2006

The paleopolyploid nature of the soybean genome: duplicate gene identification, regional sequence characterization and expression studies

Jessica Ann Schlueuter
Iowa State University

Follow this and additional works at: https://lib.dr.iastate.edu/rtd
Part of the Bioinformatics Commons, and the Genetics Commons

Recommended Citation
https://lib.dr.iastate.edu/rtd/3020

This Dissertation is brought to you for free and open access by the Iowa State University Capstones, Theses and Dissertations at Iowa State University Digital Repository. It has been accepted for inclusion in Retrospective Theses and Dissertations by an authorized administrator of Iowa State University Digital Repository. For more information, please contact digirep@iastate.edu.
The paleopolyploid nature of the soybean genome: duplicate gene identification, regional sequence characterization and expression studies

by

Jessica Ann Schlueter

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

DOCTOR OF PHILOSOPHY

Major: Genetics

Program of Study Committee:
Randy C. Shoemaker, Co-major Professor
Volker Brendel, Co-major Professor
Jonathan Wendel
Patrick Schnable
E. Charles Brummer
Phillip Dixon

Iowa State University
Ames, Iowa
2006
INFORMATION TO USERS

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleed-through, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

UMI Microform 3229122
Copyright 2006 by ProQuest Information and Learning Company.
All rights reserved. This microform edition is protected against unauthorized copying under Title 17, United States Code.

ProQuest Information and Learning Company
300 North Zeeb Road
P.O. Box 1346
Ann Arbor, MI 48106-1346
This is to certify that the doctoral dissertation of

Jessica Ann Schlueter

has met the dissertation requirements of Iowa State University

Signature was redacted for privacy.

Co-major Professor

Signature was redacted for privacy.

Co-major Professor

Signature was redacted for privacy.

For the Major Program
Dedicated to:

My husband Shannon who has always encouraged me to do the best I could with honesty and integrity and to my son Harmon who is the reason I smile every day.
TABLE OF CONTENTS

CHAPTER 1. GENERAL INTRODUCTION
   Introduction 1
   Dissertation Organization 3
   Literature Review 5
   References 24

CHAPTER 2. MINING EST DATABASES TO RESOLVE EVOLUTIONARY EVENTS IN MAJOR CROP SPECIES
   Abstract 38
   Introduction 39
   Materials and Methods 40
   Results and Discussion 43
   Acknowledgments 51
   References 51
   Appendices: Additional Figures 63

CHAPTER 3. SEQUENCE CONSERVATION OF HOMELOGOUS BACS AND EXPRESSION OF HOMELOGOUS GENES IN SOYBEAN (*GLYCINE MAX* L. MERR.)
   Abstract 67
   Introduction 68
   Materials and Methods 71
   Results 76
   Discussion 86
   Acknowledgments 93
   References 93
   Appendix 119
CHAPTER 1. GENERAL INTRODUCTION

Introduction

Polyploidy is widespread in the plant kingdom, with estimates of 70 to 100% of angiosperms and upwards of 95% of pteridophytes having a polyploid history (Masterson, 1994; Lockton and Gaut, 2005). Most of the major crop species belong to this group (wheat, corn, potato, soybean, peanut, sugar cane, alfalfa, coffee, cotton, etc.; Wendel, 2000) as well as model species, Arabidopsis, rice, Medicago and Lotus (TAGI 2000; Vision et al., 2000; Lynch and Conery, 2000; Simillion et al., 2002; Blanc et al., 2003; Yu et al., 2003; Goff et al., 2002; Young et al., 2005; Shoemaker et al., 2006). While some of these species are more recent polyploids or neopolyploids, most are ancient polyploids or paleopolyploids with their genomes having undergone cyclic rounds of duplication and diploidization (Schlueter et al., 2004; Blanc and Wolfe, 2004). Most plant genomes have likely experienced polyploidy followed by major genome rearrangements that allow the genome to revert to a near diploid state (Stebbins, 1966). The general belief in the existence of a simple diploid plant species is likely a false one, with most all plant genomes having undergone extensive duplication and reshuffling.

Genome duplication has been implicated as a driving force in the evolution and expansion of eukaryotic genomes (Stebbins, 1950; Ohno, 1970, Lewis, 1980; Petrov, 2001). Most plant genes are members of gene families, indicating that gene duplication is widespread in the origin and formation of diverse gene functions (Wendel 2000; Adams and Wendel, 2005). Gene and genome duplication is also seen as a mechanism for the creation of genetic diversity and as a source of new genes and gene functions as well as leading to silenced genes or pseudogenes (Ohno, 1970; Pickett and Meeks-Wagner, 1995; Long et al., 2003).

Studies in soybean (Glycine max L. Merr.) have suggested that the genome has experienced cyclic rounds of this process of genome duplication and diploidization (Shoemaker et al., 1996; Lee et al., 1999; Lee et al., 2001). Cytogenetic studies have shown
that the 'diploid' *Glycine* have 2n=40 chromosomes while other papilionoids have 2n=10 or 11 suggesting at least one large-scale genome duplication (Lackey, 1980). While soybean functionally behaves as a diploid, haploid genome analysis have suggested that soybean is a diplodized ancient tetraploid (Hadley and Hymowitz, 1973). Evidence for at least two distinct sub-groups in soybean multigene families also support this hypothesis (Nielson et al., 1989). Mapping studies in soybean were able to identify homeologous regions as well as nested duplications within the genome suggesting multiple rounds of duplication (Shoemaker et al., 1996; Lee et al., 1999; Lee et al., 2001). Further, these mapping studies revealed that more than 90% of the RFLP probes hybridized to two or more genomic fragments, suggesting conserved duplicated regions (Shoemaker et al., 1996).

In addition to mapping studies, analysis of BAC-end sequences has suggested that the retained duplicate regions of the soybean genome still share sequence homeology (Marek et al., 2001; Foster-Hartnett et al., 2002). Similarly, hybridization based approaches showed fairly extensive sequence homeology between RFLP anchored paralogous BACs (Yan et al., 2003; Yan et al., 2004). Fluorescent in situ hybridization studies (FISH) using BACs as probes has shown cross-hybridization of BACs to more than one chromosome (Pagel et al., 2004). These results suggest that conserved duplicated segments in the soybean genome may retain sequence similarity.

These studies have laid the foundation to ask questions as to the structural and functional evolution of the paleopolyploid soybean genome. This dissertation is an investigation of retained duplicate genes identified from ESTs as well as structural/functional divergence and conservation of homeologous regions in the soybean genome. While whole genome sequence is not available for many plant genomes, it is possible to utilize EST resources and BAC-based sample sequencing to gain knowledge about the organization, structural and functional evolutionary history of a genome. The purpose of my doctoral research is to identify and describe preserved duplicated genes, sequence and characterize homeologous regions anchored by duplicate genes, and study the possible functional difference between conserved homeologs to gain an understanding of the mechanisms and processes of gene and genome evolution in the paleopolyploid *Glycine max*.
Dissertation Organization

This dissertation is organized into five chapters. Chapter 1 contains a literature review documenting the current state of research in polyploidy and gene duplication, fates of duplicate genes, genome restructuring and diploidization, the paleopolyploid nature of the soybean genome, comparative genomics of duplicated genomes, and gene expression in polyploids as they relate to this dissertation. Chapters 2, 3 and 4 each consist of a manuscript, published, accepted for publication or in preparation for publication, respectively. Chapter 5 is a summary of conclusions reached during the course of this dissertation research as well as recommendations for further research in these areas.

Chapter 2, entitled “Mining EST databases to determine evolutionary events in major crop species” was published in volume 47, 2004 issue of Genome (Schlueter et al., 2004). Potential duplicate genes were identified from the extensive EST collection in soybean as well as for maize, barley, sorghum, rice, tomato, potato, and Medicago. An analysis of synonymous distances between conserved duplicates allowed the authors to cluster them into groups with similar synonymous distances. Under the assumption of a molecular clock, dates were assigned to major genome duplication events. Using a ratio of nonsynonymous to synonymous distances, the authors showed that selection pressures between duplicate genes are generally negative. Additionally, Pearson correlation coefficients were calculated between pairs to measure how gene expression patterns have changed between duplicate pairs, and relative to their evolutionary distances. Contributions from co-authors include: Phillip Dixon was instrumental in determining the mixtures of normal distributions as well as text, Cheryl Granger began this analysis prior to moving forward with her career, David Grant provided support for the Pearson correlation coefficients for tomato and potato, and Lynn Clark and Jeff J. Doyle both provided significant background knowledge and writing of the evolutionary history of the grasses and legumes, respectively. All other analysis and writing was done by Jessica Ann Schlueter with editorial addictions from Randy Shoemaker.

Chapter 3, entitled “Sequence conservation of homeologous BACs and expression of homeologous genes in Glycine max” has been accepted to Genetics and is currently under revision. While homology between genomes in the legumes has been studied, homeology
between the genomes contained within a paleopolyploid had not been observed. Utilizing a pair of duplicated genes from Schlueter et al., 2004, we identified BACs (bacterial artificial chromosomes) representing homeologous regions in soybean that were subsequently sequenced to investigate the evolution and sequence conservation between duplicate regions in a paleopolyploid. The authors found that sequence conservation between homeologs in soybean is stronger than expected with most genes conserved in both order and orientation. An interrupted network of synteny to Arabidopsis and Medicago was identified between these regions. Finally, a subset of these identified duplicate genes were used to study how expression patterns between paleohomeologs have diverged across a variety of developmental stages and tissues. BAC sequencing was performed by Brian Scheffler as well as methods text. Visualization of the gene structures done with the xGDB system implemented by Shannon D. Schlueter. All other analysis and writing was done by Jessica Ann Schlueter with editorial changes from Randy Shoemaker.

Chapter 4, entitled “The FAD2 gene family of soybean: insights into the structural and functional divergences of a paleopolyploid genome” has been prepared for submission to The Plant Genome. This study looks at four BACs anchored by ω-6 fatty acid desaturase genes (FAD2). High sequence homeology was observed between two BACs, while the other two BACs had only the FAD2 genes in common. Expression patterns between the FAD2 genes was investigated specifically to look temperature dependent increase in expression in developing pods to correlated with temperature dependent increases in linoleic acid. Only one gene, FAD2-2C, showed evidence for increase in expression in developing pods from cool conditions. Sequencing of BACs was performed by Iryna F. Sanders, Shweta Deshpande, Jing Yi, Bruce Roe, and Brian Scheffler. Bruce Roe and Brian Scheffler provided methods text. Shannon Schlueter provided the xGDB system as an annotation tool. All other analysis and writing was done by Jessica Ann Schlueter with editorial changes from Randy Shoemaker.
Literature Review

Polyploidy and gene duplication

Gene duplication can occur by a variety of mechanisms: duplication of regions or segments of chromosomes, tandem duplication, reverse-transcriptase mediated duplication, and whole genome (polyploidy) duplication (Schmidt, 2002; Bennetzen, 2002; Wendel, 2000; Lawton-Rauh, 2003). Regional duplications, often called dispersive processes, can occur through abnormal crossing over events. Tandem duplications are frequently the result of replication slippage or transposon activity (Bennetzen, 2002; Lawton-Rauh, 2003). While single gene or regional duplication is seen in all plant species, whole genome duplication or polyploidy, has likely played the largest role in the evolution of plant genomes (Lawton-Rauh, 2003).

Gene duplication, arising from region-specific duplication or genome-wide polyploidization, is a prominent feature in plant genome evolution. All currently characterized eukaryotic genomes exhibit evidence of gene duplications, which continues to shape genomic diversity (Lynch, 2002). Most genes are members of multigene families, indicative of the prevalence of gene duplication in the origin and formation of diverse gene functions (Wendel, 2000). Although found across eukaryotic lineages, gene duplication appears to occur at an elevated rate in plants (Muller, 1925; Mable, 2004). The high incidence of gene duplication in plants could be due to its potential impact on genetic diversity and adaptation (Lawton-Rauh, 2003). Differential patterns of gene silencing following polyploidy may provide the genetic context to facilitate speciation (Werth and Windham, 1991). In Arabidopsis thaliana, 65% of all genes belong to gene families, 37.4% of which have at least five genes in their families.

Regional events of gene duplication include local duplications and dispersive mechanisms (Schmidt, 2002). Local duplications occur through transposable element activity or replication slippage resulting in tandem duplications whereas dispersive processes result in the transfer of genes or chromosomal segments to more distant chromosomal locations. Dispersive processes include single gene duplications from active transposable elements and
duplications of chromosomal regions of various sizes through aberrant crossing over events (inversion, translocation, unequal recombination, non-reciprocal translocation) (Bennetzen, 2002). A significant proportion of *A. thaliana* genes (17%) are arranged as tandemly repeated segments indicating that regional duplication and persistence are common in *A. thaliana*, and perhaps in most plant species (TAGI, 2000).

Studies focusing on duplicated genes and genome segments in *Arabidopsis* have provided insights into its evolutionary history. In 1994, Kowalski et al. found that single copy markers in *Brassica oleracea* mapped to duplicate regions in the *Arabidopsis* genome. Comparative genomics between *Arabidopsis* and soybean (Grant et al., 2000; Lee et al., 2001) and *Arabidopsis* and tomato (Ku et al., 2000) showed strong evidence for segmental and possible whole genome duplication in *Arabidopsis*. Blanc et al. (2000) suggested that *Arabidopsis* might be a degenerate tetraploid. Analysis of the whole *Arabidopsis* genome sequence using dot-plots indicated that upwards of 60% of the genome is duplicated and probably more than one large-scale genome duplication has occurred, although the exact number of rounds is still debated (TAGI, 2000; Vision et al., 2000; Lynch and Conery, 2000; Simillion et al., 2002; Blanc et al., 2003). As data from the rice genome sequencing projects have been compiled, it has become apparent that the rice genome also shows extensive evidence of gene duplication, likely the result of a major genome duplication event (Goff et al., 2002; Yu et al, 2003; Vandepoele et al., 2003). Two studies considering duplicated genes identified from EST collections has found evidence that most major plant species show evidence of retained duplicate genes with similar synonymous distances indicative of major duplication events (Blanc and Wolfe, 2004a; Schlueter et al., 2004).

Polyploidy was originally termed in 1916 by Winkler to describe organisms whose genomes have a greater amount of genetic material and chromosomes than their ancestors. There exist two main paths for a polyploid event to occur; either a doubling of unreduced gametes from a single species (autopolyploid) or the hybridization of unreduced gametes or somatic doubling from two different genomes (allopolyploid; Pikaard, 2001; Lawton-Rauh, 2003). The most common mechanism for polyploid formation is the fusion of unreduced gametes containing a diploid rather than haploid chromosome number and subsequence crossing with similar individuals (Pikaard, 2001). In some cases, the diploid progenitors of an
allopolyploid can be identified and the type of polyploidy event determined, as with cotton and *Brassica* (Wendel and Cronn, 2002; Parkin et al., 2003). However, if the polyploid event is not recent or the species has undergone multiple rounds of duplication and rearrangement, determining whether the event was allo- or autopolyploid can be very complex, as with soybean (Doyle et al., 2003; Straub et al., 2006).

The prevalence of polyploids in the plant kingdom suggests some sort of advantage over diploids; this is likely due to a heterotic effect generated by homeologous gene combinations (Comai, 2000). For example, within cultivated strawberries (*Fragaria. X ananassa* Duch.), octaploid strawberries were found to be morphologically favored over diploid strawberries (Bringhurst and Voth, 1984). Additionally, Bringhurst and Voth found that the octaploid strawberries were highly polymorphic whereas the diploids were not; the polyploids were likely able to adapt to more diverse environments with greater ease (1984). This observation was in agreement with that of Stebbins (1971) in that polyploids seem to tolerate more extreme climates better than diploids.

Polyploids often are formed on multiple occasions from the same or similar diploid progenitors, i.e. recurrent formation (Soltis and Soltis, 2000; Soltis and Soltis, 1999). Recurrent origins of a polyploid are seen as “the rule, not the exception” (Soltis and Soltis, 1999). Doyle et al. found that there were at least six hybridizations that led to *Glycine tabacina* (1990). Isozyme studies have shown that there were multiple origins of the *Glycine tomentella* polyploid (Kollipara et al., 1994). Similar examples are seen with *Heuchera grossulariifolia*, *Draba norvigica*, *Tragopogon miscellus* and *T. mirus* (Segraves et al., 1991; Brochmann and Elven, 1992; Cook et al., 1998). Recurrent formation may account for the large array of genetic diversity found within polyploid species (Soltis and Soltis, 1995). Further, gene flow between genetically different polyploids allows for recombination events and increased genotypes (Soltis and Soltis, 1999).

The increased genomic activity that occurs after a polyploid event often leads to difficulty in discerning ancient duplication events due to chromosomal shuffling and epigenetic silencing that arise from genomic coping mechanisms and the increased diversification among duplicate gene copies. This shuffling is referred to as diploidization.
and allows for multiple rounds of polyploidy to occur in the genomic history of a plant (Stebbins, 1966; Grant, 1981).

**Fates of duplicate genes**

There are various models for the fate of duplicated genes. The classical model predicts that after duplication, one member accumulates mutations and eventually becomes nonfunctional (Haldane, 1933). Ohno (1970) proposed that gene duplication was an efficient means to create new genes, and that whole genome duplication would enable the duplication of major biochemical pathways. Ohno proposed a model of gene duplication where, after duplication, one gene copy will retain the original function allowing the second “additional” copy to begin accumulating mutations. Since most mutations would be deleterious, they would lead to silencing of one gene copy. However, on occasion an advantageous mutation will become fixed and both gene copies will be retained (Ohno, 1970). Ohno’s model suggests that most secondary gene copies should become nonfunctional in a relatively few number of generations as determined by the mutation rate (Force et al., 1999).

Kimura extended these models by suggesting that after duplication, purifying selection is lessened in one or more of the copies allowing mutations to accumulate that would normally be selected against (1983). Most mutations would lead to pseudogene formation of one copy; occasionally a mutation would be beneficial and both copies would be retained. This model suggests that redundancy exists early in duplicate gene evolution and relaxation of selection allows the mutations to be nearly neutral in both copies. Studies done by Ohta in 1991 and 1994 found that accelerated amino acid substitutions are correlated with functional diversification of duplicate genes. These results implicated positive selection as having a role in duplicate gene fate.

In 1995, Walsh showed that if population size is large enough, even though advantageous mutations are relatively rare, there is a greater probability for a duplicate gene to acquire a new function than to be lost as a pseudogene thereby showing that the rates of gene silencing are lower than originally predicted. Further, when considering that whole genome duplications of vertebrates that happened over 250 MYA, upwards of 50% of
duplicate genes were estimated to be not only retained but have undergone functional divergence (Nadeau and Sankoff, 1997, Wagner, 1998). Using doubles and quadruples from human and mouse gene families with known chromosomal positions Nadeau and Sankoff (1997) were able to determine the rate of gene loss. Their results showed that there is a lower rate of gene loss than predicted under the classical model for duplicate gene evolution. They resolved this by postulating that after duplication more neutral mutations than deleterious mutations accumulate allowing for higher duplicate gene retention (Nadeau and Sankoff, 1997).

This classical model however lacks a mechanism through which new gene copies are retained and how they might survive the level of neutral mutations. Other studies have shown that the retention of duplicate genes is higher than expected under Ohno’s classical model of duplicate gene fate. Mapping studies between maize and sorghum showed that 72% of the probes were duplicated in maize and 38% were duplicated in sorghum (Whitkus et al., 1992). Both of these species are now considered to be paleopolyploids (Gaut and Doebley, 1997; Blanc and Wolfe 2004a; Schlueter et al., 2004). Additionally, in soybean, combined data from nine mapping populations uncovered extensive homoeologous relationships among linkage groups, with 90% of soybean RFLP probes detecting more than two fragments (Shoemaker et al. 1996). In fact, nested duplications were observed, suggesting that one of the soybean genomes involved in recent tetraploidization may have undergone duplication prior to tetraploidization (Shoemaker et al., 1996; Lee et al., 1999, Lee et al., 2001; Schlueter et al., 2004).

The higher preservation of duplicated genes than predicted under the classical model led Force et al. in 1999 to develop the duplication-degeneration-complementation (DDC) model for duplicate gene preservation. Their model incorporates the complexity of gene structure and regulation of gene expression by breaking down expression of a gene into subfunctions. Following duplication, mutations begin to accumulate across the sequence. These mutations can cause duplicate genes to begin losing subfunctions and functionally degenerating. The DDC model allows for three different outcomes for duplicate genes: 1) One copy may acquire a deleterious mutation in the coding region or in an essential regulatory region and become nonfunctional, essentially Ohno’s classical model for duplicate
gene loss. 2) One copy may acquire a mutation that allows it to gain new function and is
selectively swept to fixation a process referred to as neofunctionalization, the other outcome
possible under Ohno’s classical model. 3) Both copies begin to degenerate and lose
expression of particular subfunctions and both copies are required to compliment ancestral
function, a process called subfunctionalization. Of course, this model has an added layer of
complexity due to overlapping and embedded regulatory regions. Force et al. showed
examples of their model in zebrafish engrailed genes and in maize ZAG1 and ZMM2 genes
(1999). Altschmied et al. has also found plausible evidence for subfunctionalization if mitf
genes in teleost fish species (2002). In plants, excellent examples of subfunctionalization
following duplication in Arabidopsis are the APETALA 1 (AP1), CAULIFLOWER (CAL)
and FRUITFUL (FUL) genes. These genes are found in syntenic regions of the Arabidopsis
genome and are likely the result of a large-scale duplication (Prince and Pickett, 2002).
Studies in cotton have also seen potential evidence for subfunctionalization (Adams et al.,
2003)

Large-scale genomics project have allowed estimates to be derived as to the
preservation of duplicate genes as well as the fate of these duplicate genes (Lynch and
Conery, 2000). In C. elegans, S. cerevisiae, and D. melanogaster, synonymous mutations
were used to determine the rate of duplicate gene loss post duplication. By estimating the
half-life of duplicate genes, their results suggest that in the span of 50 million years more
than 90% of genes that were duplicated are lost. From the ratio of synonymous to
nonsynonymous distances, Lynch and Conery suggested that early in duplicate gene
evolution the pairs experience a brief period of relaxed selection or even near-neutrality
allowing for accelerated evolution (2000).

Determination of synonymous and nonsynonymous distances allowed Kondrashov et
al. to study the selection pressures acting on preserved duplicated genes (2002). Using
relative-rate tests, they were able to determine in Archaea, Bacteria, S. cerevisiae, C. elegans,
D. melanogaster, and Mammalia that most paralogs evolved under similar levels of purifying
selection. However, they did find that duplicated genes evolved faster than unduplicated
genes (orthologs) with similar levels of divergence. This was explained as being a result of
either relaxation of purifying selection and/or positive selection, a similar result to those of Lynch and Conery (2000; Kondrashov et al., 2002).

After duplication, there may be a short-term fitness advantage for duplicated genes (Kondrashov et al., 2002). A relaxation of purifying selection is likely the main mechanism behind the acceleration of evolution after duplication. Those duplicates that are retained aid in the adaptation to various environmental conditions, primarily through a protein dosage effect. Kondrashov et al. (2002) suggested a two-stage evolutionary model of gene duplication: 1) Immediately after duplication and during the early phase of their evolution, paralogs are retained and are subject to purifying selection because of the gene duplications are likely to proved a long-term advantage by enabling the creation of new functions. 2) Later in duplicate gene evolution, gene duplications are likely to proved a long-term advantage by enabling the creation of new functions.

Recently, Francino (2005) has taken the DDC concept and combined it with the results of Kondrashov et al., (2002). He suggests that the development of new gene functions is in response to the environment. A new situation arises, the original genes try to interact and something seems to work, so that particular product (gene) is amplified in expression and/or copy number until a mutation arises that actually does fit the new situation. As a result, it is fixed in the population. These amplifications and mutations are the result of positive selection. This is called the adaptive radiation model. Through gene duplication, the secondary copies allow for the retention of the initial product and allowing the secondary copy to be free of the purifying selection on the original copy. There are three predictions of this model: 1) the evolution of new functions after short bursts of gene amplification and gene fixation in response to specific selection pressures; 2) an initial period of positive selection on the paralogous gene, including those that eventually become pseudogenes; and 3) generation of numerous pseudogenes and eventual pseudogene loss accompanying the successful establishment of a new gene function or group of related functions. The fixation probabilities of gene duplication and loss during rapid adaptive periods will be very different from the long-term rates estimated from genome comparisons or from the silent substitutions between fixed duplicate pairs (Lynch and Conery, 2000). An example that fits this model is the olfactory receptor family in humans. This is a very large family (>1000 members) with
over 60% pseudogenes. Copy number and functionality is polymorphic between humans and some copies have high nucleotide substitution indicative of positive selection, suggesting ongoing adaptive radiation (Francino 2005).

**Genome restructuring and diploidization**

After a polyploidy event, duplicated regions begin to diverge from one another at both the sequence and chromosomal levels either through mutational or epigenetic means such that the polyploid becomes genetically diploidized (Stebbins, 1966; Grant, 1981; Pickett and Meeks-Wagener, 1995). Diploidization is likely a response to the stress or “genomic shock” experienced by a plant while in a polyploid state (Stebbins, 1966; McClintock, 1984). Allopolyploids have been shown to undergo numerous physical changes ranging from DNA sequence elimination, heterochromatin expansion, and reciprocal chromosome segment translocations and inversions, all thought to have a role in diploidization (Pikaard, 2001). Additionally, diploidization is not simply chromosomal/structural in nature, it also involves the diploidization of gene expression (Leipoldt and Schmidtke, 1982). In other words, RNA content in a diploidizing tetraploid is thought to be reduced to the level of the related diploids (Leipoldt and Schmidtke, 1982). On a genic level, diploidization involves the silencing of one copy or a divergence leading to a change in function of a copy (Pickett and Meeks-Wagner, 1995).

Following polyploidy there seems to be a genome-wide removal of some but not all of the redundant genomic material. It has been suggested that “differential gene loss” after a major duplication event may be responsible for much of the differences between closely related plants (Adams and Wendel, 2005). Diploidization at the chromosomal level is caused by additions, deletions, mutations, and rearrangements that rapidly inhibit non-homologous pairing of chromosomal tetravalents (Ohno, 1970). The primary effect of diploidization is the switch from tetrasomic to disomic inheritance in meiosis (Wolfe, 2001). Gilles and Randolph (1951) observed a reduction of quadrivalents and an increase in bivalents in synthetic autotetraploid maize stocks studied across a 10-year period. Allotetraploids derived from 4N maize and 4N teosinte as well as maize autotetraploids were used by Shaver to observe
macro-rearrangements of chromosomes in polyploids (1963). Shaver found that in the allotetraploid, there is increased preferential pairing that would influence diploidization (1963).

Diploidization was first observed in *Nicotiana tobacum* (Clausen, 1941). Clausen created an artificial amphidiploid of *Nicotiana tobacum*. Utilizing both the natural and synthetic *Nicotiana*, he was able to study chromosome behavior in response to polyploidization. He observed widespread changes such as genomic deficiencies in the natural *Nicotiana* that allowed it to appear and behave as a diploid. These results were a first glance at the process of diploidization (Clausen, 1941).

In 1958, Riley and Chapman identified a gene in polyploid wheat, *Phl*, which is involved in diploid chromosome pairing. Expression of *Phl* was found to be associated with a reduction in the number of quadrivalents and trivalents in meiosis (Riley and Chapman, 1958). This gene contributes to diploidization by inhibiting multivalent formation and suppression of homeologous pairing (Riley and Chapman, 1958).

Genomic reorganization often occurs rapidly after polyploidy and is extensive in most polyploids (Soltis and Soltis, 1999). Utilizing both nuclear and mitochondrial RFLP probes against synthetic polyploids in *Brassica*, Song et al. (1995) found an acceleration of genomic changes in the polyploid after formation, specifically chromosomal rearrangements. Although they considered that the observed banding patterns may have resulted from DNA methylation pattern changes or gene conversion, there was not overwhelming evidence to support these hypotheses. This led them to conclude that they were observing structural rearrangements leading to different RFLP banding patterns (Song et al., 1995). They concluded that rapid genomic changes might allow for greater genetic variation and possibly an acceleration of evolution in polyploid progeny (Song et al., 1995). Interestingly, some of the changes that were observed had a directionality bias toward the paternally inherited genome. Comparative mapping studies showed that when compared to the diploid progenitors, the natural *Brassica* allopolyploids showed extensive genomic rearrangements (Lagercrantz and Lydiate, 1996). These results along with those of Song et al. (1995) suggested that when the diploid progenitors are more divergent, the resulting genomic changes seen in the polyploid would be greater (Soltis and Soltis, 1999). Schranz and Osborn
have shown that resynthesized *Brassica napus* lines show extensive variation for traits in seemingly genetically homozygous lineages selected for one trait that differentiated between them (2004). This study utilized the synthetic polyploid lines from the study of Song et al. (1995). Schranz and Osborn (2004) showed the significant role of genetic and environmental interactions in the evolution of polyploids.

Results similar to those in *Brassica* have been observed in wheat; after allopolyploidy, sequence elimination is rapid and nonrandom for specific sequences. This results in a divergence of homeologous chromosomes (Feldman et al., 1997). Their analysis was done by comparing the pattern of loss of specific genomic sequences in hexaploid wheat to those of synthetic hexaploid wheat. They found the sequence elimination pattern was similar in wild and cultivated hexaploid wheat as well as synthetic hexaploid wheat. They were able to exclude methylation and intergenomic recombination as a means for genomic change since the observed changes were well conserved across many different lines (Feldman et al., 1997). These results suggested that the genomic changes after polyploidy might provide the physical basis for the diploid-like meiotic behavior of polyploid wheat (1997). Work looking at genomic rearrangements in tobacco in a synthetic polyploid showed that not only do genomic changes occur rapidly, within the first few generations, but that these changes are repeatable and consistent between synthetic and naturally occurring polyploids (Skalická et al., 2005).

It should be noted however that in synthetic *Gossypium* polyploids the rapid changes observed by Song et al. (1995) and Feldman et al. (1997) were not observed (Cronn et al., 1999; Liu et al., 2001). When comparing 16 loci in the diploid progenitors and both copies in the allopolyploid, there was little to no interaction between loci with little to no rate differences either (Cronn et al., 1999). Additionally, these results suggested that in cotton, duplicated genes evolve independently in the polyploid and at the same rate as the diploid progenitors (Cronn et al., 1999). There was evidence however for changes in repeated sequence copy number and activation of transposable elements in synthetic cotton polyploids (Liu et al., 2001). Similar results of polyploid genomic stability were found with *Tragopogon* (Pires et al., 2004) and *Spartinia* (Baumel et al., 2002). Comparison of these results with
those of Song et al. (1995) and Feldman et al. (1997) show that the concept of rapid genomic change may not hold true for all plant species.

A recent study looking at preserved paralogs from Arabidopsis revealed that some gene classes, such as transcription and signal transduction, may have been preferentially retained while other gene families have been preferentially lost (Blanc and Wolfe, 2004b). Seoighe and Gehring (2004) have similarly suggested that there is a functional bias of genes involved in transcriptional regulation and signal transduction to have been preserved in Arabidopsis following successive paleopolyploid events and that if a gene is retained after one round of duplications, that it is more likely to continue being retained in further duplications. This suggests that the birth and death process of gene duplication and rediploidization is not a random process (Seoighe and Gehring (2004).

