripts between the NILs revealed a number of transcription factor binding sites and seed-specific motifs (unpublished data). A regulatory factor expressed in the high oil/low protein line may activate higher oil synthesis or accumulation pathways. This is consistent with the greater abundance of candidate gene transcript accumulation seen in LoPro (Table 3). Inheritance of a *G. soja* allele that does not allow for expression or accumulation of the high oil gene could account for the low oil and high protein phenotype in HiPro.

Control of protein and oil accumulation in the seed occurs at many different levels and is likely influenced by more than one gene. Of the candidates genes identified in this study, any combination could be responsible for the observed change in protein and oil and phenotypes conditioned by the alleles of the LG I QTL. Other protein/oil QTLs have been identified in QTL mapping studies, but the LG I QTL is of great interest because its additive effect on seed protein and oil is the largest of any QTL identified to date. The models presented here are compatible with the role of additional genes and pathways as well as mixed models for control of seed protein and oil. Resources that include the availability of additional recombinants and the use of markers established in this study will allow for further demarcation of the QTL region. Further studies are being conducted on additional mapping populations to dissect the relationship between protein and oil levels, and functional studies are under way to identify and validate the role(s) of candidate genes in the accumulation of protein and oil in the seed.
MATERIALS AND METHODS

Physical mapping of the QTL region

The QTL flanking SSRs from the Chung et al. (2003) study, Satt239 and Satt496, as well as three other SSR markers (Sat_174, Sat_219 and Satt700) in the vicinity of the putative QTL region were used to PCR screen multi-dimensional pools of the soybean ‘Williams 82’ and the ‘Fairbault’ BAC libraries. BAC clones were end-sequenced using M13 forward and reverse primers at the Iowa State University DNA sequencing and synthesis facility. The BAC libraries were then re-screened by PCR using primers designed from BAC-end sequences and the BAC contigs were extended by chromosome walking. BACs were fingerprinted using restriction enzymes EcoRI and AccI and BAC overlap was confirmed by FPC 4.6.4 [61]. BAC overlap was verified by PCR using primers from BAC-end sequences. A minimal tiling path of BACS were identified and subsequently sequenced.

BAC sequencing and assembly

BAC DNA was isolated by plasmid midi-prep (Qiagen, Valencia, CA). Random sheared BAC DNA was size selected for 2-3 Kb and subcloned onto vector PCR4Blunt-TOPO using TOPO shotgun subcloning kit (Invitrogen). The recombinant plasmids were transformed into competent TOP10 E. coli cells by electroporation. Transformants were isolated on LB plates containing Kanamycin. Subclones were sequenced using M13 forward and reverse primers at the Iowa State University DNA sequencing and synthesis facility. Vector trimming, removal of poor quality reads and sequence assembly was carried out using the program SeqManII (DNASTAR, Inc) using default parameters with a minimum match
percentage of 95% for sequence assembly. Contigs were ordered based on the positions of
the reverse and forward reads of the same subclones. Sequence gaps were filled either by
complete sequencing of the subclones that spanned the gaps or by PCR amplification across
the gap using BAC DNA followed by complete sequencing of the PCR products.

Demarcation of the QTL region

The BAC sequences were aligned to the 8X genome sequence (ftp.jgi-psf.org/pub/JGI_data/Glycine_max/assembly/) by BLASTN, [62]). All the BAC sequences
showed the best match to chromosome 20 (LG I). Based on the fine map of the QTL region
[32] the putative QTL region was positioned on chromosome 20 between Satt239 and
Satt354. Additional SSRs were identified from within the putative QTL region and tested for
polymorphism between lines A81-356022 and PI468916. All the polymorphic SSRs were
amplified from ‘Williams 82’ (the reference genotype for which the genome sequence is
available) to verify that the primers were amplifying products of expected sizes and therefore
were targeted to the QTL region. Further, the polymorphic markers from within the QTL
region were screened for segregation in the BC₅ population P-C609-45-2 that segregates for
only the 3 cM region surrounding the QTL [32]. This SSR analysis identified the
recombination break points for a more precise positioning of the QTL region.

Plant growth and experimental design

In order to minimize uncontrolled environmental conditions, two NILs, LoPro (low
protein) and HiPro (high protein), were grown in growth chambers at the University of
Minnesota. These NILs (LD04-15146 and LD04-15154) are from the population P-C609-45-
2-2 that was developed from a BC$_3$F$_5$ plant from the P-C609-45-2 population heterozygous for the QTL region [32]. Soybeans were grown in the growth chamber at a photoperiod of 14/10 and thermocycle of 23°C/16°C. Daylength and temperatures were monitored to mimic Illinois field growing conditions. Contrasting NILs were planted in staggered pairs, and three biological replicates were conducted following a complete random design. Each replicate was harvested at the same time of day on different days and consisted of seed samples at four developmental stages pooled from three plants. Samples were harvested from the NILs in parallel and flash frozen in liquid nitrogen before storage at -80°C. Stage one corresponded to 25-50 mg seed, stage two >50-100 mg, stage three >100-200 mg, stage four >200-300 mg, stage five (mature seed).

**Seed protein and oil analysis**

The NILs were grown to maturity, and seed from both genotypes was harvested at each of the four stages. Seed was harvested from the final mature seed stage and all samples were analyzed for protein and oil at the Agricultural Experiment Station chemical laboratories at the University of Missouri-Columbia (UMC). Soybean seed was weighed before and after freeze-drying and then submitted to UMC for laboratory analysis. A combustion method by LECO (AOAC Official Method 990.03, 2006) was used to analyze protein concentration in the soybean seed samples. Oil levels were determined by lipid extraction and drying to constant weight (USDA Chemistry Laboratory Guidebook, 1986).
RNA isolation

Seed was ground with liquid nitrogen by mortar and pestle. Total RNA was isolated by a modified Trizol (Invitrogen) protocol and then digested with on-column RNase-free DNase (Qiagen) and purified by RNeasy column (Qiagen). RNA quality was evaluated by gel electrophoresis, spectrophotometer, and Bioanalyzer® chip.

Microarray preparation and processing

Labeling of total RNA was performed by Qiagen Target prep robot at the Biomedical Image Processing Facility at the University of Minnesota. Both hybridization of biotin-labeled cRNA to the Soy Affymetrix® GeneChip and scanning of the microarray by GeneChip 3000 scanner were performed at the same facility according to standard Affymetrix® procedures.

Affymetrix® GeneChip data processing and analysis

The Soy Affymetrix® Genechip, containing greater than 37,500 probesets and representing 35,611 soybean transcripts (www.Affymetrix.com, [63]) was used to assess gene expression. Microarray data were analyzed using Expressionist Pro software from Genedata Inc. Raw data in the form of .CEL files from the Affymetrix® GeneChip were uploaded to the platform and the robust microarray analysis (RMA) algorithm was used to condense and normalize the data against all genes. The detection quality was set to a value of one to ensure that all probe sets were considered. MAS5.0 data condensation and normalization were performed for comparison purposes. Either a $q$ value or FDR was computed for each $P$ value [35, 64]. Differentially expressed gene lists were produced at
false discovery rates estimated at 5% or less. Data sets obtained in this study were deposited in the Array-Express database [64] and the Plant Expression database (PLEXdb) [65, 66].

**SFP identification**

Single Feature Polymorphisms (SFPs) were identified using a method [67] based on the Li-Wong model [50, 66]. This method compares the relative probe intensities of each of the 11 probes on the Affymetrix® GeneChip between genotypes. Statistical analysis of the probe affinity difference was calculated from feature intensity of the perfect match (PM) probes. Briefly, given the raw intensity (S) of each feature (probe) determined by the gene expression level (I), the affinity (A) between the target transcript and the probe, and random error (E) [66, 68, 69, 70], the equation can be modeled as \( A_{tij} + E_{tij} = S_{tij} - I_i \). Here, \( S_{tij} \) is the raw PM intensity and \( I_i \) is derived from the RMA expression value of each gene for the designated genotype (t), probe set (i), and probe (j), where \( E_{t1ij} \approx E_{t2ij} \), since E is an independent identically distributed error with a mean of zero. The Bioconductor Affymetrix® package was used to extract PM intensity and to calculate RMA expression, and the Bioconductor Siggenes package was used to evaluate all probe sets.

**Gene annotation**

Genes were annotated using the Affymetrix® GeneChip Soybean Genome Array Annotation (www.soybase.org/AffyChip) from SoyBase and The Soybean Breeder’s Toolbox in conjunction with Affymetrix annotations from the HarvEST soy assembly website (http://www.harvest-web.org). Unannotated genes were scanned by BLASTX and TBLASTX at an E value cutoff of \( 10^{-4} \). The UniProt protein database [54], the Pfam protein
database [71], the *Arabidopsis thaliana* genome database (TAIR, www.arabidopsis.org), and the *Medicago* genome database were used for annotation purposes. TAIR gene ontology (GO) and GO slim annotations [72] were provided for each *Arabidopsis* match [73]. BLASTP results with an E value of less than $10^{-10}$ were used to describe gene sequences referenced on the 8x soybean genome.

*Statistical analysis of gene ontology and expression*

The consensus sequences of the ~37,000 soybean genes on the soybean Affymetrix Genechip were compared to the most recent release of predicted genes in the Arabidopsis genome (TAIR v. 8, www.arabidopsis.org) using TBLASTX (E<10^{-4}, [62]). The top Arabidopsis gene was used to query the Arabidopsis gene ontology (TAIR ATH_GO_GOSlim.20080308, www.arabidopsis.org). A database was created linking each Affymetrix probe to the most similar Arabidopsis gene (E<10^6) and its corresponding gene ontology information (Gene Ontology Consortium, 2000). Custom perl scripts were used to mine the database for the GO slim annotations of the differentially expressed genes of interest.

To determine if particular GO slim categories were overrepresented in our expression data, the number of genes matching each GO slim category was determined. This procedure was repeated to determine the number of genes matching each GO slim category for all the soybean consensus sequences represented on the chip. For each GO slim category, Fisher's exact test [37] was used to compare the number of expressed genes in the GO slim category, the number of genes not differentially expressed in the GO slim category, the number of differentially expressed genes outside the GO slim category, and the number of genes not
differentially expressed and outside the GO slim category. To correct for oversampling, a Bonferroni correction [36] was used to adjust the two-tail probability $P$ value. The $P$ value obtained using Fisher's exact test was multiplied by the total number of GO categories represented on the soybean Affymetrix® chip. Only $P$ values more significant than .05 after Bonferroni correction are reported. Further, only GO Slim categories that were significantly over represented in the expression data are reported.

**qRT-PCR analysis**

Quantitative RT-PCR was performed and analyzed using the Applied Biosystem Real-Time PCR system. Gene-specific primers spanning 150 bp were designed using Primer Express® software (Applied Biosystems). Gene-specific actin primers were used for control and calculation purposes. Template cDNA was synthesized from total RNA using a reverse transcription cDNA synthesis kit (Invitrogen). Reactions with no reverse transcriptase were performed as controls. RT-PCR was performed in three replicates in a 96-well plate using SYBR Green (BioRad) at 35 cycles. Results were calculated using the comparative $C_T$ method to evaluate gene expression in LoPro vs. HiPro with respect to the actin control at each stage.

**High-throughput deep sequencing by Illumina®**

Total RNA from stages one through four of LoPro and HiPro was converted to double stranded cDNA at the National Center for Genome Resources (NCGR). Adapters were ligated to the ends of random primed cDNA and bridge amplified to create a digital expression library. Clonal Single Molecule Array™ flow cells were generated in a fully
automated process by Illumina® Station. Sequencing-by-Synthesis was carried out using four nucleotides with reversible fluorophore and termination properties under the Illumina® Genome Analysis System.

**Sequencing data processing and analysis**

To process the data for analysis, data files were mirrored to an off-instrument computer using the Illumina® platform to perform image analysis, base-calling, and per-base confidence scores. Individual transcript tags were identified, counted, and scored for uniqueness. Sequence reads were aligned against the 8X soybean genome sequence using GMAP [52]. The maximum allowed mismatch was set to two base pairs per read to align each 36-49 bp read. A web-based software system, Alpheus (NCGR), provided comprehensive detection of variants in sequence reads. The maximum allowed mismatch was set to two base pairs per read to align each 36 bp read. Significant differences in gene expression between the two genotypes were calculated by finding the probability of observing a difference in LoPro or HiPro in a sample size N. For a $P$ value of .001, at least 62 unique reads were required to trust a four-fold change in expression. The required sample size for identifying four-fold differences at this $P$ value was estimated from 100,000 draws from two simulated pools of genes. For these count comparisons, sample sizes were scaled so that the pools were of similar size.

**ACKNOWLEDGEMENTS**

This work was supported by the U.S. Department of Agriculture, Agricultural Research Service Current Research Information System, North Central Soybean Research
Program (WBS # 26-6222-0183-005) and National Science Foundation (Grant # DBI-0501877). We are grateful for initial funding from the Minnesota Soybean Research and Promotion Council and funding from the United Soybean Board. We would also like to acknowledge the use of resources at the MSI Supercomputing Institute at the University of Minnesota.

REFERENCES


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


Fig. 1
Fig. 1: Demarcation of the LG I QTL region. A. The genetic map of LG I (Soybean Linkage Map, 2006) showing the markers that mapped close to the QTL region. B. The fine map of the QTL (Nichols et al., 2006) where the QTL was positioned to the segregating region between markers Satt239 and ACG9b. The fixed region to the left of Satt239 is fixed for the PI468916 alleles. The fixed region to the right of ACG9b is fixed for the A81-356022 alleles. C. Physical map of the QTL region where BACs were anchored to the SSR markers Satt239, Sat700, Sat_174, Sat_219 and Satt496. The BACs included in the physical map are GM_WBa008L20, GM_WBa0082N17, GM_WBa0010E08, GM_WBa0048N22, GM_WBa0089018, GM_WBa0045F22, GM_WBa0054D13, GM_WBa0049I09, GM_WBa0009G07, GM_WBa0038H09, and GM_WBa0100B10 from left to right. The BACs shown as bold lines were sequenced. The BACs shown as thin lines were not sequenced, only BAC-end sequences were generated. D. Demarcation of the QTL region on chromosome 20 (Gm20) using additional SSR markers. The new SSR markers were named ssrptql_1 through ssrptql_42 as with their ascending position on chromosome 20. The region between the non-segregating marker Sat_174 and the segregating marker ssrptql_4 contained the recombination point on left. The region between the segregating marker ssrptql_37 and the non-segregating marker ssrptql_38 contain the recombination point on the right side. E. The QTL region highlighted on chromosome 20. The dark oval shape represents the position of the centromere.
Fig. 2: Phenotypic evaluation of NILs. A. Different stages of the developing soybean seed are shown. Stages 1-4 correspond to the seed fill stages that were harvested for phenotypic evaluation and also for gene expression profiling in this study. Stage 1 = 25-50 mg seed. Stage 2 = >50-100 mg seed. Stage 3 = >100-200 mg seed. Stage 4 = >200-300 mg seed. B. Crude protein profiles graphed on a w/w% dry matter basis for the different stages of developing seed (Stages 1-4) and the final mature soybean seed (Stage 5). Protein profiles are graphed for both the low protein line (LoPro) and the high protein line (HiPro). C. Crude oil profiles graphed on a w/w% dry matter basis for the different stages of developing seed (Stages 1-4) and the final mature soybean seed (Stage 5).
Fig. 3: Top differentially accumulated transcripts between NILs detected by microarray.  
A. Scatter plot of probesets (x) from one-way ANOVA of combined stages from line LoPro vs. line HiPro highlighted six probesets with greater than 4-fold change expression values. Dashed diagonal lines represent 2-fold, 5-fold, and 10-fold change borders in either direction.  
B. The six probesets from (A) are represented individually. Expression value is graphed as a function of stage within each genotype. Error bars represent three replicates. Note that the six probesets correspond to a total of three genes represented by two probesets each.
Fig. 4: Evaluation of top differentially accumulated transcripts between NILs detected by microarray.  

A. Single feature polymorphism (SFP) evaluation of the probesets for the three genes selected from Fig. 4. Plots show the log intensity of the affinity difference between line A and line B for each probe of the representative 11 member probeset for each gene.  

B. Quantitative reverse transcriptase-polymerase chain reaction (RT-PCR) was performed in triplicate for each of the three genes. Transcript level fold changes were compared between low protein and high protein lines with reference to the actin control at each seed fill stage.
Fig. 5: Location of candidate genes at the LG I protein QTL region in the soybean genome. A. Genes with differentially accumulated transcripts identified by Affymetrix® Soy Genechip and Illumina® high-throughput sequencing at the LG I protein QTL region. B. The location of differentially accumulated transcripts found by N-way ANOVA are mapped onto the 20 soybean chromosomes. A high-density cluster of transcripts is found at the LG I protein QTL region on chromosome 20.
Table 1. Overrepresented gene categories with transcript accumulation changes during seed fill.