The paleopolyploid nature of the soybean genome

Glycine Wild. is a member of the papilionoid Leguminosae tribe Phaseoleae, subtribe Glycininae. The north Asian subgenus Soja comprises the cultigen G. max (L.) Merr. (soybean) and its wild progenitor (G. soja; Doyle et al., 2003). Most genera of the Phaseolae have a genome complement of 2n=22, suggesting that soybean may have been derived from a diploid ancestor (n=11) that underwent aneuploid loss to n=10 and subsequent tetraploidization followed by diploidization (Lackey, 1980). Soybean (Glycine max (L.) Merrill) has a chromosome number of 2n=40 with a haploid genome size of 1115 Mbp compared with Arabidopsis (145 Mbp) and rice (415 Mbp) (Arumuganathan and Earle 1991). The physical to genetic distance ratio in soybean could be up to 400Kb/cM (Song et al, 2004). It has been reported that 40 to 60% of the soybean genome is made up of repetitive sequences (Goldberg 1978; Gurley et al., 1979). Heterochromatic DNA in pachytene chromosomes accounts for about 35% of the genome and appears to be clustered in distinct regions (Singh and Hymowitz, 1988). Sixty to seventy percent of the low copy sequences are thought to have interspersed repetitive elements of 0.3 to 0.4 kb in size (Gurley et al., 1979). Further, work of Polzin et al., 1998 showed by hybridization of genomic DNA to clones that only 30% of a genic region was repetitive suggesting that repetitive DNA may be clustered
and not dispersed through the genome. Marek et al., (2001) conducted a genome survey analysis, where mapped RFLP and SSR markers were used to identify BAC clones across the genome. Analysis of the BAC-end sequences suggested that gene organization in soybeans seems to follow the pattern where genes are more or less evenly distributed over long genomic stretches (Marek et al., 2001). FISH analysis with a highly repetative BAC shows that a large fraction of the repetitive sequence in soybean is concentrated in the pericentromeric regions (Lin et al., 2005).

Haploid genome studies have suggested that soybean is a diploidized ancient tetraploid (Hadley and Hymowitz, 1973). Study of relationships within soybean gene families has long supported this hypothesis (Lee and Verma 1984; Hightower and Meagher 1985; Grandbastien et al., 1986; Nielsen et al., 1989; Shoemaker et al., 2002). Shoemaker et al., (1996) compared the relative positions of RFLP probes across nine different mapping populations of soybean and reported that the soybean genome is highly duplicated. 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 they observed that each chromosome segment is duplicated on an average 2.55 times. This high level of duplication can be explained as, besides the tetraploidization, soybean genome would have undergone an additional round of genomic duplication for large parts of the genome (Shoemaker et al., 1996). In fact, nested duplications were observed, suggesting that one of the soybean genomes involved in recent tetraploidization may have undergone additional duplication prior to tetraploidization (Shoemaker et al., 1996; Lee et al., 1999; Lee et al., 2001). A study of 256 duplicated genes identified with EST sequences showed that soybean has undergone at least two major rounds of duplication at approximately 14.5 and 45 million years ago (Schlueter et al., 2004).

Whether the soybean genome is the results of an allo- or autotetraploid event is still not resolved. Utilizing the duplicated chloroplast-expressed glutamine synthetase gene in soybean Doyle et al. (2003) has shown that the duplication occurred subsequent to the divergence of *Glycine* from the extant Glycininae. The phylogeny of ncpGS further suggests that this duplication is the result of either an autopolyploid event or simple gene duplication (Doyle et al., 2003). However, this study was done with only one gene sequence and while
suggestive of autopolyploidy does not eliminate the possibility of allopolyploidy. A more recent study investigating the question of allo- versus autopolyploid in soybean utilized a low-copy nuclear gene for phylogenetic analysis and further supports an autopolyploid event (Shannon et al., 2006).

A large scale BAC-end sequencing project allowed Marek et al., (2001) to study the microsynteny between duplicated soybean regions identified by RFLP or BAC-end probes. On average, each RFLP probe identified 2.9 homoeologous regions. Of 34 contigs between duplicated regions, a total of 22 contigs (64.7%) showed microsynteny between potential paralogous regions (Marek et al., 2001). While only a sample of the genome, this number suggests that retained duplicated regions in soybean still share structural synteny.

Foster-Hartnett et al. (2002) utilized BACs to study homeologous regions in soybean. Twenty-eight BACs spanning a total of 1 Mb p spread throughout a 10 cM segment of soybean linkage group-G were identified with RFLP or PCR-based markers as well as fifty-nine homeologous BACs in seventeen contigs elsewhere in the genome. To compare homeologous regions they utilized DNA fingerprinting, DNA hybridization, and BAC-end sequence analysis (Foster-Hartnett et al., 2002). Fingerprinting patterns between homeologous BACs were sufficient enough to distinguish between loci. Cross-hybridization between BACs showed that 46% of the contigs showed extensive conserved microsynteny. Homoeologous BACs also demonstrated 98% sequence identity in several short segments (BAC-end sequences) that were compared in detail (Foster-Hartnett et al., 2002).

Similarly, by utilizing a hybridization-based approach, microsynteny was further found between paralogous genomic regions in soybean originally identified from RFLP probes (Yan et al., 2003). Yan et al. (2003) found that 86.5% of the 37 duplicated BAC contigs exhibited hybridization-based microsynteny. Restriction maps between eight soybean BAC contigs identified by RFLP’s show extensive genomic structure conservation between homoeologous regions in soybean (Yan et al., 2004). Chromosome level homeology has been seen with BAC-based FISH to soybean chromosomes; more often than not a BAC will hybridize to another chromosome (Walling et al., 2006). These results combined with those of Yan et al. (2003, 2004), Foster-Hartnett et al. (2002), and Marek et al. (2001) suggest that
conserved duplicated segments in the soybean genome retain highly similar genome and sequence organization.

Two recent studies looking at BAC sequences identified from duplicated genes suggests that while the soybean genome is a diploidized paleopolyploid, an astounding amount of sequence is conserved (Schlueter et al., in press; Schlueter et al., unpublished results). These results seem to follow the predictions of selective sequencing and mapping studies that there is still strong conservation in the coding regions of the soybean genome.

Comparative genomics of duplicated genomes

Colinearity between most plant species has traditionally been shown with mapping studies using probes from one species in another species. In the grasses, these studies have suggested that colinearity is highly conserved (Ahn and Tanksley, 1993; Devos et al., 1994; Gale and Devos, 1998; Devos and Gale, 2000). The sequencing of the rice genomes, both indica and japonica, has provided a new resource to the grass community for comparative analysis (Goff et al., 2002; Yu et al., 2002). What has come to light is that the conservation observed at the genetic map level is not necessarily conserved at the sequence level (Bennetzen and Ramakrishna, 2002; Feuillet and Keller, 2002; Bennetzen and Ma 2003). In other words, colinearity does not necessarily reflect microcolinearity. Sequence similarity in orthologous regions between species has recently been utilized across the grasses to further understand microcolinearity (Tarchini et al., 2000; Iliac et al., 2003). By studying a microcolinear region from many species, it has been possible to gain insight in how sequence evolution has changed since the major grass radiation (Iliac et al., 2003).

Colinearity of the sh2-homologous region in maize, rice, and sorghum suggests that although gene arrangement is conserved across these three species, the intergenic DNA content varies greatly (Chen et al., 1997). Another example of such study is the Adh1 locus across the grasses. Adh1 exists in duplicate, Adh2, and was likely duplicated prior to the grass radiation (Gaut et al., 1999). The Adh1 genes from maize and sorghum are not colinear with those in rice, but showed interrupted colinearity with deletions, insertions and translocations of genes between the species (Tarchini et al., 2000). Analysis of this region also suggests that
the expansion of the Adh1 region in maize occurred in the last 3-6 million years likely caused by retrotransposon insertions (Tarchini et al., 2000; SanMiguel et al., 1998).

Ilic et al. (2003) expanded the study of the Adh1 orthologous regions to include the Adh1-homoeologous region in maize, a documented paleotetraploid (Gaut and Doebley, 1997; Schlueter et al., 2004). Homeologous genomic segments in maize have been found to cover approximately 60-82% of the genome (Gaut, 2001). Iliac et al.’s analysis revealed several genic rearrangements, and demonstrated different evolutionary trajectories of this region in maize, sorghum, and rice. Sequence comparison of homoeologous regions in maize showed local genomic structure of the diploid progenitors of maize, as well as the nature and timing of numerous subsequent genomic rearrangements. Between these homoeologous Adh1 regions, only four predicted genes/gene fragments were colinear in order and orientation. Sequence similarity between the homoeologous regions allowed the detection of small fragments of residual gene sequences. In maize, over 40% of the total genes in the two homoeologous regions were deleted. Further, there did not seem to be preferential deletion from one subgenome over the other subgenome. As expected by the mobile nature of the maize genome, in both regions transposable element insertions and deletions were widespread. It appears from this analysis that the maize genome has undergone extensive rearrangements and gene loss while rediploidizing (Ilic et al., 2003). This was the first published studies of homoeologous sequences in a rediploidized monocot genome.

To additionally homeologous regions have been sequenced since Ilic’s analysis; the lg2/lrs1 loci and the Orp loci (Langham et al., 2004; Ma et al., 2005). Although the homeologous Adh1 regions retained four predicted genes/gene fragments, the lg2/lrs1 loci and the Orp loci, only the duplicated gene that anchored each region was retained. It appears from these analyses that the maize genome have undergone extensive rearrangements, transposable element insertions, and gene loss after duplication (Ilic et al., 2003; Langham et al., 2004; Ma et al., 2004). Conversely, an analysis of homologous CesA1 regions in cotton, a relatively recent allotetraploid, found extensive genic as well as intergenic sequence conservation with variation only in small insertion and deletions and transposable elements (Grover et al., 2004). Much as with maize and cotton, analyses of homeologous regions in soybean have provided insights into the organization of a paleopolyploid genome (Schlueter
et al., in press). Surprisingly, sequencing of homeologs in soybean has shown that there is still much sequence retention although the last duplication event is thought to have occurred at a similar time as maize (Schlueter et al., in press).

An analysis of the bzl region in two maize lines (McC and B73) has revealed high levels of deletions in this region between the two lines (Fu and Dooner, 2002). Interestingly, analysis of this same region in several maize inbreds indicated polymorphism for these deletions. That is, some haplotypes still contained genes that were deleted from other haplotypes. By hybridizational criteria, all of the genes that were sometimes missing from the bzl region still had at least one copy elsewhere in the maize genome (Fu and Dooner, 2002). These results suggest that the bzl region is undergoing the same gene-loss events that were detected in Adhl orthologous segments, but that the deleted state is not fixed near bzl. Their analysis demonstrated that within a single species, sequence microcolinearity might not be conserved at the same locus.

Evidence for conserved microsynteny in the dicots has been found between soybean, Medicago, and Arabidopsis thaliana (Grant et al., 2000; Yan et al., 2003) Utilizing RFLP sequences mapped to the soybean genetic map, Grant et al. was able to show conserved syntenic regions between soybean and Arabidopsis (2000). Further, by reconstructing putative ancestral chromosomes between the two species, evidence for a major genome duplication in Arabidopsis was found (Grant et al., 2000). A hybridization strategy using both whole BACs and probes derived from soybean BACs was able to show approximately 54% microsynteny between soybean BAC contigs and Medicago (Yan et al., 2003). In addition, 86.5% of the contigs showed conservation within the soybean genome, further evidence for the paleopolyploid nature of the soybean genome. Yan et al. also found that two of their probes provided evidence for a conserved ancient duplication in both soybean and Medicago and most likely predates the divergence of the two species (2003). Similar results were found by an analysis of EST's in soybean and Medicago (Schlueter et al., 2004; Pfeil et al., 2005). A phylogenetic approach used by Pfeil et al. determined that the ancient duplication in soybean was shared between soybean and Medicago and likely with all of the legumes approximately 50 MYA. (2005). Significant marker colinearity with EST-based markers was seen between Medicago truncatula and Medicago sativa (Choi et al., 2004).
Comparative mapping studies in the legumes have shown that there is some conservation at the genomic level across species. Within Phaseoleae, mungbean (*Vigna radiata*), cowpea (*Vigna unguicalata*) exhibit conserved linkage blocks (Menancio-Hautea et al., 1993). Similarly, within the Vicieae tribe, pea and lentil (*Lens culinaris*) exhibit a conserved genetic structure (Weeden et al, 1992). Gualtieri et al, (2002) analyzed microsynteny between pea and *Medicago* at the SYM2 locus, and concluded that the gene content and order are largely conserved between the two species at this region. They also showed that the microsynteny is conserved throughout a significant portion of the corresponding linkage group in both species.

Boutin et al, (1995) conducted a comparative mapping study among soybean, mungbean and cowpea using 219 DNA probes derived from these species. Genome conservation between soybean and mungbean or commonbean restricted to short and dispersed linkage blocks. Colinearity among these species was evident, but many conserved linkage blocks exhibited large differences in genetic distance between the markers. It appears that across the legume species, synteny is high among the closely related species and the degree of synteny decreases with increasing phylogenetic distance. Sequence comparisons revealed a high level of genome conservation between the *Medicago* and *Lotus japonicus* whereas lower levels of conservation between *Medicago* and soybean (Choi et al, 2004).

All these reports provide evidence of conservation in genome macrostructure among the legume species. But the level of conservation is not always perfect. Often, structural divergence including insertion or deletion of gene or groups of genes disrupts the conserved genome macrostructure. (Choi et al, 2004) and the colinearity may be apparent only over small chromosome segments (Paterson et al, 1996).

*Gene expression in polyploids*

Relatively little is known about the functional consequences and evolutionary importance of expression changes after genome doubling, although previous studies predict that one of the consequences is a partitioning of ancestral function between the two
duplicates (Force et al., 1999; Lynch and Connery, 2000; Lynch et al., 2001). Galitski et al. (1999) approached the question of gene expression changes related to ploidy levels by utilizing oligoarrays in *Saccharomyces cerevisiae*. They found that there were particular genes identifiable whose expression relative to total gene expression either increased or decreased in response to ploidy levels. The identified genes were assumed to be ploidy-regulated genes. Two mechanisms were thought to play a role in the transcriptional response to ploidy changes: 1) an overall increase in gene number per cell causing transient homologous pairing and 2) an increase in the DNA content per cell affecting the import and nuclear concentration of regulatory proteins (Galitski et al., 1999). In Arabidopsis, a recent study utilizing microarrays and synthetic allopolyploids found evidence for differential expression and that this expression is nonadditive (Wang et al., 2006).

Studies done in hexaploid and tetraploid wheat suggest intergenomic suppression of expression after polyploidy (Galili and Feldman, 1984). Protein bands were seen in the tetraploid that were not observed in the hexaploid. These bands were attributed to suppression by the D genome in the hexaploid of particular genes usually expressed in the A and/or B genomes. This suppression was reestablished in a synthetic hexaploid as well suggesting that this process occurs rapidly after polyploidy and is directional, i.e. can be repeated experimentally (Galili and Feldman, 1984).

Duplicated genomes have also been shown to exhibit dosage effects in the level of expression of duplicated genes. In general, as ploidy increases, the levels of proteins also increase (Guo et al., 1996). A study in maize using a series of plants with increasing ploidy levels (monoploid, diploid, triploid and tetraploid) followed 18 genes and their relative levels of expression. In general, as there are more gene copies, the overall expression in the cell increases. However, they also found evidence of dosage effects, dosage compensation, and down-regulation of a gene as dosage increased (Guo et al., 1996). Polyploidy allows for a greater range of variation in expression levels in genes showing allele-dosage effects (Osborn et al., 2003). This larger range of variation is another possible explanation for the preservation of duplicate genes over time (Osborn et al., 2003; Lynch and Conery, 2000).

Another effect of polyploidy on gene expression is at the level of regulatory elements. Following polyploidy, there is not only an increase in gene number, but in the regulatory
network itself. Regulatory factors from each progenitor genome interact and affect the resulting expression (both promoters and inhibitors of gene expression) (Osborn et al., 2003). This is amplified when considering interactions between transcription factors that are dimers, trimers, etc. As discussed above, models for the fates of duplicated genes now consider changes in regulatory elements as having a strong role in functional differentiation (Force et al., 199).

The use of synthetic polyploids provides an excellent resource to study the immediate consequences of polyploidy as well as comparative to the natural polyploid. A synthetic allopolyploid between *Arabidopsis thaliana* and *Cardaminopsis arenosa*, supposed to reconstruct *Arabidopsis suecica*, found that of 700 genes examined using cDNA-AFLP techniques 20 were suppressed (Comai et al., 2000). There was evidence for activation of genes in the tetraploid, although less frequently than suppression of expression. This suggested to them that approximately 0.4% of the genes in the allotetraploid are silenced. As expected, the phenotypes of the allopolyploid showed a wide range between the parental phenotypes. Lee and Chen (2001) using cDNA-AFLP in the same synthetic allotetraploid *Arabidopsis* were able to show silencing of one diploid progenitor gene in the polyploid. Their results implicated epigenetic mechanisms for gene silencing in a polyploid.

Similar screens in have been performed in hexaploid wheat, *Triticum aestivum*, as well as the diploid, *Aegilops tauschii*, and tetraploid, *T. turgidum*, progenitors using cDNA-AFLP techniques (He et al., 2003). Their results showed reduced expression of genes in the hexaploid relative to the diploid and tetraploid individuals, often the products of silencing of one homeolog versus the other (He et al., 2003). They also found evidence for a few genes where expression was activated in the hexaploid. Replication of these experiments in different synthetic hexaploids as well as natural hexaploids showed that these gene expression alterations are not random and likely associated with the increase in genomic complexity (He et al., 2003).

A study by Adams et al., (2003) aimed to identify genes in allotetraploid cotton as well as in the synthetic allotetraploid that showed evidence for silencing, biased expression, or tissue-specific expression. They found evidence of reciprocal organ specific silencing of one gene copy versus another with a bias towards one homoeolog versus another. These
results support subfunctionalization or partitioning of ancestral function in a polyploid species. They did not however find evidence for preferential accumulation of transcripts from one genome versus the other. Adams et al. (2003) implicated epigenetic changes as a major factor for the observed gene expression differences. Because cotton is a relatively recent polyploid, it was unlikely that mutational forces had played a strong role in the observed expression differences (Adams et al., 2003). Further work in cotton has shown that expression differences or silencing between duplicated genes is often organ-specific (Adams et al., 2004).

Another similar study looked at the changes in gene expression in flowers via differential display during the early stages of autopolyploidy in *Paspalum notatum*, or bahia grass (Martelotto et al., 2005). They found out of the 9614 transcripts they surveyed, only 48 were shown to have differing expression between the diploid and the tetraploid. This constituted ~0.49% of the genes altering their expression following polyploidy (Martelotto et al., 2005).

A study of recent duplicated genes in Arabidopsis by Blanc and Wolfe (2004b) found that more than half of the pairs had significantly different expression patterns. They found evidence that perhaps some of the duplicate genes underwent “concerted divergence”, where duplicate genes evolve together in metabolic networks. They also found evidence that over half of the retained recent duplicates have some sort of functional change.

REFERENCES:


Mable, B. K., 2004. ‘Why polyploidy is rarer in animals than plants’ myths and mechanisms. Biological Journal of the Linnean Society. 82: 453-466.


CHAPTER 2. MINING EST DATABASES TO RESOLVE EVOLUTIONARY EVENTS IN MAJOR CROP SPECIES

A paper published in *Genome*¹

Jessica A. Schlueter², Phillip Dixon³, Cheryl Granger⁴, David Grant⁵, Lynn Clark⁶, Jeff J. Doyle⁷, Randy C. Shoemaker⁸

ABSTRACT

Utilizing plant EST collections, we obtained 1,392 potential gene duplicates across eight plant species: Zea mays, Oryza sativa, Sorghum bicolor, Hordeum vulgare, Solanum tuberosum, Lycopersicon esculentum, Medicago truncatula, and Glycine max. We estimated the synonymous and nonsynonymous distances between each gene pair and identified two to three mixtures of normal distributions, corresponding to one to three

¹ Reprinted with permission of *Genome*, 2004, 47, 868-876.

² Interdepartmental Genetics, Department of Genetics Developmental and Cellular Biology, Iowa State University, Ames, IA, 50011.

³ Department of Statistics, Iowa State University, Ames, IA, 50011.

⁴ Edwin O. Smith High School, Storres, CT, 06268

⁵ USDA-ARS, Corn Insect and Crop Genetics Research Unit, and Department of Agronomy, Iowa State University, Ames, IA, 50011.

⁶ Department of Ecology Evolution and Organismal Biology, Iowa State University, Ames, IA, 50011.

⁷ Department of Plant Biology, Cornell University, Ithaca, NY, 14853

⁸ USDA-ARS, Corn Insect and Crop Genetics Research Unit, and Department of Agronomy, Iowa State University, Ames, IA, 50011. Telephone 515-294-6233, FAX 515-294-2299, and e-mail rcsshoe@iastate.edu, corresponding author.
rounds of genome duplication in each species. Within the Poaceae we found a conserved duplication event among all four species at approximately 50-60 million years ago (Mya); an event that probably occurred prior to the major radiation of the grasses. In the Solanaceae, we found evidence for a conserved duplication event approximately 50-52 Mya. A duplication in soybean occurred approximately 44 Mya and a duplication in Medicago about 58 Mya. Comparing synonymous and nonsynonymous distances allowed us to determine that most duplicate gene pairs are under purifying, negative selection. We calculated Pearson correlation coefficients to provide us with a measure of how gene expression patterns have changed between duplicate pairs, and compared this across evolutionary distances. This analysis showed that some duplicates seemed to retain expression patterns between pairs while others showed uncorrelated expression.

INTRODUCTION

Gene and genome duplication have long been accepted as driving forces in the evolution of eukaryotic genomes (Stebbins 1966; Ohno 1970). Upwards of 80% of all angiosperms likely have polyploid origins (Masterson 1994). Genome duplication acts as a mechanism for the creation of genetic diversity through genome expansion, gene silencing, and creation of new gene functions (Pickett and Meeks-Wagener 1995; Petrov 2001). Until now, a global picture of the timing of genome duplications in the evolutionary histories of major plant species has been lacking.

Duplicate gene analysis allows us to gain an understanding of the processes of genome evolution and the evolutionary trajectories of various plant species (Gaut and Doebley 1997). A limitation of previous analyses has been the lack of large sequence sets and a commonality of data sources across numerous plant species. While whole genome sequence would be ideal, large-scale expressed sequence tag (EST) projects in maize (Lunde et al. 2003), sorghum (http://funegen.org/Sorghum.htm), rice (Ewing et al. 1999), barley (Michalek et al. 2002), tomato (Van der Hoeven et al. 2002), potato (Ronning et al. 2003), soybean (Shoemaker et al. 2002), and Medicago (Gyorgyey et al. 2000) represent an untapped resource for dissection of gene families and evolutionary studies of genomes.
This study utilized EST datasets to identify duplicated genes and gain insight into the evolutionary events and selective forces acting upon four Poaceae, two Solanaceae, and two Fabaceae. Utilizing large datasets minimizes a major problem in estimating coalescence times. It is widely accepted that the rate of nucleotide substitution can vary from gene to gene (Wray 2001). This can cause skewing of estimates of coalescence times when one or a few genes are used to estimate the genetic distance between segments of a genome. However, by drawing upon a large number of gene sequences representing a wide range of functions and genomic locations, the accuracy of divergence estimates increases. Determination of evolutionary distances between large numbers of pairs of duplicated genes allowed us to estimate when these polyploidization events may have occurred as well as patterns of duplication between related species. We also compared EST expression profiles between pairs of duplicated genes to gain insights in changes into gene expression in duplicated genes.

MATERIALS AND METHODS
IDENTIFICATION OF POTENTIAL DUPLICATES

Tentative contigs (TC's) from The Institute for Genome Research (TIGR) for Zea mays, Oryza sativa, Sorghum bicolor, Solanum tuberosum, Lycopersicon esculentum, Medicago truncatula, and Glycine max assembled on May 13, 2002 and Hordeum vulgare assembled on May 13, 2003, were obtained from the TIGR genome database (Quackenbush et al. 2000). Open reading frames (ORF's) were found for each TC consensus sequence with the program getorf (Rice et al. 2000). A perl script was written to select the longest ORF, of all possible ORFs, for each TC consensus sequence.

A similarity search of each species-specific longest ORF TC set against itself was done using a tblastx search algorithm (Altschul et al. 1990). The program MuSeqBox (Xing and Brendel 2000) tabularized the BLAST output, and a perl script was written to parse all potential paralogs. The criteria used were: 1) Query and subject coverage must be equal to or greater than 80%; 2) The subject identified must not be the query itself (a identifying a); 3) The subject and query were required to be reciprocal (if a identified b, then b must identify a); 4) Only pairs that fit the criteria were considered (i.e., if a triplet,
quadruplet etc. fit all criteria, it was not considered). Using this criterion, the average accepted E-value was $6 \times 10^{-9}$.

Nucleotide differences within a species as a result of variation among cultivars were minimized by developing genotype-specific contigs (TC's). The most prominent genotype within each EST collection was used. For *Zea mays*, the genotype used was cv. B73; for *Oryza sativa*, cv. Nipponbare; for *Sorghum bicolor*, cv. BTx623; for *Hordeum vulgare*, cv. Barke; for *Solanum tuberosum*, cv. Kennebac; and for *Lycopersicon esculentum*, cv. TA496. Contigs composed of all acceptable genotypes were unchanged, and the TIGR consensus sequence was used in subsequent analyses. Any contigs from which EST's were removed were recontigged with CAP3 (Huang and Madan 1999) and the parameters originally used by TIGR, 40 bp overlap and 95% similarity (Quackenbush et al. 2000). The consensus sequences for the genotype-specific contigs were used in subsequent analysis. Finally, any contig containing a single EST, and its paralogous contig, were eliminated from further analysis.

The longest putative ORF was obtained for each genotype-specific consensus sequence as described above. DNA alignments between pairs were done using virtual translation with the Clustalx method, default parameters (MegAlign DNASTAR, Inc.). Alignments were cropped to the largest sequence overlap, or between start and stop codons. For pairs for which no alignment was found, a blastx with both consensus sequences against the NCBI non-redundant database was used to determine the correct frame (Altschul et al 1990). That ORF was then used for alignment as previously described.

Putative annotation of each gene pair was accomplished by searching against the NCBI non-redundant database using the tblastx algorithm (Altschul et al. 1990) and an E-value of $1 \times 10^{-10}$.

EVOLUTIONARY DISTANCES AND MIXTURES OF NORMAL DISTRIBUTIONS

Synonymous (Ks) and nonsynonymous (Kn) distances as well as their standard deviations (sd) were calculated using the method of Goldman and Yang (1994) as implemented in PAML (Yang 1997). Since the sd of each Ks was not constant (larger
values generally having larger sd’s), a natural log (Ln) transformation of Ks values was used to normalize the standard deviations. For determination of mixture of normal distributions, Ks values less than 0.05 were not used (Kondrashov et al. 2002). Ks values greater than 1.0 were considered to have reached a saturation point where multiple substitutions may have occurred at a synonymous site and were not included (Lynch and Conery 2000).

The Ln-transformed Ks values were modeled as a mixture of k normal distributions (McLachlan and Peel 2000). We considered models from k=1 through k=5 components. Each component was allowed a different mean, variance, and probability of membership in that component. The membership probabilities were constrained to sum to 1. The model for k components has 3k-1 parameters.

Parameters were estimated by maximum likelihood (ML). The log likelihood function has many false optima in which one component represents a single observation. The variance for that component converges to 0 and the log likelihood function converges to infinity. We therefore constrained the variance of each component to be greater than 0.0001 to avoid false optima.

The log likelihood function was numerically maximized using Splus (MathSoft Inc., 1998). We used 100 different sets of starting values, randomly chosen from the space of reasonable parameter values. Solutions with a component that represented a single observation were rejected. The ML estimates were obtained from the non-rejected solution with the largest log likelihood.

The number of components was chosen by a series of log likelihood ratio tests comparing models for k components to models for k+1 components. The distribution of the log likelihood ratio statistic for this test is non-standard because of the constraints on the component variances. For all number of components, P-values for these tests were computed using the 5 d.f. Chi-square distribution. The accepted number of components was determined with statistical significance of p ≤ 0.05. The number of components chosen by hypothesis testing was generally consistent with those from Akaike Information Criteria (AIC) and Bayesian Information Criteria (BIC) model selection criteria (data not
shown). The AIC and BIC model selection criteria choose a model without a hypothesis test.

Genes were classified using the maximum posterior probability rule. For each gene, the probability of membership in each component was computed from the parameter estimates and the ln Ks value. The classification of each gene was the component with the largest posterior probability.

For graphical representation of the data, the normal distributions fit to the ln-transformed data were back-transformed, and the normal distribution density functions became ln-normal density functions. For determining coalescence estimates, the median of the back-transformed data was used. Using the median minimizes the impact of skewness in the synonymous distance distributions.

PEARSON CORRELATION

Clustering of libraries and contigs with respect to expression profiles was done as described by Ewing et al. (1999) and Shoemaker et al. (2002). All EST's that comprised each TIGR TC were used to determine counts of EST's from each EST library for each contig. Data were normalized using percentage counts (EST count/library size) instead of raw EST data (Shoemaker et al. 2002). Pearson correlation coefficients (PCC's) for expression profile comparisons between duplicate contigs were calculated using Excel (Microsoft Corporation).

RESULTS AND DISCUSSION
IDENTIFICATION OF POTENTIAL PARALOGS

A total of 1,392 duplicated genes were identified across the eight studied species (Table 1). Synonymous distances (Ks) between duplicated gene pairs across all species ranged from 0.004 ± 0.004 to 5.22 ± 49.1. Nonsynonymous distances (Kn) ranged from 0.0 ± 0.0 to 0.6783 ± 0.0639. This wide range is indicative of our methods identifying both more recent duplicate genes as well as ancient preserved duplicate genes. By using the translated blast to search for similarity between EST consensus sequences we identify more divergent paralogs, those containing numerous synonymous substitution, than a
simple nucleotide-to-nucleotide blast would have allowed. Constraining our search parameters to those alignments having 80% subject and 80% query coverage greatly reduces the possibility of identifying small motifs is reduced. Additionally, our strategy was designed to eliminate any sequence difference between cultivars that would result in skewed evolutionary distances. This was based on major differences that have been identified in maize in the \textit{bz} genomic region between two cultivars McC and B73 (Fu and Dooner, 2002).

**EVOLUTIONARY DISTANCES AND MIXTURES OF DISTRIBUTIONS**

Our assumption was that groups of gene pairs with similar synonymous distances within a species correspond to temporally correlated duplication events, such as would be produced by large-scale segmental duplications or polyploidy. We further assumed that synonymous distances characteristic of these groups would be represented as normal distributions, and each normal distribution then corresponded to a duplication event. Multiple duplication events resulted in mixtures of normal distributions.

We found evidence across all species for more than one normal distribution. Maize synonymous distances demonstrated a trimodal distribution, whereas all other species demonstrated bimodal distributions (Figure 1). The number of distributions, the \text{ln}-transformed mean, standard deviation of the distribution, the median, and the number of genes associated with each distribution are shown in Table 2.

A few caveats of distance measures should be addressed. First, an overestimation of 'time since divergence' can occur due to asymmetrically bounded variables (non-elastic boundary at the present and an elastic boundary at the past) (Rodriguez-Trelles et al. 2002). Second, reduction in the precision of estimates of divergence can also be caused by variation in rates of nucleotide substitution (Wray 2001) and rates of substitution may vary with functional differentiation, something often associated with gene duplication (Ohta 1994; Pickett and Meeks-Wagner 1995). Third, low numbers of EST's in contigs can result in incomplete coverage of the coding region and can result in biased distance measures. Finally, ancient duplications become saturated at synonymous sites. This leads to potentially skewed estimates of synonymous and nonsynonymous substitutions among
duplicated genes and may contribute to inaccuracies in the estimation of divergence times (Ohta 1994). Several of these sources of experimental error were minimized. Utilizing EST contigs with large numbers of EST's increased the probability of obtaining the full coding region. Also, as the number of gene sequences increased, as with large datasets such as ours, the accuracy of the divergence estimates also increases (Nei et al. 2001). Finally, by bounding our search to Ks values less than one, we reduced the effects of saturation in more ancient duplications.

COALESCENCE ESTIMATES

If synonymous substitutions are selectively neutral and accumulate approximately linearly with respect to time, then the relative age-distribution of duplicated genes may be inferred from the observed distribution of synonymous distances (Gaut and Doebley 1997; Lynch and Conery 2000). We assumed that each ln-normal distribution is representative of groups of gene pairs evolving at a similar rate, and that rate is approximately equal across all paralogs. Coalescence estimates were determined using estimated synonymous substitution rates of $6.5 \times 10^{-9}$ substitutions per synonymous site per year for the Poaceae (Gaut et al. 1996), and $6.1 \times 10^{-9}$ substitutions per synonymous site per year for the Solanaceae and Fabaceae (Lynch and Conery 2000). Coalescence estimates correspond to the point when groups of gene pairs began evolving independently.

GRASSES

Our analysis of 101 pairs of maize genes yielded three normal distributions with coalescence estimates of approximately 10, 20 and 49 million years ago (Mya) (Figure 1). Our results support the segmental allotetraploid model proposed by Gaut and Doebley (1997) and Gaut et al. (2000). The divergence of the diploid progenitors probably occurred 20 Mya and the allotetraploid event 10 Mya. Evidence of gene duplication in maize has been well documented with multiple-copy RFLP's showing conserved order on homeoeologous chromosomes (Helentjaris et al. 1988). More than 72% of the loci in maize are estimated to exist in duplicate (Ahn and Tanksley 1993).
Bimodal distributions in sorghum correspond to two duplication events at approximately 11 and 50 Mya. Previous work found 38% of RFLP probes hybridize to more than one locus (Whitkus et al. 1992), and it has been suggested that sorghum is a paleopolyploid (Levy and Feldman 2002). An estimate, based upon two genes, placed the divergence between sorghum and maize approximately 16.5 Mya (Gaut and Doebley 1997). However, considering the likely geographical distribution of maize and sorghum or their immediate progenitors during that time period (Grass Phylogeny Working Group 2001), the event approximately 11 Mya in sorghum was more likely independent of the 10 Mya event in maize.

Similarly, both barley and rice show bimodal distributions (Figure 1). Although the earlier distributions in barley and rice are statistically significant, these distributions were not considered to be indicative of single duplication events. These distributions had large standard deviations, greater than 0.6 (Table 2). We believe that these broad distributions more likely reflect numerous independent duplications across time, and not a single major genome event. The more ancient peaks in barley and rice, with smaller standard deviations, yielded coalescence estimates of 60 and 52 Mya, respectively.

Our results show evidence for both segmental and whole genome duplications in rice, as previously proposed (McCouch et al. 1988; Nagamura et al. 1995; Goff et al. 2002; Yu et al. 2002, Salse et al. 2002; Vandepoele et al. 2003). Although the rice genome has been considered a simple diploid, there is a growing amount of evidence that its genome has undergone multiple duplication events (McCouch et al. 1988; Nagamura et al. 1995; Goff et al. 2002; Yu et al. 2002, Salse et al. 2002; Vandepoele et al. 2003).

The barley genome is almost double the size of maize, suggesting that duplication has played a role in the evolution of its genome. However, comparative RFLP mapping between barley and rice showed conservation of single copy sequence suggesting that the barley genome probably did not have a major duplication event after divergence from rice (Saghai Maroof et al. 1996). Our results support this suggestion.

The genera represented here belong to three major groups of grasses (pooids: barley; panicoids: maize, sorghum; and ehrhartoids: rice), all of which belong to the major radiation in the grasses (BEP + PACCAD clade) (Grass Phylogeny Working Group 2001). In all of these species, there is evidence for a major duplication event that occurred
approximately 50 - 60 Mya (Figure 1). Fossil data suggest a probably origin for the grass family 55 - 70 Mya, with the major radiation of the family occurring in the mid-tertiary, probably 15 - 25 Mya (during the Miocene) based on molecular clock estimates and macrofossils (Grass Phylogeny Working Group 2001, Kellogg 2001). The simplest explanation is that the duplication event occurred in the most recent common ancestor of the pooid, panicoid, and ehrhartoid grasses, but it could have occurred very early in the diversification of the family, perhaps even in the ancestor of all grasses.

SOLANACEAE

Our results suggest tomato is likely a paleopolyploid, with a large-scale genome duplication having occurred approximately 52 Mya (Figure 1). Because of its large standard deviation we do not use the more recent peak for coalescence estimation (Table 2). Estimates in tomato have shown that upwards of 50% of the genes exist as members of a gene family (Van der Hoeven et al. 2002, Bernatzky and Tanksley 1986). Previous studies also suggested that tomato is an ancient polyploid (Van der Hoeven et al. 2002, Ku et al. 2000).

Kennebec, the potato cultivar chosen for this study, is a cultivated, tetraploid outcrossing variety (Bonierbale et al. 1988). Our finding of a relatively recent grouping of duplicated genes corresponding to a major duplication event about 13 Mya was thus anticipated. This date, presumably, is the time at which the average gene shifted from tetrasomic to disomic inheritance in tetraploid potato. Our results coincide with previous studies suggesting that multiple rounds of duplication have occurred in the evolution of potato (Gebhardt et al. 2003).

The more ancient duplication in potato, 50 Mya, most likely is shared with tomato, 52 Mya, and represents a polyploid event early in the evolution of the Solanaceae. Colinearity is well conserved at the genetic map level between tomato and potato (Bonierbale et al. 1988, Tanksley et al. 1992). A previous divergence estimate between tomato and potato at 11.6 ± 3.6 Mya was obtained from members of the actin gene family using nonsynonymous distances (Kn) (de Sa and Drouin 1996). Our results, using
synonymous distance estimates from a larger sample set suggest similarly that tomato and potato diverged some point prior to 13 Mya (Figure 1).

**LEGUMES**

Two major duplication events at 15 Mya and 44 Mya in soybean were found in our analysis (Figure 1). Study of relationships within soybean gene families has long suggested that soybean is an ancient polyploid (Shoemaker et al. 2002, Lee and Verma 1984, Hightower and Meagher 1985, Grandbastien et al. 1986, Nielsen et al. 1989). Hybridization-based genetic maps are consistent with these findings (Shoemaker et al. 1996, Lee et al. 1999, Lee et al. 2001). Combined data from nine mapping populations uncovered extensive homoeologous relationships among linkage groups, with 90% of soybean RFLP probes detecting more than two fragments (Shoemaker et al. 1996). In fact, nested duplications were observed, suggesting that one of the soybean genomes involved in recent tetraploidization may have undergone duplication prior to tetraploidization (Shoemaker et al. 1996; Lee et al. 1999, Lee et al. 2001). Doyle et al. recently suggested that the soybean genome duplication was likely an autoploidy event but may have been an allopolyploid event (Doyle et al. 2003).

Our results also suggest that Medicago underwent a major genome duplication event approximately 58 Mya. This is likely prior to the divergence of the Medicago and soybean lineages (Wojciechowski 2003). If true, then this event should be seen in the soybean genome as well, and the duplication event estimated at 44 Mya is the likely candidate. We cannot rule out skewing of estimated coalescence times due to lineage-specific rate differences between soybean and Medicago. If the events in soybean and Medicago that occurred 44-58 Mya were a single event, then the difference in the rate of Ks is 1.3 times greater in Medicago than in soybean.

When gene trees are constructed from triplet and quadruplet sequences in soybean and Medicago, topologies predicted from a shared ancient event are found more frequently than are patterns expected from independent duplications (Pfeil, B.E., Schlueter, J.A., Shoemaker, R.C., and Doyle, J.J., unpublished data). Regardless of whether or not the ancient event was shared with soybean, Medicago is most likely an ancient polyploid.
Since the recent ‘event’ at 22 Mya in Medicago has a somewhat large standard deviation, (Table 2) it may not actually represent a major event but the result of numerous chromosomal duplications spread across time. This would be consistent with the general belief that Medicago has not undergone a recent whole-genome duplication event, but segmental duplications (Frendo et al. 2001).

**POST-DUPLICATION SELECTION**

Nonsynonymous to synonymous substitution (Kn/Ks) ratios > 1 suggest that genes are under positive selection with nucleotide changes affecting the amino acid sequence. Conversely, a Kn/Ks ratio < 1 indicates negative, purifying selection. Ks’s that are approximately equal to Kn’s are indicative of neutrality. Our results (Figure 2) are consistent with those observed for a number of eukaryotic species including *H. sapiens, M. musculus, D. melanogaster*, etc. in that most duplicated genes appear to be under purifying selection (Kondrashov et al. 2002; Lynch and Conery 2000). Genes under seemingly positive selection included two nodulation genes, two lipid transfer genes, a putative serpin, two malate dehydrogenases, an anaerobically induced early gene, a zinc finger protein gene, a DNA binding protein gene, and ten unknown genes with no similarity to the NCBI nonredundant database.

**DUPLICATE GENE EXPRESSION**

Redundancy caused by gene duplication often creates a situation in which selective constraints on both gene copies are temporarily lessened early after duplication (Kondrashov et al. 2002). During this period, both genes are free to accumulate mutations. Thus, gene duplication is often considered to be a major source of genetic diversity (Wendel et al. 2000).

To further understand how expression of duplicated genes has changed over time, Pearson correlation coefficients (PCC’s) between duplicate pairs of legume and Solanaceae genes were calculated. Duplicated gene pairs showing identical expression profiles will have a PCC of 1 whereas a PCC of −1 indicates the genes are showing completely opposite expression profiles. A PCC of 0 suggests uncorrelated expression of
the two genes. Across all four species we found a wide range of PCC values (Figure 3). What is intriguing is that PCC values of 1 can be found in all four species. Those duplicated genes have not diverged in their expression profiles yet both copies have been retained. However, most of the gene pairs gave PCC values ≥ -0.3. If we consider Ks as a measure of time we see a general trend of PCC values moving away from 1, purifying selection, towards 0, neutrality. This reflects a gradual change in gene expression over time probably indicative of sub- or neofunctionalization (Force et al. 1999). However, even after tens of millions of years of divergence a surprising number of duplicated genes still retain relatively similar patterns of expression.

A comparison of nonsynonymous distances (Kn) to PCC values suggests that even though the amino acid structure of the duplicated genes is changing, expression patterns often remain similar. Interestingly, four gene pairs in soybean and one in tomato seem to maintain a strong positive relationship of expression pattern with large Kn distance (Figure 3). In soybean, these genes show similarity to isoflavone reductase, dynein light chain, peroxidase, and an unknown gene; in tomato the gene was similar to a glycosyltransferase.

When we further examine Kn values and PCC values in the context of the relative ages of duplicates, an interesting trend appears: PCC values of ancient duplicates are intermixed with those of more recent duplicates (Figure 4). This coincides with our observation of strong negative selection among duplicated genes and also suggests that newly duplicated genes accumulate nonsynonymous substitutions and then seemingly stabilize.

Our observations agree with Lynch and Conery (2000) and Kondrashov et al. (2002) suggesting that duplicated genes experience a brief period of relaxed selection immediately after duplication. The duplicates surviving this initial process then remain relatively stable. Further, it has been suggested that during this relaxation in selection pressure, duplicated genes affect the fitness of the organism, leading to a temporary evolutionary advantage (Kondrashov et al. 2002). The efficiency of purifying selection has been estimated to increase approximately 10-fold after the initial relaxed selection (Lynch and Conery 2000). As a result, it has been hypothesized that the relaxation in
purifying selection is the main driving force of duplicate gene evolution, while positive selection plays a lesser role (Kondrashov et al. 2002).

Our results show that genome duplication has played a prominent role in the evolution of the Poaceae, Solanaceae, and Fabaceae. In all species we found evidence for putative segmental and whole genome duplications. Our analysis has shown that most duplicated genes are under negative, purifying selection. We do see evidence for some genes under positive selection, however they are few and not conserved across species. Large public EST databases and Pearson correlation coefficients provided us with a mechanism to examine gene expression of duplicate genes. Analyses such as this take advantage of the resources available to answer evolutionary questions that we have previously been unable to answer. As more sequence becomes available, our understanding of how genomes evolve will progress.

ACKNOWLEDGEMENTS
This work was supported by a North Central Soybean Research Program grant. This is a contribution of the Corn Insect and Field Crop Genetics Research Unit, U.S. Department of Agriculture (USDA) – Agricultural Research Service, Midwest Area, Ames, IA 50011.

REFERENCES


Table 1. Characteristics of identified duplicated genes.

<table>
<thead>
<tr>
<th>Genotype*</th>
<th>Number of paralogs</th>
<th>Percent genotype†</th>
<th>Total ESTs‡</th>
<th>Avg. Length, bp§</th>
<th>Avg. Syn. dist.¶</th>
<th>Avg. Nonsyn. dist.¶</th>
</tr>
</thead>
<tbody>
<tr>
<td>Soybean</td>
<td>275</td>
<td>Williams +</td>
<td>55%</td>
<td>236,461</td>
<td>473</td>
<td>1.105</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Williams82</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Medicago</td>
<td>308</td>
<td>Jemalong</td>
<td>91%</td>
<td>164,304</td>
<td>619</td>
<td>1.342</td>
</tr>
<tr>
<td>Tomato</td>
<td>238</td>
<td>TA496</td>
<td>70%</td>
<td>155,054</td>
<td>655</td>
<td>1.328</td>
</tr>
<tr>
<td>Potato</td>
<td>107</td>
<td>Kennebac</td>
<td>53%</td>
<td>78,934</td>
<td>495</td>
<td>0.9120</td>
</tr>
<tr>
<td>Maize</td>
<td>101</td>
<td>B73</td>
<td>32%</td>
<td>148,429</td>
<td>378</td>
<td>0.9548</td>
</tr>
<tr>
<td>Sorghum</td>
<td>138</td>
<td>BTx623</td>
<td>99%</td>
<td>81,190</td>
<td>554</td>
<td>1.794</td>
</tr>
<tr>
<td>Rice</td>
<td>59</td>
<td>Nipponbare</td>
<td>86%</td>
<td>103,340</td>
<td>373</td>
<td>1.353</td>
</tr>
<tr>
<td>Barley</td>
<td>166</td>
<td>Barke</td>
<td>30%</td>
<td>308,901</td>
<td>611</td>
<td>1.442</td>
</tr>
</tbody>
</table>

*The genotype selected for in the identification of paralogs.
†Percent of a particular genotype within entire EST collection at time of analysis.
‡Total ESTs used in TIGR TC development.
§Average alignment length of duplicate gene pair.
¶Average synonymous or nonsynonymous distance between pairs. Standard error is in parentheses.
<table>
<thead>
<tr>
<th></th>
<th>Number</th>
<th>Ln mean (std. dev.)</th>
<th>Median of distribution*</th>
<th>Divergence estimate (Mya)^</th>
<th># genes under peak^+</th>
</tr>
</thead>
<tbody>
<tr>
<td>Soybean</td>
<td>2</td>
<td>-1.671 (0.457)</td>
<td>0.188</td>
<td>14.5</td>
<td>60</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-0.6152 (0.289)</td>
<td>0.540</td>
<td>41.6</td>
<td>101</td>
</tr>
<tr>
<td>Medicago</td>
<td>2</td>
<td>-1.337 (0.597)</td>
<td>0.263</td>
<td>20.2</td>
<td>85</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-0.3425 (0.205)</td>
<td>0.710</td>
<td>54.6</td>
<td>81</td>
</tr>
<tr>
<td>Tomato</td>
<td>2</td>
<td>-1.285 (0.650)</td>
<td>0.277</td>
<td></td>
<td>32</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-0.4583 (0.248)</td>
<td>0.632</td>
<td>46.2</td>
<td>98</td>
</tr>
<tr>
<td>Potato</td>
<td>2</td>
<td>-1.876 (0.427)</td>
<td>0.153</td>
<td>12.5</td>
<td>22</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-0.5106 (0.267)</td>
<td>0.600</td>
<td>48.6</td>
<td>38</td>
</tr>
<tr>
<td>Maize</td>
<td>3</td>
<td>-2.056 (0.221)</td>
<td>0.128</td>
<td>9.83</td>
<td>18</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-1.364 (0.177)</td>
<td>0.256</td>
<td>19.6</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-0.4430 (0.213)</td>
<td>0.642</td>
<td>49.3</td>
<td>11</td>
</tr>
<tr>
<td>Sorghum</td>
<td>2</td>
<td>-1.953 (0.255)</td>
<td>0.142</td>
<td>10.9</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-0.4203 (0.240)</td>
<td>0.657</td>
<td>50.5</td>
<td>25</td>
</tr>
<tr>
<td>Rice</td>
<td>2</td>
<td>-1.979 (0.653)</td>
<td>0.138</td>
<td></td>
<td>9</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-0.3826 (0.247)</td>
<td>0.682</td>
<td>52.5</td>
<td>20</td>
</tr>
<tr>
<td>Barley</td>
<td>2</td>
<td>-1.142 (0.610)</td>
<td>0.319</td>
<td></td>
<td>41</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-0.2466 (0.181)</td>
<td>0.781</td>
<td>60.1</td>
<td>34</td>
</tr>
</tbody>
</table>

*The median was determined from back-transformation of the Ln-mean.
^Divergence estimates were calculated from the median of the distribution.
^The number of pairs under each distribution.
Figure 1. Diagrammatic representation of the determined ln-normal distributions in all eight studied species. The y-axis is the density function for ln-normal distributions. Each distribution represents a cluster of genes.
Figure 2. Scatter plot of synonymous distance versus nonsynonymous distance for all duplicate pairs in all eight species. The line shows neutral selection, where synonymous distance equals nonsynonymous distance. All points above the line show positive selection and all points below the line show negative selection.
Figure 3. Scatter plot of synonymous and nonsynonymous distances against Pearson correlation coefficients for soybean, Medicago, tomato and potato. Each point represents a duplicate gene pair. The circled pairs are discussed in the text.
Figure 4. Scatter plots of nonsynonymous distance against Pearson correlation coefficients.
The figure legends refer to the median of distributions for each determined distribution.
Appendices: Additional Figures
Appendix 1. Histograms of the percentage of duplicate gene pairs (secondary y-axis) versus synonymous distance between pairs (x-axis) for Zea mays, Sorghum bicolor, Hordeum vulgare and Oryza sativa. Overlying the histograms are Ln-normal distributions of the histograms. The density of the ln-normal function under each curve is shown on the primary y-axis.
Appendix 2. Histograms of the percentage of duplicate gene pairs (secondary y-axis) versus synonymous distance between pairs (x-axis) for the Solanaceae; *Solanum tuberosum* and *Lycopersicon esculentum*. Overlying the histograms are Ln-normal distributions of the histograms. The density of the Ln-normal function under each curve is shown on the primary y-axis.
Appendix 3. Histograms of the percentage of duplicate gene pairs (secondary y-axis) versus synonymous distance between pairs (x-axis) for the legumes; *Medicago truncatula* and *Glycine max*. Overlying the histograms are Ln-normal distributions of the histograms. The density of the ln-normal function under each curve is shown on the primary y-axis.
CHAPTER 3: SEQUENCE CONSERVATION OF HOMEOLOGOUS BACS 
AND EXPRESSION OF HOMEOLOGOUS GENES IN SOYBEAN (GLYCINE MAX 
L. MERR) 

A paper accepted by Genetics

Jessica A. Schlueter¹, Brian E. Scheffler², Shannon D. Schlueter¹ and Randy C. Shoemaker³

ABSTRACT

The paleopolyploid nature of the soybean genome was investigated by sequencing homeologous BAC clones anchored by previously identified duplicate N-hydroxycinnamoyl/benzoyltransferase (HCBT) genes. Genetic map locations of these BACs on linkage groups C1 and C2 show an RFLP marker A_059 syntenic between these regions. Annotation of the 173,747 bp and 98,760 bp BACs showed that gene conservation in both order and orientation is high between homeologous regions with only one gene insertion or loss and local tandem duplications differing between the regions. The nucleotide sequence conservation extends beyond the genic regions as well, probably due to conserved regulatory sequences. Most of the homeologs appear to have a role in either transcription/DNA binding or cellular signaling; this suggests a potential preference for retention of duplicate genes with particular functions in a paleopolyploid. Extended comparison to Medicago truncatula and Arabidopsis thaliana demonstrated a network of synteny with conserved genes interrupted by blocks with no synteny. Reverse transcriptase-PCR analysis of the retained homeologs in soybean showed that in the tissues sampled, most homeologs have not diverged greatly in

¹ Department of Genetics, Developmental and Cellular Biology, Iowa State University, Ames, IA 50011

² USDA-ARS MSA Genomics Laboratory, Stoneville, MS 38776

³ USDA-ARS-CICGR, Ames, IA 50011, corresponding author, e-mail: rcsshoe@iastate.edu.

Sequence data from this article have been deposited with the EMBL/GenBank Data under accession nos. DQ336954 and DQ336955.
their expression profiles, with four exceptions. Tandemly duplicated clusters of HCBT genes could be identified in *Medicago* and *Arabidopsis* and a phylogeny suggests that the tandem duplications occurred independently after speciation. These results are the first sequenced-based analysis of homeologous regions in soybean, a diploidized paleopolyploid.

**INTRODUCTION**

Gene duplication, arising from region-specific duplication or genome-wide polyploidization, is a prominent feature of genome evolution. Gene and genome duplication has been shown to provide morphological and fitness advantages, create genetic redundancy, expand genome size, and provide a source for forming diverse/novel gene functions (Wendel 2000). Although found across most eukaryotic lineages, gene duplication appears to occur at an elevated rate in plants with up to 100% of all angiosperms having a polyploid or paleopolyploid history (Masterson 1994; Lockton and Gaut 2005). The high incidence of gene duplication in plants is probably due to its impact on genetic diversity and adaptation (Lawton-Rauh 2003).

Increased evidence of paleopolyploidy in plants once thought to be purely diploid has come to light in recent years. Comparative mapping studies as well as genome sequencing efforts have revealed that both *Arabidopsis* and rice are actually paleopolyploids. (TAGI 2000; Vision et al. 2000; Lynch and Conery 2000; Simillion et al. 2002; Blanc et al. 2003; Yu et al. 2002; Goff et al. 2002). EST-based analyses of several plant genomes have also revealed evidence for large-scale genome duplications in a wide range of genera (Schlueter et al. 2004; Blanc and Wolfe 2004a).

Further evidence of paleopolyploidy in soybean has grown in the last few years. *Glycine max* (L.) Merr. (soybean) and its wild progenitor (*Glycine soja*) are members of the papilionoid Leguminosae tribe Phaseoleae, subtribe Glycininae, subgenus *Soja* (Doyle et al. 2003). While most genera of the Phaseoleae have a genome complement of $2n = 22$, soybean has a chromosome number for $2n = 40$ suggesting that soybean may have been derived from a diploid ancestor ($n = 11$) that underwent aneuploid loss to $n = 10$ and subsequent tetraploidization followed by diploidization (Hadley and Hymowitz 1973; Lackey 1980). Studies of soybean gene families have also suggested that soybean is a paleopolyploid (Lee
and Verma 1984, Hightower and Meagher 1985, Grandbastien et al. 1986, Nielsen et al. 1989). Combined data from nine mapping populations uncovered extensive homeologous relationships among linkage groups, with 90% of soybean RFLP probes detecting more than two fragments (Shoemaker et al. 1996). In many cases nested duplications were observed, suggesting at least two rounds of duplication and diploidization (Shoemaker et al. 1996; Lee et al. 1999; Lee et al. 2001). Utilizing duplicate chloroplast-expressed glutamine synthetase genes, Doyle et al. (2003) postulated that the most recent duplication might be the result of autopolyploid event, but that it was not completely clear.

The most compelling evidence to date has been generated from an analysis of duplicate genes identified from expressed sequence tags (ESTs). Large numbers of these conserved duplicate gene pairs with similar levels of divergence from one another allowed the identification of at least two major genome duplications (Schlueter et al. 2004; Blanc and Wolfe 2004a). These gene pairs are referred to as homeologs and not paralogs, here and throughout, since they most likely resulted from a polyploidy event and not single gene duplications. The coalescence times of these events was estimated to be approximately 14.5 and 41.6 million years ago (Schlueter et al. 2004). A phylogenetic based approach has shown that the most ancient duplication in soybean was likely shared with *Medicago* (Pfeil et al. 2005). BAC hybridization, BAC-end sequencing, and fingerprinting studies have suggested that the conserved homeologous regions within the soybean genome still retain upwards of 46 - 86.5% structural identity (Marek et al. 2001; Foster-Hartnett et al. 2002; Yan et al. 2003; Yan et al. 2004). Recently, integration of the soybean genetic and physical map using FISH has shown an identifiable conserved homeologous region between two chromosomes (Walling et al. 2006).

While these previous studies foreshadow the sequence level conservation in homeologous regions in soybean, no studies to date have actually sequenced and characterized any of these duplicated regions. In maize however, three separate homeologous regions have been studied, the *Adhl* loci, the *lg2/lrs1* loci, and the *Orp* loci (Illic et al. 2003; Langham et al. 2004; Ma et al. 2005). Between the homeologous *Adhl* regions, only four predicted genes/gene fragments were retained in both regions; for the *lg2/lrs1* loci and the *Orp* loci, only the duplicated gene that anchored each region was retained. It appears from
these analyses that the maize genome have undergone extensive rearrangements, transposable element insertions, and gene loss after duplication (Ilic et al. 2003; Langham et al. 2004; Ma et al. 2004). Conversely, an analysis of homologous CesA1 regions in cotton, a relatively recent allotetraploid, found extensive genic as well as intergenic sequence conservation with variation only in small insertion and deletions and transposable elements (Grover et al. 2004). Much as with maize and cotton, an analysis of homeologous regions in the soybean genome provides insights into the evolutionary forces that have shaped the genome after duplication.

Genetic colinearity and orthology has traditionally been identified with comparative mapping. While marker colinearity is highly conserved in the grasses (Ahn and Tanksley 1993; Devos and Gale 2000), it is less so in the legumes. However, various comparative mapping studies have identified conserved syntenic blocks (Menancio-Hautea et al. 1993; Boutin et al. 1995; Choi et al. 2004; Zhu et al. 2005). Syntenic blocks have been identified between soybean and Arabidopsis as well (Grant et al. 2000; Lee et al. 2001). A common theme to all of these studies is that synteny is high among the closely related species and decreases with larger phylogenetic distance. A BAC-based hybridization study between soybean and Medicago estimates 54% microsynteny (Yan et al. 2003).

Recently an analysis of three megabases of soybean sequence showed unusually high synteny to two Medicago chromosomes with upwards of 75% gene colinearity (Mudge et al. 2005). What has come to light in comparative studies in the grasses is that the conservation observed at the genetic map level is not necessarily observed at the sequence level (Bennetzen and Ramakrishna 2002; Feuillet and Keller 2002). In other words, genetic colinearity does not necessarily reflect sequence microcolinearity.

In this paper, we report the first sequence-based analysis of homeologous regions in soybean. Two soybean bacterial artificial chromosomes (BACs) representing homeologous regions of the genome were fully sequenced and annotated. Unlike the maize homeologous regions, analysis of the soybean sequences provides evidence of higher than expected genic and non-genic sequence conservation in a paleopolyploid genome. Comparative analysis of these regions to Medicago and Arabidopsis show a complex network of synteny. A conserved block of tandemly duplicated N-hydroxycinnamoyl/benzoyltransferase (HCBT)
genes was found in each soybean BAC. Based upon synonymous and nonsynonymous
distances, these genes seem to be evolving more rapidly than the other genes in both regions. Interestingly, clusters of HCBT genes are also found in *Medicago* and *Arabidopsis* and appear to have independently undergone tandem duplication in each species as supported by the resulting gene tree. Expression screens of the soybean homeologs showed that while cases of expression changes between homeologs can be identified, most of the duplicate genes show the same expression patterns.

**MATERIALS AND METHODS**

**Duplicate BAC selection:** A BLASTN-based similarity search with default parameters (Altschul *et al.*, 1990) was performed with previously identified duplicate gene pairs (Schlueter *et al.* 2004) against all genetically mapped RFLP sequences at NCBI. The duplicate pair corresponding to TIGR tentative contigs, TC104546 and TC114014 (Quackenbush *et al.* 2000) identified a *Phaseolus vulgaris* RFLP probe pBng181 (AZ044940, AZ044941) with e-value’s of $1 \times 10^{-59}$ and $1 \times 10^{-43}$ respectively. These TC’s correspond to N-hydroxycinnamoyl benzoyltransferase genes as annotated by Schlueter *et al.* (2004).

The consensus sequence for each TC was used to obtain a DNA alignment using virtual translation with the Clustalx method, default parameters (MegAlign, DNASTAR, Inc.). Homeolog-specific PCR primers were designed using Oligo 6.82 (Molecular Biology Insights, Cascade, CO) The primers were tested against *Glycine max* cultivar Williams 82 using a DNA Engine Gradient Cycler from MJ Research (Watertown MA). PCR reactions were 10 µl in volume and contained 1.1 X MasterAmp 2X PCR PreMix B (Epicentre, Madison, WI), 0.11 µM each primer, 50 ng Williams 82 DNA, and 0.1375 U of Taq DNA polymerase (Invitrogen, Carlsbad CA). PCR cycling conditions were 94° for 2 min, 35 cycles of 94° for 45 sec, annealing temperature for 30 sec, 72° for 45 sec, followed by a final extension of 72° for 3 min. Products were gel purified and subsequently sequenced at the DNA Synthesis and Sequencing Facility (Iowa State University, Ames IA) to verify homeolog specificity. The primer sequences for HCBT copy 1 (TC144014) were U, 5’ TGG TGC TGC AAT CTC TGA AGG T 3’ and L, 5’ GGA TTG GAC TTA GAA ACA GCA T
3', and for HCBT copy 2 (TC104546) primer sequences were U, 5' CAA ACC ATA ATG CCA GTG CT 3' and L, 5' TTG TAT CCG GTG AAA GAC AG 3'.

The Williams 82 G. max BAC library (Marek and Shoemaker 1997) was PCR screened using the previously described conditions. One BAC was identified for each primer pair, gmw1-74i13 for HCBT copy 1 (TC144014) and gmw1-52d3 for HCBT copy 2 (TC104546). BAC DNA was isolated using a Plasmid Midi kit (Qiagen) and insert size was determined by Not I digest and CHEF gel electrophoresis. Digest conditions were 3 μl of BAC DNA, 1X NEBuffer 3, and 0.8 U of Not I (New England Biolabs). BAC-end sequences were obtained using M13 forward and reverse primers at the DNA Synthesis and Sequencing Facility (Iowa State University, Ames IA).

**BAC sequencing and assembly:** BAC DNA was randomly sheared using a nebulizer (Invitrogen) and size selected for 2-4 kilobases (Kb) on a 1% low-melt agarose gel. Sheared DNA was phosphatase-treated, blunt-end repaired and cloned into the vector pCR4Blunt-TOPO (TOPO shotgun subcloning kit, Invitrogen). The recombinant plasmids were transformed into TOP10 E.coli cells by electroporation and selected on LB plates containing Kanamycin. For gmw1-52d3 a second subclone library of 7-9 Kb fragments was also made as described above.