<table>
<thead>
<tr>
<th>Genotype</th>
<th>Stage comparison</th>
<th>Trend</th>
<th>GO term</th>
<th>GO description</th>
<th># of genes</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>LoPro</td>
<td>1 to 4</td>
<td>Decrease</td>
<td>GO:0009773 BP</td>
<td>photosynthetic electron transport in photosystem I</td>
<td>7</td>
<td>0.000875203</td>
</tr>
<tr>
<td>LoPro</td>
<td>1 to 4</td>
<td>Decrease</td>
<td>GO:0015979 BP</td>
<td>photosynthesis</td>
<td>12</td>
<td>0.003820291</td>
</tr>
<tr>
<td>LoPro</td>
<td>1 to 4</td>
<td>Decrease</td>
<td>GO:0016168 MF</td>
<td>chlorophyll binding</td>
<td>7</td>
<td>0.005275654</td>
</tr>
<tr>
<td>LoPro</td>
<td>1 to 4</td>
<td>Increase</td>
<td>GO:0009250 BP</td>
<td>glucan biosynthetic process</td>
<td>5</td>
<td>1.90236E-05</td>
</tr>
<tr>
<td>HiPro</td>
<td>1 to 3</td>
<td>Decrease</td>
<td>GO:0003777 MF</td>
<td>microtubule motor activity</td>
<td>9</td>
<td>0.003221062</td>
</tr>
<tr>
<td>HiPro</td>
<td>1 to 3</td>
<td>Increase</td>
<td>GO:0045735 MF</td>
<td>nutrient reservoir activity</td>
<td>9</td>
<td>0.000003806</td>
</tr>
<tr>
<td>HiPro</td>
<td>1 to 4</td>
<td>Decrease</td>
<td>GO:0007018 BP</td>
<td>microtubule-based movement</td>
<td>17</td>
<td>1.074E-07</td>
</tr>
<tr>
<td>HiPro</td>
<td>1 to 4</td>
<td>Decrease</td>
<td>GO:0005975 BP</td>
<td>carbohydrate metabolic process</td>
<td>31</td>
<td>0.011237334</td>
</tr>
<tr>
<td>HiPro</td>
<td>1 to 4</td>
<td>Decrease</td>
<td>GO:0003777 MF</td>
<td>microtubule motor activity</td>
<td>22</td>
<td>0</td>
</tr>
<tr>
<td>HiPro</td>
<td>1 to 4</td>
<td>Decrease</td>
<td>GO:0016538 MF</td>
<td>cyclin-dependent protein kinase regulator activity</td>
<td>9</td>
<td>0.000351653</td>
</tr>
<tr>
<td>HiPro</td>
<td>1 to 4</td>
<td>Decrease</td>
<td>GO:0004356 MF</td>
<td>glutamate-ammonia ligase activity</td>
<td>5</td>
<td>0.00133227</td>
</tr>
<tr>
<td>HiPro</td>
<td>1 to 4</td>
<td>Decrease</td>
<td>GO:0004553 MF</td>
<td>hydrolase activity, hydrolyzing O-glycosyl compounds</td>
<td>18</td>
<td>0.00676422</td>
</tr>
<tr>
<td>HiPro</td>
<td>1 to 4</td>
<td>Increase</td>
<td>GO:0019915 BP</td>
<td>sequestering of lipid</td>
<td>6</td>
<td>0.002788199</td>
</tr>
<tr>
<td>HiPro</td>
<td>1 to 4</td>
<td>Increase</td>
<td>GO:0045735 MF</td>
<td>nutrient reservoir activity</td>
<td>12</td>
<td>0.000549571</td>
</tr>
</tbody>
</table>

Gene ontology (GO) categories that are overrepresented in the NIL genotypes LoPro and HiPro are shown next to the stages compared and the direction of trend changes. For *Stage comparison: 1 to 4* and *Trend: decrease*, the trend of transcript accumulation decreases from stage one to stage four. The GO term identifier is indicated along with the functional categorization. BP = biological process. MF = molecular function. # of genes = number of genes represented by the Affymetrix® Soy Genechip with transcript accumulation changes.
Table 2. Differentially accumulated transcripts between LoPro and HiPro identified by Affymetrix® Soy Genechip.

<table>
<thead>
<tr>
<th>#</th>
<th>Affy ID</th>
<th>LoPro</th>
<th>HiPro</th>
<th>LoPro/HiPro</th>
<th>P-Value</th>
<th>FDR</th>
<th>Uniprot Desc</th>
<th>E value</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>GmaAffx.74372.1.S1_at</td>
<td>2478</td>
<td>502</td>
<td>4.94</td>
<td>2.62823E-12</td>
<td>1.98174E-08</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Gma.1680.1.S1_x_at</td>
<td>289926</td>
<td>63810</td>
<td>4.54</td>
<td>5.51455E-15</td>
<td>1.03952E-10</td>
<td>Hypothetical protein</td>
<td>6E-24</td>
</tr>
<tr>
<td>4</td>
<td>Gma.1680.1.S1_at</td>
<td>134237</td>
<td>30039</td>
<td>4.47</td>
<td>1.05377E-14</td>
<td>1.32427E-10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>GmaAffx.49130.1.S1_at</td>
<td>10163</td>
<td>842</td>
<td>12.06</td>
<td>4.37642E-13</td>
<td>4.12489E-09</td>
<td>no hits</td>
<td>na</td>
</tr>
<tr>
<td>6</td>
<td>GmaAffx.55722.1.S1_at</td>
<td>12024</td>
<td>5126</td>
<td>2.35</td>
<td>1.0066E-10</td>
<td>4.74372E-07</td>
<td>Hypothetical protein</td>
<td>2E-27</td>
</tr>
<tr>
<td>7</td>
<td>GmaAffx.47978.1.S1_at</td>
<td>3007</td>
<td>1378</td>
<td>2.18</td>
<td>2.30998E-05</td>
<td>0.045836044</td>
<td>Putative phosphatase</td>
<td>3E-75</td>
</tr>
</tbody>
</table>

Each Affymetrix® probeset identifier (ID) is shown with corresponding normalized expression values for LoPro and HiPro and the ratio of mean LoPro divided by mean HiPro values. Criteria for the list: P-Value < .001, FDR < .05, and fold-change > 1.5. The Uniprot description for each Affymetrix® ID is accompanied by the expect value for the alignment.
Table 3. Differentially accumulated transcripts between LoPro and HiPro identified by Illumina® high-throughput sequencing.

<table>
<thead>
<tr>
<th>#</th>
<th>Comparison</th>
<th>Sequence ID</th>
<th>LoPro</th>
<th>HiPro</th>
<th>Ch</th>
<th>start</th>
<th>end</th>
<th>strand</th>
<th>BlastP Description</th>
<th>E-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Overall</td>
<td>Glyma20g18880</td>
<td>51</td>
<td>0</td>
<td>Gm20</td>
<td>26510968</td>
<td>26513359</td>
<td>-</td>
<td>Mov34-1</td>
<td>2e-21</td>
</tr>
<tr>
<td>2</td>
<td>Stage 3</td>
<td>Glyma20g18880</td>
<td>15</td>
<td>0</td>
<td>Gm20</td>
<td>26510968</td>
<td>26513359</td>
<td>-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Stage 4</td>
<td>Glyma20g18880</td>
<td>18</td>
<td>0</td>
<td>Gm20</td>
<td>26510968</td>
<td>26513359</td>
<td>-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Overall</td>
<td>Glyma20g19620</td>
<td>268</td>
<td>106</td>
<td>Gm20</td>
<td>27706435</td>
<td>27707431</td>
<td>+</td>
<td>no hits &lt; e-10</td>
<td>na</td>
</tr>
<tr>
<td>5</td>
<td>Stage 3</td>
<td>Glyma20g19620</td>
<td>259</td>
<td>95</td>
<td>Gm20</td>
<td>27706477</td>
<td>27706935</td>
<td>-</td>
<td>no hits &lt; e-10</td>
<td>na</td>
</tr>
<tr>
<td>6</td>
<td>Stage 4</td>
<td>Glyma20g19680</td>
<td>24</td>
<td>2</td>
<td>Gm20</td>
<td>27899125</td>
<td>27899596</td>
<td>-</td>
<td>Hsp22.5</td>
<td>5e-72</td>
</tr>
<tr>
<td>7</td>
<td>Overall</td>
<td>Glyma20g19680</td>
<td>24</td>
<td>0</td>
<td>Gm20</td>
<td>27899125</td>
<td>27899596</td>
<td>-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Stage 5</td>
<td>Glyma20g21030</td>
<td>61</td>
<td>12</td>
<td>Gm20</td>
<td>29984895</td>
<td>29986397</td>
<td>+</td>
<td>Putative ammonium transporter AMT1</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>Overall</td>
<td>Glyma20g21030</td>
<td>36</td>
<td>4</td>
<td>Gm20</td>
<td>29984895</td>
<td>29986397</td>
<td>+</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Stage 1</td>
<td>Glyma20g21040</td>
<td>40</td>
<td>8</td>
<td>Gm20</td>
<td>29984951</td>
<td>29986210</td>
<td>-</td>
<td>no hits &lt; e-10</td>
<td>na</td>
</tr>
<tr>
<td>11</td>
<td>Stage 1</td>
<td>Glyma20g21040</td>
<td>27</td>
<td>4</td>
<td>Gm20</td>
<td>29984951</td>
<td>29986210</td>
<td>-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>Overall</td>
<td>Glyma20g21080</td>
<td>13</td>
<td>0</td>
<td>Gm20</td>
<td>30044891</td>
<td>30045091</td>
<td>+</td>
<td>ATP synthase D chain</td>
<td>4e-17</td>
</tr>
<tr>
<td>13</td>
<td>Stage 1</td>
<td>Glyma20g21140</td>
<td>76</td>
<td>0</td>
<td>Gm20</td>
<td>30178277</td>
<td>30182887</td>
<td>-</td>
<td>no hits &lt; e-10</td>
<td>na</td>
</tr>
<tr>
<td>14</td>
<td>Stage 3</td>
<td>Glyma20g21140</td>
<td>13</td>
<td>0</td>
<td>Gm20</td>
<td>30178277</td>
<td>30182887</td>
<td>-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>Stage 3</td>
<td>Glyma20g21140</td>
<td>38</td>
<td>0</td>
<td>Gm20</td>
<td>30178277</td>
<td>30182887</td>
<td>-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>Stage 4</td>
<td>Glyma20g21140</td>
<td>15</td>
<td>0</td>
<td>Gm20</td>
<td>30178277</td>
<td>30182887</td>
<td>-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>Overall</td>
<td>Glyma20g21540</td>
<td>32</td>
<td>0</td>
<td>Gm20</td>
<td>30873569</td>
<td>30873806</td>
<td>+</td>
<td>Putative uncharacterized protein</td>
<td>2e-21</td>
</tr>
<tr>
<td>18</td>
<td>Stage 3</td>
<td>Glyma20g21540</td>
<td>13</td>
<td>0</td>
<td>Gm20</td>
<td>30873569</td>
<td>30873806</td>
<td>+</td>
<td></td>
<td>2e-21</td>
</tr>
<tr>
<td>19</td>
<td>Stage 3</td>
<td>Glyma20g21780</td>
<td>121</td>
<td>36</td>
<td>Gm20</td>
<td>31386550</td>
<td>31389333</td>
<td>+</td>
<td>Ethylene receptor</td>
<td>0</td>
</tr>
<tr>
<td>20</td>
<td>Stage 3</td>
<td>Glyma20g22170</td>
<td>142</td>
<td>46</td>
<td>Gm20</td>
<td>32098751</td>
<td>32103750</td>
<td>+</td>
<td>Glutamyl-tRNA synthetase</td>
<td>0</td>
</tr>
<tr>
<td>21</td>
<td>Stage 1</td>
<td>Glyma20g22650</td>
<td>9</td>
<td>42</td>
<td>Gm20</td>
<td>32589230</td>
<td>32589715</td>
<td>+</td>
<td>no hits &lt; e-10</td>
<td>na</td>
</tr>
</tbody>
</table>
**Supplementary Table 1.** Primer sequences of newly developed SSR primers form within the QTL region.

<table>
<thead>
<tr>
<th>ssr_name</th>
<th>forward_primer</th>
<th>reverse_primer</th>
<th>Segregation status</th>
</tr>
</thead>
<tbody>
<tr>
<td>ssrpqtl_1</td>
<td>TTC TCA CAC ACT TAA CCA</td>
<td>TAA ATC AGG AGG AAC TAT A</td>
<td>no</td>
</tr>
<tr>
<td>ssrpqtl_2</td>
<td>CTC TAG ATC TGG CAT TTC ACT TAT A</td>
<td>GCA TTG GTC ATA ATC TTG TCAG TGA</td>
<td>no</td>
</tr>
<tr>
<td>ssrpqtl_3</td>
<td>GCC AAG TAT CCA TAC TTA T</td>
<td>GTT AGA TCT GAT GAG CTG TTA</td>
<td>no</td>
</tr>
<tr>
<td>ssrpqtl_4</td>
<td>TGGTATGATTTACATAGTATGATTTTATGTTT</td>
<td>ACCTTACATCAACATCAATTTAGGGCTC</td>
<td>yes</td>
</tr>
<tr>
<td>ssrpqtl_5</td>
<td>TCCTGTGATGATATATAGTATGTTTCAATCATGAAG</td>
<td>GAAATGAGAGCATATTCAATACAAAGATGA</td>
<td>yes</td>
</tr>
<tr>
<td>ssrpqtl_6</td>
<td>GCTAAATGGGATCAGATTTTTATTGTTT</td>
<td>ACGAAAATATCATCATATCAATTCTCTTCTTT</td>
<td>yes</td>
</tr>
<tr>
<td>ssrpqtl_7</td>
<td>CACATCCCTAAGGGTCAAAAATAGGAA</td>
<td>GCACCTAGGTTGGCTAAAATCCACA</td>
<td>yes</td>
</tr>
<tr>
<td>ssrpqtl_8</td>
<td>GCATTTATATATTACTATTATTCGATGTATTT</td>
<td>CAATAAACATTTCATTTAAATCAGTCTACA</td>
<td>yes</td>
</tr>
<tr>
<td>ssrpqtl_9</td>
<td>GAGGAGATAGCTAGGGAAAAATATGTTG</td>
<td>GGAAGATTTCAAGATGACACGAAGACTAAGACA</td>
<td>yes</td>
</tr>
<tr>
<td>ssrpqtl_10</td>
<td>TCTGATTATAGGAAACAAATGACTGGA</td>
<td>GGACGCCATTTCTCTTGTGGA</td>
<td>yes</td>
</tr>
<tr>
<td>ssrpqtl_11</td>
<td>ATCACTGTCGTCCTCGCGTT</td>
<td>AAAATTATGCTAGTAAAGAGAAGAGATGCGA</td>
<td>yes</td>
</tr>
<tr>
<td>ssrpqtl_12</td>
<td>AAGCTCTGGACTCTCCTCAT</td>
<td>AACCATTTCATCGTACCAC</td>
<td>yes</td>
</tr>
<tr>
<td>ssrpqtl_13</td>
<td>ATT GCA CAG GTA TTG ACG</td>
<td>TAT TGC AGT ATA ATG AA</td>
<td>yes</td>
</tr>
<tr>
<td>ssrpqtl_14</td>
<td>CCC ATC ATT AGA CTA GC</td>
<td>AGT TTT ATT GAT TGG TAG TAT A</td>
<td>yes</td>
</tr>
<tr>
<td>ssrpqtl_15</td>
<td>CCT CTA TGG TAG CT T AA T</td>
<td>GAT ATC TAA TTA GCT AGC ACA</td>
<td>yes</td>
</tr>
<tr>
<td>ssrpqtl_16</td>
<td>GAA GAA TTT CCG TGC CT</td>
<td>GTG GGG AAT TTA TCA GTG CTG</td>
<td>yes</td>
</tr>
<tr>
<td>ssrpqtl_17</td>
<td>GAA TAT TTG TCA CCG ATA GT</td>
<td>GAG TAA GAG ATT TCA CGT TTG</td>
<td>yes</td>
</tr>
<tr>
<td>ssrpqtl_18</td>
<td>ACA TTT GCT TAG GCC AAT ATA C</td>
<td>CTA ATCA CAT ATA TTG CCA TTG AAG TAA</td>
<td>yes</td>
</tr>
<tr>
<td>ssrpqtl_19</td>
<td>ATC ACC AAT GGT AGG CTC G</td>
<td>AAG TTT CTT TCC TTA GCT CCC</td>
<td>yes</td>
</tr>
<tr>
<td>ssrpqtl_20</td>
<td>GTGTGGTTGAAAGCCATGCT</td>
<td>TCACACAGCAGACACCTTTGA</td>
<td>yes</td>
</tr>
<tr>
<td>ssrpqtl_21</td>
<td>CCCCTTCCTCACAGATCTCAA</td>
<td>TGCCATTTTGAGATCCCTTTTC</td>
<td>yes</td>
</tr>
<tr>
<td>ssrpqtl_22</td>
<td>CAAAGTTGACAGGCCACTAA</td>
<td>ACCCCCAGATGATAGCTTTCC</td>
<td>yes</td>
</tr>
</tbody>
</table>