Subclones were sequenced at the USDA-ARS Mid-South-Area Genomics Laboratory using M13 forward and reverse primers on an ABI3730XL with BigDye3.1. Base calling, vector trimming and contig assembly was done using SeqMan II starting with a match size of 12 bp, a maximum gap of 70 bp, and a minimum match percentage of 95% and decreasing the minimum match percentage to 90% as necessary to merge contigs (DNAStar, Inc.). The complete sequence of gmw1-52d3 was obtained using this method.

Closing of gaps on gmw1-74i13 was accomplished by two methods. In all cases but one, clone pairs spanned a gap and complete sequencing of that clone closed the gap. The last gap was closed using PCR primers designed from adjacent contigs. The PCR product was gel-purified, subcloned into TOPO TA vector (Invitrogen), transformed into TOP10 E.coli cells (Invitrogen) and sequenced with M13 forward and reverse primers. All gap-closing sequencing was done at the DNA Synthesis and Sequencing Facility (Iowa State University).
**Genetic mapping of BACs:** To genetically map each BAC, the corresponding sequences were manually scanned for di- and tri-nucleotide repeats of at least 7 base pairs in length. Primer pairs flanking the potential SSR markers were designed using Oligo 6.82 (Molecular Biology Insights, Cascade, CO) and tested against various soybean parents of mapping populations. PCR reactions were 10 µl in volume and contained 1 X PCR buffer, 1.5 mM magnesium chloride, 5 mM dNTPs, 0.5 µM each primer, 50 ng *Glycine max* parental DNA, and 0.025 U of Taq DNA polymerase (Invitrogen, Carlsbad CA). PCR cycling conditions were 94° C for 2 min, 35 cycles of 94° for 45 sec, 60° for 30 sec, 72° for 45 sec, followed by a final extension of 72° for 3 min. Resulting bands were run on either a 3% agarose 1 x TAE gel for larger (greater than 250 bp) products or 6% polyacrylamide 0.5 x PBE gel for smaller fragments.

Testing of parental lines revealed one SSR marker from each BAC that was polymorphic in the mapping population *Glycine max* A81-356022 X *Glycine soja* PI 468.916 (Diers et al. 1992; Shoemaker et al. 1996). The SSR from gw1-74i13 corresponds to a TAA/TAT repeat found from 99536 bp to 99593 bp with primer pair sequences of U, 5' AGG AAG CTG CTT TAC AAC GTC 3' and L, 5' CAA AGC GTC CAT ACC AAA GTC A 3' and a resulting PCR product of 682 bp in Williams 82. The SSR from gmw1-52d3 corresponds to a TA repeat found from 84195 to 84213 bp with primer pair sequences of U, 5' AGT CAT CGA ATA AAC ATA G 3' and L, 5' AGT AAA AAC TTG AAA TTG G 3' and a resulting PCR product of 150 bp. Genetic relationships between these SSRs and the established map were determined using MapMaker with a minimum lod score of 3.0 (Diers et al. 1992; Lander et al. 1987).

**Sequence Analysis and annotation:** Gene prediction was done using a combination of ab initio and EST-alignment based methods. For ab initio predictions, Genscan with *Arabidopsis thaliana*-based parameters (Burge and Karlin 1997), FgeneSH with *Medicago truncatula*-based parameters (http://www.softberry.com) and GeneMark.hmm with A. *thaliana*-based parameters (Lukashin and Borodovsky 1998) were run. For EST verified structure prediction, GeneSeqer at PlantGDB (Schlueter et al. 2003) was used to align both soybean ESTs and other plant putatively unique transcripts (PUTs; Dong et al. 2005) to the BAC sequences. These EST alignments can be viewed in an xGDB-based database at
http://haldane.agron.iastate.edu/GmaxGDB (Schlueter et al. unpublished results; xgdb.sourceforge.net). Each predicted gene was subjected to a BLASTP query of the NCBI nr database with default parameters to assign a putative function. An e-value threshold of $1 \times 10^{-10}$ was used to assign putative function. Also, through BLASTP, any conserved motifs in predicted genes were identified.

AVID global pair-wise alignments, with default parameters, were done between homeologous BACs to produce a VISTA plot to visualize nucleotide identity between sequences (Frazer et al. 2004). The percent identity and similarity between genes both intra- and inter-chromosomally was calculated using the program WATER (gap penalty of 10; extension penalty of 0.2; EMBOSS). Synonymous and nonsynonymous distances were calculated using PAML (Yang 1997) with the same parameters as Schlueter et al. (2004).

Putative retroelements were initially identified from ab initio gene predictions that were most similar by BLAST based annotation to polyprotein sequences. Both BLASTN and TBLASTX were performed against the TIGR repeat databases (www.tigrblast.tigr.org/euk-blast/index.cgi?project=plant.repeats). Potential LTR retrotransposons were searched for using LTR_STRUC, default parameters (McCarthy and McDonald 2003). Putative MITEs were identified using FINDMITE with direct repeats of TA, TAA, TAT, TTA and GCA, a terminal inverted repeat (TIR) length of 12, allowing 3 mismatches, repeat filtering, and a minimum and maximum MITE length of 50 and 750 respectively (Tu 2001). Soybean-specific repetitive sequences identified from soybean BAC-end sequencing projects (Marek et al. 2001) were also searched using BLASTN. RepeatMasker was run as well utilizing Repbase (Smit, AFA & Green, P RepeatMasker at http://ftp.genome.washington.edu/RM/RepeatMasker.html). A self-BLASTN of each BAC further identified putative repetitive regions as well as tandem duplications.

Putative homologous regions from Medicago truncatula, Lotus japonicus and Arabidopsis thaliana were identified by performing BLAST searches against the available sequences from Lotus, Medicago, and Arabidopsis from NCBI (http://www.ncbi.nlm.nih.gov). Medicago annotations were viewed using TIGR’s Medicago Gbrowse (http://www.tigr.org). Gene structures for Medicago were taken from TIGR but
further confirmed with Geneseqer as described above. Gene structures from *Arabidopsis* were taken directly from AtGDB (Schlueter *et al.* 2003; Xhu *et al.* 2001).

The gene tree for tandemly duplicated N-hydroxycinnamoyl/benzoyltransferase (HCBT) genes was obtained using PAUP (Sinaur Associates, Massachusetts). A multiple sequence alignment for all HCBT genes was generated using MUSCLE, default parameters (Edgar 2004). The alignment was then imported into MacClade (Sinaur Associates, Massachusetts) and large gaps were manually removed from the alignment. That alignment was then imported into PAUP (Sinaur Associates, Massachusetts) and a parsimony-based tree was generated with manual rooting with *Arabidopsis* using TBR branch swapping and 1,000 iterations.

**mRNA expression of retained homeologs:** Alignments between each retained gene pair, or gene family, were performed using ClustalX, default parameters (MegAlign, DNASTAR, Inc.) Primer pairs were designed for each gene to be specific for only one gene copy as described above for BAC identification. Additionally, when possible each primer pair was designed to flank an intron as an internal control. Each primer pair was tested by PCR against *Glycine max* cultivar Williams 82 genomic DNA to verify the homeolog specificity of each product as described above for BAC identification.

Greenhouse grown soybean tissue was collected from a range of organs and developmental stages of *Glycine max* cultivar Williams 82. For each time point, tissue was taken from at least three independent plants. Tissue for cotyledons, roots, and furled unifoliate were collected 3 days after emergence (DAE). Unfurled unifoliates tissue was 4 DAE. Another sample of cotyledons and roots were taken at 7 and 8 DAE respectively. Furled trifoliolate was collected 11 DAE and unfurled trifoliolate at 15 DAE. Flowers and pods were taken at 60 and 76 DAE, respectively. All tissue was flash frozen with liquid nitrogen. mRNA was extracted and purified from frozen tissue using the RNeasy Plant Mini Kit (Qiagen). Extracted mRNA was further treated using DNA-free DNase treatment and removal kit (Ambion) and quantified using a NanoDrop ND-1000 spectrophotometer (Wilmington, DE).

Reverse transcriptase-PCR screens were conducted across all tissues with the above primers. Reactions were 25 µl in volume and done using the SuperScript One-Step RT-PCR
with Platinum taq (Invitrogen) containing 1 X reaction mix, 5 μM each primer, 150 ng of RNA, and 1 μL of Platinum taq DNA polymerase (Invitrogen, Carlsbad, CA). PCR cycling conditions were 42°C for 45 min, 94°C for 4 min, 35 cycles of 94°C for 45 sec, annealing temperature for 30 sec, 72°C for 45 sec, followed by a final extension of 72°C for 5 min. RT-PCR products were run out on a 1% agarose, 1 X TAE, ethidium bromide gel. Controls for RT-PCR reactions included a “minus” reverse transcriptase reaction to test for genomic DNA contamination, a water template reaction to test for reagent contamination and a tubulin primer pair positive control as designed by Graham et al. (2003). All RT-PCR reactions were done with two to three independent biological replicates. These reactions provide a positive/negative screen for presence of a transcript in a particular tissue.

RESULTS

Identification, assembly and mapping of soybean homeologous regions: Shotgun sequencing of BACs gmw1-74i13 and gmw1-52d3 yielded 5,088 and 3,896 sequence reads respectively. Of the 3,896 sequence reads from gmw1-52d3, 2,072 were from the small fragment library and 1,824 were from the large fragment library. For gmw1-74i13, sequence assembly yielded three major contigs. Sequencing entirely across clone pairs closed two gaps; the final gap was closed by PCR amplification across the gap and subsequent sequencing. For gmw1-52d3, assembly was first done with the smaller library. Addition of the larger library to the assembly closed all gaps on gmw1-52d3. The overall assembled sequence length of gmw1-74i13 was 173,747 bp with an average coverage of 13x and gmw1-52d3 was 98,760 bp with an average coverage of 16x.

Simple sequence repeats (SSRs) were identified from each BAC by scanning the sequence for di- and tri-nucleotide repeats of at least 7 repeats in length; 14 SSRs for gmw1-74i13 and eight SSRs for gmw1-52d3 were observed. Linkage analysis placed gmw1-74i13 on linkage group C1 and gmw1-52d3 on linkage group C2 of the Glycine max A81-356022 X Glycine soja PI 468.916 mapping population (Lander et al. 1987; Diers et al. 1992; Shoemaker et al. 1996; Figure 1). These two markers combined with the RFLP A_059 seem to show a portion of these linkage groups that are syntenic and likely homeologous. Based
upon the *Glycine max* cultivar Forrest physical map, this region anchored by A_059_1 on linkage group C1 identified a higher than average number of BACs suggesting this region would be an excellent candidate to look for homeologous regions (Wu *et al.* 2004; Schultz *et al.* 2006).

It was surprising that neither of these BACs mapped to the pBng181 locus on A2 (www.soybase.org) given that marker showed high similarity to the HCBT EST contig sequences. However, another region of A2 has previously been shown to have syntenic markers with C2 based on RFLP mapping (Shoemaker *et al.* 1996). Therefore, it is likely then that pBng181 corresponds to another HCBT locus or cluster located on linkage group A2.

**Sequence annotation of potential genes:** The three ab initio gene prediction programs used (GenemarkHMM, Lukashin and Borodovsky 1998; Genscan, Burge and Karlin 1997 and FgeneSH, http://www.softberry.com) yielded somewhat similar gene structures, but varied enough that EST alignment-based gene structure confirmation using GeneSequer (Schlueter *et al.* 2003) was warranted. On average, 65% of the predicted gene structures had exon coverage with a soybean EST and a total of 62% of the intron junctions were confirmed with EST support. A total of 28 genes were predicted between both BACs with gmwl-74i13 containing 18 genes and gmwl-52d3 containing 10 genes. The ab initio predictions, EST alignments and resulting predicted gene structures can be visualized at http://haldane.agron.iastate.edu/GmaxGDB. Between the two homeologous regions, 9 genes appear mutually retained (Table 1; Figure 2).

The average gene density of gmwl-74i13 is one gene every 9.1 Kb and gmwl-52d3 is one gene every 9.9 Kb. This is slightly less dense than a previous estimate of one gene every 8 Kb (Young *et al.* 2003) and significantly less dense than the most recent estimate of one gene every 5.8-6.7 Kb (Mudge *et al.* 2005). Soybean genic regions appear to be half as dense as *Arabidopsis thaliana* (one gene per 4.5 Kb, TAGI 2000), almost half as dense as *Oryza sativa* (one gene every 5.3 Kb, www.tigr.org), and a third less dense than both *Medicago truncatula* (one gene per 6.7 kb, Young *et al.* 2005) and *Lotus japonicus* (one gene every 6.3 Kb, Young *et al.* 2005). The average GC content of each BAC is similar to one another with gmwl-74i13 being 32.27% and gmwl-52d3 being 31.85%.
Annotation of each gene sequence was done with BLASTP searches using translated coding sequence against the nonredundant database (nr) at NCBI. Further information pertaining to the annotation of each gene can be found in Appendix A. The coding regions of predicted genes (Table 1; Figure 2) range in size from a partial 567 bp WOX4 homeobox-leucine zipper transcription factor protein to a 2454 bp gene similar to an Arabidopsis thaliana expressed protein. The intron/exon structure of genes on both BACs varied widely from a cluster of 3 to 6 single exon HCBT genes to a membrane-like protein containing 11 or 12 exons, varying by one exon between the BACs.

Predicted gene 1 on BAC gmwl-74i13 encodes a WOX4-like homeobox-leucine zipper transcription factor protein. Homeobox genes have been implicated across diverse organisms as being involved in the regulation of various developmental processes, but most specifically with cell fate in embryonic patterning in Arabidopsis (Haecker et al. 2004). This gene family contains a characteristic amino acid sequence downstream of the homeodomain called the WUS box (Haecker et al. 2004). The gmwl-74i13 translated sequence contains an exact match to this conserved sequence, LFELFPLH, from Arabidopsis thaliana WOX4, as found from AtGDB (Schlueter et al. 2003; Xhu et al. 2001).

The second predicted gene poly-A binding protein (PABP) is most like a PABP from Daucus carota (Genbank accession AAK30205.1; supplemental table 1). PABPs are generally associated with binding to the poly(A) tails of mRNAs and providing stability and increasing translational efficiency. Specifically in D. carota PABP shows differential expression in the development of somatic embryos (Lin et al. 2003). PABP proteins are a diverse multigene family in plants usually containing characteristic RNA recognition motifs (RRMs) on their N terminal ends, each supposed to have various RNA-binding affinities (Le et al. 2000). The gmwl-74i13 PABP (full length copy) encodes four RRMs at the 5’ end of the mRNA. The gmwl-52d3 PABP is at the end of the BAC and only encodes the last exon of PABP, not enough of the coding sequence to identify any RRMs. Characterized PABP’s in Arabidopsis seem to be expressed in a tissue-specific manner and the observed high levels of divergence in the C-terminal end of each protein have been suggested to be responsible for these expression differences (Le et al. 2000).
Putative 474 and 492 amino acid membrane-like proteins are downstream of the predicted PABP proteins on gmw1-74i13 and gmw1-52d3, respectively. Both of these proteins are most like putative membrane-like proteins from Arabidopsis and Oryza genomic sequence. Beyond this similarity, no other conserved or characterized features could be identified for these genes.

In both BACs the subsequent genes are clusters of 3-6 tandemly duplicated N-hydroxycinnamoyl/benzoyltransferase (HCBT) genes. HCBT functions as the first step of phytoalexin biosynthesis by catalyzing the reaction of anthranilate and benzoyl-CoA to N-benzoylanthranilate or dianthramide B (Yang et al. 1997). The accumulation of these dianthramide phytoalexins has been associated with plant response to a pathogen attack (Yang et al. 1997). Three independent HCBT genes were first cloned in Dianthus caryophyllus L. (Yang et al. 1997). Much like 8 of the 9 observed HCBT genes in soybean, the carnation HCBT genes were found to lack introns. An HCBT gene has also been identified in a sequenced BAC containing a resistance gene cluster in wheat (Brooks et al. 2002) and from a sequenced marker linked to fusarium resistance in chickpea (Benko-Iseppon et al. 2003).

Following the HCBT gene clusters in both soybean BACs are genes with characteristic C3HC4 zinc-finger domains. On gmw1-74i13 the zinc-finger gene is tandemly duplicated and shares 95% nucleotide identity with one another. This tandem duplication is not found of the homeologous BAC. Zing-finger motifs are a large family of genes that are generally cystine rich and contain DNA-binding domains allowing them to function as transcription factors (Sun et al. 2001). Continuing in the same direction, a putative heat shock factor (HSF) follows in both soybean regions. Heat shock factors, much like zinc-finger proteins, are generally transcription factors associated with stress response pathways and as such contain DNA binding domains (Nover et al. 1998). Both heat shock genes show similarity by BLAST (1e\(^{-37}\)) to characterized HSF domains (pfam00447).

On both BACs, the next genes encode a protein with basic/helix-loop-helix (bHLH) domains. These proteins are part of a diverse superfamily of transcription factors involved in a wide range of cellular networks (Toledo-Ortiz et al. 2003). Following this gene on gmw1-74i13 is a gene highly similar to an expressed Arabidopsis thaliana protein with no known
function (At5g42920). BAC gmw1-52d3 also contains a colinear copy of this gene, but it is located after a remorin gene found only on gmw1-52d3 (Figure 1). The gmw1-52d3 copy is also located at the end of the BAC sequence resulting in a truncated product with only four exons (Table 1).

A putative remorin gene found only on gmw1-52d3 lies between the predicted bHLH and truncated *A. thaliana*-like expressed protein (Figure 2). This gene is the result of either gene loss in gmw1-74i13 or an insertion event in gmw1-52d3. This soybean remorin contains the characteristic coiled-coil domain on the C-terminal end of the protein as well as a conserved C-terminal tetrapeptide. This sequence is thought to be a potential site for isoprenylation, a modification associated with membrane interaction (Bariola *et al.* 2004). Previous research has shown that remorins form oligomer filaments similar in size to actins and in vivo these filaments/oligomers are membrane-associated. Based upon localization studies, remorins are thought to have a role in cytoskeleton and/or membrane skeleton (Bariola *et al.* 2004). Although most remorins are thought to have a proline rich N-terminal region, the gmw1-52d3 remorin has a reduced percent of proline residues relative to previously characterized proteins.

The gmw1-74i13 BAC sequence extends for another 30 Kb beyond the *A. thaliana*-like expressed protein and contains four more predicted genes (Table 1; Figure 2). The first gene encodes a 418 amino acid bZIP/ABA-responsive transcription factor. The translated protein sequenced showed similarity to a *Phaseolus vulgaris* bZIP (AAK39132) that was identified from ethylene-treated leaf abscission zones (Whiteclaw and Tucker, unpublished results).

The second gene on the extended region of gmw1-74i13 is a three exon defective chloroplasts and leaves protein (DCL). Previous work shows that DCL is required for chloroplast development, palisade cell morphogenesis and embryogenesis as well as rRNA maturation (Bellaoui and Gruissem 2004). Further, GFP localization studies showed that DCL is targeted to plastids suggesting this gene is a nuclear encoded chloroplast gene (Bellaoui and Gruissem 2004).

The third gene on the extended region of gmw1-74i13 is a 4 exon gene containing an aromatic rich protein family/oligoketide cyclase domain in the C-terminal region of the protein.
(Genbank CDD 5696 and 12219; pfam PF03364). Interestingly, this family of proteins is involved in antibiotic biosynthesis in bacteria. This catalytic domain surely has a different function in soybean, but still retains enough homology to be identifiable as an oligoketide cyclase domain.

The final gene on gmwl-74il3 is a truncated PPR-domain containing gene. These are a large family of genes in plants and numerous PPR genes have been implicated with fertility restoration activity and RNA stabilization (Bentolila et al. 2002). Much like DCL proteins, most PPR proteins are targeted to the plastids (Williams and Barkan 2003). All proteins in this family contain a pentatricopeptide (PPR) repeat approximately 35 amino acids long and is repeated numerous times across the peptide. The gmwl-74i13 PPR-like gene contains at least three characteristic PPR repeats, with the rest of the protein sequence being unknown.

Only the last three genes on gmwl-74i13 contained potential localization signals when analyzed with TargetP (Emanuelsson et al. 2000; Nielsen et al. 1997) and Predotar (Small et al. 2004). TargetP predicts the first 45 amino acids of DCL and the first 78 amino acids of the aromatic rich protein are localization signals to the chloroplast with scores of 0.732 and 0.840, respectively (Emanuelsson et al. 2000; Nielsen et al. 1997). Similarly, the PPR-domain contains an endoplasmic reticulum localization signal with a probability of 0.63 (Small et al. 2004).

**Analysis of repetitive elements:** Various methods for identifying probable transposons/retroelements were employed including similarity searches to the TIGR repeat database (www.tigr.org) and a self-BLAST of each BAC to identify long terminal repeats (LTRs). No full-length LTR-retrotransposon elements could be identified in either BAC sequence. However, three very degenerate polyprotein like sequences were found based upon BLAST similarity searches to the nr database (Figure 2). Two polyproteins were found on gmwl-74i13; the first just 3' of the WOX4-like gene; the second contained within the first intron of the first zinc finger protein (Figure 2). The last degenerate polyprotein was found just 3' of the heat shock transcription factor on BAC gmwl-52d3 (Figure 2). None of the degenerate polyproteins were conserved between the two BACs.

While LTR-retrotransposons were not observed, analysis of the sequences using FINDMITE yielded many potential miniature inverted transposable elements (MITEs; Tu
MITEs are commonly found in gene rich region (Feschotte et al. 2002). Although they are considered class 2 transposable elements, they do not encode a transposase, are generally no larger than 600 base pairs in length and the sequence between the terminal inverted repeats (TIRs) is not well conserved. Most classification of MITEs has been achieved by TIR and target site duplication (TSD) similarity (Feschotte et al. 2002). Determining the validity of the predictions of these elements is difficult due to the variability among elements and between the TIRs. In soybean, there have been no characterized MITEs to compare against again compounding the difficulty of identifying and verifying MITEs.

**Comparison of homeologous soybean regions:** When the two BACs are compared, the results are very striking in that nine genes shared between BACs gmw1-74i13 and gmw1-52d3 are conserved in both order and orientation. These results are very different from what has previously been observed in maize, a paleotetraploid (Illic et al. 2003; Langham et al. 2004; Ma et al. 2004), and are in fact more like those of cotton, a much more recent polyploid (Grover et al. 2004). There are, however, discernable differences between the homeologous soybean regions (Figure 2). BAC gmw1-52d3 contains one gene, remorin, which has either been lost in gmw1-74i13 or inserted into gmw1-52d3. Gmwl-74i13 contains an extra tandemly duplicated zinc-finger protein as well as three additional copies of HCBT. All full-length genes contain the same number of introns and exons. Gmwl-74i13 gene 5, the first HCBT gene, appears to be a fragmented copy with four exons whereas all other HCBT genes are single exon. A frameshift mutation is found in gmw1-52d3 gene 7, the last HCBT gene on this BAC, leading to a stop codon and truncation of the resulting protein. Additionally, the BAC-ends of gmw1-52d3 are each in the middle of a gene; consequently, the 5’ PABP gene on gmw1-52d3 is truncated and likely missing 8 exons and gene 14, an *Arabidopsis*-like expressed protein is truncated and missing 6 exons (Table 1; Figure 2).

When we consider the composition of the genes shared between the homeologous regions, we find that the total length (exons plus introns) of each retained gene is relatively similar. The average nucleotide identity between homeologous coding regions is 89.8%; and the resulting amino acid identity and similarity is 88% and 90.7%, respectively, not including
the HCBT genes. The HCBT genes themselves have an average nucleotide identity of 75.1%, markedly less than that of the other homeologs. This is discussed further below.

Figure 2 shows a VISTA based identity plot between the two regions. Not surprisingly, those regions corresponding to genic sequence are the most highly conserved with upwards of 95% nucleotide identity. This conservation can also be seen in the intronic regions with somewhat less identity. What is striking is the conservation of non-coding sequence (CNS) between the genes. Although some of the conservation of intergenic sequence may be due to promoter elements and transcription factor binding sites, there is more than anticipated. It would be expected that some of the promoter regions would have some sequence conservation if the genes are still similarity expressed. What is interesting as well is that just downstream of most of these genes, there is still a relatively high (greater than 75%) level of sequence retention. While there is extensive intergenic sequence conservation, the intergenic distance, however, is not as well conserved. It appears that gmwl-74i13 contains 16785 base pairs more DNA in the intergenic regions; that is 16.5% more noncoding DNA. The average intergenic distance for gmwl-74i13 is 8,486 base pairs and for gmwl-52d3 is 9,450.

The larger intergenic distance is also reflected in the total length of the homeologous overlapping regions; for gmwl-52d3 it is the full 98,760 bp and a corresponding 127,461 bp for gmwl-74i13. If we assume that the region of duplication was the result of an autopolyploid event, then the homeologs just after duplication would have been identical and allowed to accumulate mutations over approximately 14 million years (Schlueter et al. 2004). As a result, there has either been an expansion of the gmwl-74i13 region or a contraction of the gmwl-52d3 region. The prevalence of tandem duplications on gmwl-74i13 relative to gmwl-52d3 supports the hypothesis that the gmwl-74i13 region has undergone expansion.

Synonymous and nonsynonymous distance measures were calculated using PAML between the retained duplicate genes. The average synonymous distance of the non-tandemly duplicated retained homeologs (genes 2, 3, 10, 12 and 13) was 0.149 suggesting that this region was duplicated approximately 12.2 million years ago. The tandemly duplicated genes were not considered in this estimate because were probably tandemly duplicated after the polyploid event (or loss of a tandem duplication after polyploidy). Based upon this evidence,
these BACs likely represent homeologous segments that have been retained since the most recent (14 Mya) genome duplication in soybean (Schlueter et al. 2004).

**A proposed network of synteny with* Arabidopsis thaliana *and* Medicago truncatula:** Of major interest to the legume community is the ability to utilize the developing sequences from the model legumes, *Medicago truncatula* and *Lotus japonicus*, as well as from *Arabidopsis thaliana* for research in *Glycine max*. Annotated genes from both soybean BACs were used to search for syntenic blocks in *Medicago*, *Lotus* and *Arabidopsis* by BLAST, with default parameters, against the available sequence from *Medicago*, *Lotus* and *Arabidopsis*. No syntenic blocks were observed between soybean and *Lotus*. For each putatively syntenic gene from *Medicago* and *Arabidopsis*, the corresponding regions were scanned for fragmented blocks of synteny. Using this method allowed a network of synteny between the three species to be observed (Table 2, Figure 3).

The major syntenic blocks identified in *Medicago* correspond to two different BACs; AC148486 that maps to linkage group 1, and AC144431 that maps to linkage group 5 ([http://www.Medicago.org](http://www.Medicago.org)). The first BAC, AC148486, shares five genes colinear in order and orientation with gmwl-52d3 (Table 2, Figure 3). Four other genes that are not found in the syntenic soybean regions, however, interrupt the syntenic block. It is these genes that allowed the identification of the network of synteny to the second *Medicago* BAC. Gene 18, an S-ribonuclease binding protein and gene 19, lipase 3, while not found on the soybean BACs are conserved between AC148486 and AC144431 with the order of these genes flipped on AC144431. Beyond gene 20 on AC144431, there are not other genes that are syntenic between these BACs.

Three chromosomes in *Arabidopsis* were found to correspond to the two soybean BACs (Table 2; Figure 3). A region of chromosome one starting from At1g45180, a zinc-finger protein, to At1g45230, DCL protein, contains in total eleven genes, of which five are colinear in order and orientation with gmw1-52d3 and gmw1-74i13. Much as with the *Medicago* segments, the synteny is broken with other genes not found in the soybean sequences. The synteny between *Arabidopsis* and the soybean BACs appears to be stronger in this region than the synteny between *Arabidopsis* and *Medicago*. What is also interesting is that this *Arabidopsis* segment seems to combine characteristics from both regions,
suggesting either that the colinearity between the regions predates the soybean duplication or that the *Arabidopsis* segment is only truly colinear with gmw1-52d3 and that the region beyond gmw1-52d3 contains genes 25 and 26 as well.

The final two syntenic chromosomes in *Arabidopsis* contain only N-hydroxycinnamoyl/benzyoltransferase genes in clusters just as is observed in soybean (Figure 3). The first cluster on chromosome 5 consists of At5g07850, At5g07860, and At5g07870; further downstream on chromosome 5 is At5g67150 and At5g67160. On chromosome 3 is another cluster of three HCBT genes, At3g50270, At3g50280, and At3g50300. Although no other genes in the *Arabidopsis* chromosomes are syntenic in these regions, it is surprising to find the tandem duplications of these HCBT genes across species. Additionally in *Arabidopsis*, three other copies of HCBT were identified (At5g42830, At2g39980, and At5g01210) that were not found in tandemly amplified clusters. Similarly, in *Medicago* we identified one cluster of three HCBT genes on BAC AC122728 with a PABP gene showing identity to the gmw1-47i13 PABP gene (Figure 3). No other HCBT genes were identified from the available *Medicago* sequence.

**Tandem duplication of N-hydroxycinnamoyl/benzyoltransferase:** A characteristic of both gmw1-74i13 and gmw1-52d3 is a conserved cluster of N-hydroxycinnamoyl/benzyoltransferase (HCBT) genes. These genes show a number of characteristics that set them apart from the surrounding genes. First, most all of these genes are single exon genes; with the exception of gmw1-74i13 gene 4, HCBT 1, which contains 4 exons and is a fragmented copy. The gmw1-74i13 HCBT clustered genes are 60-80% identical at the nucleotide level and 58-88% similar at the amino acid level. Similarly, the gene cluster on gmw1-52d3 is 58-80% identical at the nucleotide level and 58-85% similar at the amino acid level (Table 3). Between gmw1-74i13 and gmw1-52d3 the nucleotide identity and amino acid similarity is slightly higher ranging from 58-92% and 58-96%, respectively. Interestingly, gmw1-74i13 gene 5 (HCBT 2) and gmw1-52d3 gene 6 (HCBT 2) are 96% similar at the amino acid level (Table 3).

When synonymous and nonsynonymous distances are calculated between all HCBT copies, an intriguing trend appears (Table 3, Figure 4). Relative to the other homeologous genes, most all of the HCBT genes seem to have larger synonymous and nonsynonymous
distances, both within BACs and between BACs. This suggests that the HCBT gene have evolved at a faster rate than the other genes in this region. Again, gmw1-74i13 gene 5 and gmw1-52d3 gene 6 differ from this trend and have much smaller synonymous and nonsynonymous distances as is seen on Figure 4.

RT-PCR analysis of retained homeologous genes: To better understand the functional evolution of these regions, 22 RT-PCR primer pairs were designed that differentiated between each retained homeolog (Table 4). Ten different tissue types were chosen to look at a variety of organs and developmental stages. Negative controls confirmed that the mRNA samples were free of genomic DNA contamination. Tubulin was used as a positive control to verify the integrity of each mRNA sample.

These results demonstrate that 20 of the 22 predicted homeologs are expressed. The first HCBT gene on gmw1-74i13 (gene 4) and the last HCBT gene on gmw1-52d3 (gene 7) showed no evidence of expression. Only gene 13, homeologous bHLH proteins and three HCBT genes show any evidence for differential expression between homeologs. Gene 13 on gmw1-74i13 shows no expression in unfurled unifoliate whereas the gmw1-52d3 copy is expressed in that tissue. The first HCBT gene on gmw1-52d3 (gene 4) and the second HCBT gene on gmw1-74i13 (gene 5) both appear to only be expressed in the below-ground portion of the plant. Gene 9 on gmw1-74i13 was detected in furled unifoliate, furled trifoliate, flowers and pods (Figure 4). All other HCBT genes appeared to be constitutively expressed in all tissues sampled. These results are surprising considering that these regions were duplicated approximately 14 mya (Schlueter et al. 2004). We anticipated finding more evidence for subfunctionalization or neofunctionalization. These results suggest that the differences in expression may be more quantitative than qualitative.