Segregation status refers to the segregation of the markers in the P-C609-45-2 population.
| ssrpqtl_23 | GCCGAACAAATCAATGACA | TGCAAAAGGACCATGTGATGT | yes |
| ssrpqtl_24 | CATCCCTCCCCCTAAACATA | TGGAATTGGCTTGTCGCCTTA | yes |
| ssrpqtl_25 | TGCAAGATCTGTACACGGTA | GAAATTTGCTTCTTCAACGCTTT | yes |
| ssrpqtl_26 | AAACAAATTCCAGGCAAGGAA | AAGTTTTGCAGCAAGCAT | yes |
| ssrpqtl_27 | CCC CAC TAT TCC TAC TTC TAC CAG | CAA AAT TAA CCT CCC AGC ACT TGT G | yes |
| ssrpqtl_28 | CAT TAA AAA TAT CAA TAG AAA TGT G | TGT TTT TGT ACT ATC CAT CG | yes |
| ssrpqtl_29 | TTG AAA ATT GAT AAT CGA TCT GAT | GCA GCA GAG AGT TTT GTA GAG TTA | yes |
| ssrpqtl_30 | CAG TTT ATT TGA TAT TAC TAC TG | TAA AAC TAA CAC TAA CGT ATC | yes |
| ssrpqtl_31 | ATC ACC ATA ATC AAT TAA ACG | TGT TTT GCT GAT ATG GGT AG | yes |
| ssrpqtl_32 | GTA AGA TTT GAA TTA TGC TGC ATA T | ATT CCA TCA AAT AAA AGT TGT TAC G | yes |
| ssrpqtl_33 | GGG TGA ATT ATA CTA TAT TAC | TTG ATT ACA GAA TAA CCA | yes |
| ssrpqtl_34 | GAA ATT CAT GTC CAG AAC | TAC ATC ATA TAC GGT ATG AGA | yes |
| ssrpqtl_35 | TAC CAA TAA TAT TGA GGT T | TAT ATA GAA AAC AAT CAA GTG | yes |
| ssrpqtl_36 | TTT GCA TAT TTG TTA CGT GCT | TAG TGA TCT ATA CCG GTT AT | yes |
| ssrpqtl_37 | AGA GTA GCA GTA ACG C | CAT TCC AGA AAA TTT CGA TAC | yes |
| ssrpqtl_38 | ATG ATA AAG AGA AAA AGT GAT AGA T | AAG ACA TCA ATG TAC CTC ATC ATT A | no |
| ssrpqtl_39 | AAC CTT TTA AAT AAG ATA AGA | AGG GTT TTA TAT TAT GAA G | no |
| ssrpqtl_40 | ATA AGC ACA AAA GGT CCA AGG | ATA GTC CAA GAA ACA ACG GAT | no |
| ssrpqtl_41 | TAT TTA TAC CTA GGG TAC TCA | GGC TCA TCC GTT ATG TA | no |
| ssrpqtl_42 | GAC AAG AAA TTC TGG ATG GC | TGAATGGATTTGAGAAGCGGA | no |
CHAPTER THREE

STRUCTURAL AND TRANSCRIPTIONAL DIVERGENCE OF A
MAJOR SEED PROTEIN QTL REGION AND ITS HOMOELOGOUS
REGION IN SOYBEAN

A manuscript to be submitted to *Genome*

Bindu Joseph\(^1\), Yung-Tsi Bolon\(^2\), Steven Cannon\(^3\), Brian Diers\(^4\), Andrew Farmer\(^5\),
Michelle Graham\(^3\), Gregory May\(^2\), James Specht\(^6\), Nathan Weeks\(^3\), Carroll Vance\(^2\),
Randy Shoemaker\(^3\)

ABSTRACT

Polyploidy is a major source of evolutionary innovation and genetic diversity in
plants. It is often followed by genome rearrangements and functional divergence of
duplicates. Duplicate genes are preserved primarily through subfunctionalization where the
ancestral gene function is partitioned between duplicated genes. There has been in-depth
analysis of expression divergence of duplicated genes in Arabidopsis, measured as alterations
in spatiotemporal profiles as well as changes in transcript abundance, utilizing expression
data from a range of tissues/developmental stages. Similar studies are lacking in many plant
species. In this paper, we describe the comparative analysis of an approximately 8.5 Mbp
region surrounding a major seed protein/oil QTL on soybean chromosome 20 with its

\(^1\)Department of Agronomy, Iowa State University, Ames, IA 50011, USA; \(^2\)USDA-ARS,
Plant Research Unit, St. Paul, MN 55108, USA; \(^3\)USDA-ARS, Corn Insects and Crop
Genetics Research Unit, Ames, IA 50011, USA; \(^4\)University of Illinois, 1101 West Peabody
Dr., Urbana, IL 61801, USA; \(^5\)National Center for Genome Resources, Santa Fe, NM 87505,
USA; \(^6\)Department of Agronomy, University of Nebraska, Lincoln, NE 68583, USA.
homoeologous region on chromosome 10, to understand the structural and expression divergence. Consistent with the pericentromeric location of the regions on the corresponding chromosomes, we observed a major portion of the region is gene poor (retrotransposon rich) transitioning to a more gene rich euchromatic region. Synteny was poor within the gene poor regions while gene rich regions were highly syntenic and colinear. The synonymous distance (Ks) values indicated the homoeologues are likely to be products of the most recent genome duplication in soybean, 10-14 MYA. Duplicated genes belonged primarily to functional categories of stress response, protein metabolism and DNA binding related genes and are under purifying selection. This study analyzed the expression divergence of duplicated genes in soybean, utilizing digital expression data from seed transcriptome. We observed significant differences in transcript abundance from duplicated genes in the developing seed. Expression subfunctionalization is supported by the tissue specificity in expression determined from soybean EST data. However, expression divergence was not correlated to synonymous distances (Ks) or nonsynonymous distances (Ka).

INTRODUCTION

Polyploidy is ubiquitous among flowering plants (Wendel, 2000) with estimates of up to 70% of angiosperms having undergone at least one round of genome duplication during their evolution (Leitch and Bennett, 1997; Vision et al., 2000; Blanc and Wolfe, 2004). The ‘genomic shock’ caused by polyploidization often leads to dynamic and massive genome rearrangements and gene loss (Song et al., 1995). Levy and Feldman (2002) proposed an evolutionary phase of rapid genetic and epigenetic changes immediately after polyploidization followed by a phase characterized by slow changes in DNA sequence and
rearrangement in later generations. However, in synthetic cotton polyploids the genome changes following polyploidization were negligible (Liu et al., 2001).

Gene duplicates created either by polyploidization or by regional genomic events like tandem duplications or dispersed duplications are subject to evolutionary divergence. The evolutionary fates of duplicated genes include functional redundancy, pseudo-gene formation, subfunctionalization and neofunctionalization (Force et al., 1999; Lynch and Force, 2000). Subfunctionalization in the expression of duplicated genes now referred to as ‘expression subfunctionalization’ (reviewed in Doyle et al., 2008), is the partitioning of ancestral expression domains among duplicated genes.

Expression divergence of duplicated genes has been studied in depth in cotton (Adams et al., 2003; Hovav et al., 2008) and Arabidopsis (Haberer et al., 2004; Ganko et al., 2007; Kliebenstein et al., 2008). At the genic level, expression of homoeologues in polyploid cotton indicated unequal contribution to the transcriptome. This variation in expression relates to the genomic dominance when divergent genomes are combined into one nucleus (Adams et al., 2003; Hovav et al., 2008). It was concluded that hybridity and the dosage imbalance during polyploid formation could trigger the expression subfunctionalization (Doyle et al., 2008). With regard to the timing of expression divergence, the homoeologue bias in expression appears immediately after genome merger and is maintained following genome doubling in cotton (Flagel et al., 2008), acting more quickly than classic subfunctionalization. This could preserve a large number of homoeologues from mutational decay (Adams et al., 2003; Rapp and Wendel, 2005). On the contrary, Arabidopsis paleo-homoeologues from genome duplications show less expression subfunctionalization
compared to other duplicates and may not be central to the preservation of homoeologues (Blanc and Wolfe, 2004; Casneuf, 2006; Ganko et al., 2007). Analysis of Arabidopsis duplicated genes indicated the pattern of expression divergence vary between different classes of duplicates, say tandem duplicates, and segemental duplicates. Tandem duplicates exhibited increased variation in gene expression in Arabidopsis compared to segemental duplicates (Kliebenstein et al., 2008).

Even though expression subfunctionalization of homoeologues following polyploidization appears to be widespread, studies in legumes are limited. Here we explore the expression divergence of homoeologues in soybean using high throughput transcriptome data. Soybean is a diploidized polyploid, with at least two rounds of whole genome duplications at approximately 15 and 44 MYA (Schlueter et al., 2004). We focus our analysis to an 8.4 Mbp region around a major seed protein/oil QTL on chromosome 20 (LG I) and its homoeologous region on chromosome 10 (LG O). Both regions fall within pericentromeric regions on respective chromosomes. The structural and transcriptional divergence of these homoeologous regions was analyzed. Results indicated that both regions act as hot spots for retrotransposon accumulation. Homoeologues showed evidence of differential transcript accumulation in developing seed as well as evidence of tissue specific expression divergence in a range of tissues studied using EST data.

**MATERIALS AND METHODS**

**Identification of homoeologous regions**

The soybean seed protein/oil QTL region on chromosome 20 (LG I) was defined as previously described (Bolon et al., in preparation). The predicted gene models
(www.soybase.org) within the QTL region (excluding splice variants and transposable
element-like gene models) were used for a BLASTN (Altschul et al., 1990) search against the
soybean genome sequence (www.soybase.org). Besides the self match, the next best match
for the majority of the predicted genes was to chromosome 10 (LG O). BLASTN of the
predicted coding sequences within the QTL region against the chromosome 10 genomic
sequence identified the syntenic region corresponding to the region on chromosome 20. The
gene models in the region of synteny on chromosome 10 were considered as homoeologues.
The wgvista (Couronne et al., 2003) option of the vista tools for sequence alignment (Frazer
at al., 2004) was used to visualize the extent of synteny between the two regions using repeat
masked sequences available at ftp.jgi-
psf.org/pub/JGI_data/Glycine_max/Glyma1/assembly/masked/.

*Estimation of Ks, Ka and expression divergence*

The synonymous (Ks) and the nonsynonymous (Ka) distances of the homoeologues
were determined using PAML (Yang, 1997) based on codon alignment of the aligned protein
sequences generated by PAL2NAL (Suyama et al., 2006). Expression divergence of the
homoeologues was estimated using two strategies; differential transcript accumulation and
tissue specificity of expression. Illumina high throughput transcript counts of seed
transcriptome described in Bolon et al. (in preparation) were used to determine the difference
in transcript accumulation of homoeologues. Transcript counts estimated for the
homoeologues from the low protein line LoPro (Bolon et al., in preparation) at four stages of
seed development were used in this study. The counts were normalized using a scaling factor
obtained by dividing the average counts of all four stages by the counts for each stage. To account for the differences in size of the predicted coding regions of homoeologues, normalized counts per Kb of the coding region was determined. Homoeologues showing statistically significant difference in transcript accumulation were determined based on the GENMOD procedure in SAS using a negative binomial distribution.

Tissue specificity was determined from the soybean EST data. The singletons or Tentative Consensus sequences (TCs) associated with each homoeologue were identified by BLASTN alignment of the coding region against the Dana- Farber Cancer Institute (DFCI) soybean gene index (release13, July 2008) with an e-value cut off of $e^{-20}$. To ensure the homoeologue specificity of each aligned EST/TC, the following criteria were used. The best match to the corresponding homoeologous region based on BLAST alignment against the whole genome sequence with a minimum of 98% identity and 98% coverage. The tissue from which each EST (singletons or ESTs in TC) originated was determined using custom perl scripts. The Jaccard similarity index (Shoja et al., 2007) was used to score the tissue-specific expression divergence of the homoeologues. The distributions of Ks and Jaccard index were plotted using JMP 7.0.2 analysis package from SAS.

**GO annotations of the homoeologues**

The coding sequences of the homoeologues were compared to the most recent release of predicted genes in the Arabidopsis genome (TAIR v. 8, www.arabidopsis.org) using TBLASTX ($E<10^{-4}$, Altschul et al. 1997). The top Arabidopsis gene was used to query the

RESULTS

Genome organization of the homoeologous regions

The protein/oil QTL on chromosome 20 (LG I) was defined to an approximately 8.4 Mbp pericentromeric region (Bolon et al., in preparation). The corresponding homoeologous region spanned approximately 7.3 Mbp of the pericentromeric region on chromosome 10 (Figure 1). A total of 345 genes were predicted in the chromosome 20 region (excluding splice variants and transposable element like gene models) while the corresponding chromosome 10 region contained 369 predicted genes. Even though the numbers of genes were comparable in both regions, chromosome 10 region showed an overall higher gene density compared to the chromosome 20 region. The distribution of genes was not uniform across the region. For both homoeologous regions, as expected, regions closer to the centromere were rather gene poor (retrotransposon rich) while ~1 Mbp region farthest from the centromere was gene rich (Figure 2). Analysis of synteny between the regions revealed 138 out of 352 (39%) genes on chromosome 20 had a homoeologue in the corresponding region on chromosome 10 with 114 genes showing colinearity. Two of the colinear genes on chromosome 20 and one gene on chromosome 10 showed tandem duplications as well. As evident from Figure 2, synteny was minimal in the gene poor region while synteny and colinearity was extensive towards the gene rich region. Fifty eight of the 114 colinear genes were located in the ~ 7 Mbp gene poor regions while the remaining 56 colinear genes were located in the adjacent gene rich region. Only the colinear genes were considered for further
analysis because they are likely to have been derived from the same genome duplication event.

**Homoeologue bias in genome expansion**

The syntenic region on chromosome 20 was approximately 1 Mbp longer than the corresponding region on chromosome 10. This raises the question, is there a bias in genome expansion between these regions? To address this, the intergenic distances between colinear genes were used as a measure of genome expansion (Figure 3). It was evident that expansion was not strictly biased to any one of the regions; both regions showed evidence of expansion. For both regions, genome expansion was more pronounced in the gene poor regions (Figure 3). The intergenic distances as high as 905 and 791 Kb were observed in the gene poor regions of chromosome 20 and chromosome 10, respectively. However, the largest intergenic distances in the gene rich region were 94 and 86, respectively, on chromosome 20 and 10. As expected, the main cause of the variation in intergenic distances of corresponding colinear genes was accumulation of retrotransposons, eventhough non-colinear genes also could have contributed to the difference in intergenic distance in a few intervals (Data not shown).

**Nucleotide divergence of homoeologues**

The synonymous (Ks) and nonsynonymous (Ka) substitutions in the protein coding regions were estimated for the homoeologues. The distribution of Ks values (Ks<1.5) is shown in Figure 4. Upwards of 90% of the genes showed Ks values less than 0.6 suggesting these duplicated genes could have originated from the most recent genome duplication in
soybean, 10-14 MYA (Schlueter et al., 2004). The Ka/Ks ratios ranged from 0.001 to 99 with 98% of the homoeologous pairs with ratio <1.

**Expression divergence of the homoeologues**

Expression divergence of the homoeologues was studied in two ways; differential transcript abundance and tissue specificity. High throughput Illumina digital transcript counts from a low protein line (LoPro) at four stages of seed development (Bolon et al., in preparation) were used to compare the transcript abundance of homoeologues. Prior to the comparison of each pair of homoeologues, the average transcript count for all the genes in each region was used to see if there was a significant difference in transcript accumulation between the regions. There was no significant difference between the average transcript counts from the chromosome 20 region vs the chromosome 10 region for all four stages of seed development. This means that a homoeologue difference in transcript accumulation, if any, is not due to a regional effect.