DISCUSSION

Stability of homeologous soybean regions: Prior to our findings, previous detailed analyses of homeologous regions in a paleopolyploid have been limited to maize (Illic et al. 2003; Langham et al. 2004; Ma et al. 2004). These combined results found that gene content is relatively unstable between homeologous regions and that reciprocal deletions have let to
the retention of only one copy in each homeologous region. This suggests that during
diploidization, natural selection worked such that only one gene copy was retained. Further
studies of BAC-end sequences in maize show that while tandemly amplified genes are
conserved, there is a surprising lack of retained homeologous genes suggesting that during
rediploidization maize has experienced significant gene loss through various means such as
mutation and transposition (Messing et al. 2004). In light of those studies, our results are
strikingly different. In fact, they more closely resemble those of homologs in cotton, a
relatively recent polyploid (Grover et al. 2004). We find an almost gene for gene retention
between two soybean BACs as well as fairly high nucleotide identity in noncoding regions
(Figure 2). These results confirm what was suspected nearly a decade ago and observed by
further hybridization and restriction mapping studies of homeologous regions; the soybean
genome has retained extensive conserved homeologous sequence (Shoemaker et al. 1996;
Marek et al. 2001; Yan et al. 2003; Foster-Hartnett et al. 2002; and Yan et al. 2004).

Both maize and soybean have been shown to be paleopolyploids with their last major
duplication approximately 11 million years and 14 million years ago, respectively (Schlueter
et al. 2004; Gaut and Doebley 1997). However, while the homeologous regions of maize
have diverged greatly during rediploidization, those in soybean remain relatively stable. The
evolutionary histories of maize and soybean need to be considered to explain these
differences. The maize genome is well documented to have experienced a large amplification
of transposable elements though to have shuffled the rediploidizing genome (SanMiguel and
Bennetzen 1998; Zhang and Peterson, 1999; Lai and Hannah 2005). Conversely, there has
been little to no evidence for a similar transposon explosion in the soybean genome. An
analysis of soybean sequence around a cyst nematode resistance gene showed that only 3%
of the predicted genes were transposable elements (Mudge et al. 2005). Similarly, the
apparent lack of identifiable recent transposable element insertions in either gmw1-74i13 or
gmw1-52d3 seems to support this explanation.

The apparent lack of recent LTR insertions as well as the relative gene density
suggests that these homeologous soybean regions are relatively gene rich. Mudge et al.
(2004) proposed that the soybean gene space may be limited to as little as 24% of the
genome based upon a Poisson distribution of RFLP probes to soybean BAC pools. Combined
with this, our results also suggest that gene rich regions of soybean may not be hotspots for recent retroelement insertions. This follows the results seen by Lin et al. (2005) using FISH that the gene space in soybean will be restricted to non-repetitive regions of the genome.

If we consider the annotations of the genes in the homeologous regions, a trend begins to appear. Four of the nine homeologs are thought to be transcription factors; WOX4, the zinc-finger genes, HSF, and bHelix-loop-helix protein. Additionally, the HCBT genes are implicated with a role in disease response signaling and the synthesis of phytoalexins. Blanc and Wolfe (2004b) found that in Arabidopsis duplicated genes encoding proteins involved in transcription or signal transduction are preferentially retained, while only one copy of genes involved in DNA repair are kept (more likely to be silenced). Our results in soybean support their hypothesis. Most of the genes identified in the homeologous BACs are in some way involved in either signal transduction or binding of DNA as a transcription factor. This could account for the greater than expected gene retention between these regions. Further, this suggests that there may be clustering of signaling genes or transcription factors within the soybean genome. While we could find no evidence of coexpression of these genes, it is possible that they may function in a response pathway.

Our results suggest that the soybean genome still retains much of its polyploid nature at the sequence level (Figure 2). While there may be a bias towards the types of genes retained in the homeologous BACs analyzed, these findings show that retention of duplicate genes in soybean is quite common between homeologous regions. Paleopolyploids generally are those species that have over millions of years undergone a switch from tetrasomic inheritance to disomic inheritance but still retain characteristics of a polyploid in the form of duplicate genes (Wolfe 2001). Wolfe suggested that within a single species, there may be a mixture of diploid and tetraploid loci and that this rediploidization probably does not happen simultaneously for all chromosomes (2001). The higher than expected retention of duplicate genes in homeologous regions suggests that in soybean, the process of diploidization is a slow and ongoing process, and that many duplicated regions are not just retained but retained at a high level.

Sequence retention in these soybean homeologous regions is not only seen in the coding regions, but in the noncoding regions as well. Following a duplication event, genes
may undergo several fates; retained function, subfunctionalization, neofunctionalization, or silencing (Force et al. 1999). After the period of relaxed selection, those duplicates that have survived the birth and death process of duplication are under purifying negative selection and mutations in the coding regions tend to be silent (Kondrashov et al. 2002; Schlueter et al. 2004). However, many of the mutations that lead to duplicate gene retention and functional changes are not within the coding region, but upstream in regulatory sequences (Force et al. 1999). When duplicated genes or even homologs can be considered, it is possible to identify conserved non-coding sequences (CNSs) that are potential transcription factor binding sites (Lockton and Gaut 2005). When we consider the elevated sequence conservation between the soybean homeologs (Figure 2) we find that while there have been some changes in the non-coding regions of these BACs, much of the non-coding sequence is retained. The ratio of synonymous to nonsynonymous distances for all of the homeologs shows that while some genes may be evolving faster, all genes are under negative/purifying selection (Figure 4). In other words, there is some selective constraint that is retaining both copies of homeologs.

**Plastid insertions into genomic sequence:** The last two genes of BAC gmw1-74i13, DCL (gene 17) and the oligo ketide gene (gene 18), seem to encode localization signals at the N-terminal of the resulting protein that target them to the chloroplast. Likely, these two genes have been transferred from the chloroplast genome to the nuclear genome. Conversely, these genes may have evolved to encode a domain that targets the product to the chloroplast, although the first scenario seems the most likely. While the corresponding homeologous sequences are not yet available, it would be interesting to see if this chloroplast genome transfer is conserved between regions and if it occurred before or after the most recent polyploid event. Similarly, the truncated PPR-like gene on the end of BAC gmw1-74i13 (gene 19) also appears to have a localization signal targeting the endoplasmic reticulum. While most PPR proteins are thought to have a role in mitochondrial signaling, this copy seems to be targeted elsewhere. Perhaps, in the process of diploidization this copy has assumed a new function.

**Synteny among the legumes:** Comparisons of the soybean homeologous regions in soybean to *Medicago* and *Arabidopsis* suggest that synteny between these species may not be well conserved (Table 3; Figure 3). No synteny with the available *Lotus japonicus* sequences
could be identified. This may simply be due to the ongoing process of genome sequencing and the regions of interest simply not being available yet, or the syntenic blocks may not be conserved enough to be detected. Although many of the genes that are conserved between species have the same order and orientation, the synteny is fragmented between all three species. It is interesting that even between *Medicago* and *Arabidopsis* in these regions, the gene conservation is still fragmented. This suggests that there has been extensive divergence in genome structure and sequence after divergence of these species. These findings are contradictory to those of Mudge *et al.* (2005) who found a region of soybean that showed hypersynteny to both *Medicago* and *Arabidopsis*. While microsynteny may be used to identify orthologous regions from one legume species to another, based upon our findings, regions of microsynteny may not always be found. This observation will likely confound the use of *Medicago* and a model species for identification of orthologous genes in soybean. Some cases may be straightforward as was seen by Mudge *et al.* (2005), but this will not always be the case as our observations of synteny have shown.

**Tandemly duplicated HCBT genes:** Both segmental and tandem duplications have been shown in plants to have a role in the evolution of large gene families such as the GRAS gene family in rice and *Arabidopsis* and the NIN-like proteins in *Arabidopsis*, rice and *Lotus japonius* (Tian *et al.* 2004; Schauser *et al.* 2005). In soybean alone, the chalcone synthase gene has been shown to exist in four copies near the I locus (Clough *et al.* 2003). We identified a tandemly amplified cluster of HCBT genes in both homeologous soybean BACs (Figure 2). Each BAC appeared to contain one HCBT gene copy that was either truncated, as gene 7 on gmwl-52d3, or fragmented, as gene 4 on gmwl-74i13, suggesting these copies many no longer encode a functional protein (Table 1).

In comparison to the synonymous and nonsynonymous distance measures between other homeologs, the HCBT genes have accumulated more mutations, both silent and non-silent (Table 3, Figure 4). These larger HCBT synonymous and nonsynonymous distance measures both between BACs and within BACs suggest that these tandemly amplified genes are for some reason undergoing more rapid evolution that the surrounding genes. When synonymous distance versus nonsynonymous distance is plotted, this finding is even more apparent. Although all of the HCBT genes appear to be under negative or purifying selection,
there appears to be a slight relaxation in this pressure that has allowed these genes to mutate more than the surrounding homeologs. Perhaps, this is a result of the proposed function of HCBT in disease response (Yang et al. 1997; Brooks et al. 2002; Benko-Iseppon et al. 2003). Previous work has shown that genes involved in disease response can have regions of a gene that are either under positive selection or have accumulated a number of mutations allowing plasticity in response to various pathogens (Graham et al. 2002). By being tandemly amplified, the sheer number of these genes allows more mutations to accumulated in both the coding and upstream promoter regions allowing for a broader response to pathogen attack.

Further, previous studies of HCBT proteins have suggested that conserved cystine residues may allow the formation of disulfide bridges thereby allowing for the formation of dimers (Yang et al. 1997). By having multiple combinations of dimers, this also would permit a broader pathogen response.

Interestingly, in *Medicago truncatula* and *Arabidopsis thaliana* HCBT genes also appeared to be tandemly amplified in all three genomes. In *Medicago*, only three copies of HCBT were identified and clustered on one BAC, AC122728. Similarly, in *Arabidopsis*, three clusters of HCBT genes could be identified; two independent clusters of three genes and two genes respectively on chromosome 5 (At5g07850, At5g07860, At5g07870, At5g67150, At5g67160) and one cluster of three genes on chromosome 3 (At3g50270, At3g280, At3g50300). It should be noted that virtually none of the surrounding genes in either *Medicago* or *Arabidopsis* showed any synteny to the soybean sequences.

Conserved tandem duplications of HCBT genes in all three species suggest that there is a biologically relevant function for these clusters. This raised the question whether these clusters were formed independently in each species; in other words, if these tandem duplications occurred prior to or post species divergence. A parsimony based phylogenetic tree with these clustered HCBT sequences suggests that the tandem duplication of these genes likely occurred after speciation. Figure 6 shows that genes from one species almost always form single clades, indicative of independent duplication events. One except to these findings is gmwl-74i13 gene 8 (HCBT 5) that falls basal to both the soybean groups and *Medicago* group. This is likely due to the wide divergence in sequence observed both within a species and between species. Further analysis revealed that even in rice, clusters of these
tandem duplicated HCBT genes can be identified (results not show). Although the reason for these independent tandem amplifications have occurred, we can speculate that it may again be due to the role of HCBT genes in disease response pathways.

**Homeologous gene expression:** Finally, we wished to better understand the higher than expected levels of gene conservation in these soybean BACs by investigating the expression of retained homeologs. Through RT-PCR screens (Figure 5) the changes in expression across a variety of tissues from one region of the genome versus the other was studied. Although we expected to find evidence for subfunctionalization between homeologs, we found that expression differences were quite limited. Only gene 17 (bHLH) and three HCBT genes show any major evidence for differences in expression between homeologs. These findings paired with the elevated sequence retention in the non-coding regions between homeologs suggests retention transcription factor binding sites and thereby retention of expression patterns.

The duplication-degeneration-complementation model (DDC) suggested by Force *et al.* proposed that the retention of duplicated genes after polyploidy is the result of changes allowing either new gene function or compartmentalized gene function (1999). Although most of our genes don’t show evidence for subfunctionalization, the HCBT genes themselves seem to fit with this model. It is also possible that many of the changes were not observed because they are more quantitative than qualitative. An example of this is gmw1-74i13 HCBT3 (gene 6; Figure 5). Even though this gene is expressed in all tissues sampled, there are intensity differences in the bands that are observed between biological replicates, and different between HCBT gene copies. Additionally, the tissues we sampled may not be representative of those showing changes in expression between homeologs. Of particular interest is the putative root specific expression of gmw1-74i13 HCBT 1 (gene 5) and gmw1-52d3 HCBT 2 (gene 5; Figure 5). Almost all other HCBT genes have expression in all tissues sampled, except gmw1-74i13 HCBT 6 (gene 9). This suggests that either the two homeologs with root-specific expression have independently developed this expression through the loss of regulatory elements, or that the other genes have become more broadly expressed relative to these copies.
This study provides us with our first glimpse at a higher than expected genic and inter-genic sequence conservation in the paleopolyploid soybean genome. Not only was sequence conservation higher than expected, this was carried over into the limited expression differences observed between homeologs. Although the soybean homeologs are very similar, only a limited network of synteny could be identified between soybean, *Medicago* and *Arabidopsis*. Further studies of homeologous regions in soybean are warranted to further understand the evolutionary history of this paleopolyploid genome.

Acknowledgments: The authors would like to thank Mary Duke and Xiaofen Liu for their sequencing expertise. We would also like to thank Corrinne Grover for her knowledge and help annotating genes and identifying transposable elements and Jody Hayes for her skills in mapping SSR markers. Names are necessary to report factually on the available data; however, the USDA neither guarantees nor warrants the standard of the product, and the use of the name by the USDA implies no approval of the product to the exclusion of others that may also be suitable. This article is a contribution of the Corn Insect and Crop Genetics Research Unit (USDA-ARS). J.A. Schlueter was supported by a grant from the United Soybean Board.

REFERENCES:


Brudno, M., Do, C., Cooper, G., Kin, M. F., Davydov, E., et al., 2003 LAGAN and Multi-LAGAN: Efficient tools for large scale multiple alignment of genomic DNA. Genome Res 13:721-731


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


Table 1. Predicted gene features of homoeologous BACs gmw1-74i13 and gmw1-52d3

<table>
<thead>
<tr>
<th>Gene Putative function</th>
<th>Total Length&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Identity Exons</th>
<th>Number Exons</th>
<th>Length exons</th>
<th>Number Introns</th>
<th>Length introns</th>
<th>Length&lt;sup&gt;b&lt;/sup&gt; aa</th>
<th>Identity aa</th>
<th>Similarity aa</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 WOX4 protein (homeobox-leucine zipper transcription factor protein)</td>
<td>1016</td>
<td>-</td>
<td>3</td>
<td>567</td>
<td>-</td>
<td>449</td>
<td>189</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>2 Poly(A)-binding protein</td>
<td>4128 225</td>
<td>96.0</td>
<td>9</td>
<td>1 1893</td>
<td>8</td>
<td>-2235</td>
<td>630 75</td>
<td>97.3</td>
<td>98.6</td>
</tr>
<tr>
<td>3 Membrane-like protein</td>
<td>4104 4191</td>
<td>96.9</td>
<td>12</td>
<td>12 14221476</td>
<td>11</td>
<td>11 26822715</td>
<td>474 492</td>
<td>97.5</td>
<td>97.7</td>
</tr>
<tr>
<td>4 N-hydroxycinnamoyl/benzoyl transferase-like 1</td>
<td>1226</td>
<td>*&lt;sup&gt;c&lt;/sup&gt;</td>
<td>4</td>
<td>807</td>
<td>3</td>
<td>419</td>
<td>269</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>5 N-hydroxycinnamoyl/benzoyl transferase-like 2</td>
<td>1383 1368</td>
<td>*</td>
<td>1</td>
<td>1 1383 1368</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>461 456</td>
<td>*</td>
</tr>
<tr>
<td>6 N-hydroxycinnamoyl/benzoyl transferase-like 3</td>
<td>1461 1374</td>
<td>*</td>
<td>1</td>
<td>1 1461 1374</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>487 458</td>
<td>*</td>
</tr>
<tr>
<td>7 N-hydroxycinnamoyl/benzoyl transferase-like 4</td>
<td>1410 483</td>
<td>*</td>
<td>1</td>
<td>1 1410 483</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>470 161</td>
<td>*</td>
</tr>
<tr>
<td>8 N-hydroxycinnamoyl/benzoyl transferase-like 5</td>
<td>1398</td>
<td>-</td>
<td>*</td>
<td>1 -1398</td>
<td>-</td>
<td>0</td>
<td>-</td>
<td>466</td>
<td>*</td>
</tr>
<tr>
<td>9 N-hydroxycinnamoyl/benzoyl transferase-like 6</td>
<td>1386</td>
<td>-</td>
<td>*</td>
<td>1 -1386</td>
<td>-</td>
<td>0</td>
<td>-</td>
<td>462</td>
<td>*</td>
</tr>
<tr>
<td>10 Zinc finger (C3HC4-type) 1</td>
<td>8157</td>
<td>-</td>
<td>92.5</td>
<td>4 -2073</td>
<td>-</td>
<td>3</td>
<td>-6084 -691</td>
<td>88.3</td>
<td>90.8</td>
</tr>
<tr>
<td>11 Zinc finger (C3HC4-type) 2</td>
<td>7685 6015</td>
<td>92.3</td>
<td>5</td>
<td>5 2043 2043</td>
<td>4</td>
<td>4</td>
<td>56423972 681 708</td>
<td>88.5</td>
<td>90.9</td>
</tr>
<tr>
<td>12 Heat shock transcription factor 31</td>
<td>1190 1174</td>
<td>95.3</td>
<td>2</td>
<td>2 1092 1092</td>
<td>1</td>
<td>1</td>
<td>98 82</td>
<td>364 364</td>
<td>94.5</td>
</tr>
<tr>
<td>13 bHelix-loop-helix transcription factor</td>
<td>795 777</td>
<td>80.7</td>
<td>1</td>
<td>1 795 777</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>264 259</td>
<td>77.9</td>
</tr>
</tbody>
</table>
Table 1. (continued)

<table>
<thead>
<tr>
<th></th>
<th>Total Length</th>
<th>Identity Exons</th>
<th>Number Exons</th>
<th>Length Exons</th>
<th>Length Exons</th>
<th>Length Intron</th>
<th>Number Intron</th>
<th>Length Introns</th>
<th>Length Introns</th>
<th>Length aa</th>
<th>Identity aa</th>
<th>Similarity aa</th>
</tr>
</thead>
<tbody>
<tr>
<td>14</td>
<td>Remorin-like protein</td>
<td>-3885</td>
<td>-</td>
<td>7</td>
<td>1581</td>
<td>6</td>
<td>2304</td>
<td>527</td>
<td>7</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>Arabidopsis-like expressed protein</td>
<td>4684 3065</td>
<td>96.7</td>
<td>8</td>
<td>2406 1452</td>
<td>7</td>
<td>2278 1613</td>
<td>802 484</td>
<td>96.2</td>
<td>98.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>bzip transcription factor</td>
<td>3757</td>
<td>-</td>
<td>4</td>
<td>1254</td>
<td>3</td>
<td>2503</td>
<td>418</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>DCL protein</td>
<td>4452</td>
<td>-</td>
<td>3</td>
<td>636</td>
<td>2</td>
<td>3816</td>
<td>212</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>Aromatic-rich family protein</td>
<td>4088</td>
<td>-</td>
<td>4</td>
<td>756</td>
<td>3</td>
<td>3332</td>
<td>252</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>PPR-repeat containing protein</td>
<td>2092</td>
<td>-</td>
<td>4</td>
<td>1749</td>
<td>3</td>
<td>343</td>
<td>583</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Average</td>
<td>92.9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>91.5</td>
<td>93.7</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*a* Total length based upon nucleotide exons plus introns from translation start to stop.

*b* Amino acid, aa

*c* See Table 4 for identity and similarity information for the HCBT genes, *
Table 2. Syntenic genes between soybean, *Medicago* and *Arabidopsis*

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>WOX4 protein (homeobox-leucine zipper transcription factor protein)</td>
<td>present</td>
<td>present</td>
<td>present</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Poly(A)-binding protein</td>
<td>present</td>
<td>present</td>
<td>present</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Membrane-like protein</td>
<td>present</td>
<td>present</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>N-hydroxycinnamoyl/benzoyl transferase-like 1</td>
<td>present</td>
<td>present</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>N-hydroxycinnamoyl/benzoyl transferase-like 2</td>
<td>present</td>
<td>present</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>N-hydroxycinnamoyl/benzoyl transferase-like 3</td>
<td>present</td>
<td>present</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>N-hydroxycinnamoyl/benzoyl transferase-like 4</td>
<td>present</td>
<td>present</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>N-hydroxycinnamoyl/benzoyl transferase-like 5</td>
<td>present</td>
<td>present</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>N-hydroxycinnamoyl/benzoyl transferase-like 6</td>
<td>present</td>
<td>present</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Unknown protein</td>
<td>present</td>
<td>present</td>
<td>present</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>Zinc finger (C3HC4-type) 1</td>
<td>present</td>
<td>present</td>
<td>present</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>Zinc finger (C3HC4-type) 2</td>
<td>present</td>
<td>present</td>
<td>present</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>Heat shock transcription factor 31</td>
<td>present</td>
<td>present</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>Arabidopsis expressed protein</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>Glycosyl hydrolase NADH dehydrogenase subunit</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>fragment</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>bHelix-loop-helix transcription factor</td>
<td>present</td>
<td>present</td>
<td>present</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>S-ribonuclease binding protein 1</td>
<td>present</td>
<td>present</td>
<td>present</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>Lipase 3</td>
<td>present</td>
<td>present</td>
<td>present</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>Remorin-like protein</td>
<td>present</td>
<td>present</td>
<td>present</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>DC1 domain-containing protein</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>WW domain-containing protein</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>Arabidopsis-like expressed protein</td>
<td>present</td>
<td>present</td>
<td>present</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>Arabidopsis hypothetical protein</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>bzip transcription factor</td>
<td>present</td>
<td>present</td>
<td>present</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>DCL protein</td>
<td>present</td>
<td>present</td>
<td>present</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>Aromatic-rich family protein</td>
<td>present</td>
<td>present</td>
<td>present</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>PPR-repeat containing protein</td>
<td>present</td>
<td>present</td>
<td>present</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 3. Identity, similarity, synonymous and nonsynonymous distances of N-hydroxycinnamoyl/benzoyl transferase (HCBT) genes from gmw1-52d3 and gmw1-74i13

<table>
<thead>
<tr>
<th></th>
<th>52d3-5</th>
<th>52d3-6</th>
<th>52d3-7</th>
<th>74i13-4</th>
<th>74i13-5</th>
<th>74i13-6</th>
<th>74i13-7</th>
<th>74i13-8</th>
<th>74i13-9</th>
</tr>
</thead>
<tbody>
<tr>
<td>52d3-5</td>
<td>-</td>
<td>80.3a</td>
<td>78.3b</td>
<td>59.1</td>
<td>53.0</td>
<td>63.8</td>
<td>80.3</td>
<td>80.3</td>
<td>83.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>85.7c</td>
<td></td>
<td>59.1</td>
<td>55.4</td>
<td>61.2</td>
<td>85.3</td>
<td>79.5</td>
<td>80.5</td>
</tr>
<tr>
<td>52d3-6</td>
<td>0.134c</td>
<td>0.168</td>
<td>0.149</td>
<td>58.2</td>
<td>52.4</td>
<td>65.2</td>
<td>91.4</td>
<td>82.3</td>
<td>83.1</td>
</tr>
<tr>
<td></td>
<td>0.461e</td>
<td>0.523</td>
<td>0.466</td>
<td>58.4</td>
<td>52.6</td>
<td>56.5</td>
<td>91.1</td>
<td>79.4</td>
<td>78.5</td>
</tr>
<tr>
<td>52d3-7</td>
<td>0.152</td>
<td>0.151</td>
<td>0.136</td>
<td>62.9</td>
<td>52.6</td>
<td>60.8</td>
<td>59.0</td>
<td>65.0</td>
<td>58.1</td>
</tr>
<tr>
<td></td>
<td>0.603</td>
<td>0.510</td>
<td>0.362</td>
<td>59.6</td>
<td>58.7</td>
<td>52.6</td>
<td>50.9</td>
<td>59.7</td>
<td>51.1</td>
</tr>
<tr>
<td>74i13-4</td>
<td>0.122</td>
<td>0.044</td>
<td>0.137</td>
<td>0.137</td>
<td>0.303</td>
<td>0.044</td>
<td>0.244</td>
<td>0.643</td>
<td>0.817</td>
</tr>
<tr>
<td></td>
<td>0.449</td>
<td>0.218</td>
<td>0.428</td>
<td>0.428</td>
<td>0.384</td>
<td>0.218</td>
<td>0.449</td>
<td>0.643</td>
<td>0.817</td>
</tr>
<tr>
<td>74i13-5</td>
<td>0.123</td>
<td>0.152</td>
<td>0.184</td>
<td>0.184</td>
<td>0.303</td>
<td>0.218</td>
<td>0.449</td>
<td>0.643</td>
<td>0.817</td>
</tr>
<tr>
<td></td>
<td>0.522</td>
<td>0.374</td>
<td>0.469</td>
<td>0.469</td>
<td>0.303</td>
<td>0.218</td>
<td>0.449</td>
<td>0.643</td>
<td>0.817</td>
</tr>
<tr>
<td>74i13-6</td>
<td>0.104</td>
<td>0.128</td>
<td>0.095</td>
<td>0.095</td>
<td>0.119</td>
<td>0.244</td>
<td>0.542</td>
<td>0.432</td>
<td>0.357</td>
</tr>
<tr>
<td></td>
<td>0.396</td>
<td>0.362</td>
<td>0.244</td>
<td>0.244</td>
<td>0.119</td>
<td>0.357</td>
<td>0.396</td>
<td>0.542</td>
<td>0.432</td>
</tr>
<tr>
<td>74i13-7</td>
<td>0.143</td>
<td>0.188</td>
<td>0.204</td>
<td>0.204</td>
<td>0.124</td>
<td>0.357</td>
<td>0.395</td>
<td>0.432</td>
<td>0.357</td>
</tr>
<tr>
<td></td>
<td>0.516</td>
<td>0.437</td>
<td>0.530</td>
<td>0.530</td>
<td>0.124</td>
<td>0.357</td>
<td>0.395</td>
<td>0.432</td>
<td>0.357</td>
</tr>
<tr>
<td>74i13-8</td>
<td>0.119</td>
<td>0.143</td>
<td>0.170</td>
<td>0.170</td>
<td>0.129</td>
<td>0.470</td>
<td>0.470</td>
<td>0.545</td>
<td>0.470</td>
</tr>
<tr>
<td></td>
<td>0.409</td>
<td>0.415</td>
<td>0.450</td>
<td>0.450</td>
<td>0.129</td>
<td>0.470</td>
<td>0.470</td>
<td>0.545</td>
<td>0.470</td>
</tr>
<tr>
<td>74i13-9</td>
<td>0.119</td>
<td>0.143</td>
<td>0.170</td>
<td>0.170</td>
<td>0.129</td>
<td>0.470</td>
<td>0.470</td>
<td>0.545</td>
<td>0.470</td>
</tr>
<tr>
<td></td>
<td>0.409</td>
<td>0.415</td>
<td>0.450</td>
<td>0.450</td>
<td>0.129</td>
<td>0.470</td>
<td>0.470</td>
<td>0.545</td>
<td>0.470</td>
</tr>
</tbody>
</table>

a Percent nucleotide identity, grey shaded boxes
b Percent protein identity, grey shaded boxes
c Percent protein similarity, grey shaded boxes
d Nonsynonymous distance, unshaded boxes
e Synonymous distance, unshaded boxes
Table 4. Soybean homeolog specific primers for RT-PCR

<table>
<thead>
<tr>
<th>Primer name</th>
<th>Primer sequence (5'-3')</th>
<th>Primer Length&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Primer Location&lt;sup&gt;b&lt;/sup&gt;</th>
<th>Size&lt;sup&gt;c&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>gmwl-52d3</td>
<td>GAT AAG CTG GAG CAT GAA AC</td>
<td>20</td>
<td>68-86</td>
<td>152</td>
</tr>
<tr>
<td>PABP-U</td>
<td>TGC TGC TGT GTA ACT TTT CTC</td>
<td>21</td>
<td>198-218</td>
<td></td>
</tr>
<tr>
<td>PABP-L</td>
<td>TAG ATA AGC TGG AGC ATG AAG</td>
<td>21</td>
<td>13418-13438</td>
<td>154</td>
</tr>
<tr>
<td>gmwl-74i13</td>
<td>TGC TGC TGT GTA ACT TTC TTC</td>
<td>21</td>
<td>13551-13571</td>
<td></td>
</tr>
<tr>
<td>PABP-U</td>
<td>AAA GAA GAC TTC AAT AAC AGC A</td>
<td>22</td>
<td>8212-8233</td>
<td>361</td>
</tr>
<tr>
<td>PABP-L</td>
<td>AAG GGC TGT GGA AGA CTT AG</td>
<td>20</td>
<td>7756-7775</td>
<td></td>
</tr>
<tr>
<td>memb-U</td>
<td>AGT GAG AAA GAA GAC TTC AAG A</td>
<td>22</td>
<td>23124-23145</td>
<td>363</td>
</tr>
<tr>
<td>gmwl-52d3</td>
<td>AGG GCT GTG GAA GAC TTG G</td>
<td>19</td>
<td>22668-22686</td>
<td></td>
</tr>
<tr>
<td>memb-L</td>
<td>AGA GTC AAA CCA AAT ATG CAA C</td>
<td>22</td>
<td>14270-14291</td>
<td>416</td>
</tr>
<tr>
<td>gmwl-74i13</td>
<td>TGA CTT GGA TGG ACA ACA ATG</td>
<td>21</td>
<td>14665-14685</td>
<td></td>
</tr>
<tr>
<td>memb-U</td>
<td>AGG TTG TTC CAT GAA CCA CTC</td>
<td>21</td>
<td>24132-24152</td>
<td>485</td>
</tr>
<tr>
<td>gmwl-74i13</td>
<td>TTG TTT CTG CAC TCA CTC TAT TA</td>
<td>23</td>
<td>24594-24616</td>
<td></td>
</tr>
<tr>
<td>HCBT1-U</td>
<td>GCC TTC TCT TCA AGA AGC CAG</td>
<td>21</td>
<td>26010-26030</td>
<td>298</td>
</tr>
<tr>
<td>HCBT1-L</td>
<td>CAA CGA TTG TCG ATG GGC AC</td>
<td>20</td>
<td>26240-26439</td>
<td></td>
</tr>
<tr>
<td>gmwl-74i13</td>
<td>CAT GAT GGT CAC ACC ATG CTC</td>
<td>21</td>
<td>26482-26502</td>
<td>357</td>
</tr>
<tr>
<td>HCBT3-U</td>
<td>CGT GGC TCC ACC ACT CAC TG</td>
<td>20</td>
<td>27006-27025</td>
<td></td>
</tr>
<tr>
<td>HCBT2-U</td>
<td>TAT CCT TTG GCT GGC CGA T</td>
<td>19</td>
<td>33885-33903</td>
<td>376</td>
</tr>
<tr>
<td>gmwl-74i13</td>
<td>AAA TGG GAA CAT CAG TTT CGT</td>
<td>21</td>
<td>34240-34260</td>
<td></td>
</tr>
</tbody>
</table>

<sup>a</sup> Length of primer in base pairs

<sup>b</sup> Location of primer within BAC sequence, relative to 5' end of BAC

<sup>c</sup> Size of PCR amplicon
<table>
<thead>
<tr>
<th>Primer name</th>
<th>Primer sequence (5'-3')</th>
<th>Primer Length</th>
<th>Primer Location</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>gmwl-74i13</td>
<td>AAC CCA AGA CCC TCC CTC G</td>
<td>19</td>
<td>39153-39171</td>
<td>509</td>
</tr>
<tr>
<td>HCBT3-U</td>
<td>CTT TTG CTT TCA GTT TTG CTA TG</td>
<td>23</td>
<td>39639-39661</td>
<td></td>
</tr>
<tr>
<td>HCBT3-L</td>
<td>ATG ACC CTC CCT CTT ATT C</td>
<td>19</td>
<td>41083-41101</td>
<td>319</td>
</tr>
<tr>
<td>HCBT4-U</td>
<td>AGA GCA TGA GAT TGA GAT TGA</td>
<td>21</td>
<td>41381-41401</td>
<td></td>
</tr>
<tr>
<td>HCBT4-L</td>
<td>TGC TAC TGT TCC CAT TTC ACG</td>
<td>18</td>
<td>47929-47949</td>
<td>478</td>
</tr>
<tr>
<td>HCBT5-U</td>
<td>CGC TTT GTC GTT TTG GTT CG</td>
<td>20</td>
<td>48387-48406</td>
<td></td>
</tr>
<tr>
<td>HCBT5-L</td>
<td>ACA TGA TGT TCC CAT TTC GC</td>
<td>20</td>
<td>55798-55817</td>
<td>408</td>
</tr>
<tr>
<td>HCBT6-U</td>
<td>TCA AGC AAT TTC CCT GAG GT</td>
<td>20</td>
<td>56186-56205</td>
<td></td>
</tr>
<tr>
<td>gmwl-52d3</td>
<td>GTT GGA GAG GCA AGG TGA T</td>
<td>19</td>
<td>39339-39357</td>
<td>477</td>
</tr>
<tr>
<td>znfngr-U</td>
<td>TGT TTC AAG TGT TTC AAT ACA G</td>
<td>22</td>
<td>43454-43475</td>
<td></td>
</tr>
<tr>
<td>gmwl-52d3</td>
<td>GAG CAT CAA TCA TTT CTT TCT A</td>
<td>22</td>
<td>76441-76462</td>
<td>336</td>
</tr>
<tr>
<td>znfngr-L</td>
<td>ACA AAT GGG ACA CAG ATT CTT A</td>
<td>22</td>
<td>76973-76994</td>
<td></td>
</tr>
<tr>
<td>gmwl-74i13</td>
<td>AAA CTC TTG AAA CAG AGA AAG C</td>
<td>22</td>
<td>93005-93026</td>
<td>181</td>
</tr>
<tr>
<td>znfngr1-U</td>
<td>AAA TGG GAC ACA GAT TCT TTT</td>
<td>21</td>
<td>93428-93268</td>
<td></td>
</tr>
<tr>
<td>gmwl-74i13</td>
<td>AAC ATA CCA ACT GGT GGA TGA T</td>
<td>22</td>
<td>62292-62313</td>
<td>483</td>
</tr>
<tr>
<td>znfngr2-U</td>
<td>GAA GTG TTG TAG TTG GCT GC</td>
<td>20</td>
<td>62837-62856</td>
<td></td>
</tr>
<tr>
<td>gmwl-74i13</td>
<td>AGA CAT ACC AAC TGT TGG ACG AA</td>
<td>23</td>
<td>109474-109496</td>
<td>484</td>
</tr>
<tr>
<td>HSF-U</td>
<td>ACG GAA GTG TTG TAG TTG GC</td>
<td>20</td>
<td>110036-110055</td>
<td></td>
</tr>
<tr>
<td>gmwl-52d3</td>
<td>ACT TGC CAG CAT CCA CTA C</td>
<td>19</td>
<td>81566-81584</td>
<td>520</td>
</tr>
<tr>
<td>bHLH-U</td>
<td>GGA GCT TGG TAA GGT TTA CTA</td>
<td>21</td>
<td>82065-82085</td>
<td></td>
</tr>
<tr>
<td>Primer name</td>
<td>Primer sequence (5'-3')</td>
<td>Primer Length(^a)</td>
<td>Primer Location(^b)</td>
<td>Size(^c)</td>
</tr>
<tr>
<td>---------------</td>
<td>------------------------</td>
<td>---------------------</td>
<td>----------------------</td>
<td>------------</td>
</tr>
<tr>
<td>gmwl-74i13 bHLH-U</td>
<td>CAT AGC CCG AGT GAA ACA T</td>
<td>19</td>
<td>126213-126231</td>
<td>495</td>
</tr>
<tr>
<td>gmwl-74i13 bHLH-L</td>
<td>CAT ATT CTC CAT GTG GCG A</td>
<td>19</td>
<td>126689-126707</td>
<td></td>
</tr>
<tr>
<td>gmwl-52d3 exp-U</td>
<td>ACA GCA AGG CTT CCG TGG A</td>
<td>19</td>
<td>95443-95461</td>
<td>371</td>
</tr>
<tr>
<td>gmwl-52d3 exp-L</td>
<td>GGC ATC GTT CGA CAA GAC G</td>
<td>19</td>
<td>95871-95889</td>
<td></td>
</tr>
<tr>
<td>gmwl-74i13 exp-U</td>
<td>GCC TCC GTA CAG AAC ATT G</td>
<td>19</td>
<td>137566-137584</td>
<td>361</td>
</tr>
<tr>
<td>gmwl-74i13 exp-L</td>
<td>CGT CGT TGG ACA AGA AGC</td>
<td>18</td>
<td>137984-138001</td>
<td></td>
</tr>
</tbody>
</table>
Figure 1. Map position of BACs gmw1-74i13 and gmw1-52de3 on the *Glycine max* A81-356022 X *Glycine soja* PI 468.916 linkage groups C1 and C2 (Diers *et al.* 1992). Mapping of the BACs was based upon SSRs found in each BAC sequence. Lines show homeologous relationships between these linkage groups.
Figure 2. Gene positions on homeologous soybean BACs gmw1-74i13 and gmw1-52d3. Each colored block arrow on the line represents a gene and grey boxes between genes show homeologs. All gene numbering corresponds to Table 1. The plots above and below the gene locations are VISTA plots showing the relative nucleotide identity between the two BACs. The light purple boxes on top of the VISTA plots correspond to exon positions.
Figure 3. Diagrammatic representation of a network of synteny between two soybean homeologous regions, two *Medicago truncatula* BACs and *Arabidopsis thaliana*. Each colored arrow-block represents a gene as numbered in Tables 1 and 3. Grey bars between genes show the syntenic genes. The scale is in kilobases.
Figure 4. A plot of synonymous distance versus nonsynonymous distance for all soybean homeologous gene pairs. All red dots correspond to N-hydroxycinnamoyl/benzoyltransferase (HCBT) gene alignments while blue dots are for all other homeologs between soybean BACs gmw1-74i13 and gmw1-52d3.
A

B

gm1-l-52d3-bHLH (gene 17)
gm1-l-74i13-bHLH (gene 17)
gm1-l-52d3-HCBT1 (gene 5)
gm1-l-74i13-HCBT2 (gene 5)
gm1-l-74i13-HCBT3 (gene 6)
gm1-l-74i13-HCBT6 (gene 9)

C

gm1-l-52d3-PABP (gene 2)
gm1-l-74i13-PABP (gene 2)
gm1-l-52d3-membrane (gene 3)
gm1-l-74i13-membrane (gene 3)
gm1-l-52d3-HCBT1 (gene 5)
gm1-l-52d3-HCBT2 (gene 6)
gm1-l-52d3-HCBT3 (gene 7)
gm1-l-74i13-HCBT1 (gene 4)
gm1-l-74i13-HCBT2 (gene 5)
gm1-l-74i13-HCBT3 (gene 6)
gm1-l-74i13-HCBT4 (gene 7)
gm1-l-74i13-HCBT5 (gene 8)
gm1-l-74i13-HCBT6 (gene 9)
gm1-l-52d3-Zinc finger (gene 12)
gm1-l-74i13-Zinc finger1 (gene 11)
gm1-l-74i13-Zinc finger2 (gene 12)
gm1-l-52d3-HSF (gene 13)
gm1-l-74i13-HSF (gene 13)
gm1-l-52d3-bHLH (gene 17)
gm1-l-74i13-bHLH (gene 17)
gm1-l-52d3-Expressed (gene 23)
gm1-l-74i13-Expressed (gene 23)
Figure 6. Parsimony based gene tree of the tandemly duplicated HCBT genes from soybean, *Medicago* and *Arabidopsis*. The resulting tree had a length of 1750, a consistency index (CI) of 0.7754, and a homoplasy index (HI) of 0.2246, a CI excluding uninformative characters of 0.7562, a HI excluding uninformative characters of 0.2438, and retention index of 0.7179, and a rescaled consistency index of 0.5567.
Appendix
### Appendix 1. Annotation of gmw1-74i13 and gmw1-52d3

<table>
<thead>
<tr>
<th>Gene</th>
<th>Putative function</th>
<th>BLASTP identified homeolog</th>
<th>Positives score a</th>
<th>E-value a</th>
</tr>
</thead>
<tbody>
<tr>
<td>gmw1-74i13-1</td>
<td>WOX4 protein (homeobox-leucine zipper transcription factor protein)</td>
<td>Arabidopsis NP_175145.2 (251 aa)</td>
<td>137/206</td>
<td>2e-44</td>
</tr>
<tr>
<td>gmw1-74i13-2</td>
<td>Poly(A)-binding protein</td>
<td>Daucus carota AAK30205.1 (658 aa)</td>
<td>511/625</td>
<td>0.0</td>
</tr>
<tr>
<td>gmw1-74i13-3</td>
<td>Membrane-like protein</td>
<td>Oryza sativa BAD29095.1 (481 aa)</td>
<td>348/431</td>
<td>2e-18</td>
</tr>
<tr>
<td>gmw1-74i13-4</td>
<td>N-hydroxycinnamoyl/benzoyl transferase-like 1</td>
<td>Arabidopsis NP_199097.1 (450 aa)</td>
<td>90/123</td>
<td>4e-36</td>
</tr>
<tr>
<td>gmw1-74i13-5</td>
<td>N-hydroxycinnamoyl/benzoyl transferase-like 2</td>
<td>Arabidopsis NP_196402.2 (456 aa)</td>
<td>335/458</td>
<td>6e-142</td>
</tr>
<tr>
<td>gmw1-74i13-6</td>
<td>N-hydroxycinnamoyl/benzoyl transferase-like 3</td>
<td>Arabidopsis NP_199097.1 (450 aa)</td>
<td>323/442</td>
<td>1e-152</td>
</tr>
<tr>
<td>gmw1-74i13-7</td>
<td>N-hydroxycinnamoyl/benzoyl transferase-like 4</td>
<td>Arabidopsis NP_199097.1 (450 aa)</td>
<td>313/442</td>
<td>4e-145</td>
</tr>
<tr>
<td>gmw1-74i13-8</td>
<td>N-hydroxycinnamoyl/benzoyl transferase-like 5</td>
<td>Arabidopsis NP_199097.1 (450 aa)</td>
<td>314/448</td>
<td>8e-137</td>
</tr>
<tr>
<td>gmw1-74i13-9</td>
<td>N-hydroxycinnamoyl/benzoyl transferase-like 6</td>
<td>Arabidopsis NP_199097.1 (450 aa)</td>
<td>329/441</td>
<td>8e-157</td>
</tr>
<tr>
<td>gmw1-74i13-11</td>
<td>Zinc finger (C3HC4-type) 1</td>
<td>Arabidopsis NP_199108.1 (691 aa)</td>
<td>359/730</td>
<td>4e-64</td>
</tr>
<tr>
<td>gmw1-74i13-12</td>
<td>Zinc finger (C3HC4-type) 2</td>
<td>Arabidopsis NP_199108.1 (691 aa)</td>
<td>338/693</td>
<td>3e-59</td>
</tr>
<tr>
<td>gmw1-74i13-17</td>
<td>bHelix-loop-helix transcription factor</td>
<td>Arabidopsis NP_201507.1 (241 aa)</td>
<td>127/200</td>
<td>5e-26</td>
</tr>
<tr>
<td>gmw1-74i13-23</td>
<td>Arabidopsis-like expressed protein</td>
<td>Arabidopsis NP_974873.1 (819 aa)</td>
<td>589/789</td>
<td>0.0</td>
</tr>
<tr>
<td>gmw1-74i13-25</td>
<td>bzip transcription factor</td>
<td>Phaseolus vulgaris AAK39132.1 (415 aa)</td>
<td>375/422</td>
<td>7e-144</td>
</tr>
<tr>
<td>gmw1-74i13-26</td>
<td>DCL protein</td>
<td>Lycopersicon esculentum AAC49433.1 (224 aa)</td>
<td>108/116</td>
<td>3e-56</td>
</tr>
</tbody>
</table>

a Positives and E-values determined from BLASTP report from NCBI
<table>
<thead>
<tr>
<th>Gene</th>
<th>Putative function</th>
<th>BLASTP identified homeolog</th>
<th>Positives score</th>
<th>E-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>gmw1-74i13-27</td>
<td>Aromatic-rich family protein</td>
<td>Oryza sativa BAD802071.1 (257 aa)</td>
<td>154/175</td>
<td>2e-76</td>
</tr>
<tr>
<td>gmw1-74i13-28</td>
<td>PPR-repeat containing protein</td>
<td>Arabidopsis NP_178563.1 (582 aa)</td>
<td>315/547</td>
<td>2e-118</td>
</tr>
<tr>
<td>gmw1-52d3-2</td>
<td>Poly(A)-binding protein</td>
<td>Nicotiana tabacum AAF66823.1 (649 aa)</td>
<td>52/57</td>
<td>2e-16</td>
</tr>
<tr>
<td>gmw1-52d3-3</td>
<td>Membrane-like protein</td>
<td>Oryza sativa BAC29095.1 (481 aa)</td>
<td>346/439</td>
<td>3e-114</td>
</tr>
<tr>
<td>gmw1-52d3-4</td>
<td>N-hydroxycinnamoyl/benzoyl transferase-like 1</td>
<td>Arabidopsis NP_199097.1 (450 aa)</td>
<td>324/442</td>
<td>3e-151</td>
</tr>
<tr>
<td>gmw1-52d3-5</td>
<td>N-hydroxycinnamoyl/benzoyl transferase-like 2</td>
<td>Arabidopsis NP_199097.1 (450 aa)</td>
<td>318/442</td>
<td>2e-140</td>
</tr>
<tr>
<td>gmw1-52d3-6</td>
<td>N-hydroxycinnamoyl/benzoyl transferase-like 3</td>
<td>Arabidopsis NP_199097.1 (450 aa)</td>
<td>109/229</td>
<td>9e-31</td>
</tr>
<tr>
<td>gmw1-52d3-10</td>
<td>Zinc finger (C3HC4-type) 1</td>
<td>Arabidopsis NP_175132.1 (645 aa)</td>
<td>376/719</td>
<td>1e-74</td>
</tr>
<tr>
<td>gmw1-52d3-13</td>
<td>Heat shock transcription factor 31</td>
<td>Glycine max CAA878079.1 (306 aa)</td>
<td>297/307</td>
<td>3e-142</td>
</tr>
<tr>
<td>gmw1-52d3-17</td>
<td>bHelix-loop-helix transcription factor</td>
<td>Arabidopsis NP_190602.1 (231 aa)</td>
<td>92/104</td>
<td>2e-26</td>
</tr>
<tr>
<td>gmw1-52d3-20</td>
<td>Remorin-like protein</td>
<td>Arabidopsis NP_973976 (555 aa)</td>
<td>246/379</td>
<td>2e-62</td>
</tr>
<tr>
<td>gmw1-52d3-23</td>
<td>Arabidopsis-like expressed protein</td>
<td>Arabidopsis NP_974837.1 (819 aa)</td>
<td>378/463</td>
<td>2e-162</td>
</tr>
</tbody>
</table>
CHAPTER 4:
THE FAD2 GENE FAMILY OF SOYBEAN: INSIGHTS INTO THE STRUCTURAL AND FUNCTIONAL DIVERGENCE OF A PALEOPOLYPLOID GENOME

A paper to be submitted to The Plant Genome

Jessica A. Schlueter¹, Iryna F. Sanders², Shweta Deshpande², Jing Yi², Bruce A. Roe², Shannon D. Schlueter¹, Brian E. Scheffler³ and Randy C. Shoemaker⁴

ABSTRACT

The ω-6 fatty acid desaturase (FAD2) gene family in soybean consists of at least five members in four regions of the genome. These desaturases are responsible for the conversion of oleic acid to linoleic acid. Bacterial artificial chromosomes (BACs) corresponding to these loci were sequenced to investigate both structural and functional conservation between paralogous loci. Sequence comparisons between the regions showed that the soybean genome is a mosaic with some regions retaining high sequence conservation in both the genic and intergenic regions while others have only the FAD2 genes in common. Genetic linkage analysis of these BACs using SSRs showed that two BACs with high sequence identity mapped to linkage groups I and O; two linkage groups that have other syntenic markers between them. Another BAC mapped to linkage group L that has previously been shown to have homeology with another chromosome in the genome. The last BAC contained significant repetitive sequence and was unable to be mapped by SSRs. Reverse transcriptase-PCR analysis of the five FAD2 genes showed that the newly identified FAD2-2B and FAD2-

¹ Department of Genetics, Developmental and Cellular Biology, Iowa State University, Ames, IA 50011
² Department of Chemistry and Biochemistry, University of Oklahoma, Norman, OK 73019
³ USDA-ARS MSA Genomics Laboratory, Stoneville, MS 38776
⁴ USDA-ARS-CICGR, Ames, IA 50011, Corresponding author: USDA-ARS-CICGR, Ames, IA 50011. E-mail: rcsshoe@iastate.edu
2C copies are the best candidates for temperature dependent expression changes in developing pod tissue. Semi-quantitative RT-PCR confirmed these results with FAD2-2C showing upwards of an eight-fold increase in expression in developing pods grown in cooler conditions relative to those grown in warm conditions. The implications of these results suggest a candidate gene for decreasing the levels of linoleic acid in developing pods grown in cooler climates.

INTRODUCTION

Cultivated soybeans (*Glycine max* L. Merr) provide greater than two thirds of the oil consumed in the United States and 80 percent worldwide (Mattson et al. 2004). Soybean seed oil generally consists of varying amounts of oleic (12.2 - 22.4%) and linoleic acid (37.6 – 60.2%) depending on the variety (http://ars-grin.gov). Other traditional oil-seed crops, however, have much higher levels of oleic acid such as sunflower (10 – 88%; GRIN, 2006, peanut (36 - 67%; Patel et al. 2004) and canola (60 – 70%; Scheffler et al. 1997).

Traditionally, oils for human consumption are preferred to have lower levels of polyunsaturated fatty acids such as linoleic and linolenic acid. Increased levels of linoleic acid often leads to autooxidation and undesirable flavor changes such as green beany- or grass- flavored and later, as the flavor reversion progresses, fishy-flavored oil (Warner et al. 1994). This poses a great problem to shelf-stability of soybean oils since soybean seed oil is high in linolenic acid. To improve stability, partial and selective hydrogenation during oil processing is carried out to lower the linolenic acid concentration; this process however introduces undesirable trans-double bonds. Conversely, vegetable oils that are higher in oleic acid are prized for having greater shelf-stability and increased structural integrity at higher cooking temperatures (Warner et al. 1994). Vegetable oils high in oleic acid are also desirable because they are supposed to have improved nutritional benefits such as lowering low-density lipoprotein (LDL) cholesterol (Mattson and Grundy 1985; Grundy 1986).

Most polyunsaturated fatty acids, up to 90% in non-photosynthetic tissues, are synthesized through the 18:1 desaturase in the endoplasmic reticulum (Miquel and Browse, 1994). The gene encoding microsomal ω-6 desaturase enzyme, fatty acid desaturase-2 (FAD2) was originally characterized in *Arabidopsis thaliana* and associated with the
conversion of 18:1 oleic acid to 18:2 linoleic acid by inserting a double bond at the 12th carbon in the fatty acid hydrocarbon chain (Okuley et al. 1994).

In soybean, two copies of microsomal ω-6 desaturase, FAD2-1 and FAD2-2 have been cloned (Heppard et al. 1996). FAD2-1 seemed to be expressed primarily in developing seeds, and FAD2-2 was expressed in both vegetative tissues and developing seeds, but at lower levels in seeds than FAD2-1. There is a temperature dependent relationship between oleic and linoleic acid concentration in soybean where lower temperatures lead to an increase in polyunsaturated fatty acids, such as linoleic acid, and a decrease in unsaturated and monounsaturated fatty acids, such as oleic acid (Neidleman, 1987; Thompson, 1993; Thomas et al. 2003; Tang et al. 2005). The transcript levels of FAD2-1, FAD2-2 and FAD6 (the plastidal ω-6 desaturase) did not appear to increase at low temperatures in developing pods as would be expected if they were responsible for the increased levels of linoleic acid.

Recently, EST based searches by Tang et al. identified an additional copy of FAD2-1 in soybean (2005). An analysis of the EST libraries from which the FAD2 genes were identified found that all FAD2-1 ESTs were from seed-related libraries and FAD2-2 ESTs came from a variety of tissues as previously reported (Tang et al. 2005). Analysis showed that FAD2-1B (the newly identified FAD2-1 copy) was more represented in mRNA pools of developing seeds than the FAD2-1A copy. Looking at the protein stability of the ω-6 desaturase FAD2-1 enzymes at varying temperatures, they found two domains that might be responsible for FAD2-1A enzyme instability at higher temperatures. Further, they identified a serine residue that is phosphorylated during soybean seed development and might down regulate enzyme activity (Tang et al. 2005). Another study using a soybean mutant with a deletion of one of FAD2-1 gene (a mid-oleate line) suggested that modifying genes might have an important influence on oleate content (Alt et al. 2005).

The existence of multiple FAD2 gene copies in soybean is not surprising given that the soybean genome is a paleopolyploid. Evidence for this has been based upon cytogenetic studies (Hadley and Hymowitz 1973; Lackey 1980; Walling et al., 2006), soybean gene family studies (Lee and Verma 1984, Hightower and Meagher 1985, Grandbastien et al., 1986, Nielsen et al., 1989), genetic mapping studies (Shoemaker et al., 1996; Lee et al., 1999; Lee et al., 2001), BAC hybridization, end sequence and fingerprinting (Marek et al.,
2001; Foster-Hartnett et al., 2002; Yan et al., 2003; Yan et al., 2004) and EST based analyses that identified at least two major genome duplications (Schlueter et al. 2004; Blanc and Wolfe 2004a). Most recently, the first study looking at the sequence conservation in homeologous regions found that there is strikingly strong conservation of both gene order and orientation as well as sequence conservation in the non-coding regions (Schlueter et al., in press).

The FAD2 genes provide an excellent resource to further study the evolutionary dynamics of a paleopolyploid genome. In this paper, we report the identification of two more FAD2 gene copies (FAD2-2B and FAD2-2C) from the EST collections. Four soybean BACs representing the five FAD2 gene copies were sequenced. Surprisingly, only the two BACs anchored by FAD2-1A and FAD2-1B showed high levels of homeology. The other two BACs had only the FAD2-2 genes in common. In fact, the BAC containing FAD2-2C is a highly repetitive BAC containing numerous retrotransposon insertions. Since there are now five FAD2 gene copies known, we also evaluated the spatial and temporal expression of each of these genes in developing seeds at both warm and cool temperatures as well as other vegetative tissues.

MATERIALS AND METHODS

FAD2 BAC selection: Identification of all putative FAD2 gene copies was done utilizing the FAD2-1A sequence (L43920) in a TBLASTX search against all soybean ESTs using default parameters (Altschul et al. 1990). All ESTs were aligned into contigs using Sequencher v. 4.5 with default parameters (Gene Codes Corp., MI). Four FAD2 copies were identified. PCR primers were designed to distinguish between copies using Oligo 6.82 (Molecular Biology Insights, Cascade, CO) and tested by PCR against Glycine max cultivar Williams 82 using a DNA Engine Gradient Cycler from MJ Research (Watertown MA). PCR reactions were 10 µl in volume and contained 1.1 X MasterAmp 2X PCR PreMix B (Epicentre, Madison, WI), 0.11 µM each primer, 50 ng Williams 82 DNA, and 0.1375 U of Taq DNA polymerase (Invitrogen, Carlsbad CA). PCR cycling conditions were 94° for 2 min, 35 cycles of 94° for 45 sec, annealing temperature for 30 sec, 72° for 45 sec, followed by a final extension of 72° for 3 min. The resulting products were gel-purified and
subsequently sequenced at the DNA Synthesis and Sequencing Facility (Iowa State University, Ames IA) to verify homeolog specificity. The primer sequences for FAD2-1A were U, 5’ CCA AAG TGG AAG TTC AAG GGA 3’ and L, 5’ CT'T CCT AGA GGG TTG TTT AAG 3’; for FAD2-1B, U, 5’ GCC AAA GGT GAA ATT CAG CAG 3’ and L, 5’ CTT CCT AGA GGG TTG TTC AGG 3’; for FAD2-2A (and FAD2-2B) U, 5’ GTC CCA TTT ACT CTG ACC GA 3’, and L, 5’ ACC GAT AAT ACT CTC CCA AGA 3’; and for FAD2-2C U, 5’ TCT AAA TAC CTT AAC AAT CCT C 3’, and L, 5’ GCC TTG ACA AAT GGA GTC TC 3’.

Once primers specificity was verified, the Williams 82 *G. max* BAC library (Marek and Shoemaker 1997) was PCR screened using the previously described conditions to identify FAD2 BACs. BAC DNA was isolated using a Plasmid Midi kit (Qiagen, Valencia CA) and identities were verified with PCR as described above. BAC-end sequences were obtained using M13 forward and reverse primers at the DNA Synthesis and Sequencing Facility (Iowa State University, Ames IA).

**BAC sequencing and assembly:** Three BACs, gmwl-45m6, gmwl-15k6 and gmwl-11j16 were sequenced at the University of Oklahoma. The detailed procedures for cloned, large insert genomic DNA isolation, random shot-gun cloning, fluorescent-based DNA sequencing and subsequent analysis were used as described previously (Bodenteich et al. 1993; Pan et al., 1994; Roe et al., 1996; Chissoe et al., 1995; Roe 2004). Fifty microgram portions of purified BAC DNA were randomly sheared, blunted, kinase treated, and gel purified. Fragments 2-4 Kb in size were ligated into Smal-cut bacterial alkaline phosphatase (BAP)-treated pUC18 (Pharmacia, NJ) and transformed by electroporation into *E. coli* strain XL1BlueMRF competent cells (Stratagene, La Jolla CA).

Sequencing reactions were performed as described (Roe et al., 1996; Chissoe et al., 1995; Roe 2004) using the Amersham ET Terminator kit (US-81070) or Applied Biosystem BigDyes version 3.1 (4336921) and were resolved by electrophoresis on an ABI 3730 Capillary DNA sequencer. After based calling with the ABI Analysis software, the analyzed data was assembled using Phred and Phrap (Ewing et al., 1998; Ewing and Green 1998). Overlapping sequences and contigs were analyzed using Consed (Gordon et al., 1998). Gap closure and proofreading was performed using either custom primer walking or using PCR
amplification of the region corresponding to the gap in the sequence followed by sequencing directly using the amplification or nested primers, or by sub-cloning into pUC18 and cycle sequencing with the universal primers and PCRs with Ideaza-dGTP replacing dGTP (Roe, 2004) or by rolling circle amplification (Detter et al. 2002) were necessary to obtain at least three-fold coverage for each base.

BAC DNA for gmwl-105h23 was randomly sheared using a nebulizer (Invitrogen) and size selected for 2-4 kilobases (Kb) on a 1% low-melt agarose gel. Sheared DNA was phosphatase-treated, blunt-end repaired and cloned into the vector pCR4Blunt-TOPO (TOPO shotgun subcloning kit, Invitrogen). The recombinant plasmids were transformed into TOP10 E.coli cells by electroporation and selected on LB plates containing kanamycin. Subclones were sequenced at the USDA-ARS Mid-South-Area Genomics Laboratory using M13 forward and reverse primers on an ABI3730XL with BigDye 3.1. Base calling, vector trimming and contig assembly was done using SeqMan II starting with a match size of 12 bp, a maximum gap of 70 bp, and a minimum match percentage of 98% and decreasing the minimum match percentage to 90% as necessary to merge contigs (DNAStar, Inc.). Gaps were closed using PCR primers designed from adjacent contigs. The PCR product was gel-purified, subcloned into TOPO TA vector (Invitrogen), transformed into TOP10 E.coli cells (Invitrogen) and sequenced with M13 forward and reverse primers. All gap-closing sequencing was done at the DNA Synthesis and Sequencing Facility (Iowa State University).

**Genetic mapping of BACs:** To genetically map each BAC, the corresponding sequences were scanned for SSRs as previously described (Schlueter et al. in press). Testing of parental lines revealed polymorphisms in the mapping population *Glycine max* A81-356022 X *Glycine soja* PI 468.916 (Diers et al. 1992; Shoemaker et al., 1996). The SSR primer sequences are as follows: for gw1-15k6 U, 5’ ACT TGC CAG TTG CCA CAT TTG 3’ and L, 5’ AGC CAA TAT CTG CCA AGG TC 3’, for gmwl-105h23 U, 5’ ACT TGG GGT TGC TTG ATA ACA AAT 3’ and L, 5’ TAC AAG TTT TGG GCG TTG TAT TCC 3’ and for gmwl-11j16 U, 5’ CTT TAT GTT TGC TTT GGA ATC 3’ and L, 5’ CAG ATG TAT ATT CAG AAC TAG A 3’. No polymorphic SSR markers could be found for gmwl-45m6. Genetic relationships between these SSRs and the established map were determined using MapMaker with a minimum lod score of 3.0 (Diers et al., 1992; Lander et al., 1987).
Sequence Analysis and annotation: Gene prediction was done using a combination of \textit{ab initio} and similarity-based methods displayed by xGDB system (Schlueter et al., unpublished results). For \textit{ab initio} prediction, Genscan with \textit{Arabidopsis thaliana}-based parameters (Burge and Karlin 1997), FgeneSH with \textit{Medicago truncatula}-based parameters (http://www.softberry.com) and GeneMark.hmm with \textit{A. thaliana}-based parameters (Lukashin and Borodovsky 1998) were run. Each BAC sequence had all soybean ESTs aligned as well as all plant putatively unique transcripts (PUTs) using Geneseqer at PlantGDB (Schlueter et al., 2003; Dong et al. 2005) For each gene, the structure was predicted using the best available EST alignments, then the PUT alignments, and then relying on the \textit{ab initio} predictions utilizing the User Contributed Annotation system as part of xGDB (Schlueter et al., 2005; Wilkerson et al., unpublished results). Each predicted gene was subjected to a BLASTP query of the NCBI nr database with default parameters to assign putative function as well as to identify conserved domains within the protein (Altschul et al. 1990).

Alignment of the homeologous BACs was done using shuffle-LAGAN (Brudno et al., 2003) to perform global pair-wise alignments, with default parameters and anchored by predicted gene regions, to produce a VISTA plot to visualize sequence identity between sequences (Frazer et al., 2004). The nucleotide and protein percent identity and similarity between genes both intra- and inter-chromosomally was calculated using the program WATER (gap penalty of 10; extension penalty of 0.2; EMBOSS). Synonymous and nonsynonymous distances were calculated using PAML, with default parameters (Yang 1997).