The normalized transcript counts per Kb estimated for four stages of seed development was used to determine the homoeologue difference in transcript abundance. The normalized transcript counts per Kb for all consecutive homoeologue pairs are shown in Figure 5. Twelve homoeologue pairs, for which both copies did not show evidence of expression, were excluded from transcript analysis. Thirteen pairs showed expression of only one of the copies. Out of the remaining 89 homoeologue pairs for which both copies were expressed in the seed, 36 pairs showed a significant difference in transcript counts. Out of the
36 showing significant difference, 22 pairs (61%) showed higher transcript accumulation from the homoeologue on chromosome 10. The fold change in transcript counts and the p-values for the 89 homoeologue pairs are listed in Suppl. Table 1. The fold change in transcript abundance was plotted against the Ks (Figure 6) and Ka values (Data not shown). There was no evidence of correlation ($r = 0.197$, p-value $= 0.0673$) between the expression divergence and the rate of synonymous or non-synonymous substitutions.

The spatial divergence in expression also was measured using EST data. EST libraries were grouped into six tissue categories; roots, hypocotyls/leaves, seeds/pods, somatic embryo and cotyledons. The evidence of homoeologue expression in these tissues was examined. The Jaccard similarity index was used to score the tissue specificity of homoeologue expression. This index ranges from 0 to 1, where 0 means complete divergence in tissue specificity and 1 complete overlap of tissues of expression. The distribution of the index showed evidence of no overlap of tissues of expression to complete overlap of tissues in which the homoeologues are expressed (Figure 7). However, the homoeologue group with a score of 1, exhibiting complete overlap of tissues, was the largest group.

**GO annotations**

The Arabidopsis Gene Ontology (GO) terms were used to determine the functional categories of duplicated genes. Functional categories of the majority of the homoeologues remain unknown (Table 1). Among the genes for which a GO slim annotation for biological process could be assigned, genes involved in protein metabolism and stress response were predominant, followed by genes involved in transcription and transport (Table 1). With
regard to molecular function, DNA/RNA binding and hydrolase categories were predominant followed by genes involved in protein binding, nucleotide binding, kinases and transporter categories.

**DISCUSSION**

This study compares the evolutionary divergence of a genomic region around a major seed protein QTL region on chromosome 20 and its homoeologous region on chromosome 10. The homoeologous relationship between chromosome 20 and 10 was reported previously (Schlueter et al., 2007). The rates of synonymous substitutions (Ks) of genes encompassed within these regions indicate these regions were likely to have been created from the most recent genome duplication event in soybean around 10 - 14 MYA (Schlueter et al., 2004). The location of these homoeologous regions towards the end of pericentromeric regions transitioning to euchromatic regions provides a unique opportunity to compare the evolutionary divergence of gene poor (retrotransposon rich) and gene rich regions. Utilizing high throughput digital expression data from developing seed, we attempted to gain insights into the expression subfunctionalization of homoeologues in soybean.

The expansion of the intergenic regions of colinear genes indicates that the pericentromeric portions of both homoeologous regions act as hot spots for retrotransposon accumulation. However, the chromosome 20 region has a net gain of approximately 1 Mbp DNA suggestive of a preferential accumulation of retrotransposons. It should be noted that chromosome 10 (50.9Mbp) is larger than chromosome 20 (46.7Mbp) (Figure 1). Therefore, even though there seemed to be an expansion of the genome in the region around the QTL on chromosome 20, there could be other genomic regions where chromosome 10 could have
gained DNA relative to chromosome 20. The relative efficiency of genome expansion and contraction processes (Devos et al., 2002; Bennetzen et al., 2005) could vary between genomic regions.

The extent of synteny in the gene poor (pericentromeric) region was minimal in contrast to the extensive synteny and colinearity observed for the gene rich region. The disruption of colinearity observed in the pericentromeric portions could be due to gene loss or gain. Given the abundance of retrotransposons, it is likely that gene loss or gain could be common due to retrotransposition as well as illegitimate or intra-strand recombinations brought about by LTR retrotransposons (Devos et al., 2002; Ma et al., 2004). Which of the two chromosomes experienced gene loss or gain could not be inferred without comparison to another legume diverged from a common ancestor with soybean.

Functional divergence of duplicated genes is a major source of evolutionary novelty (Ohno, 1970). We analyzed the expression divergence of homoeologues in this region in terms of the difference in transcript abundance and tissue specificity of homoeologue expression. High throughput transcript sequencing allowed the detection of low abundant transcripts as well as duplicate specific transcripts without the noise from cross hybridization. The data represents absolute counts of transcripts, unlike array based platforms where transcript abundance is relative (t’Hoen et al., 2008). Our results indicate that 40% of the homoeologue pairs with both copies expressed in the seed, showed significant difference in transcript abundance. Homoeologue bias in expression was evident in both chromosomes where 22 out of 36 (61%) showed higher transcript accumulation from the homoeologue on chromosome 10 while for others the opposite was observed. The extreme of this bias, where
only one of the homoeologues was expressed also was observed for 13 cases. The tissue specificity in expression further supported the observation of expression subfunctionalization of homoeologues. Given the range of tissues studied, there evidence of complete redundancy in expression to complete subfunctionalization. Similar homoeologue bias in expression is well documented in cotton following polyploidy (Adams et al., 2003; Hovav et al., 2008).

There was no evidence of correlation between expression divergence and Ks or Ka in our study. A number of studies in recent years have shown either no correlation or a positive correlation between expression divergence and sequence divergence. Gu et al. (2002) found a significant positive relationship between the divergence in expression and number of synonymous (K_S) and nonsynonymous (K_A) substitutions per site in yeast, though the relationship held only for K_A < 0.3. This contradicted an earlier study, where no significant correlation between expression and protein sequence divergence was observed (Wagner, 2000).

Expression divergence of duplicated genes measured in a range of tissues in Arabidopsis showed no relationship between divergence in expression profile and the number of synonymous substitutions per site (K_s), while a negative relationship was observed with the number of nonsynonymous substitutions per site (K_a) among segmental duplicates. It was concluded that there appears to be a coupling between the rate of expression divergence and the rate of nonsynonymous substitution for a gene pair, presumably through covariation in the strength of functional constraint (or positive selection) among gene pairs (Ganko et al., 2007). A lack of correlation between sequence divergence of protein coding regions and expression observed in our study may indicate that the divergence in expression could be
attributed to changes in regulatory elements or epigenetic changes. The expression of homoeologues expected to be similar soon after duplication. Then, over time the expression patterns are expected to diverge, possibly due to changes in regulatory elements as proposed by the DDC (Duplication-Degeneration-Complementation model of subfunctionalization (Force et al., 1999). Haberer et al. (2004) noted that tandem and segmental duplicate gene pairs had divergent expression in Arabidopsis even when they shared many similar cis-regulatory sequences changes to a small fraction of cis-elements could be sufficient for neofunctionalization or subfunctionalization. However, expression divergence of homoeologues immediately following polyploidization was observed in cotton even before any sequence divergence was observed (Adams et al., 2003).

One consistent theme across many studies of duplicate gene evolution is the similarity of the functional categories of genes retained as duplicates (Seoighe and Gehring, 2004; Blanc and Wolfe, 2004; Paterson et al., 2006). In agreement with previous studies, our study showed that the retained duplicates were predominantly transcription factors or genes involved in stress response or signaling. In a study of Arabidopsis duplicates, Kliebenstein et al. (2008) concluded that gene duplication provides raw material for evolution, but the biology of the pathway determines the likelihood that the duplicates are maintained to evolve alternate functionality. In secondary metabolic pathways where variation is required, duplicates are more likely to evolve diverse expression patterns, whereas in primary metabolic pathways where variation may be detrimental, may select against the presence of duplicated genes (Kliebenstein et al., 2008).
Close to 50% of the duplicated gene pairs identified in our study are located in retrotransposon rich regions. This would mean that selection against the loss of these duplicate pairs could be the main reason for the retention of these duplicates. Having duplicate copies of genes involved in signaling or stress responses could add to the adaptation and therefore purifying selection could play a role in the retention of these duplicates. Ka/Ks ratio <1 for majority of the duplicated pairs further supports this argument. Duplicated genes evolve under purifying selection acting with similar strength on both duplicated copies of genes (Lynch and Corney, 2000). In a study of duplicated genes in bacterial, archaeal, and eukaryotic genomes, duplicated genes that achieve fixation in the population increase fitness when present in two or more copies in the genome and are subject to purifying selection from the moment of duplication (Kondrashov et al., 2002).

Our analysis was centered on the QTL homoeologous regions on recently duplicated homeologous chromosomes 20 and 10. Further analysis of duplicated genes at the genome level, utilizing expression data from a range of tissues or developmental stages, would facilitate better understanding of the evolution of duplicated genes in a paleopolyploid like soybean.

ACKNOWLEDGEMENTS

The authors are grateful to the North Central Soybean Research Program (WBS # 26-6222-0183-005) and National Science Foundation (Grant # DBI-0501877) for financial support for this project.
REFERENCES


Figure 1. Chromosomal location of the homoeologous regions. The grey shaded region along the chromosome represents the pericentromeric region. The Black oval region shows the position of the centromere. Black rectangles show the homoeologous regions.
**Figure 2.** Synteny between the homoeologous regions. The vista plot of synteny using repeat masked genomic sequences of the homoeologous regions. 8.4Mbp region on chromosome 20 was used as the reference region. The positions of the predicted gene models on chromosome 20 are indicated on top of the synteny plot. The percent scales on the right side of the plot indicate the percent identity. The peaks with blue color are conserved genic regions while peaks with pink color are conserved intergenic regions. Synteny for the first 4.1 Mbp on chromosome 20 is shown in this figure.
**Figure 2 continued.** The synteny for the 4.1-8.3 Mbp is shown here.
Figure 3. Intergenic expansion measured as distance between colinear genes on both chromosomes. A. Expansion in the gene poor (retrotransposon rich) region. B. Expansion in the gene rich region.
Figure 4. The distribution of the synonymous distances (Ks < 1.5) of the duplicated genes.
Figure 5. Transcript counts of the homoeologue pairs. The transcript counts are normalized and per Kb of the coding region. The homoeologue pairs are in the consecutive order as they are in the region. From left to right the position of the genes are farther from the centromere.
Figure 6. The plot of Ks against fold change in transcript abundance of homoeologues. Homoeologues with Ks >1 are not included in this analysis. Pearson correlation coefficient ($r = 0.197$, $p$ value = 0.0673) was used to estimate the correlation.
Figure 7. The distribution of Jaccard similarity index measuring the tissue specificity in expression of homoeologues.
Table 1. Predominant GO slim categories of duplicate genes.

<table>
<thead>
<tr>
<th>GO slim category</th>
<th>Type</th>
<th>Number of duplicate gene pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Protein metabolism</td>
<td>BP</td>
<td>12</td>
</tr>
<tr>
<td>Response to biotic/abiotic stress</td>
<td>BP</td>
<td>11</td>
</tr>
<tr>
<td>Transport</td>
<td>BP</td>
<td>6</td>
</tr>
<tr>
<td>Transcription</td>
<td>BP</td>
<td>5</td>
</tr>
<tr>
<td>Cell organization or biogenesis</td>
<td>BP</td>
<td>4</td>
</tr>
<tr>
<td>Unknown</td>
<td>BP</td>
<td>65</td>
</tr>
<tr>
<td>DNA/RNA binding</td>
<td>MF</td>
<td>10</td>
</tr>
<tr>
<td>Hydrolase activity</td>
<td>MF</td>
<td>10</td>
</tr>
<tr>
<td>Protein binding</td>
<td>MF</td>
<td>7</td>
</tr>
<tr>
<td>Nucleotide binding</td>
<td>MF</td>
<td>6</td>
</tr>
<tr>
<td>Kinase activity</td>
<td>MF</td>
<td>5</td>
</tr>
<tr>
<td>Transporter activity</td>
<td>MF</td>
<td>5</td>
</tr>
<tr>
<td>UnKnown</td>
<td>MF</td>
<td>55</td>
</tr>
</tbody>
</table>

BP-Biological Process; MF- Molecular Function.

Only the GO categories with five genes or more are listed.
**Supplementary Table 1. Transcript fold change, Ks, Ka and Ka/Ks ratios of homoeologue pairs.**

<table>
<thead>
<tr>
<th>Gene ID chr_20</th>
<th>Gene ID chr_10</th>
<th>Fold change</th>
<th>Std. error</th>
<th>P-value</th>
<th>Ks</th>
<th>Ka</th>
<th>Ka/Ks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Glyma20g17550</td>
<td>Glyma10g23810</td>
<td>1.0185</td>
<td>0.2278</td>
<td>0.9346</td>
<td>0.0901</td>
<td>0.0197</td>
<td>0.2187</td>
</tr>
<tr>
<td>Glyma20g17560*</td>
<td>Glyma10g23840</td>
<td>3.8</td>
<td>1.7484</td>
<td>0.0037</td>
<td>0.1635</td>
<td>0.082</td>
<td>0.5014</td>
</tr>
<tr>
<td>Glyma20g17960</td>
<td>Glyma10g24060</td>
<td>0.9813</td>
<td>0.2742</td>
<td>0.9461</td>
<td>0.1123</td>
<td>0.0306</td>
<td>0.2723</td>
</tr>
<tr>
<td>Glyma20g18290*</td>
<td>Glyma10g24340</td>
<td>5.6667</td>
<td>1.6325</td>
<td>&lt;.0001</td>
<td>0.1161</td>
<td>0.016</td>
<td>0.138</td>
</tr>
<tr>
<td>Glyma20g18440</td>
<td>Glyma10g24350</td>
<td>1.3438</td>
<td>0.4061</td>
<td>0.3282</td>
<td>0.1664</td>
<td>0.0161</td>
<td>0.097</td>
</tr>
<tr>
<td>Glyma20g18480</td>
<td>Glyma10g24360</td>
<td>0.4842</td>
<td>0.3719</td>
<td>0.7681</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Glyma20g18550*</td>
<td>Glyma10g24360</td>
<td>0.3982</td>
<td>0.1775</td>
<td>0.0388</td>
<td>0.346</td>
<td>0.1719</td>
<td>0.4967</td>
</tr>
<tr>
<td>Glyma20g18620</td>
<td>Glyma10g24420</td>
<td>0.4375</td>
<td>0.1853</td>
<td>0.051</td>
<td>0.1227</td>
<td>0.0482</td>
<td>0.3929</td>
</tr>
<tr>
<td>Glyma20g18870</td>
<td>Glyma10g24540</td>
<td>0.8729</td>
<td>0.0948</td>
<td>0.2107</td>
<td>0.1611</td>
<td>0.014</td>
<td>0.0866</td>
</tr>
<tr>
<td>Glyma20g18890</td>
<td>Glyma10g24550</td>
<td>1.3617</td>
<td>0.3426</td>
<td>0.2197</td>
<td>0.0855</td>
<td>0.0157</td>
<td>0.1841</td>
</tr>
<tr>
<td>Glyma20g18900</td>
<td>Glyma10g24570</td>
<td>1.5</td>
<td>0.8173</td>
<td>0.4568</td>
<td>0.1601</td>
<td>0.0678</td>
<td>0.4234</td>
</tr>
<tr>
<td>Glyma20g18970*</td>
<td>Glyma10g24580</td>
<td>3.5294</td>
<td>1.0392</td>
<td>&lt;.0001</td>
<td>0.187</td>
<td>0.0923</td>
<td>0.4935</td>
</tr>
<tr>
<td>Glyma20g18980</td>
<td>Glyma10g24590</td>
<td>1.2412</td>
<td>0.1953</td>
<td>0.1696</td>
<td>0.1123</td>
<td>0.0108</td>
<td>0.0961</td>
</tr>
<tr>
<td>Glyma20g19000*</td>
<td>Glyma10g24620</td>
<td>2.583</td>
<td>0.5179</td>
<td>&lt;.0001</td>
<td>0.2659</td>
<td>0.0052</td>
<td>0.0195</td>
</tr>
<tr>
<td>Glyma20g19200*</td>
<td>Glyma10g24630</td>
<td>2.16</td>
<td>0.8472</td>
<td>0.0496</td>
<td>0.1694</td>
<td>0.0129</td>
<td>0.0762</td>
</tr>
<tr>
<td>Glyma20g19210</td>
<td>Glyma10g24650</td>
<td>0.962</td>
<td>0.2231</td>
<td>0.8674</td>
<td>0.2504</td>
<td>0.1013</td>
<td>0.4046</td>
</tr>
<tr>
<td>Glyma20g19640</td>
<td>Glyma10g25440</td>
<td></td>
<td>0.134</td>
<td>0.0258</td>
<td>0.1922</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Glyma20g19660</td>
<td>Glyma10g25480</td>
<td>1.3182</td>
<td>0.2895</td>
<td>0.2084</td>
<td>0.5561</td>
<td>0.2324</td>
<td>0.4179</td>
</tr>
</tbody>
</table>