Putative retroelements were initially identified from \textit{ab initio} gene predictions that were most similar by BLAST based annotation to polyproteins. Both BLASTN and TBLASTX were performed against the TIGR repeat databases (www.tigrblast.tigrorg/euk-blast/index.cgi?project=plant.repeats). Potential LTR retrotransposons were searched for using LTR_STRUC, with an intensity parameter of 1 (McCarthy and McDonald 2003). Each BAC was fragmented into 1 Kb pieces and a BLASTN, default parameters, performed against one another to further identify putative LTRs (Altschul, 1990). Putative MITEs were identified using FINDMITE with direct repeats of TA, TAA, TAT, TTA and GCA, a
terminal inverted repeat (TIR) length of 12, allowing 3 mismatches, repeat filtering, and a minimum and maximum MITE length of 50 and 750 respectively (Tu 2001). RepeatMasker was run utilizing Repbase (Smit, AFA & Green, P RepeatMasker at http://ftp.genome.washington.edu/RM/RepeatMasker.html).

**Transcript accumulation of homeologous genes, specifically FAD2: RT-PCR**

primers for homeologs retained between BACs gmw1-15k6 and gmw1-105h23 as well as for FAD2-2 genes on BACs gmw1-11j16 and gmw1-45m6 were designed. Where possible, primers were designed to flank an intron as an internal control. Additionally, primer pairs were designed to be homeolog specific. Each primer pair was tested by PCR against *Glycine max* cultivar Williams 82 genomic DNA to verify the homeolog specificity. PCR products were gel purified and subsequently sequenced at the DNA Synthesis and Sequencing Facility (Iowa State University).

Soybean tissue was collected from developing pods from the soybean cultivar Williams 82. All plants were grown in growth chambers with a day/night temperature cycle of 32/28 or 18/12°C and light/dark cycle of 12/12 hours. Developing pod tissue was collected at 6-10, 13, 17, 21 and 26 days after flowering (DAF) in both the warm and cool growth chambers as previously done by Heppard et al. (1996). Tissue for cotyledons, roots, and furled unifoliate were collected 3 days after emergence (DAE). Unfurled unifoliates tissue was collected 4 DAE. Samples of cotyledons and roots were taken at 7 and 8 DAE respectively. Furled trifoliolate was collected 11 DAE and unfurled trifoliolate at 15 DAE. Flower tissue was collected from soybean plants grown in the greenhouse at approximately 60 DAE. For each tissue collected at each time point, samples were taken from at least three independent plants. All tissue was flash frozen with liquid nitrogen and stored at ~80°C for subsequence RNA extraction.

mRNA was extracted and purified from frozen tissue using the RNeasy Plant Mini Kit (Qiagen). This kit yielded approximately 35-65 μg of RNA from a single extraction using 100 mg of tissue. Extracted mRNA was treated using DNA-free DNase and removal kit (Ambion) to remove any DNA contamination and quantified using a NanoDrop ND-1000 spectrophotometer (Wilmington, DE).
Non-quantitative RT-PCR was done using SuperScript One-Step RT-PCR system (Invitrogen) across all tissues with the primers designed and tested as discussed above. PCR conditions and controls were as previously described (Schlueter et al., in press). Tubulin was used as a positive control (Graham et al., 2002). All RT-PCR reactions were done with two or three independent biological replicates. These reactions provide a positive/negative screen for presence of a transcript in a particular tissue.

Semi-quantitative real-time PCR was performed with all five of the FAD2 genes as well as their alternatively spliced constructs. Reactions were performed using Stratagene’s Brilliant qRT-PCR kit (La Jolla, CA) and the Stratagene Mx3000P thermocycler. Reactions were 25 uL in volume as suggested by the manufacturer with 2mM MgCl₂, 0.2 uM of each primer, 0.625 uL of a 1:700 dilution of Sybr Green, and 0.1875 uL of a 1:1000 dilution of reference dye. Cycling parameters were 42° C for 45 min for reverse transcription, 95° for 10 min to deactivate the reverse transcriptase, 40 cycles of 95° for 45 sec, annealing temperature for 30 sec, and 72° for 45 sec, followed by a dissociation curve reading progressing from 55° to 95° C to verify a single product. The Stratagene analysis system established a threshold fluorescence level where product fluorescence levels were statistically higher than background fluorescence; this threshold level is referred to as the Ct value, the cycle at which the samples fluorescence is above threshold. To ensure fluorescence levels were valid, a passive reference dye was added to each sample and each sample was run in triplicate. Further, two biological replicates were run to verify the biological significance of the observed expression patterns. To be considered differentially expressed, the warm pods and the cool pods at the same time point had to differ in where they crossed the fluorescence threshold by more than 1 cycle.

RESULTS

Genetic map positions and sequence organization of FAD2 regions: Shotgun sequencing of BACs gmw1-15k6, gmw1-105h23, gmw1-11j6 and gmw1-45m6 yielded approximately 497 kb of soybean genomic sequence anchored by microsomal ω-6 fatty acid desaturase genes (FAD2). PCR-based screens of the Glycine max Williams 82 BAC library
identified two BACs each for FAD2-1A, FAD2-1B, and FAD2-2A. However, for FAD2-2, only one BAC was identified. To maximize overlap between regions, the largest BAC for each gene copy was sequenced. Genetic mapping of these BACs positioned gmwl-15k6 and gmwl-105h23 on linkage groups O and I, respectively (Figure 1). Map positions were determined based upon the Glycine max A81-356022 X Glycine soja PI 468.916 (Diers et al. 1992; Shoemaker et al., 1996) mapping population. Gmw1-15k6 was placed 10.1 cM from Sat_105 and 10.3 cM from Satt270. BAC gmwl-105h23 mapped to 26.9 cM from Sat_109 25.7 cM from Mng085_1, and 37.9 cM from A664_1. Figure 1 shows their placement relative to markers in the soybean composite map and illustrates the markers that are synteny between these linkage groups (Song et al., 2003; http://soybase.org). The placement of gmwl-15k6 on the composite map was problematic. Mng085_1 has been placed only 1.13 cM from Sat_109 and moved further up the linkage map compared to the Glycine max A81-356022 X Glycine soja PI 468.916 map. Likely, this relocation is due to the small population size of the Max X soja cross whereas the composite map is based upon 5 large mapping populations (Song et al., 2003). As a result, gmwl-15k6 was placed on the composite map relative to Sat_109 and A664_1 instead of Mng085_1. Previous mapping studies have shown that linkage group O and linkage group I have numerous synteny markers suggestive of homeologous regions (Shoemaker et al. 1996). There do not appear to be any oil QTLs associated with markers in these regions of the linkage groups although linkage group I does have an oil QTL near the region (Figure 1; http://soybase.org). BAC gmwl-11j16 mapped to linkage group L, 24.3 cM from peG488_1 and 23.0 cM from Satt156. Ten different SSRs were tested from BAC gmwl-45m6, and none show polymorphisms in any of the mapping populations screened. In all likelihood, gmwl-45m6 falls in a region of the genome with low recombination frequency since none of the SSR markers identified polymorphisms. All mapped BACs showed linkage to only the group with which they were placed.

Gene predictions for these four BACs identified a total of 57 genes (Figure 2; Table 1). The structure of each gene was predicted first based upon EST alignments to the genomic sequence, then by putatively unique transcripts (PUTs; Dong et al. 2005) from other plants and finally based upon ab initio predictions. Of all predicted genes, 65.5% had exons that were supported by EST alignments, similar to the EST support of genes predicted from a
previous study (65%; Schlueter et al., in press). All of these predicted gene structures as well as the EST support and *ab initio* predictions are available at http://haldane.agron.iastate.edu/GmaxGDB. The average gene size was approximately 956 bp, with introns and exons generally being the same size. The gene density in these regions seems to vary greatly from 1 gene every 6.7 and 7.95 kb for BACs gmwl-15k6 and gmwl-105h23, respectively. Whereas gmwl-11j6 is somewhat less compacted with 1 gene every 7.77 kb. Gmwl-45m6 has only 1 gene every 19.2 kb containing fewer genes than the other BACs. These predictions, not considering gmwl-45m6, seem to fall in the range of previous gene density predictions in soybean ranging from 1 gene per 5.8 to 9.9 kb (Young et al., 2003, Mudge et al., 2005, Schlueter et al., in press).

All of the translated protein sequences of these genes showed some similarity to proteins in the NCBI non-redundant database. Table 1 lists the annotation of predicted genes as well as the top BLASTP hit, the number of positives and e-values for the top BLASTP hit. Thirteen genes showed similarity to unknown or predicted genes in the database with no known functions or domains. While these BACs were chosen for their containing FAD2 genes it appears that in these regions there are no other genes known to be involved in oil biosynthesis. It is possible, however, that some of the transcription factor-like genes or unknown genes may have a role in oil biosynthesis. There appeared to be only two cases of tandem duplication, the WD-domain containing genes on gmwl-45m6 and the FAD2-2 genes (A and B) on gmwl-11j16 (Figure 2).

Based upon the EST alignments, FAD2-1A and FAD2-1B both seem to have some sort of alternative spliced construct. For both FAD2-1A (gmwl-15k6) and FAD2-1B (gmwl-105h23,) the alternative spliced construct moves the start position by increasing (or decreasing) the intron size, but the first exon only includes the start codon so alternative splicing does not change the resulting protein. On BAC gmwl-15k6, the vesicle-associated membrane protein has six possible alternative transcripts, with both intron retention and exon loss. Similarly, on gmwl-15k6, the ribonuclease HII and the unknown protein most like a *Cicer arietinum* gene (CAD31715.1) both show possible cases of intron retention. The gmwl-105h23 copy of melvionate disphosphate decarboxylase has two possible transcripts; the most common transcript has a fourth intron while the other two possible constructs lack...
that intron as well as having another possible splicing acceptor site. Gmwl-105h23 pollen-specific protein also shows a possible intron retention truncating the resulting protein. All of these alternatively spliced constructs can be viewed at http://haldane.agron.iastate.edu/GmaxGDB.

Only degenerate sequences with similarity to retroelement reverse transcriptases could be identified on gmwl-1lj16 and gmwl-105h23. The degenerate retroelement sequence on gmwl-1lj16 is most like a gag-pol polyprotein previously identified from Glycine max (AAC64917.1), although the BLAST-based alignment was relatively low. Gmwl-105h23 has a degenerate Tyl-copia like reverse transcriptase at approximately 112,250-113,000 bp. Interestingly, as an artifact of the subcloning and sequencing processes BAC gmwl-105h23 has an E. coli transposon insertion in intron 16 of an unknown gene most like Arabidopsis NP_197637.1. Gmwl-1-15k6 contains only one retrotransposon that is inserted between the ferredoxin gene and an unknown protein most like Oryza sativa NP_915582.1 This retroelement is nearly identical to a previously identified gypsy-like retroelement in soybean (AAO23078.1) with an e-value of 0.0 and 1379/1554 positives. The expected structure of a gypsy element with the reverse transcriptase (RT) preceding the integrase (INT) is conserved in this retrotransposon. The LTRs for this element are only 363 bp in length, but are highly similar to one another suggesting a recent insertion.

BAC gmwl-45m6 is very different from the other three in that the lower gene density is likely due to the numerous repetitive elements that can be identified on this BAC. A BLASTX search of known reverse transcriptase sequences against the BAC identified seven putative Diaspora Ty3-gypsy like elements. The program LTR_STRUC, with a stringency setting of 1, identified one LTR-retrotransposon of over 15706 bp in length. When searching for the location of the LTRs, an interesting trend occurred, eleven different regions of gmwl-45m6 showed high sequence identity to that LTR sequence. What this means is that this particular LTR-retrotransposon has inserted numerous times across this region. Additionally, a putative CACTA element fragment was identified from 57875-59365 (just 3' of the FAD2-2C gene copy) based upon a TBLASTX similarity search against the TIGR Brassicacea repeat sequences; the region was most similar to an Arabidopsis thaliana CACTA1, with an e-value of 0.0, but only a small region of the CACTA element was conserved.
Although putative MITEs (miniature inverted transposable elements) were search for using FINDMITE (Tu 2001), the very nature of these elements makes determining their authenticity difficult. As with Schlueter et al. (in press) hundreds of MITEs were predicted between the four BACs. Since MITEs are often found in gene rich regions, some of these are likely to be true MITEs, but verification of these remains difficult (Feschotte et al., 2002).

**Comparison of FAD2 homeologous regions:** Although these regions were chosen because a small gene family anchored them, only two of the four BACs have any other genes retained beyond the FAD2 genes (Figure 2). The homeology between gmw1-105h23 and gmw1-15k6 is extremely strong as shown in Figure 3, a VISTA based plot showing sequence identity between the two regions. Between the regions, there is one major inversion from the F1F0-ATPase inhibitor-like gene to the pollen-specific gene on gmw1-105h23 with respect to gmw1-15k6. Structurally, this is the only major difference with all of the genes then conserved in both order and orientation in these regions. Although it is difficult to see in figure 2 due to the high gene density in these regions, between genes there is some sequence similarity (greater than 75%) in close proximity to the coding regions likely due to promoter elements and transcription factor binding sites (Figure 3). Not surprisingly, there is much stronger sequence conservation in introns than outside of the genes. The full-length retroelement on gmw1-15k6 BAC is not found on gmw1-105h23. Also, there are large regions between genes that have less than 50% sequence identity; in other words, they are not conserved at all.

Gene structures themselves between these regions are also well conserved. Much as with previous findings (Schlueter et al., in press) the gene length between each BAC is preserved. Between each gene pair, the number of exons and introns for most genes are identical. Exceptions to this are the gmw1-105h23 ribonuclease HII gene that appears to be truncated with respect to the gmw1-15k6 gene; the gmw1-15k6 F1F0-ATPase inhibitor-like gene appears to have an included intron that the gmw1-105h23 copy does not show evidence for; four exons on gmw1-105h23 unknown gene most like *Oryza sativa* NP_915582.1 are not predicted in the gmw1-15k6 copy; one exon on the gmw1-105h23 unknown gene most like *Arabidopsis* NP_197637.1 is not predicted in the gmw1-15k6 copy; and the alternative splicing events as discussed above. The gmw1-15k6 copy of melvionate diphosphate
decarboxylase contains five more exons relative to the gmw1-105h23 copy, although this is most likely due to truncation of the gmw1-105h23 copy since it is at the end of the BAC. The average nucleotide identity between the coding regions of gmw1-15k6 and gmw1-105h23 is 90.7%; and the resulting amino acid identity and similarity is 88.9% and 90.4%, respectively (Table 2). These findings are relatively similar to that of Schlueter et al. (in press), although the nucleotide identity is slightly higher between these BACs.

Since we are interested in the paleopolyploid nature of the soybean genome, synonymous and nonsynonymous distance measures were calculated using PAML between homeologs on gmw1-15k6 and gmw1-105h23. Two studies utilizing different methods have predicted that the soybean genome underwent major duplications approximately 14.5 and 45 million years ago (MYA; Schlueter et al., 2004) or approximately 4.5 and 16.1 MYA (Blanc and Wolfe, 2004a). The predicted date based upon the average synonymous distance of homeologs between gmw1-15k6 and gmw1-105h23 (0.1061) gives a coalescence estimate of approximately 8.70 MYA (Table 2). This value fails to fall in either predicated range from previous studies.

**The FAD2 gene family:** When only the FAD2 genes are considered, some interesting trends emerge. On average, the nucleotide identity between the FAD2 gene copies is 76.48% but ranges from 66.6 to 95.8%, the average protein identity is 77.94 with a range of 67.4 to 95.9% and the average protein similarity is 85.12% with a range of 72.1 to 96.9%. Not surprisingly, FAD2-1A (gmw1-15k6) and FAD2-1B (gmw1-105h23) are most similar to one another. What are interesting are the high nucleotide identity, 95.8% as well as protein identity, 95.9% and similarity, 96.6% between the overlapping regions of FAD2-2B (gmw1-11j16) and FAD2-2C (gmw1-45m6). There has been surprising conservation in these FAD2-2 coding regions with seemingly nothing else in common between the BACs.

Again, not surprisingly, the FAD2-1A and FAD2-1B genes gave a divergence estimate of approximately 8.59 MYA, approximately the average of all homeologs on those BACs. However, when these two closely related copies are compared to the FAD2-2 genes, synonymous distances suggest that they diverged from 105 to 277 MYA. When just the FAD2-2 genes are compared to one another, it appears that the tandem duplication on gmw1-11j16 occurred about 28 MYA and that FAD2-2B is more closely related to the FAD2-2C
(gmwl-45m6) with a coalescence estimate of 10 MYA. This is reflected as well in the sequence identity discussed above. Similarly, FAD2-2A and FAD2-2C distances suggest a divergence at 31.9 MYA. This estimate is relatively close to that of the tandem duplication on gmwl-1 lj 16. One possibility is that FAD2-2B was tandemly duplicated, the region itself was duplicated (leading to the FAD2-2C BAC) and that region underwent extensive loss/rearrangement/retroelement insertion. Another hypothesis would be that FAD2-2C is the result of a single gene insertion into a region of the genome that is very repetitive, or that FAD2-2C or FAD2-2B are closely related due to gene conversion. Taken together, these results suggest that the duplication that generated the FAD2-1 and FAD2-2 genes is quite ancient. Ratios of nonsynonymous to synonymous distances showed that all FAD2 genes are under purifying selection.

Expression of FAD2 in developing pod tissue: Of particular agronomic interest is the functional divergence between the FAD2 genes, especially in developing pod tissues grown at both warm and cool temperatures. Previous work has looked at the expression of FAD2-1A and FAD2-2B in both vegetative tissues and developing seeds (Heppard et al., 1996). These results suggested that FAD2-1A is expressed specifically in seeds and that FAD2-2B is expressed in all tissues, but at lower levels in seeds than FAD2-1. Further, when looking for temperature dependent expression differences with either FAD2 copy, none were observed (Heppard et al. 1996). This was surprising since oleic acid is converted to linoleic acid in developing pods grown at cooler temperatures there is an increase in linoleic acid. This study has identified three more FAD2 genes that might have a role in the temperature dependent accumulation of linoleic acid in developing pods.

Developing pod tissue was harvested at several developmental stages as was previously done (Heppard et al. 1996). Two samples of 6-10 DAF pods were used to account for the tissue being harvested in a range of days. In addition, tissue for cotyledons, roots, and furled unifoliate (3 days after emergence, DAE), unfurled unifoliates (4 DAE), cotyledons and roots (7 and 8 DAE respectively), furled trifoliolate (11 DAE), unfurled trifoliolate (15 DAE), and flowers (60 DAE). RT-PCR primers were designed for all five copies of FAD2 as well as for the two possible alternative transcripts (FAD2-1A and FAD2-1B). Where possible, the primers were designed across the intron in the 5’ UTR to account for any
possible DNA contamination in the samples. The primer sequences for these genes can be found in Table 3. Tubulin primers were used as a positive control and to verify the integrity of each mRNA sample.

Reverse-transcriptase PCR screens were performed across all pod tissues as well as the vegetative tissues with all FAD2 constructs to determine where each FAD2 gene’s transcripts accumulate. These results can be seen in Figure 4. The FAD2-2A (gmw1-11j16) primer pairs did not amplify any RNA samples, although the positive Williams 82 genomic DNA control amplified (data not shown). This suggests that FAD2-2A is either not expressed in tissue surveyed in this experiment, or that this gene copy is no longer functional. The FAD2-1A and FAD2-1B alternatively spliced transcripts both seem to be biologically relevant and both expressed, verifying the predicted structures. These screens were also performed with a number of the homeologs from gmw1-15k6 and gmw1-105h23. These results can been seen in Appendix 1 and 2. Similar to the results of Schlueter et al. (in press) some genes show different expression patterns where others do not. Also, some of the genes do not appear to be expressed in the surveyed tissues.

Neither FAD2-1A and FAD2-1B, including the alternatively spliced transcripts, seem to have any temperature dependent expression in developing pods. These results confirm that there does not appear to be increased expression of FAD2-1A in response to cooler temperatures (Heppard et al. 1996). In fact, there appears to be a possible delayed expression of FAD2-1A and FAD-1B in developing pods grown in cooler conditions. What was surprising was the amplification of both FAD2-1A and FAD2-1Balt, the alternatively spliced FAD2-1B structure, in vegetative tissues when these copies were originally thought to be expressed only in developing pods. FAD2-1B has transcript accumulation in 3 DAE furled unifoliolates and 4 DAE unfurled unifoliolates. FAD2-1A shows transcript accumulation in 7 days after emergence (DAE) cotyledons, 8 DAE roots, and 11 DAE furled trifoliolates (Figure 4). As expected, FAD2-2B is expressed in all tissues surveyed, but the intensity of the bands in developing pods from cooler tissues suggests that FAD2-2B may be a candidate for increased expression of FAD2 in developing pods at cooler temperatures leading to increased levels of linoleic acid (Figure 4). FAD2-2C seems to show very similar expression
patterns to FAD2-2B. This is not surprising given that FAD2-2B and FAD2-2C are very similar in their coding regions as discussed above.

Given the results of the traditional RT-PCR screens, semi-quantitative reverse-transcriptase PCR (qRT-PCR) experiments were performed. The objective was to determine if significant fold changes in expression could be identified at each time point for developing pods grown under warm conditions versus those grown under cool conditions. The number of cycles to reach threshold fluorescence or the Ct value was determined based upon the point at which the FAD2 amplicon fluoresced at a higher level than background fluorescence. QRT-PCR screens were run with all six FAD2 primer pairs that showed expression in developing pods (Figure 4). Only the developing pod tissue was used in these experiments since we were only looking for changes in expression between different environmental conditions for each independent gene. The results seemed to somewhat verify what was expected based upon the RT screens (Table 4). In order to be considered a significant difference between warm pod tissue and cool pod tissue at a particular time point, the change in Ct value must be greater than 1 cycle. A difference of one cycle at a particular time point between the warm pod tissue and the cool pod tissue translates into a 2-fold change in expression. Nothing under 2-fold significance was considered. Both FAD2-1A and FAD2-1B seem to have significant changes in expression in warm developing pods versus cool developing pods, although not in the desired direction. In both cases, expression of FAD2 increases in the warm pod tissues relative to the cool pod tissues (Table 4). In 13 DAF pods, there was a 12.4-fold and 8.4-fold increase in expression of FAD2-1A and FAD2-1B, respectively, in the warm pods relative to the cool pods (Table 4). Similarly, this is seen in the 17 DAF pods with 14.2-fold and 11.2-fold increase in expression of FAD2-1A and FAD2-1B in warm pods relative to cool pods. At 26 DAF, FAD2-1A and FAD2-1B genes seem to have 5.8-fold and 3.4-fold increase in expression in warmer tissues.

Although FAD2-2B seemed a good candidate for temperature dependent increase in expression based upon the RT-PCR screens, the semi-quantitative RT-PCR showed no significant changes in expression between the warm developing pods and the cool developing pods. FAD2-2C is the only ω-6 desaturase that shows increases in expression in the cool
developing pods. Table 4 shows that at 6-10 DAF, 13 DAF, 19 DAF, 21 DAF and 26 DAF fairly significant fold changes in expression are observed ranging from 2.6 to 8.0.

DISCUSSION

**Map locations relative to QTLs:** Genetic mapping of gmwl-15k6 and gmwl-105h23 places these BACs on linkage groups I and O in the *Glycine max* A81-356022 X *Glycine soja* PI 468.916 mapping population (Diers et al. 1992). Based upon comparison of the soybean composite maps for these linkage groups, there are numerous syntenic markers suggesting these are homeologs (Figure 1). The strong association of markers between the linkage groups makes these regions excellent candidates to further study the evolution of a duplicated genome. The mapping of gmwl-11j16 to linkage group L is significant in that this linkage groups has been shown through FISH analysis with BACs to chromosomes that seven BACs along linkage L had secondary hybridization signals on another chromosome (Walling et al., 2006). This suggests that there may be another region on that chromosome that is homeologous to gmwl-11j6 and that there may be more copies of FAD2 in the soybean genome.

Given the map positions of these genes, and their relative locations on the soybean composite maps, QTLs spanning these regions were identified using SoyBase (http://soybase.org). Although these regions are anchored by FAD2 genes, there are no oil QTL at these map positions. However, all three linkage groups O, I, and L do contain oil QTLs; on LG-O it is approximately 70 cM away from gmwl-15k6 (Panthee et al. 2005), on LG-I it is approximately 14 cM away from gmwl-105h23 (Diers et al., 1992) and on LG-L it is about 24 cM away from gmwl-11j16 (Fasoula et al., 2004). Gmwl-11j16 seems to map within a Sclerotinia stem rot resistance QTL (data not shown; Arhana et al., 2001), while gmwl-15k6 is approximately 8 cM from Sclerotinia QTLs. Both gmwl-11j16 and gmwl-15k6 map near pod maturity QTLs (Specht, 2001; Wang et al., 2003). Gmwl-105h23 maps to a region of linkage group I with iron efficiency QTLs (Lin et al, 1997; Lin et al., 2000).
Although these BACs do not map to regions with oil QTLs, the existence of the pod maturity QTL is promising. The role of FAD2 in oil biosynthesis of seeds and in seed development makes these genes excellent candidates for pod maturity. Also of great interest is the ferredoxin gene on gmw1-105h23 that maps directly to an iron efficiency QTL. This gene is certainly an excellent candidate for iron chlorosis studies. Significant efforts were made to map gmw1-45m6 but none of the SSR markers identified from the BAC showed any polymorphism in the soybean genetic mapping populations. Current models of the soybean genome suggest that repetitive regions are generally heterochromatic and that euchromatic and heterochromatic regions are distinct (Polzin et al. 1998). Gmwl-45m6 contains many repetitive sequences, and if it follows the proposed model of the soybean genome is likely on the edge of heterochromatic sequence.

**Paleopolyploid genome structure:** Paleopolyploids are ancient polyploids that are undergoing the switch from tetrasomic to disomic inheritance in a process called diploidization. It has been previously suggested that during this ongoing process of diploidization there are both diploid and tetraploid loci within the genome (Wolfe 2001). Most plant species are now considered to be paleopolyploids (Lockton and Gaut, 2005; Blanc and Wolfe 2004a; Schlueter et al., 2004; Masterson 1994). The first study looking at homeologous regions in soybean suggested that duplicated regions in this paleopolyploid genome have higher than expected gene retention and that particularly in soybean the process of diploidization is quite slow (Schlueter et al., in press). Two BACs support previous findings that the soybean genome has retained extensive conserved homeologous sequence (Schlueter et al., in press). However, the two BACs show very different evolutionary dynamics within the soybean genome.

Between gmw1-15k6 and gmw1-105h23 we observe genic sequence retention in both order and orientation (Figure 2). Gene structure variation between the BACs may be due to the lack of EST support for gene structure and the reliance on ab initio predictions. When we compare these findings with those of Schlueter et al. (in press), these two BACs have an even higher level of sequence retention with no tandem duplications or gene losses as previously observed. The coalescence estimate between these BACs is 8.40 MYA (Table 2); a much more recent coalescence estimate than the 12.2 MYA for the HCBT duplicated BACs.
Further, this divergence estimate does not fall in either predicted duplication times (Schlueter et al., 2004; Blanc and Wolfe 2004a). It is possible that this duplication is the result of a segmental or aneuploidy event that happened independently of the paleopolyploidy events.

Current plant genome structure models have suggested that genes are arranged in two major ways. In smaller genomes such as *Arabidopsis thaliana* the genes are fairly evenly spaced along the chromosomes as "beads on a string" (TAGI, 2000). Conversely, the maize genome has been predicted to have genes existing in islands separated by repetitive sequence (SanMiguel and Bennetzen, 1998; Messing et al., 2004), although this model for the maize genome has recently been called into question with sample sequencing (Haberer et al., 2005). When all four BAC are considered, very different genome structures are seen. There are two cases of densely packed genes with almost no retroelement insertions; one BAC has a gene density similar to previously sequenced soybean BACs (Figure 2; Schlueter et al., in press; Mudge et al., 2005); finally, the last region has very few genes with large regions of retroelement insertions and is nearly twice as less dense as the other FAD2 BACs. BAC gmw1-45m6 shows very different structural dynamics than the other BACs sequenced. On average, there are nearly 2 to 3 times fewer genes per kb than in the other sequenced BACs. The genes in this region also seem to be located in more of a cluster surrounded by numerous repetitive sequences much as in the gene island model. Although there is little to no evidence for a transposon explosion in the soybean genome, this BAC would suggest that some regions might be hotspots for retroelement insertions. These varying gene densities suggest that the soybean genome does not fit a single model for genome organization and appears to be a mosaic. Although most of the gene space is probably most like gmw1-11j16, gmw1-15k6 and gmw1-105h23, there are likely parts of the genome that are more like gmw1-45m6 with sparse genes interspersed with repetitive elements. Further, gene space may not be restricted to non-repetitive regions as has previously been suggested (Lin et al., 2005). This needs to be considered as soybean genome sequencing strategies are developed as selective sequencing of proposed gene rich regions may miss a number of genes in the less densely packed regions of the genome.
Functional divergence of FAD2 genes: The FAD2 genes provide an excellent model for looking at functional divergence in a paleopolyploid genome. Previous studies have identified these genes as encoding ω-6 fatty acid desaturases that catalyze the conversion of 18:1 oleic acid to 18:2 linoleic acid in soybean (Okuley et al., 1994; Heppard et al., 1996). This conversion is of particular interest in soybean due to the growing desire for low-linoleic soybean oil (Thomas et al. 2003). Generally the structure of the five sequenced FAD2 genes in soybean shows that they are relatively similar to one another but form two distinct subgroups, the FAD2-1’s and FAD2-2’s. Originally, FAD2-1 was thought to be expressed only in developing pod tissue whereas FAD2-2 was more of a housekeeping gene. This partitioning of function is very characteristic of possible fates of duplicate genes in a paleopolyploid. Following duplication, paralogs may either retain original gene function, subfunctionalize, neofunctionalize (obtain a new function), or be silenced (Force et al., 1999). In this case, FAD2 in soybean may have become subfunctionalized, with partitioning of expression of one copy to just pods. The identification of three more copies of FAD2 raises the question as to the expression of each gene; is FAD2-1 only expressed in pod tissues. Reverse-transcriptase PCR screens using FAD2 primers showed that FAD2-1, both A and B as well as alternative transcripts, do not appear to be expressed only in pods. While their expression is primarily in developing pod tissues, both FAD2-1A and FAD2-1Balt showed expression in other tissues. These tissues also represent different developmental stages of young seedling tissues; it appears that FAD2-1Balt may be expressed earlier in the seedling followed by expression of FAD2-1A.

EST-based gene structure prediction identified at least two functional cases of alternative splicing in FAD2 genes, FAD2-1A and FAD2-1B. Alternative splicing during expression is another means for functional divergence (Su et al., 2005). Based upon the similar structure of the alternative spliced products, it appears that the FAD2-1 genes retain their alternatively spliced constructs after duplication. Conversely, FAD2-2 ESTs did not show any evidence for alternative splicing although FAD2-2 has been tandemly duplicated on gmw1-11j16. This shows that even within a gene family the mechanisms of increasing the material for functional divergence may not be the same. Reverse-transcriptase results show slight differences between the alternatively spliced constructs in both cases. The FAD2-1A
copy (containing the longer intron) is expressed in vegetative tissue. FAD2-1Balt (containing
the smaller intron) is the FAD2 copy that shows transcript accumulation in vegetative tissues.
This appears to be the only difference between the alternative spliced constructs. Similarly,
when semi-quantitative RT-PCR reactions were performed, no differences could be
identified in the developing pod tissue between the alternatively spliced products. In both
cases, when looking for changes in fold expression at each time point between developing
pods from warm conditions and cool conditions, the fold change was the same for both
constructs of FAD2-1A and FAD2-1B.