The fold change in transcript abundance, Standard error of the fold change, and the p-values are listed along with the Ks, Ka, and Ka/Ks ratios. Fold change >1 indicates higher transcript accumulation from the homoeologue on chromosome 10; <1 indicates higher transcript accumulation from the homoeologue on chromosome 20. The list includes the 89 homoeologue pairs with both copies expressed. The genes with the asterisk are the homoeologue pairs with significant difference (P <0.05) in transcript abundance.
| Glyma20g19670 | Glyma10g25510 | 0.9362 | 0.2455 | 0.8014 | 0.131 | 0.0495 | 0.3781 |
| Glyma20g19970* | Glyma10g25620 | 1.963 | 0.536 | 0.0135 | 0.1381 | 0.0329 | 0.2386 |
| Glyma20g19980 | Glyma10g25630 | 1.2473 | 0.396 | 0.4864 | 0.1227 | 0.0102 | 0.083 |
| Glyma20g20010* | Glyma10g25670 | 1.902 | 0.4618 | 0.0081 | 0.1182 | 0.0088 | 0.0748 |
| Glyma20g20040 | Glyma10g25690 | 0.8837 | 0.4556 | 0.8105 | 0.3439 | 0.1115 | 0.3243 |
| Glyma20g20050 | Glyma10g25700 | 0.5694 | 0.213 | 0.1323 | 0.0566 | 0.0225 | 0.3978 |
| Glyma20g20070* | Glyma10g25710 | 6.75 | 3.4558 | 0.0002 | 0.0911 | 0.0204 | 0.2335 |
| Glyma20g20180* | Glyma10g25760 | 0.1333 | 0.0815 | 0.001 | 0.135 | 0.027 | 0.2004 |
| Glyma20g20280 | Glyma10g25790 | 1 | 0.3407 | 1 | 0.1897 | 0.0165 | 0.0869 |
| Glyma20g20600* | Glyma10g26240 | 2.0588 | 0.7355 | 0.0432 | 0.1722 | 0.0262 | 0.152 |
| Glyma20g20630 | Glyma10g26380 | 1.307 | 0.4066 | 0.3894 | 0.1387 | 0.0129 | 0.0929 |
| Glyma20g20890 | Glyma10g26470 | 1.1176 | 1.0056 | 0.9016 | 0.1028 | 0.029 | 0.2818 |
| Glyma20g20910 | Glyma10g26530 | 0.7826 | 0.3022 | 0.5255 | 0.5035 | 0.2394 | 0.4755 |
| Glyma20g20920* | Glyma10g26640 | 1.5391 | 0.2665 | 0.0124 | 0.0796 | 0.0257 | 0.323 |
| Glyma20g21030* | Glyma10g26690 | 3.4167 | 1.0825 | 0.0001 | 0.3724 | 0.0459 | 0.1232 |
| Glyma20g21190 | Glyma10g26790 | 0.7787 | 0.1031 | 0.0589 | 0.1426 | 0.0001 | 0.001 |
| Glyma20g21230 | Glyma10g26820 | 0.4536 | 0.242 | 0.1384 | 0.2399 | 0.0002 | 0.001 |
| Glyma20g21330 | Glyma10g26870 | 1.0965 | 0.2571 | 0.6945 | 0.1196 | 0.0079 | 0.0658 |
| Glyma20g21380 | Glyma10g26990 | 1.3333 | 0.4087 | 0.348 | 0.1031 | 0.0573 | 0.5561 |
| Glyma20g21430 | Glyma10g27010 | 0.7976 | 0.1755 | 0.3041 | 0.1509 | 14.936 | 99 |
| Glyma20g21440 | Glyma10g27020 | 1.8571 | 1.3111 | 0.3806 | 0.0973 | 0.0234 | 0.2403 |
| Glyma20g21510* | Glyma10g27570 | 1.8578 | 0.4029 | 0.0043 | 0.1391 | 0.0108 | 0.0776 |
| Glyma20g21520 | Glyma10g27610 | 0.78 | 0.1884 | 0.3036 | 0.16 | 0.0147 | 0.0921 |
| Glyma20g21600 | Glyma10g27640 | 0.6522 | 0.62 | 0.653 | 0.0656 | 0.0159 | 0.2424 |
| Glyma20g22050* | Glyma10g28040 | 3.7222 | 1.355 | 0.0003 | 0.1157 | 0.0194 | 0.1676 |
| Glyma20g22070 | Glyma10g28070 | 1.1613 | 0.3115 | 0.5772 | 0.0815 | 0.0153 | 0.1882 |
| Glyma20g22090* | Glyma10g28080 | 0.6856 | 0.1118 | 0.0206 | 0.077 | 0.0142 | 0.1838 |
| Glyma20g22110* | Glyma10g28090 | 2.9189 | 1.0957 | 0.0043 | 0.7915 | 0.42 | 0.5306 |
| Glyma20g22120 | Glyma10g28100 | 0.9726 | 0.2386 | 0.9099 | 0.1296 | 0.0294 | 0.2264 |
| Glyma20g22130* | Glyma10g28110 | 1.6593 | 0.3134 | 0.0073 | 0.0941 | 0.0155 | 0.1644 |
| Glyma20g22160 | Glyma10g28170 | 1.0566 | 0.3058 | 0.8491 | 0.1102 | 0.0273 | 0.2474 |
| Glyma20g22170 | Glyma10g28180 | 1.2621 | 0.2828 | 0.299 | 0.0931 | 0.0157 | 0.1684 |
| Glyma20g22190 | Glyma10g28200 | 1.1837 | 0.2832 | 0.481 | 0.1498 | 0.0424 | 0.2831 |
| Glyma20g22200 | Glyma10g28220 | 1.1333 | 0.4396 | 0.7469 | 0.0821 | 0.0197 | 0.2405 |
| Glyma20g22210 | Glyma10g28230 | 1.0278 | 0.4088 | 0.9451 | 0.0804 | 0.029 | 0.3608 |
| Glyma20g22230 | Glyma10g28250 | 0.962 | 0.3012 | 0.9016 | 0.1094 | 0.0553 | 0.5051 |
| Glyma20g22270* | Glyma10g28270 | 22.26 | 6.9818 | <.0001 | 0.1201 | 0.0996 | 0.8291 |
| Glyma20g22280* | Glyma10g28290 | 0.2645 | 0.0702 | <.0001 | 0.0722 | 0.044 | 0.6096 |
| Glyma20g22290* | Glyma10g28300 | 0.1936 | 0.0379 | <.0001 | 0.0995 | 0.008 | 0.0807 |
| Glyma20g22300 | Glyma10g28310 | 0.7636 | 0.2058 | 0.3171 | 0.1065 | 0.0001 | 0.001 |
| Glyma20g22380 | Glyma10g28350 | 0.2917 | 0.2586 | 0.1645 | 0.125 | 0.0602 | 0.482 |
| Glyma20g22400 | Glyma10g28370 | 1.1481 | 0.2663 | 0.5515 | 0.0967 | 0.0178 | 0.1842 |
| Glyma20g22420* | Glyma10g28380 | 0.3663 | 0.1392 | 0.0082 | 0.4003 | 0.2245 | 0.5609 |
| Glyma20g22470* | Glyma10g28430 | 4 | 1.1287 | <.0001 | 1.0466 | 0.0371 | 0.0355 |
| Glyma20g22510* | Glyma10g28460 | 0.3333 | 0.1218 | 0.0026 | 0.1357 | 0.0277 | 0.2043 |
| Glyma20g22530* | Glyma10g28470 | 1.7174 | 0.3861 | 0.0162 | 0.1275 | 0.0306 | 0.2401 |
| Glyma20g22550 | Glyma10g28490 | 0.6122 | 0.1734 | 0.0833 | 0.14 | 0.009 | 0.064 |
| Glyma20g22560 | Glyma10g28510 | 1.6545 | 0.9038 | 0.3567 | 1.0024 | 0.6374 | 0.6358 |
| Glyma20g22600* | Glyma10g28530 | 0.6269 | 0.1297 | 0.024 | 0.0811 | 0.0088 | 0.1082 |
| Glyma20g22610 | Glyma10g28540 | 0.7586 | 0.1965 | 0.2862 | 1.0518 | 0.3846 | 0.3657 |
| Glyma20g22630 | Glyma10g28560 | 0.6667 | 0.1438 | 0.0602 | 0.1404 | 0.0231 | 0.1642 |
| Glyma20g22660 | Glyma10g28580 | 1.3396 | 0.4375 | 0.3706 | 0.1082 | 0.0346 | 0.3196 |
| Glyma20g22680 | Glyma10g28590 | 0.9218 | 0.131 | 0.5666 | 0.0608 | 0.0024 | 0.0397 |
| Glyma20g22690 | Glyma10g28600 | 1.25 | 0.4467 | 0.5323 | 0.0511 | 0.0233 | 0.4553 |
| Glyma20g22700* | Glyma10g28610 | 3.1 | 1.6257 | 0.031 | 0.2186 | 0.0237 | 0.1084 |
### Supplementary Table 1. Continued.

<table>
<thead>
<tr>
<th>Protein1</th>
<th>Protein2</th>
<th>Value1</th>
<th>Value2</th>
<th>Value3</th>
<th>Value4</th>
<th>Value5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Glyma20g22780</td>
<td>Glyma10g28640</td>
<td>0.9048</td>
<td>0.2212</td>
<td>0.6823</td>
<td>0.072</td>
<td>0.0228</td>
</tr>
<tr>
<td>Glyma20g22790*</td>
<td>Glyma10g28650</td>
<td>0.3404</td>
<td>0.1074</td>
<td>0.0006</td>
<td>0.2229</td>
<td>0.2113</td>
</tr>
<tr>
<td>Glyma20g22800</td>
<td>Glyma10g28660</td>
<td>5.5</td>
<td>4.9862</td>
<td>0.0601</td>
<td>0.4974</td>
<td>0.2702</td>
</tr>
<tr>
<td>Glyma20g22810*</td>
<td>Glyma10g28670</td>
<td>0.0303</td>
<td>0.0173</td>
<td>&lt;0.0001</td>
<td>0.2522</td>
<td>0.0521</td>
</tr>
<tr>
<td>Glyma20g22820</td>
<td>Glyma10g28690</td>
<td>1.3333</td>
<td>0.4845</td>
<td>0.4286</td>
<td>0.0894</td>
<td>0.0224</td>
</tr>
<tr>
<td>Glyma20g22830*</td>
<td>Glyma10g28700</td>
<td>0.6038</td>
<td>0.1482</td>
<td>0.0398</td>
<td>0.3163</td>
<td>0.1515</td>
</tr>
<tr>
<td>Glyma20g22840*</td>
<td>Glyma10g28710</td>
<td>0.0952</td>
<td>0.0587</td>
<td>0.0001</td>
<td>4.4267</td>
<td>17.6</td>
</tr>
<tr>
<td>Glyma20g22850*</td>
<td>Glyma10g28730</td>
<td>0.446</td>
<td>0.1253</td>
<td>0.004</td>
<td>0.1307</td>
<td>0.038</td>
</tr>
<tr>
<td>Glyma20g22860</td>
<td>Glyma10g28740</td>
<td>0.963</td>
<td>0.5545</td>
<td>0.9477</td>
<td>0.1239</td>
<td>0.052</td>
</tr>
<tr>
<td>Glyma20g22870*</td>
<td>Glyma10g28760</td>
<td>0.3077</td>
<td>0.0983</td>
<td>0.0002</td>
<td>0.3405</td>
<td>0.2461</td>
</tr>
<tr>
<td>Glyma20g22880</td>
<td>Glyma10g28770</td>
<td>1.3333</td>
<td>0.8152</td>
<td>0.638</td>
<td>0.134</td>
<td>0.0339</td>
</tr>
<tr>
<td>Glyma20g22910</td>
<td>Glyma10g28800</td>
<td>1.1935</td>
<td>0.258</td>
<td>0.413</td>
<td>0.1033</td>
<td>0.0394</td>
</tr>
<tr>
<td>Glyma20g22920*</td>
<td>Glyma10g28810</td>
<td>2.3448</td>
<td>0.5864</td>
<td>0.0007</td>
<td>0.1025</td>
<td>0.0335</td>
</tr>
<tr>
<td>Glyma20g22930*</td>
<td>Glyma10g28810</td>
<td>4.8571</td>
<td>1.5063</td>
<td>&lt;0.0001</td>
<td>0.1969</td>
<td>0.0918</td>
</tr>
<tr>
<td>Glyma20g22960*</td>
<td>Glyma10g28810</td>
<td>6.1818</td>
<td>2.1501</td>
<td>&lt;0.0001</td>
<td>0.1541</td>
<td>0.0398</td>
</tr>
</tbody>
</table>
CHAPTER FOUR

RETROTRANSPOSONS WITHIN SYNTENIC REGIONS BETWEEN
SOYBEAN AND *MEDICAGO TRUNCATULA* AND THEIR
CONTRIBUTION TO LOCAL GENOME EVOLUTION

A manuscript accepted pending revisions by *The Plant Genome*

Bindu Joseph, Jessica A. Schlueter, Jianchang Du, Michelle A. Graham, Jianxin Ma and
Randy C. Shoemaker

ABSTRACT

Comparative genome analyses have described the extent of macro and microsynteny among closely related legumes. The organization of the intergenic regions within syntenic blocks and the involvement of retrotransposons in the evolution of these regions have not been studied in detail. In this paper, retrotransposon rich (gene-poor) and retrotransposon poor (gene-rich) soybean regions showing synteny with *Medicago truncatula* were analyzed to understand the role of retrotransposons in the evolution of syntenic regions. The majority of the intact retroelements were inserted 2 million years ago (MYA). The abundance and the
types of retrotransposons vary in homoeologous soybean regions. The retrotransposon rich region showed local genome expansion in soybean compared to *Medicago*. The retrotransposon poor region showed local genome expansion in *Medicago* suggesting that local genome sizes do not always reflect the global genome size difference between soybean and *Medicago*. One unique observation was the three-fold expansion of an Ubiquitin Specific Protease (UBP12) gene in soybean due to multiple intronic retroelement insertions. Yet, a full-length transcript from the soybean UBP12 gene was confirmed. This study also uncovered a sample of the less explored non-LTR retrotransposons in soybean and their presence in the introns of genes.

INTRODUCTION

Plant genomes vary considerably in their nuclear DNA content, yet this variation is not correlated with an increase in biological complexity (Thomas, 1971). Variation in genome size is attributed to repetitive DNA (Flavell et al., 1974). Transposable elements constitute a major portion of the repetitive DNA of plant genomes, and LTR-retrotransposons represent the most abundant class of transposable elements contributing significantly to genome size variation (San Miguel et al., 1996; Vicient et al., 1999). Species-specific amplification of retrotransposons is attributed to the many-fold differences in genome size among closely related species within the genera *Oryza*, *Vicia*, and *Gossypium* (Piegu et al., 2006; Neumann et al., 2006; Hawkins et al., 2006).

However, this ‘one way ticket to genomic obesity’ (Bennetzen and Kellogg, 1997) is often interrupted by the forces of genome contraction. Retrotransposon-mediated genome expansion is counterbalanced by the processes of unequal homologous recombination and
illegitimate recombination that can remove retrotransposon DNA from the genome (Vitte and Panaud, 2003; Shirasu et al., 2000; Devos et al., 2002; Ma et al., 2004). A whole genome study of 41 retrotransposon families in rice indicated a rapid turnover of retroelements in the genome with a half-life of less than 3 MY (Vitte et al., 2007). Transcriptional silencing of retroelements also has been suggested as a mechanism for regulation of retrotransposon activity (Hirochika et al., 2000). Similarly, polyploidization (Wendel, 2000), another factor in genome expansion, is followed by genome rearrangements (Illic et al., 2003; Song et al., 1995) including deletion of duplicated genes, gene silencing and functional divergence of duplicated genes with one of the consequences being diplodization.

Many studies have focused on the mechanisms of local genome evolution based on comparative sequence analysis of orthologous genomic regions (Illic et al., 2003; Tikhonov et al., 1999; Swigonova et al., 2005; Ma et al., 2005). These studies indicated that gene deletions, insertions, duplications, gene movement, gene conversions along with transposable element accumulations contribute to the evolution of genomic regions. Illic et al., (2003) compared the Adh-I region of sorghum, rice, and maize and concluded that maize had an unstable genome, rice the most stable and the sorghum genome had medium stability. Similar studies were conducted in cotton where diploid A and D genomes with a two-fold difference in genome size were compared in the CesA and AdhA regions. The A and the D genomes showed no evidence of genome expansion or contraction at the CesA region. However, the AdhA region was expanded in the larger A genome by the accumulation of retroelements. In addition, the AdhA region in the smaller D genome showed twice as many small deletions compared to the A genome. (Grover et al., 2004; Grover et al., 2007).
Comparative studies among legume genomes indicate a pattern similar to grasses in that they possess conserved genomic regions (Boutin et al., 1995; Yan et al., 2003). However, colinearity is restricted to small intervals and decreases as the phylogenetic distance increases (Choi et al., 2004). Recently, a closer look at the synteny among *Glycine, Medicago, Lotus,* and *Arabidopsis* at the *HCBT* gene region revealed a network of synteny with fractionation due to gene loss, addition and rearrangements (Schlueter et al., 2008).

Although comparative studies among legumes revealed the extent and quality of synteny at the gene level (Mudge et al., 2005; Cannon et al., 2006), a detailed analysis of the role of retrotransposons in the evolution of the syntenic regions among legumes is lacking. This paper details a comparative sequence analysis of two soybean genomic regions and their homoeologues along with the corresponding syntenic regions in *Medicago*. These soybean genomic regions were chosen because they represent regions with different gene densities on the same homoeologous chromosomes. The four soybean segments studied showed distinct patterns of retrotransposon accumulation; two were hot spots for retrotransposon accumulation and had undergone genome expansion compared to *Medicago* while the other regions showed minimal retrotransposon accumulation in soybean and exhibited evidences of retrotransposon mediated genome expansion in *Medicago*.

**MATERIALS AND METHODS**

**Identification of BACs and BAC sequencing**

Two homoeologous soybean regions on soybean Linkage Groups I (chromosome 20) and O (chromosome 10) analyzed in this study were designated as the Ubiquitin Specific Protease 12 (UBP12) region and the galactinol synthase region, respectively. BACs from the
LG I UBP12 region were identified by PCR screening of multi-dimensional pools of the soybean ‘Williams 82’ BAC library (Marek and Shoemaker, 1997) using the SSR marker Satt239. BAC clones were end-sequenced using M13 forward and reverse primers at the Iowa State University DNA Sequencing and synthesis Facility. The BAC library was then rescreened by PCR using primers designed from BAC-end sequences of BAC GM_WBa0010E08 and the overlapping BAC GM_WBa0048N22 was identified. BACs were fingerprinted using restriction enzymes EcoRI and AccI. BAC overlap was confirmed by FPC 4.6.4 (Soderlund et al. 1997) and verified by PCR amplification using primers from BAC-end sequences. Shotgun sequencing of the BACs and assembly was carried out as described in Schlueter et al. (2006).

**BAC sequence analysis**

The LG I UBP12 region BAC sequences were annotated using gene prediction programs FGENESH (www.softberry.com) and GeneMark (Lomsadze et al., 2005, http://opal.biology.gatech.edu/GeneMark/eukhmm.cgi). The predicted genes were searched for similarity to known proteins by BLASTP (Altschul et al., 1990) with E-value cut off of E < E^{-10} against the National Center for Biotechnology Information (NCBI) non-redundant (nr) protein database. The predicted genes were also searched for similarity to soybean ESTs (at www.soybase.org) by BLASTN (Altschul et al., 1990) using an E-value cut off of E < E^{-20}. The regions where no genes were initially predicted also were searched by BLASTX against the NCBI nr protein database to identify any additional coding regions. The preliminary annotation of the BAC sequences indicated an abundance of retrotransposon coding regions in the UBP12 region BACs, noticeably an Ubiquitin Specific Protease12 (UBP12) gene with
multiple intragenic insertions of retrotransposon elements. Therefore this region was chosen for a detailed comparative analysis with its homoeologous region as well as the orthologous region in *Medicago truncatula*. The galactinol synthase region reported previously by Schlueter et al. (2007) was chosen to represent a retrotransposon poor (gene rich) region of the soybean genome. Identification of the galactinol synthase region homoeologous BACs, BAC sequencing and annotation region were described in Schlueter et al. (2007).

**Identification of homoeologous and orthologous regions**

The homoeologous relationship between LG I and LG O was established previously based on shared RFLP markers and genes ([www.soybase.org](http://www.soybase.org), Schlueter et al., 2007). A BLASTN search of the predicted coding sequences of the LG I UBP12 region BACs against the soybean 7X genome assembly ([www.soybase.org/gbrowse/cgi-bin/gbrowse/gmax7x](http://www.soybase.org/gbrowse/cgi-bin/gbrowse/gmax7x)) identified scaffold_20 anchored to LG O as the homoeologous region. The sequence of the homoeologous UBP12 region on LG O analyzed in this study spans scaffold_20, 9311000 bp: 9520000 bp. The 8X soybean genome assembly was released in December 2008 ([http://soybase.org/gbrowse/cgi-bin/gbrowse/gmax1.01](http://soybase.org/gbrowse/cgi-bin/gbrowse/gmax1.01)) positions the LG O UBP12 region on chromosome 10, 29944500 bp: 30152000 bp. The orthologous *Medicago* regions were identified by BLAST alignment of the predicted coding sequences against the *Medicago* genome assembly release Mt2.0. The gene annotations of the LG O and the *Medicago* UBP12 regions as well as the *Medicago* galactinol synthase region were carried out as described above. For both the UBP12 and the galactinol synthase regions, only the portions of the BAC sequences showing synteny to *Medicago* are described in this study.
Identification of conserved genomic regions

The coding sequences of the predicted genes in each region were used as query for BLASTN and TBLASTX searches against the corresponding homoeologous or orthologous regions with a cut off of \( \text{E} < 10^{-5} \) to identify the conserved genes. In addition, the alignment program AVID (Bray et al., 2003) in the mVISTA DNA comparison program was used to identify conserved coding and non-coding regions.

Estimation of the time of divergence of duplicated genes

The synonymous and nonsynonymous distances were estimated using the predicted coding sequences based on the modified Nei-Gojobori method installed in MEGA 4.0 (Tamura et al., 2007). The divergence time was calculated using the Brassicaceae synonymous substitution rate of \( 1.5 \times 10^{-8} \) synonymous substitutions per site per year (Koch et al., 2000). The divergence times of the homoeologous genes in the galactinol synthase region were estimated by Schlueter et al. (2007).

Identification and classification of LTR-retrotransposons

A combination of structural analyses and sequence homology comparisons were used to identify the retrotransposons. The intact Long Terminal Repeat (LTR) elements were identified by using LTR_STRUC, an LTR-retrotransposon mining program (McCarthy and McDonald 2003), and by homology based methods previously described (Ma et al., 2004; Ma and Bennetzen, 2004). Solo LTRs and truncated elements were identified by sequence homology searches against a soybean LTR-retrotransposon database that was developed by
collecting known LTR-retrotransposons (Jiachung Du and Jianxin Ma, unpublished) and by
cracking the soybean 7X genome assembly (www.soybase.org/gbrowse/cgi-bin/gbrowse/gmax7x) using LTR_STRUC. The structures and boundaries of all of the
identified LTR-retrotransposons were confirmed by manual inspection. The LTR-
retrotransposons were classified by sequence homology comparison, and individual families
were defined by the criteria previously described (Ma and Bennetzen 2004, Nagaki et al.

**Phylogenetic analysis of retrotransposons**

Phylogenetic analysis of the retroelements was done on the Reverse Transcriptase
(RT) domain sequences of the retroelements. A few of the putative retroelements did not
have an identifiable RT domain and were not included in the phylogenetic analysis. All the
RT sequences were aligned with the RT domain sequences of known retroelements
belonging to the copia, gypsy, and non-LTR classes using Clustalw (Chenna et al., 2003)
with default parameters. Phylogenetic analysis was done with the program MEGA 4.0
(Tamura et al., 2007) using the Neighbor- Joining method. A Bootstrap test of phylogeny
was done using 5000 replicates.

**Estimation of the time of retrotransposon insertion**

For all the intact retroelements with two LTRs, the LTR sequences of the same
element were aligned by Clustalw (Chenna et al., 2003) using default parameters. The pair
wise sequence divergence was calculated using the Kimura 2 parameter model using
MEGA4.0 (Tamura et al., 2007). The time of insertion was calculated using the equation
T = D/2t where T = time of insertion, D = divergence, t = mutation rate per nucleotide site per year (Vitte et al., 2007). The Brassicaceae synonymous substitution rate per site per year of $1.5 \times 10^{-8}$ (Koch et al., 2000) was used to estimate the time of retrotransposon insertion.

**Expression analysis of homoeologous genes with intragenic retrotransposon insertions**

Two genes in the UBP12 region had intragenic insertion of the retroelements in both the homoeologous copies. To verify the expression of these genes, soybean EST data and RT-PCR using homoeologue specific primers were used. RNA was extracted from five different soybean tissues; flowers, 1 week old pods, nodules, roots and leaves of cultivar ‘Williams 82’ using RNeasy mini kit (Qiagen). All these tissues were collected at the same growth stage (2 months after planting). RNA samples were DNase treated to eliminate genomic DNA contamination using Turbo DNA-free kit (Ambion, Inc.). Two-step RT-PCR was carried out using RETROscript kit (Ambion Inc.) using SuperTaq Plus high fidelity DNA polymerase. The PCR reactions were carried out as follows; Denaturation at $94^0$ C for 2 min followed by a three step cycle of denaturation at $94^0$ C for 30 sec, annealing temperature for 30 sec followed by extension at $72^0$ C for the duration of 1min/Kb. This cycle was repeated 35 times followed by a final extension at $72^0$ C for 5 min. A negative control without the Reverse Transcriptase and a tubulin positive control were used. RT-PCR products of expected sizes based on the predicted coding sequences were gel purified and cloned into TOPO T/A cloning vector (Invitrogen) following the addition of an A overhang and sequenced to obtain the cDNA sequence. RT-PCR products from UBP12 homoeologues with sizes different from the expected sizes were also sequenced to verify the possibility of
alternately spliced products. Sequences of the RT-PCR primers used in this study are given in Suppl. Table 3.

RESULTS

Genome organization of the regions

The soybean UBP12 homoeologous regions were located on soybean LG I (Chromosome 20) and LG O (Chromosome 10). The orthologous region in Medicago was located on the BAC mth2-157I20 (GB #AC174313 version 19) on chromosome 1. The galactinol synthase region previously reported by Schlueter et al. (2007) also was mapped to soybean LG I and LG O. The corresponding orthologous region was located on Medicago BAC mth2-150O13 (GB #AC147011 version 24) but its chromosomal location is unknown. The UBP12 region had an average gene density of 1 gene in 29 Kb while the galactinol synthase region showed 1 gene every 8 Kb, a 3 fold difference.

The content and order of genes in the soybean UBP12 regions indicated that three out of the ten genes in the soybean LG I UBP12 region retained duplicated copies in the LG O UBP12 region (Figure 1). The conserved genes showed sequence similarity both in the exon as well as the intron regions. The sequence identity of the coding regions at the nucleotide level was close to 96 percent for all three pairs of conserved genes. The sequence conservation of these genes was supported by the non-synonymous to synonymous ratio ranging from 0.09 to 0.43 indicative of purifying selection. The divergence estimates for these duplicated genes ranged 4-6 MY (data not shown). The corresponding Medicago region showed evidence of two orthologous genes only (Figure 2).
The gene conservation in the soybean galactinol synthase region is described in Schlueter et al. (2007). Synteny in the soybean galactinol synthase region is high with five out of the six genes in the LG O galactinol synthase region conserved in the corresponding LG I region (Figure 2). The *Medicago* galactinol synthase region showed considerable colinearity in gene content and order. However, a block of five genes in the LG I galactinol synthase region was absent in the LG O region and in the corresponding *Medicago* region (Figure 2). This block includes a putative gene with Pentatricopeptide (PPR) repeat domains showing tandem duplications and was described previously (Schlueter et al., 2007). One of the PPR repeat containing genes also had an intragenic retrotransposon insertion. The phosphotransferase gene showed tandem duplications as well. However, the tandem copy was truncated and shows the insertion of a PPR repeat gene into the third exon. A search of these five genes elsewhere in the genome using the whole genome assembly of soybean located all five genes on LG D1a (chromosome 1). But it should be noted that the five genes on LG D1a were not contiguous but positioned on two separate blocks of two and three genes respectively with a 5 Mb region between them (data not shown). The predicted coding regions in the intervening 5 Mb showed high levels of colinearity to LG B1 (Chromosome 11) and LG K (Chromosome 9) indicating a network of synteny and fractionation (Langham et al., 2004) following multiple rounds of large scale genome duplications (Shoemaker et al., 1996; Schlueter et al., 2004).

**Retrotransposons in the UBP12 region**

The soybean UBP12 regions on LG I and LG O had more retrotransposon elements compared to the *Medicago* UBP12 region (Table 1). The 35 elements identified in the LG I
UBP12 region accounted for 50 percent of the 247.4 Kb region (Figure 1). The proportions of retroelements in the LG O UBP12 region and the Medicago UBP12 region were 28 percent and 3 percent respectively. Among the LTR elements, truncated elements represented the largest category in all the UBP12 regions. Solo LTRs, most likely, products of intrastrand homologous recombination of retroelements, was detected in both the soybean UBP12 regions but not in the Medicago region (Figure 1). Two of the solos LTRs were internal to other retroelements suggesting nested insertion of retroelements. Non-LTR retrotransposons were detected in both the soybean UBP12 regions, but not in the Medicago region.

The soybean UBP12 regions showed evidence of both intergenic and intragenic insertions. The intergenic distance between the colinear genes UBP12 and CDPK was increased seven fold in soybean, 172-173 Kb, compared to 21.9 Kb in Medicago.

**Retrotransposons in the galactinol synthase region**

The galactinol synthase region had a gene density of one gene every 8 Kb and the accumulation of retrotransposons was far less in this region compared to the retrotransposon rich (gene poor) UBP12 region (Figure 2). The LG O galactinol synthase region appeared to be free of retrotransposon insertion while the LG I region showed evidence of seven elements including one solo LTR. Unlike the Medicago UBP12 region, the Medicago galactinol synthase region showed comparable number of retrotransposon insertions to the corresponding soybean regions (Table 1). Similar to the UBP12 regions, truncated elements represented the largest category of LTR retroelements in the galactinol synthase regions. One LTR retroelement with both LTRs intact but without the flanking target site duplication (TSD), possibly a recombined element, was detected in the Medicago region. Soybean
galactinol synthase regions did not show evidence of non-LTR retroelements, yet there was evidence of one non-LTR element insertion in the *Medicago* galactinol synthase region. Again, there was evidence of nested insertion as well as intragenic insertion of retroelement into a PPR gene (Figure 2). Genome expansion measured in terms of the intergenic distance revealed opposing trends in the galactinol synthase region. Intergenic distance between the betafructofuranosidase and the putative gene of unknown function was approximately 2.5 Kb in *Medicago*. The colinear regions expanded upwards of 5 and 18 fold in the soybean LG O and LG I galactinol synthase regions, respectively. However, the adjacent colinear regions, between the unknown gene and the galactinol synthase as well as the galactinol synthase and the AtNAP intergenic regions showed expansion in *Medicago* (Figure 2).

**Intragenic insertion of retrotransposons**

One noticeable feature of the soybean UBP12 regions was the multiple intragenic insertions of retroelements (Figure 1). The LG I UBP12 region showed intragenic insertions in two genes, one being a gene similar to the Arabidopsis UBP12 gene (At5g06600) and the other gene annotated as ‘Vascular associated death1 (VAD1)’ with a GRAM domain (At1g2120). The intragenic insertions of retroelements were all within the introns. The predicted gene structure of UBP12 indicates that this gene has 32 exons (Figure 3). There are 12 retroelement insertions within the LG I UBP12 gene and 7 of these insertions were non-LTR retroelements and the remaining 5 were LTR retroelements. The retrotransposon insertions were distributed in five introns; introns 4, 15, 21, 25, and 27 with the number of insertions of 4, 3, 2, 2, and 1 respectively. One of the insertions in intron 15 was a solo LTR (Figure 3). The LG O UBP12 gene had far fewer intragenic insertions. Among the two
retroelement insertions in the LG O UBP12 gene, one belonged to the non-LTR class and the
other to the LTR copia class. Introns 4 and 21 showed insertion of one element each. The
Medicago UBP12 gene, however, showed no retroelement insertions.

Both the UBP12 homeolologues showed insertion in introns 4 and 21. Interestingly,
the insertion of a copia-like element having the same orientation of transcription in intron 21
in both homoeologues (Figure 3) indicates a possible shared insertion prior to their
divergence (Figure 3). A lack of insertion in the corresponding position in Medicago would
mean that this insertion in soybean would have happened after the Medicago/Glycine split
50 Million Years Ago (MYA) (Schlueter et al., 2004). The truncated nature of the inserted
element in both the homoeologues suggests an older insertion, further supporting the
possibility of an insertion prior to the recent genome duplication in soybean.

Strikingly, the LGI UBP12 gene with multiple insertions is 3 times larger than the
Medicago UBP12 gene (Figure 3). The expansion of the soybean LG I UBP12 gene is largely
due to the inserted retroelements, yet intron expansion also contributed. The intragenic
insertion in the VAD1 homoeologues with 18 exons was caused by a single insertion of a
non-LTR retroelement into each of the copies (Figure 1). However, based on the predicted
gene structure, the insertion was in intron 6 for the LG I VAD1 while the insertion was in
intron 2 of the LG O VAD1 (data not shown).

**Phylogenetic analysis of retroelements**

The number of retroelements with RT domain included in the phylogenetic analysis is
shown in Table 1. The known retroelements belonging to the copia, gypsy, and non-LTR
classes that were used for the phylogenetic analysis are given in Table 2. Seventeen out of
the 35 retroelements in the LG I UBP12 region had an identifiable RT domain and seven out of the 18 in the LG O UBP12 region had an identifiable RT domain. The retroelements in the *Medicago* UBP12 region did not have an RT domain but the similarity of the elements to the soybean SIRE-1 retroelement (AY205610) indicate that the truncated copy in *Medicago* UBP12 region could be a copia type LTR retroelement. A four-fold increase in the number of non-LTR elements was observed in the LG I UBP12 region compared to LG O UBP12 region; eight non-LTR elements in the LG I UBP12 region compared to two elements in the LG O region (Figure 4, Suppl. Table 2).

Only the LG I galactinol synthase region showed evidence of retrotransposon insertions and of the seven elements in this region, only two had an intact RT domain. One of these two elements was classified as copia while the other one was of the gypsy class. Among the six retroelements in the *Medicago* galactinol synthase region, only one had a RT domain and it belonged to the gypsy class. The *Medicago* region also had one non-LTR retroelement but the soybean galactinol synthase regions showed no evidence of non-LTR retroelement insertions.

The only SIRE-1 copia like element (LGO_U14) detected in the soybean regions under study is closely related to the previously reported SIRE-1 (Laten et al., 2003), confirming a highly homogenous and a recent wave of amplification of SIRE-1 elements in soybean (Laten et al., 2003, Figure 4). Further, soybean gypsy elements were more similar to the *Vicia* or pea gypsy elements than to *Medicago* gypsy elements. This is in agreement with the findings that no major soybean repeat families were shared between soybean and *Medicago* except for the rDNA (Macas et al., 2007). Soybean non-LTR elements appeared to be in two clusters with all the non-LTR elements in the UBP12 genes except one formed one
cluster. However, both the soybean non-LTR clusters were distinct from the non-LTR elements from other species included in the analysis.

**Insertion time of retrotransposons**

The insertion times of the retroelements in the soybean LG I UBP12 region ranges from 0.26 to 11.63 Million Years (MY) (Table 3). The estimated insertion times of the intact retroelements in the LG O UBP12 region, however, indicated more recent insertions; insertion of a SIRE-1 element close to 67,000 years ago to insertions as old as 4.56 MY. Insertion time could not be estimated for the *Medicago* UBP12 region elements and for either soybean galactinol synthase regions due to the lack of intact elements. For the single intact element in the *Medicago* galactinol synthase region, insertion time was estimated to be 0.13 MY. Insertion times of the majority of intact elements (77%) was within the last 2 MY, yet an insertion as old as 11.63 MY was detected in the LG I UBP12 region. This indicates that some of the insertions would have happened around the time of the recent genome duplication of soybean ~14 MYA (Schlueter et al., 2004) and that both regions act as hotspots for continued retrotransposon accumulation.

**Expression analysis of the homoeologous genes with intragenic retrotransposon insertions**

Intragenic insertion of retrotransposons into the introns of genes could alter the gene expression (Marillonnet and Wessler, 1997). Therefore, the expression pattern of the UBP12 and the VAD1 homoeologues showing intronic insertions of retrotransposons was studied. Initially the soybean gene index (release 13, July 2008) at Dana-Farber Cancer Institute
(DFCI) was searched using the putative coding sequences of these homoeologues to see if any EST sequences aligned to them with high identity. To our surprise, the LG I UBP12 with multiple intragenic insertions identified TC258329, BG790879, and TC247081 that aligned along their entire sequence with 100 %, 99 % and 97 % sequence identity, respectively. These ESTs showed less sequence identity to the LG O UBP12 coding sequence with the % sequence identities of 97, 95 and 92 respectively, suggesting that the above EST sequences are likely to be derived from the LG I UBP12 homoeologue. On the other hand, EST data supported the expression of both the homoeologues of the VAD1. ESTs CA820220 and BI497564 aligned with the LG I VAD1 with sequence identities of 100 and 98 %, respectively. The LG O VAD1 identified TC248308 with 99% identity.

RT-PCR confirmed the expression of both VAD1 homoeologues in all the tissues studied (Figure 5). For the LG I UBP12 homoeologue, RT-PCR (Figure 5) followed by cDNA sequencing confirmed the expression of a full- length transcript despite multiple intragenic insertions. For the LG I UBP12 homoeologue, three of the four RT-PCR primer pairs used for cDNA sequencing, amplified products with sizes different than expected. These RT-PCR products also were sequenced except for I2F/I2R to verify the possibility of alternate splicing. But none of these products originated from UBP12 homoeologues and were non-specific PCR products. RT-PCR results suggest not only the expression, but a possible alternate splicing also of the LG O UBP12 homoeologue (Data not shown). Partial cDNA sequencing confirmed the alternately spliced product amplified by the primer pair O2F/O2R (Figure 3) of the LG O UBP12 gene where exons 5, 6, and 7 were spliced out.
BACs reported in this study have Genbank accession numbers FJ571602 and FJ571603. cDNA sequences of the UBP12 genes have accession numbers FJ571604, FJ571605, FJ571606.

**DISCUSSION**

**Retrotransposon insertions in homoeologous regions**

The soybean genome contains up to 40-60 % repetitive DNA (Goldberg, 1978, Gurley et al., 1979). Fluorescent *in situ* Hybridization (FISH) studies indicate that soybean repeats are largely confined to the centromeric and pericentromeric regions (Lin et al., 2005). In a survey of the repetitive fraction of the soybean genome, Nunberg et al. (2006) masked the methylation-unfiltered soybean sequences with plant repeats in the TIGR Plant Repeat Database and found that, not considering the unique repeats in soybean, retrotransposons were the most abundant class of repeats.

The four soybean genomic regions studied in this paper, the UBP12 regions and the galactinol synthase regions, represent retrotransposon rich (gene poor) and retrotransposon poor (gene dense) regions respectively, on the same homoeologous linkage groups LG I (Chromosome 20) and LG O (chromosome 10). The analysis of these regions allowed us to determine whether the homoeologous regions differ with respect to the abundance, types, and organization of retrotransposons. The retrotransposon rich UBP12 region showed both old and recent insertions, whereas the galactinol synthase region showed far fewer insertions, mostly retrotransposon remnants. Even though both the LG I and the LG O UBP12 regions acted as hot spots for retrotransposon accumulation, the numbers of LTR retroelements belonging to copia or gypsy classes were comparable in both the UBP12 regions. However,
there was a preferential accumulation of non-LTR retrotransposons in the LG I UBP12 region.

The types of elements inserted also were compared between the soybean homoeologous regions. Of the 15 different families of LTR retroelements identified in the UBP12 regions, 13 showed evidence of insertion in only one of the two homoeologous UBP12 regions (Suppl. Table 2). A few of the retroelements in the LG I and LG O UBP12 regions could not be classified. Also there were two sequence gaps in the LG O region and therefore some of the retroelements could be missing from the analysis. Yet, 6 out of the 13 families mentioned above, showed insertion into the LG O UBP12 region but did not show evidence of insertion into the LG I UBP12 region.

Further, the estimated timing of insertions indicted that LG O UBP12 region had more recent insertions with the timing of insertions ranging from 0.06 MY to 4.56 MY. The LG I region had much older insertions with the range of 0.26 to 11.63 MY. For instance, the intact Gmr6 copies (Suppl. Table 2) present in both the LG I and LG O UBP12 regions appeared to be at syntenic positions. But the insertion times of the elements indicate that the LG I copy was inserted 0.26 MYA while the LG O copy was inserted 0.93 MYA. It is likely that the two copies were inserted independently at different times and happened to be at the similar position, or alternatively, inserted independently, at the same time, but diverged at different rates. Collectively, these results suggest the independent nature of the retrotransposon evolution in the UBP12 homoeologous regions. The galactinol synthase region also presented evidence of independent evolution of homoeologous regions in that only the LG I galactinol synthase region showed evidence of insertions while the LG O region showed no evidence of retroelements (Figure 2). Independent insertions of
retrotransposons were observed in homoeologous Maize Orp regions where retrotransposons in the intergenic regions of the conserved genes did not share lengths of sequence homology (Ma et al., 2005).

**The varying trends in genome expansion**

Retrotransposon amplification is considered one of the major forces of genome growth (Vitte and Bennetzen, 2006; Neumann et al., 2006; Piegu et al., 2006). The Medicago genome is ~ 500 Mbp while soybean genome is ~1110 Mbp, a two-fold difference in size. Retrotransposon accumulation in the soybean UBP12 region resulted in upwards of seven-fold increase in intergenic space between syntenic genes UBP12 and CDPK in soybean compared to Medicago; an evidence of retrotransposon mediated genome expansion in soybean (Figure 1). In fact, the expansion of the UBP12 region in soybean is not only due to expansion of the intergenic region but the expansion of the UBP12 gene itself (Figure 3). The galactinol synthase region on the other hand showed evidence of genome expansion in soybean as well as in Medicago (Figure 2). The local genome expansion in Medicago does not reflect the overall genome size difference of soybean and Medicago. Similar incongruent pattern in local and global genome size evolution was observed in the CesA region in cotton A and D genomes with two fold change in genome size (Grover et al., 2004).

In light of the recent advancements in the understanding of genome size evolution, it is now presumed that genome size is a function of both genome expansion and contraction forces (Devos et al., 2002; Bennetzen et al., 2005). Solo LTRs or intact elements without a TSD are evidence of retrotransposon removal by intra-strand recombination between LTRs of retroelements (Vitte and Panaud, 2003; Devos et al., 2002). Both soybean UBP12 regions
showed evidence of solo LTRs, indicating the removal of retroelement DNA. Further, the majority of the LTR retroelements were truncated in the UBP12 and the galactinol synthase regions. Our analysis shows locally certain genomic regions could be expanded in soybean while in other regions the Medicago genome shows expansion. The reasons for these differential patterns could be that the retrotransposon rich (gene poor) UBP12 region in soybean could tolerate accumulation of retrotransposon DNA due to less selection pressure, exceeding the capability of the removal mechanisms. The retrotransposon poor (gene rich) soybean galactinol synthase regions were selected against the accumulation of retrotransposons together with efficient removal. The efficiency of these mechanisms could vary between genomic regions (Grover et al., 2007) and with the species (Vitte and Bennetzen, 2006).

**Intragenic insertion of retroelements**

Transposable element insertions into the coding or promoter regions of genes modulate gene function by regulation of gene expression or by formation of non-functional proteins (Hori et al., 2007; Xiao et al., 2008). Insertion of transposable elements into the introns could be less deleterious yet not inconsequential. Intronic insertions could cause splicing alterations and differential transcript accumulations (Marillonnnet and Wessler, 1997; O’Connor et al., 1999; Tighe et al., 2002). RT-PCR and cDNA sequencing confirmed the expression of both UBP12 homoeologues as well as possible alternate splicing of the LG O UBP12 gene. Even though it is possible that the differential splicing could be the result of retrotransposons inserted into the introns of this gene, due to the lack of a soybean UBP12
gene without insertions for comparison, it cannot be attributed conclusively to the intronic insertions.

The LG I UBP12 gene showed 12 insertions distributed in 4 introns. The bulk of the intronic insertions in the LG I UBP12 were non-LTR retroelements. Interestingly, all the non-LTR retroelements identified in the UBP12 region were intragenic insertions (Figure 1). The distribution or the role of non-LTR elements in the soybean genome evolution is currently unknown. Their intragenic insertion could be suggestive of their affinity to insert within genes. Non-LTR elements are present at low copy number in the genomes of other plant species (Schimdt, 1999; Zhang and Wessler, 2004; Alix et al., 2005) with the Del2 LINE family in Lily the only known exception (Leeton and Smith, 1993). Non-LTR retrotransposons in a legume species, Vicia, indicated that the LINEs contribute the least to the Vicia genome (Hill et al., 2005). Fewer number of Non-LTR retroelements compared to the LTR elements (Table 1) in all the soybean genomic regions studied, suggest a similar trend in soybean.

The effect of intronic insertions of non-LTR elements is better understood in mammalian genomes, in which non-LTR retrotransposons particularly L1 LINEs are a major component of the genome. In the mouse genome, intronic L1 insertion in the antisense orientation is less deleterious to gene expression compared to sense L1 insertions and sense insertions could be subject to purifying negative selection (Chen et al., 2006). An analysis of the non-LTR insertions in the UBP12 genes (Suppl. Table 2) showed that four out of the seven non-LTR insertions in the LG I UBP12 were in the antisense and three in the sense orientation while the LG O UBP12 had only a single non-LTR insertion in the antisense orientation (Figure 3). But it should be pointed out that there were LTR retrotransposon
insertions in the UBP12 homoeologues. The single non-LTR insertion in the VAD1 was in the antisense orientation in both the homoeologues. Another observation of the L1 insertions in the introns of human genes is that the poly (A) signal of the insertions is weak and often uses downstream poly (A) signals (Moran et al., 1996). The adoption of such a mechanism helps the L1 insertions to be almost invisible. A survey of additional soybean genomic regions would verify if non-LTR elements have target site preference for insertion within genes and their role in the divergence of homoeologous genes.

Segmental or large-scale genome duplications are followed by the functional divergence of duplicate genes. Functional divergence can manifest in functional redundancy, subfunctionalization, neofunctionalization or pseudogene formation (Lynch and Force, 2000). Even though a full-length transcript was confirmed only for the LG I UBP12 gene, RT-PCR suggests expression of both the homoeologues of the UBP12 and VAD1. The UBP gene family is involved in protein deubiquitination and many subfamilies have been identified in Arabidopsis (Liu et al., 2008). VAD1 is known to be involved in pathogen response and function in the cell death control of cells in the vicinity of vascular bundles in plants (Lorrain et al., 2004). One possible explanation for the expression of both homoeologues of UBP12 could be for substrate specificity for deubiquitination and for VAD1, pathogen specificity.

The soybean UBP12 regions act as a hot spot for retrotransposon accumulation and thereby exhibit genome expansion compared to Medicago. The soybean homoeologous regions evolve independently in terms of retrotransposon accumulation in the intergenic regions and thus retrotransposons could be agents of evolutionary divergence of homoeologous regions. Further analysis at a whole genome level would help us understand the contribution of retroelements in shaping the paleopolyploid genome of soybean.
ACKNOWLEDGEMENTS

The authors are grateful to the North Central Soybean Research Program (WBS # 26-6222-0183-005) and National Science Foundation (Grant # DBI-0501877) for financial support for this project.

REFERENCES


Figure 1. The structure of the UBPI2 region.
The retrotransposons belonging to copia, gypsy, non-LTR and unknown classes are color-coded. The blue arrows represent predicted genes and the direction of the blue arrows show the direction of transcription. Gray blocks connect conserved genes. Green arrow heads are solo LTRs. The vertical arrowheads within the genes indicate the intragenic insertion of retroelements. Retroelements with a star above are intact elements. The sequence gaps in the LG O UBPI2 region shown as yellow cross marks. Additional details of the retrotransposon annotations are provided in the Suppl. Table 1. Mt-Medicago truncatula.
Figure 2. The structure of the galactinol synthase region.
The retrotransposons and genes are represented as in Figure 1. Gene annotation details for the LG I and LG O galactinol synthase regions are described in Schlueter et al. (2007). Additional details of the retrotransposon annotations are provided in the Suppl.Table 1.

Mt-Medicago truncatula.
Figure 3. The UBP12 genes showing gene expansion.

1. LG I UBP12 gene region with the intron/exon positions. The vertical arrowheads indicate the retroelements and are color coded as in Figure 1. The arrows above the retroelements indicate their direction of transcription. The size of the gene region or the predicted ORF in bp is given on the right side of each structure. The positions of the RT-PCR primers used for cDNA sequencing were indicated by reverse and forward arrows.

2. LG O UBP12 gene structure and likely alternatively spliced transcript based on the partial cDNA sequencing.

3. Mt UBP12 predicted gene structure. ORF, Open Reading Frame.
Figure 4. Phylogenetic analysis of retroelements in the UBP12 and the galactinol synthase regions.

The retroelements in the LG I, LG O or the *Medicago* UBP12 regions are designated as LGI U, LGO U or Mt U respectively, followed by a number indicating their position from left to right as shown Figure1. Similarly, the elements in the galactinol synthase region are designated as LGI G, LGO G or Mt G respectively.
Figure 5. RT-PCR of the homoeologous genes in the UBP12 region with intragenic retrotransposon insertions.

For LG I UBP12, RT-PCR products of I1F/I1R are shown. For LG O UBP12 RT-PCR products of O1F/O1R are shown.
Table 1. Classification of retrotransposons in the UBP12 and the galactinol synthase regions.

<table>
<thead>
<tr>
<th>Type</th>
<th>UBP12 region</th>
<th>Galactinolsynthase region</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>LGI</td>
<td>LGO</td>
</tr>
<tr>
<td>LTR retrotransposons</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intact elements</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Truncated elements</td>
<td>19</td>
<td>8</td>
</tr>
<tr>
<td>Non-LTR retrotransposons</td>
<td>8</td>
<td>2</td>
</tr>
<tr>
<td>Solo LTRs</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>Elements with RT domain</td>
<td>17</td>
<td>7</td>
</tr>
</tbody>
</table>
Table 2. List of known retroelements used for phylogenetic analysis.

<table>
<thead>
<tr>
<th>Genbank Accession No.</th>
<th>Class</th>
<th>Species</th>
</tr>
</thead>
<tbody>
<tr>
<td>SIRE-1</td>
<td>Copia</td>
<td><em>Glycine max</em></td>
</tr>
<tr>
<td>Tgmr</td>
<td>Copia</td>
<td><em>Glycine max</em></td>
</tr>
<tr>
<td>Tnt1</td>
<td>Copia</td>
<td><em>Nicotiana tabaccum</em></td>
</tr>
<tr>
<td>Art1</td>
<td>Copia</td>
<td><em>Arabidopsis thaliana</em></td>
</tr>
<tr>
<td>Opie-2</td>
<td>Copia</td>
<td><em>Zea Mays</em></td>
</tr>
<tr>
<td>Diaspora</td>
<td>Gypsy</td>
<td><em>Glycine max</em></td>
</tr>
<tr>
<td>TvfG20</td>
<td>Gypsy</td>
<td><em>Glycine max</em></td>
</tr>
<tr>
<td>Cyclops-2</td>
<td>Gypsy</td>
<td><em>Pisum sativum</em></td>
</tr>
<tr>
<td>Athila 4-5</td>
<td>Gypsy</td>
<td><em>Arabidopsis thaliana</em></td>
</tr>
<tr>
<td>TvfL10</td>
<td>Gypsy</td>
<td><em>Arabidopsis thaliana</em></td>
</tr>
<tr>
<td>Ta11</td>
<td>Non-LTR</td>
<td><em>Vicia faba</em></td>
</tr>
<tr>
<td>Cin4</td>
<td>Non-LTR</td>
<td><em>Zea mays</em></td>
</tr>
<tr>
<td>TA23</td>
<td>Non-LTR</td>
<td><em>Arabidopsis thaliana</em></td>
</tr>
</tbody>
</table>
Table 3. Insertion time of the retroelements.

<table>
<thead>
<tr>
<th>Name</th>
<th>Insertion time (MY)</th>
</tr>
</thead>
<tbody>
<tr>
<td>LGI_U13</td>
<td>11.63</td>
</tr>
<tr>
<td>LGI_U21</td>
<td>2.1</td>
</tr>
<tr>
<td>LGI_U24</td>
<td>0.8</td>
</tr>
<tr>
<td>LGI_U25</td>
<td>0.26</td>
</tr>
<tr>
<td>LGO_U6</td>
<td>1.16</td>
</tr>
<tr>
<td>LGO_U7</td>
<td>4.56</td>
</tr>
<tr>
<td>LGO_U12</td>
<td>0.93</td>
</tr>
<tr>
<td>LGO_U14</td>
<td>0.06</td>
</tr>
<tr>
<td>Mt_G3</td>
<td>0.13</td>
</tr>
</tbody>
</table>

*The nomenclatures of the retroelements are as described for Figure 4.*
### Supplementary Table 1. Details of retrotransposon annotations.

<table>
<thead>
<tr>
<th>Retrotransposon</th>
<th>name</th>
<th>orientation</th>
<th>Annotation</th>
<th>subclass</th>
<th>status</th>
<th>RT domain</th>
<th>Direct repeats</th>
<th>TSRs</th>
</tr>
</thead>
<tbody>
<tr>
<td>LGI_UBP12_retroelements</td>
<td>LGI_U1</td>
<td>+</td>
<td>non-LTR</td>
<td>unknown</td>
<td>yes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>LGI_U2</td>
<td>-</td>
<td>Gmr361</td>
<td>copia</td>
<td>internal remnant</td>
<td>yes</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>LGI_U3</td>
<td>+</td>
<td>non-LTR</td>
<td>unknown</td>
<td>yes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>LGI_U4</td>
<td>+</td>
<td>copia</td>
<td>truncated</td>
<td>yes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>LGI_U5</td>
<td>+</td>
<td>non-LTR</td>
<td>unknown</td>
<td>yes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>LGI_U6</td>
<td>+</td>
<td>Gmr4</td>
<td>gypsy</td>
<td>solo</td>
<td>no</td>
<td>TG/CA</td>
<td>ATTGG/ATTGG</td>
</tr>
<tr>
<td></td>
<td>LGI_U7</td>
<td>+</td>
<td>non-LTR</td>
<td>unknown</td>
<td>yes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>LGI_U8</td>
<td>-</td>
<td>non-LTR</td>
<td>unknown</td>
<td>yes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>LGI_U9</td>
<td>+</td>
<td>Gmr252</td>
<td>copia</td>
<td>truncated</td>
<td>yes</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>LGI_U10</td>
<td>-</td>
<td>non-LTR</td>
<td>unknown</td>
<td>yes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>LGI_U11</td>
<td>-</td>
<td>non-LTR</td>
<td>unknown</td>
<td>yes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>LGI_U12</td>
<td>+</td>
<td>copia</td>
<td>truncated</td>
<td>yes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>LGI_U13</td>
<td>-</td>
<td>Gmr22</td>
<td>copia</td>
<td>intact</td>
<td>yes</td>
<td>TG/CA</td>
<td>CATAT/CATAT</td>
</tr>
<tr>
<td></td>
<td>LGI_U14</td>
<td>+</td>
<td>LTR</td>
<td>truncated</td>
<td>no</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>LGI_U15</td>
<td>-</td>
<td>Gmr4</td>
<td>gypsy</td>
<td>solo</td>
<td>no</td>
<td>TG/CA</td>
<td>GTTGG/GTTGG</td>
</tr>
<tr>
<td></td>
<td>LGI_U16</td>
<td>-</td>
<td>Gmr25</td>
<td>gypsy</td>
<td>solo</td>
<td>no</td>
<td>TG/CA</td>
<td>CATAG/CATAG</td>
</tr>
<tr>
<td></td>
<td>LGI_U17</td>
<td>-</td>
<td>Gmr4</td>
<td>gypsy</td>
<td>solo</td>
<td>no</td>
<td>TG/CA</td>
<td>ATAGG/ATAAG</td>
</tr>
<tr>
<td></td>
<td>LGI_U18</td>
<td>-</td>
<td>Gmr4</td>
<td>gypsy</td>
<td>truncated</td>
<td>no</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>LGI_U19</td>
<td>+</td>
<td>Gmr6</td>
<td>copia</td>
<td>truncated</td>
<td>no</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>LGI_U20</td>
<td>+</td>
<td>Gmr9</td>
<td>gypsy</td>
<td>truncated</td>
<td>no</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>LGI_U21</td>
<td>+</td>
<td>Gmr9</td>
<td>gypsy</td>
<td>intact</td>
<td>no</td>
<td>TG/CA</td>
<td>CACAC/CACAC</td>
</tr>
<tr>
<td></td>
<td>LGI_U22</td>
<td>-</td>
<td>Gmr6</td>
<td>copia</td>
<td>solo</td>
<td>no</td>
<td>TG/CA</td>
<td>TAGGA/TAGGA</td>
</tr>
<tr>
<td></td>
<td>LGI_U23</td>
<td>-</td>
<td>Gmr9</td>
<td>gypsy</td>
<td>truncated</td>
<td>no</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>LGI_U24</td>
<td>-</td>
<td>Gmr3</td>
<td>gypsy</td>
<td>intact</td>
<td>no</td>
<td>TG/CA</td>
<td>AATCC/CATCC</td>
</tr>
<tr>
<td></td>
<td>LGI_U25</td>
<td>-</td>
<td>Gmr6</td>
<td>copia</td>
<td>intact</td>
<td>yes</td>
<td>TG/CA</td>
<td>TATTC/TATTC</td>
</tr>
<tr>
<td></td>
<td>LGI_U26</td>
<td>-</td>
<td>Gmr9</td>
<td>gypsy</td>
<td>truncated</td>
<td>no</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>LGI_U27</td>
<td>+</td>
<td>Gmr9</td>
<td>gypsy</td>
<td>truncated</td>
<td>yes</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>LGI_U28</td>
<td>+</td>
<td>Gmr9</td>
<td>gypsy</td>
<td>truncated</td>
<td>no</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

TSR-Target Site Repeat.
Supplementary Table 1. Continued.

| LGO_U29 | + | Gmr6 | copia truncated | yes |
| LGO_U30 | + | Gmr25 | gypsy truncated | yes |
| LGO_U31 | + | Gmr6 | copia truncated | no |
| LGO_U32 | - | Gmr4 | gypsy truncated | no |
| LGO_U33 | - | non-LTR | unknown | yes |
| LGO_U34 | - | Gmr364 | gypsy truncated | no |
| LGO_U35 | + | Gmr364 | gypsy ltr remnant | no |

**LGO_O_UBP12_retroelements**

<p>| LGO_U1 | _ | copia truncated | yes |
| LGO_U2 | _ | non-LTR unknown | yes |
| LGO_U3 | + | Gmr18 | copia solo | no |
| LGO_U4 | + | Gmr37 | copia solo | no |
| LGO_U5 | + | Gmr34 | gypsy truncated | no |
| LGO_U6 | _ | Gmr5 | copia intact | no |
| LGO_U7 | _ | Gmr3 | gypsy intact | no |
| LGO_U8 | _ | gypsy truncated | yes |
| LGO_U9 | + | LTR | truncated | no |
| LGO_U10 | _ | LTR | truncated | no |
| LGO_U11 | _ | LTR | truncated | no |
| LGO_U12 | _ | Gmr6 | copia intact | yes |
| LGO_U13 | + | LTR | truncated | no |
| LGO_U14 | _ | Gmr2/SIRE-18 | copia intact | yes |
| LGO_U15 | + | gypsy | truncated | yes |
| LGO_U16 | + | Gmr5 | copia solo | no |
| LGO_U17 | Gmr79 | LTR solo | no |
| LGO_U18 | + | non-LTR | unknown | yes |</p>
<table>
<thead>
<tr>
<th>mitochondrial retrotransposon</th>
<th>LTR status</th>
<th>intact LTR</th>
<th>TSD</th>
<th>sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>LGI_G7 + Gmr75</td>
<td>+</td>
<td>gypsy</td>
<td>ltr remnant</td>
<td>no</td>
</tr>
<tr>
<td>Mt_galactinolsynthase_retroelements</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mt_G1 +</td>
<td>LTR</td>
<td>intact with no TSD</td>
<td>no</td>
<td></td>
</tr>
<tr>
<td>Mt_G2 -</td>
<td>non-LTR</td>
<td>unknown</td>
<td>no</td>
<td></td>
</tr>
<tr>
<td>Mt_G3 -</td>
<td>gypsy</td>
<td>intact</td>
<td>yes</td>
<td>TG/CA</td>
</tr>
<tr>
<td>Mt_G4 -</td>
<td>LTR</td>
<td>truncated</td>
<td>no</td>
<td></td>
</tr>
<tr>
<td>Mt_G5 -</td>
<td>LTR</td>
<td>truncated</td>
<td>no</td>
<td></td>
</tr>
<tr>
<td>Mt_G6 +</td>
<td>LTR</td>
<td>truncated</td>
<td>no</td>
<td></td>
</tr>
</tbody>
</table>
Supplementary Table 2. Sequences of RT-PCR primers used for cDNA sequencing of UBP12 and expression analysis of VAD1.

<table>
<thead>
<tr>
<th>RT-PCR primer name</th>
<th>Forward primer</th>
<th>Reverse primer</th>
</tr>
</thead>
<tbody>
<tr>
<td>11</td>
<td>CACGGTTCTTCACATTCCTGAC</td>
<td>ATCCACTGTACCATGTTCTGCT</td>
</tr>
<tr>
<td>12</td>
<td>CAGACCAAGCAGAAACATG</td>
<td>ACTTTCTTCACCATCTAGTTC</td>
</tr>
<tr>
<td>13</td>
<td>ATAATGGCAATATCTATCTCCTGAA</td>
<td>TGAGTCTAATTTTGAAGATCACT</td>
</tr>
<tr>
<td>14</td>
<td>TTGACAAGAAATCCACATTCTT</td>
<td>TGGTTTCAACAGACTACCTTCTT</td>
</tr>
<tr>
<td>O1</td>
<td>CGGTTTTTGTCTTCTTACATT</td>
<td>ACTTATAGCCCCCTACAGAAAAT</td>
</tr>
<tr>
<td>O2</td>
<td>AGCCCAAGCAGAAATCCCA</td>
<td>GCTTTGACAGCTCAAGACTAAATC</td>
</tr>
<tr>
<td>O3</td>
<td>CGGTTTACAGAAATCCACATTCCG</td>
<td>GTTTTTCTTTTCTCAGGAATCTC</td>
</tr>
<tr>
<td>LGI VAD1</td>
<td>GGCGCTTCCATCCATTTGTA</td>
<td>GCCACGAAGAACCTAAAGTTCT</td>
</tr>
<tr>
<td>LGO VAD1</td>
<td>GACGCTTACATCCATTTGG</td>
<td>ATCAATATCCTCCTGTTCCACT</td>
</tr>
</tbody>
</table>
CHAPTER FIVE

GENERAL CONCLUSIONS

CONCLUSIONS

Seed protein and oil levels are important determinants of the economic value of the soybean crop (Smith, 1987). The genetic basis of these traits are complex and many QTL associated with seed protein/oil have been reported (Csanádi et al., 2001; Fasoula et al., 2004; Lee et al., 1996; Mansur et al., 1993; Panthee et al., 2005). However, the consistent detection in many populations (Csanádi et al., 2001; Diers et al., 1992; Seboldt et al., 2000; Chung et al., 2003) and across environments (Brummer et al., 1997) makes the QTL on chromosome 20 (LG I) a significant QTL for seed protein/oil and therefore is worthy of further investigation. This dissertation describes an attempt to identify the candidate genes underlying this major QTL based on their map positions. Using back cross populations, where a high protein QTL segment from *Glycine soja* was introgressed into a low protein *Glycine max* background, the QTL was defined to an approximately 8.4 Mbp pericentromeric region on chromosome 20 (LG I) flanked by SSR markers Sat_174 and ssrpqtl_38. The region was saturated with 34 additional SSR markers. These markers will allow detection of recombinations within the QTL interval to further narrow down the QTL region for the purpose of map based cloning. In addition, transcriptome analysis of developing soybean seeds from near-isogenic lines contrasting in protein and oil, identified thirteen candidate genes located within the QTL region exhibiting differential expression in the NILs.
These candidates will be evaluated by further experimentation to confirm the candidacy. However, the genes located within the QTL region that were not differentially expressed cannot be ruled out as potential candidates for the QTL due to possible regulation at the post-transcriptional level. Analysis of additional genes of interest will be undertaken to explore this possibility.

Soybean is an ancient polyploid (Shoemaker et al., 1996). In this study, the homoeologous region corresponding to the QTL region was identified. The comparative analysis of the two regions provided insights into the evolutionary divergence of these two regions. The synonymous substitution rates of the homoeologous genes indicate these regions are likely to be products of the recent genome duplication, 10-14 MYA (Schlueter et al., 2004). The pericentromeric location of the homoeologous regions allowed us to determine the divergence of homeologous regions in gene poor regions while previous studies of duplicate gene evolution were focused primarily to gene rich regions (Schlueter et al., 2006; Schlueter et al., 2007). The extent of synteny was poor in the gene poor regions while synteny was high in the adjacent gene rich regions. The conserved genes were enriched for functional categories of stress response, protein metabolism, and transcription factor related genes and appeared to be under purifying selection. Comparing these regions to other legumes sharing a common ancestor with soybean would provide more insights into the polarity and timing of the evolutionary changes in these regions.

Retrotransposons constitute a major proportion of the DNA within the QTL region. Analysis of BAC sized genomic regions around the QTL region indicated preferential accumulation of retrotransposons in chromosome 20 compared to its homoeolog, chromosome 10. The majority of the retrotransposon accumulations happened in the past 2
million years. The retrotransposon insertions contributed to genome expansion in soybean in comparison to the corresponding regions in *Medicago*. Insertions of retroransposons and the selection against the insertions could be the predominant evolutionary forces acting on these regions.

REFERENCES