The primary objective of the semi-quantitative RT-PCR experiments was to look for
fold change differences between developing pods from warm and cool conditions. Table 4
shows that FAD2-2C is the only gene that showed significant transcript increase in
developing pods from cool conditions. This indicates that FAD2-2Calt is the most likely
candidate for being the gene that increases the pool of ω-6 fatty acid desaturases at lower
temperatures leading to an increase in linoleic acid in those seeds. This gene, however, is not
the only one to show temperature dependent changes in transcript accumulation. Both FAD2-
1A and FAD2-1B (and their respective alternative transcripts) show significant fold increases
in transcript accumulation in developing pods from warm conditions. These results are
similar to those shown with Northern blot analysis by Heppard et al. (1996). This increase in
expression may be due to the proposed instability of ω-6 fatty acid desaturases at higher
temperatures (Heppard et al., 1996; Tang et al., 2005) The semi-quantitative RT-PCR results
also suggest that there may not be a single gene controlling cool-temperature dependent
increases in expression. There may actually be a progression of FAD2 gene expression or
partitioning in the developing seeds as seen in olive (Hernandez et al. 2005). Further studies
looking at the expression of each of these genes in the various compartments of developing
pods may identify varying expression patterns of the FAD2 genes. The proposed role of
modifying genes in the conversion of oleic to linoleic acid (Alt et al., 2005) is another point
that needs to be considered.
Acknowledgments: The authors would like to thank Mary Duke and Xiaofen Liu for their sequencing expertise. We would also like to thank Jody Hayes for her skills in mapping SSR markers, Terry Olsen for his MapMaker expertise, and Greg Peiffer for his assistance in identifying BACs. Names are necessary to report factually on the available data; however, the USDA neither guarantees nor warrants the standard of the product, and the use of the name by the USDA implies no approval of the product to the exclusion of others that may also be suitable. This article is a contribution of the Corn Insect and Crop Genetics Research Unit (USDA-ARS). J.A. Schlueter was supported by a grant from the United Soybean Board.

REFERENCES


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


<table>
<thead>
<tr>
<th>BAC</th>
<th>Putative function</th>
<th>BLASTP identified homeolog</th>
<th>Positives score&lt;sup&gt;a&lt;/sup&gt;</th>
<th>E-value&lt;sup&gt;a&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>gmwl-11j16</td>
<td>KH-domain containing protein</td>
<td><em>Oryza sativa</em> XP_464444.1 (300 aa)</td>
<td>183/255</td>
<td>6e-69</td>
</tr>
<tr>
<td></td>
<td>Seed maturation protein LEA 4</td>
<td>Glycine max AAG37440.1 (173 aa)</td>
<td>173/173</td>
<td>5e-38</td>
</tr>
<tr>
<td></td>
<td>Microsomal omega-6 fatty acid desaturase (FAD2-2A)</td>
<td>Glycine max BAD89862.1 (383 aa)</td>
<td>279/368</td>
<td>2e-141</td>
</tr>
<tr>
<td></td>
<td>Microsomal omega-6 fatty acid desaturase (FAD2-2B)</td>
<td>Glycine max AAB00860.1 (383 aa)</td>
<td>371/383</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td>Putative DNA polymerase delta small subunit</td>
<td>Arabidopsis AAB88640.1 (456 aa)</td>
<td>380/445</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td>Arabidopsis At4g02400-like protein</td>
<td>Arabidopsis CAB80733.1 (870 aa)</td>
<td>38/65</td>
<td>8e-06</td>
</tr>
<tr>
<td></td>
<td>Arabidopsis-like Unknown</td>
<td>Arabidopsis NP_181741.1 (137 aa)</td>
<td>77/140</td>
<td>8e-11</td>
</tr>
<tr>
<td></td>
<td>Protein phosphatase 2C</td>
<td><em>Oryza sativa</em> XP_467189.1 (387 aa)</td>
<td>316/383</td>
<td>8e-150</td>
</tr>
<tr>
<td></td>
<td>Actin</td>
<td>Arabidopsis AAF31643.1 (377 aa)</td>
<td>376/377</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td>Putative DNA polymerase delta small subunit</td>
<td>Arabidopsis AAB88640.1 (456 aa)</td>
<td>380/445</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td>Arabidopsis-like Unknown</td>
<td>Arabidopsis NP_181741.1 (137 aa)</td>
<td>77/140</td>
<td>8e-11</td>
</tr>
<tr>
<td></td>
<td>Protein phosphatase 2C</td>
<td><em>Oryza sativa</em> XP_467189.1 (387 aa)</td>
<td>316/383</td>
<td>8e-150</td>
</tr>
<tr>
<td></td>
<td>Actin</td>
<td>Arabidopsis AAF31643.1 (377 aa)</td>
<td>376/377</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td>Putative DNA polymerase delta small subunit</td>
<td>Arabidopsis AAB88640.1 (456 aa)</td>
<td>380/445</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td>Arabidopsis-like Unknown</td>
<td>Arabidopsis NP_181741.1 (137 aa)</td>
<td>77/140</td>
<td>8e-11</td>
</tr>
<tr>
<td></td>
<td>Protein phosphatase 2C</td>
<td><em>Oryza sativa</em> XP_467189.1 (387 aa)</td>
<td>316/383</td>
<td>8e-150</td>
</tr>
<tr>
<td></td>
<td>Actin</td>
<td>Arabidopsis AAF31643.1 (377 aa)</td>
<td>376/377</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td>Putative DNA polymerase delta small subunit</td>
<td>Arabidopsis AAB88640.1 (456 aa)</td>
<td>380/445</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td>Arabidopsis-like Unknown</td>
<td>Arabidopsis NP_181741.1 (137 aa)</td>
<td>77/140</td>
<td>8e-11</td>
</tr>
<tr>
<td></td>
<td>Protein phosphatase 2C</td>
<td><em>Oryza sativa</em> XP_467189.1 (387 aa)</td>
<td>316/383</td>
<td>8e-150</td>
</tr>
<tr>
<td></td>
<td>Actin</td>
<td>Arabidopsis AAF31643.1 (377 aa)</td>
<td>376/377</td>
<td>0.0</td>
</tr>
<tr>
<td>gmwl-15k6</td>
<td>FIF0-ATPase inhibitor-like protein</td>
<td>Arabidopsis CAB86017.1 (107 aa)</td>
<td>32/38</td>
<td>0.014</td>
</tr>
<tr>
<td></td>
<td>Microsomal omega-6 fatty acid desaturase (FAD2-1A)</td>
<td>Glycine max BAD89860.1 (387 aa)</td>
<td>387/387</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td>Vesicle-associated membrane protein (ATVAMP727)</td>
<td>Arabidopsis NP_190998.1 (240 aa)</td>
<td>119/126</td>
<td>2e-54</td>
</tr>
<tr>
<td></td>
<td>Coatomer delta subunit-like unknown protein</td>
<td>Arabidopsis NP_001031836.1 (527 aa)</td>
<td>455/533</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td>HHE-domain containing protein</td>
<td>Arabidopsis CAB71003.1 (363 aa)</td>
<td>210/301</td>
<td>1e-84</td>
</tr>
<tr>
<td></td>
<td>Ribonuclease HII</td>
<td><em>Oryza sativa</em> AAX92734.1 (298 aa)</td>
<td>229/269</td>
<td>3e-112</td>
</tr>
<tr>
<td></td>
<td>Hypothetical protein</td>
<td>Cicer arietinum CAD31715.1 (121 aa)</td>
<td>68/111</td>
<td>2e-15</td>
</tr>
<tr>
<td></td>
<td>Pollen-specific protein</td>
<td>Arabidopsis NP_567560.2 (106 aa)</td>
<td>82/106</td>
<td>6e-31</td>
</tr>
<tr>
<td></td>
<td>Non-photosynthetic ferredoxin</td>
<td>Citrus sinensis CAA87068.1 (150 aa)</td>
<td>118/149</td>
<td>4e-46</td>
</tr>
<tr>
<td></td>
<td>Unknown</td>
<td><em>Oryza sativa</em> NP_915582.1 (261 aa)</td>
<td>261/280</td>
<td>1e-17</td>
</tr>
<tr>
<td></td>
<td>Tubulin folding cofactor D</td>
<td>Arabidopsis AAM22960.1 (1254 aa)</td>
<td>1017/1264</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td>Viral-response protein</td>
<td>Arabidopsis NP_175919.1 (300 aa)</td>
<td>199/280</td>
<td>7e-73</td>
</tr>
</tbody>
</table>

<sup>a</sup> Positives and E-values determined from BLASTP report from NCBI
<table>
<thead>
<tr>
<th>BAC</th>
<th>Putative function</th>
<th>BLASTP identified homeolog</th>
<th>Positives score</th>
<th>E-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACT-domain containing protein</td>
<td>Arabidopsis AAM64912.1 (301 aa)</td>
<td>207/238</td>
<td>4e-92</td>
<td></td>
</tr>
<tr>
<td>Unknown</td>
<td>Arabidopsis NP_197637.1 (1180 aa)</td>
<td>649/1309</td>
<td>2e-118</td>
<td></td>
</tr>
<tr>
<td>Lipoic acid synthetase</td>
<td>Arabidopsis AAM62684.1 (393 aa)</td>
<td>281/303</td>
<td>2e-150</td>
<td></td>
</tr>
<tr>
<td>Germin-like protein</td>
<td>Arabidopsis NP_187619.1 (227 aa)</td>
<td>53/78</td>
<td>5e-16</td>
<td></td>
</tr>
<tr>
<td>Unknown</td>
<td>Arabidopsis NP_566996.1 (379 aa)</td>
<td>175/231</td>
<td>2e-76</td>
<td></td>
</tr>
<tr>
<td>Melvonate disphosphate decarboxylase</td>
<td>Hevea brasiliensis AAL18927.1 (415 aa)</td>
<td>381/417</td>
<td>0.0</td>
<td></td>
</tr>
<tr>
<td>PPR-containing protein</td>
<td>Arabidopsis NP_181058.1 (591 aa)</td>
<td>214/343</td>
<td>2e-87</td>
<td></td>
</tr>
<tr>
<td>20S proteasome</td>
<td>Glycine max AAF70292.1 (237 aa)</td>
<td>237/237</td>
<td>2e-128</td>
<td></td>
</tr>
<tr>
<td>INDETERMINATE-related protein 9</td>
<td>Zea mays AAV51392.1 (588 aa)</td>
<td>209/282</td>
<td>3e-80</td>
<td></td>
</tr>
<tr>
<td>UDP-glucosyltransferase</td>
<td>Medicago ABE81841.1 (497 aa)</td>
<td>183/378</td>
<td>4e-51</td>
<td></td>
</tr>
<tr>
<td>Pollen-specific protein-like</td>
<td>Arabidopsis CAB78861.1 (842 aa)</td>
<td>67/91</td>
<td>4e-23</td>
<td></td>
</tr>
<tr>
<td>Unknown</td>
<td>Cicer arietinum CAD31715.1 (121 aa)</td>
<td>89/127</td>
<td>1e-15</td>
<td></td>
</tr>
<tr>
<td>Ribonuclease HII</td>
<td>Arabidopsis NP_565584.1 (296 aa)</td>
<td>44/56</td>
<td>6e-10</td>
<td></td>
</tr>
<tr>
<td>HHE-domain containing protein</td>
<td>Arabidopsis CAB71003.1 (363 aa)</td>
<td>222/356</td>
<td>3e-86</td>
<td></td>
</tr>
<tr>
<td>Coatamer</td>
<td>Arabidopsis NP_001031836.1 (527 aa)</td>
<td>457/533</td>
<td>0.0</td>
<td></td>
</tr>
<tr>
<td>Vesicle-associated membrane protein</td>
<td>Arabidopsis NP_190998.1 (240 aa)</td>
<td>223/238</td>
<td>2e-113</td>
<td></td>
</tr>
<tr>
<td>Micromosomal omega-6 fatty acid desaturase (FAD2-1B)</td>
<td>Glycine max BAD89861.1 (387 aa)</td>
<td>387/387</td>
<td>0.0</td>
<td></td>
</tr>
<tr>
<td>FIF0-ATPase inhibitor-like protein</td>
<td>Arabidopsis CAB86017.1 (107 aa)</td>
<td>44/61</td>
<td>5e-04</td>
<td></td>
</tr>
<tr>
<td>Non-photosynthetic ferredoxin</td>
<td>Citrus sinensis CAA87068.1 (150 aa)</td>
<td>118/149</td>
<td>5e-45</td>
<td></td>
</tr>
<tr>
<td>Unknown</td>
<td>Oryza sativa NP_915582.1 (261 aa)</td>
<td>185/241</td>
<td>7e-83</td>
<td></td>
</tr>
<tr>
<td>Tubulin folding cofactor D</td>
<td>Arabidopsis AAM22960.1 (1254 aa)</td>
<td>976/1250</td>
<td>0.0</td>
<td></td>
</tr>
<tr>
<td>Viral-response protein</td>
<td>Arabidopsis NP_175919.1 (300 aa)</td>
<td>201/282</td>
<td>4e-76</td>
<td></td>
</tr>
<tr>
<td>Amino acid binding</td>
<td>Arabidopsis NP_196094.1 (301 aa)</td>
<td>208/238</td>
<td>6e-94</td>
<td></td>
</tr>
<tr>
<td>Unknown</td>
<td>Arabidopsis NP_197637.1 (1180 aa)</td>
<td>586/1174</td>
<td>3e-109</td>
<td></td>
</tr>
<tr>
<td>Lipoic acid synthetase</td>
<td>Arabidopsis NP_568196.1 (394 aa)</td>
<td>302/322</td>
<td>6e-162</td>
<td></td>
</tr>
<tr>
<td>BAC</td>
<td>Putative function</td>
<td>BLASTP identified homeolog</td>
<td>Positives score</td>
<td>E-value</td>
</tr>
<tr>
<td>-----------</td>
<td>--------------------------------------------------------</td>
<td>-------------------------------------------------</td>
<td>-----------------</td>
<td>---------</td>
</tr>
<tr>
<td></td>
<td>Germin-like protein</td>
<td>Arabidopsis CAB71000.1 (408 aa)</td>
<td>40/51</td>
<td>7e-13</td>
</tr>
<tr>
<td></td>
<td>Unknown</td>
<td>Arabidopsis NP_566996.1 (379 aa)</td>
<td>268/327</td>
<td>7e-131</td>
</tr>
<tr>
<td>gmw1-45m6</td>
<td>Diphosphovalonate decarboxylase-like protein</td>
<td>Arabidopsis CAB70999.1 (413 aa)</td>
<td>65/71</td>
<td>6e-33</td>
</tr>
<tr>
<td></td>
<td>PHO1-like protein</td>
<td>Arabidopsis NP_177107.2 (777 aa)</td>
<td>202/348</td>
<td>4e-57</td>
</tr>
<tr>
<td></td>
<td>Microsomal omega-6 fatty acid desaturase (FAD2-2C)</td>
<td>Glycine max BAD89863.1 (337 aa)</td>
<td>293/293</td>
<td>6e-165</td>
</tr>
<tr>
<td></td>
<td>ADP-ribosyltransferase</td>
<td>Arabidopsis AAD20677.1 (1009 aa)</td>
<td>58/101</td>
<td>2e-05</td>
</tr>
<tr>
<td></td>
<td>Hydroxyproline-rich protein</td>
<td>Medicago ABE88601.1 (694 aa)</td>
<td>282/586</td>
<td>1e-61</td>
</tr>
<tr>
<td></td>
<td>Unknown</td>
<td>Arabidopsis NP_566314.1 (241 aa)</td>
<td>159/242</td>
<td>1e-51</td>
</tr>
<tr>
<td></td>
<td>WD-domain containing protein</td>
<td>Arabidopsis AAF06037.1 (365 aa)</td>
<td>93/119</td>
<td>1e-35</td>
</tr>
<tr>
<td></td>
<td>WD-domain containing protein</td>
<td>Arabidopsis AAF06037.1 (365 aa)</td>
<td>172/230</td>
<td>6e-64</td>
</tr>
</tbody>
</table>
Table 2. Similarity between homeologous BACs gmr1-15k6 and gmr1-105h23

<table>
<thead>
<tr>
<th>Putative function</th>
<th>Nucleotide Identity</th>
<th>Protein Identity</th>
<th>Protein similarity</th>
<th>Ks†</th>
<th>Ka‡</th>
<th>Date (Mya)§</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unknown 5</td>
<td>88.4</td>
<td>81.6</td>
<td>86.2</td>
<td>0.1280</td>
<td>0.0880</td>
<td>10.5</td>
</tr>
<tr>
<td>FAD2-1</td>
<td>94.6</td>
<td>93.5</td>
<td>96.9</td>
<td>0.1048</td>
<td>0.0329</td>
<td>8.59</td>
</tr>
<tr>
<td>Vesicle-associated membrane protein</td>
<td>97.3</td>
<td>97.9</td>
<td>98.7</td>
<td>0.0749</td>
<td>0.0116</td>
<td>6.14</td>
</tr>
<tr>
<td>Coatamer</td>
<td>97.2</td>
<td>97.3</td>
<td>98.3</td>
<td>0.0611</td>
<td>0.0131</td>
<td>5.01</td>
</tr>
<tr>
<td>HHE-domain containing</td>
<td>93.4</td>
<td>93.4</td>
<td>95.2</td>
<td>0.1360</td>
<td>0.0295</td>
<td>11.1</td>
</tr>
<tr>
<td>Ribonuclease HII</td>
<td>88.2</td>
<td>79.7</td>
<td>81.4</td>
<td>0.0268</td>
<td>0.0557</td>
<td>2.20</td>
</tr>
<tr>
<td>Unknown 70</td>
<td>89.4</td>
<td>90.4</td>
<td>91.5</td>
<td>0.2361</td>
<td>0.0257</td>
<td>19.4</td>
</tr>
<tr>
<td>Pollen-specific</td>
<td>93.5</td>
<td>91.2</td>
<td>93.4</td>
<td>0.1065</td>
<td>0.0514</td>
<td>8.73</td>
</tr>
<tr>
<td>Ferredoxin</td>
<td>94.8</td>
<td>94.8</td>
<td>98.7</td>
<td>0.1345</td>
<td>0.0247</td>
<td>11.0</td>
</tr>
<tr>
<td>Unknown 110</td>
<td>90.1</td>
<td>84.3</td>
<td>82.2</td>
<td>0.0735</td>
<td>0.0337</td>
<td>6.02</td>
</tr>
<tr>
<td>B-tubulin cofactor D</td>
<td>91.6</td>
<td>88.5</td>
<td>86.8</td>
<td>0.1012</td>
<td>0.0215</td>
<td>8.30</td>
</tr>
<tr>
<td>Viral-response protein</td>
<td>93.0</td>
<td>91.3</td>
<td>93.6</td>
<td>0.1007</td>
<td>0.0275</td>
<td>8.25</td>
</tr>
<tr>
<td>ACT-domain protein</td>
<td>97.4</td>
<td>97.5</td>
<td>98.6</td>
<td>0.0557</td>
<td>0.0115</td>
<td>4.57</td>
</tr>
<tr>
<td>Unknown 130</td>
<td>69.7</td>
<td>66.8</td>
<td>69.0</td>
<td>0.1020</td>
<td>0.0533</td>
<td>8.36</td>
</tr>
<tr>
<td>Lipoic acid synthase</td>
<td>91.9</td>
<td>90.9</td>
<td>92.3</td>
<td>0.1129</td>
<td>0.0260</td>
<td>9.25</td>
</tr>
<tr>
<td>Germin-like protein</td>
<td>95.9</td>
<td>96.6</td>
<td>97.8</td>
<td>0.1207</td>
<td>0.0156</td>
<td>9.89</td>
</tr>
<tr>
<td>Unknown 160</td>
<td>69.8</td>
<td>66.8</td>
<td>68.9</td>
<td>0.1328</td>
<td>0.0123</td>
<td>10.9</td>
</tr>
<tr>
<td>Melvonate diphosphate decarboxylase</td>
<td>96.1</td>
<td>97.2</td>
<td>97.2</td>
<td>0.1013</td>
<td>0.0532</td>
<td>8.30</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td><strong>90.7</strong></td>
<td><strong>88.9</strong></td>
<td><strong>90.4</strong></td>
<td><strong>0.1061</strong></td>
<td><strong>0.0326</strong></td>
<td><strong>8.70</strong></td>
</tr>
</tbody>
</table>

† Ks, synonymous distance between retained homeologs
‡ Ka, nonsynonymous distance between retained homeologs
§ See Table 4 for identity and similarity information for the HCBT genes
<table>
<thead>
<tr>
<th>Primer name</th>
<th>Primer sequence (5'-3')</th>
<th>Primer Length</th>
<th>Primer Location</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>gmwl-15k6</td>
<td>GTT CCC AAG AGT ATA AAA CTG C</td>
<td>22</td>
<td>5553-5574</td>
<td>843</td>
</tr>
<tr>
<td>FAD2-1A U</td>
<td>CAC AAC ATC ATC AAC CCA TTG</td>
<td>21</td>
<td>6376-6396</td>
<td></td>
</tr>
<tr>
<td>FAD2-1A L</td>
<td>GAT TTC CTG AAG GCT TAG GTG</td>
<td>21</td>
<td>5717-5737</td>
<td>452</td>
</tr>
<tr>
<td>FAD2-1A_alt U</td>
<td>GAA AGC AGT GTG GTG GAA TTG</td>
<td>21</td>
<td>6149-6169</td>
<td></td>
</tr>
<tr>
<td>FAD2-1A_alt L</td>
<td>GAT ACT ACA AGC CAC TAG GCA T</td>
<td>22</td>
<td>91152-91173</td>
<td>821</td>
</tr>
<tr>
<td>FAD2-1B U</td>
<td>TGA GTG AAC GGT CAA ACC CAT A</td>
<td>22</td>
<td>91952-91973</td>
<td></td>
</tr>
<tr>
<td>FAD2-1B L</td>
<td>GAT TTC TTG AAG GCT TAT GGT AT</td>
<td>23</td>
<td>91292-91314</td>
<td>480</td>
</tr>
<tr>
<td>FAD2-1B_alt U</td>
<td>AAG GTC ATA AAC AAG ATA GGA C</td>
<td>22</td>
<td>91751-91772</td>
<td></td>
</tr>
<tr>
<td>FAD2-1B_alt L</td>
<td>CTT CCA CAA CCT TCC TCA TC</td>
<td>20</td>
<td>14964-14713</td>
<td>335</td>
</tr>
<tr>
<td>FAD2-2A U</td>
<td>GGA TTG TTT AGG TAT TTA GAT AG</td>
<td>23</td>
<td>15277-15299</td>
<td></td>
</tr>
<tr>
<td>FAD2-2A L</td>
<td>AGT ATG CTA TGG CCT TTT CTG</td>
<td>21</td>
<td>23801-23821</td>
<td>494</td>
</tr>
<tr>
<td>FAD2-2B U</td>
<td>ACC ATC CCA CCT GAC GGA C</td>
<td>19</td>
<td>24277-24295</td>
<td></td>
</tr>
<tr>
<td>FAD2-2B L</td>
<td>ACC TCA CCA TAG CCT TCT AC</td>
<td>20</td>
<td>54676-54695</td>
<td>336</td>
</tr>
<tr>
<td>FAD2-2C U</td>
<td>TAT TTA GAG TAC CAC TTG ATA CA</td>
<td>23</td>
<td>54990-55012</td>
<td></td>
</tr>
</tbody>
</table>

† Length of primer in base pairs
‡ Location of primer within BAC sequence, relative to 5' end of BAC
§ Size of PCR amplicon
<table>
<thead>
<tr>
<th>Gene</th>
<th>Tissue</th>
<th>Ct value(^\dagger)</th>
<th>Change in Ct value(^\dagger)</th>
<th>Fold change in expression(^\dagger)</th>
</tr>
</thead>
<tbody>
<tr>
<td>FAD2-1A</td>
<td>6-10 DAF</td>
<td>Warm 35.5</td>
<td>0.9</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Cool 34.6</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>6-10 DAF</td>
<td>Warm 34.5</td>
<td>-0.2</td>
<td>-12.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Cool 34.7</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>13 DAF</td>
<td>Warm 28.9</td>
<td>-6.2</td>
<td>-12.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Cool 35.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>17 DAF</td>
<td>Warm 28.1</td>
<td>-7.1</td>
<td>-14.2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Cool 35.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>19 DAF</td>
<td>Warm 30.1</td>
<td>-0.2</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Cool 30.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>21 DAF</td>
<td>Warm 28.3</td>
<td>-0.6</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Cool 28.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>26 DAF</td>
<td>Warm 27.4</td>
<td>-2.9</td>
<td>-5.8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Cool 30.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FAD2-1B</td>
<td>6-10 DAF</td>
<td>Warm 39.0</td>
<td>0.9</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Cool 38.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>6-10 DAF</td>
<td>Warm 38.8</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Cool 38.8</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>13 DAF</td>
<td>Warm 33.9</td>
<td>-4.2</td>
<td>-8.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Cool 38.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>17 DAF</td>
<td>Warm 32.5</td>
<td>-5.6</td>
<td>-11.2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Cool 38.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>19 DAF</td>
<td>Warm 35.7</td>
<td>1.1</td>
<td>2.2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Cool 34.6</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>21 DAF</td>
<td>Warm 34.5</td>
<td>0.8</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Cool 33.7</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>26 DAF</td>
<td>Warm 33.2</td>
<td>-1.7</td>
<td>-3.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Cool 34.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FAD2-2B</td>
<td>6-10 DAF</td>
<td>Warm 32.3</td>
<td>-0.4</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Cool 32.7</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\(^\dagger\) The Ct value is the threshold where product fluorescence levels are statistically higher than background fluorescence.

\(^\ddagger\) Change in Ct value is relative to developing pods in warm conditions.

\(^\S\) The fold change in expression is based upon a 2-fold change relative to one degree in the Ct value.
Table 4. Semi-quantitative reverse-transcriptase PCR results

<table>
<thead>
<tr>
<th>Gene</th>
<th>Tissue</th>
<th>Ct value(^{\dagger})</th>
<th>Change in Ct value(^{\dagger})</th>
<th>Fold change in expression(^{\dagger})</th>
</tr>
</thead>
<tbody>
<tr>
<td>6-10 DAF</td>
<td>Warm</td>
<td>33.1</td>
<td>0.7</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cool</td>
<td>32.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>13 DAF</td>
<td>Warm</td>
<td>32.3</td>
<td>0.3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cool</td>
<td>32.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>17 DAF</td>
<td>Warm</td>
<td>32.1</td>
<td>-0.2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cool</td>
<td>32.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>19 DAF</td>
<td>Warm</td>
<td>33.1</td>
<td>0.9</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cool</td>
<td>32.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>21 DAF</td>
<td>Warm</td>
<td>32.7</td>
<td>0.4</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cool</td>
<td>32.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>26 DAF</td>
<td>Warm</td>
<td>32.5</td>
<td>-0.6</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cool</td>
<td>33.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FAD2-2C</td>
<td>6-10 DAF</td>
<td>33.4</td>
<td>1.3</td>
<td>2.6</td>
</tr>
<tr>
<td></td>
<td>Warm</td>
<td>32.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cool</td>
<td>32.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6-10 DAF</td>
<td>Warm</td>
<td>32.6</td>
<td>-1.0</td>
<td>-2.0</td>
</tr>
<tr>
<td></td>
<td>Cool</td>
<td>33.6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>13 DAF</td>
<td>Warm</td>
<td>34.6</td>
<td>2</td>
<td>4.0</td>
</tr>
<tr>
<td></td>
<td>Cool</td>
<td>32.6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>17 DAF</td>
<td>Warm</td>
<td>33.9</td>
<td>0.5</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cool</td>
<td>33.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>19 DAF</td>
<td>Warm</td>
<td>36.4</td>
<td>4</td>
<td>8.0</td>
</tr>
<tr>
<td></td>
<td>Cool</td>
<td>32.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>21 DAF</td>
<td>Warm</td>
<td>35.6</td>
<td>2.7</td>
<td>5.4</td>
</tr>
<tr>
<td></td>
<td>Cool</td>
<td>32.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>26 DAF</td>
<td>Warm</td>
<td>34.8</td>
<td>2.2</td>
<td>4.4</td>
</tr>
<tr>
<td></td>
<td>Cool</td>
<td>32.6</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Figure 1. Graphical representation of linkage groups I and O based upon the soybean composite map (www.soybase.org). The BACs gmw1-15k6 (FAD2-1A) and gmw1-105h23 (FAD2-1B) mapped to these linkage groups in the Glycine max A81-356022 X Glycine soja PI 468.916 mapping populations. Their locations relative to other markers are shown. Lines between markers show syntenic markers.
Figure 2. Predicted gene structures and retrotransposon insertions on soybean BACs gmw1-11j16, gmw1-15k6, gmw1-105h23 and gmw1-45m6. Each colored block arrow on the line represents a gene and grey boxes between genes show homeologs. Black arrowheads represent full-length predicted retrotransposon elements. Grey arrowheads represent fragmented remnants of retrotransposons.
Figure 3. VISTA based identity plots between gmw1-15k6 and gmw1-105h23. The plots above and below the representative BAC structures show the relative nucleotide identity between the two BACs. The light purple boxes on top of the VISTA plots correspond to exon positions.
Figure 4. RT-PCR amplification for FAD2-1A, FAD2-1Aalt, FAD2-1B, FAD2-1Balt, FAD2-2B, and FAD2-2C. Positive control reactions used tubulin56. Plus’s indicated amplification of that transcript in a particular tissue and minus’s indicate no amplification. Tissue types are as follows: 6-10- developing pods 6-10 days after flowering (DAF); 13- developing pods 13 DAF; 17- developing pods 17 DAF; 19- developing pods 19 DAF; 21- developing pods 21 DAF; 26- developing pods 26 DAF; C3- cotyledons 3 days after emergence (DAE); R3- roots 3 DAE; FU- furled unifoliolate 3 DAE; UU- unfurled unifoliolate 4 DAE; C7- cotyledons 7 DAE; R8- roots 8 DAE; FT- furled trifoliolate 11 DAE; UT- unfurled trifoliolate 15 DAE; P- pods 76 DAE. All samples with a “w” were grown under the warm conditions and “c” for cool conditions. All samples with a –RT are controls with no reverse transcriptase for mRNA amplification, but still contain Taq DNA polymerase. The final sample in each set of reactions is a Williams 82 DNA based positive control.
Appendices
Appendix 1. Homeolog primers from gmwl-15k6 and gmwl-105h23

<table>
<thead>
<tr>
<th>Primer name</th>
<th>Primer sequence (5'-3')</th>
<th>Primer Length</th>
<th>Primer Location</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>gmwl-15k6 coatomer U</td>
<td>TTTTGGATTGACTTCTGATGTC</td>
<td>22</td>
<td>18054-18075</td>
<td>402</td>
</tr>
<tr>
<td>gmwl-15k6 coatomer L</td>
<td>AAGTGATAGGATCGGTAAGTTG</td>
<td>22</td>
<td>17673-17694</td>
<td></td>
</tr>
<tr>
<td>gmwl-105h23 coatomer U</td>
<td>CGCTTTATGGCAACTGATACTT</td>
<td>21</td>
<td>100085-100106</td>
<td>463</td>
</tr>
<tr>
<td>gmwl-105h23 coatomer L</td>
<td>GGCAAGTACAACCTTGATCAGA</td>
<td>21</td>
<td>99643-99663</td>
<td></td>
</tr>
<tr>
<td>gmwl-15k6 ferredoxin U</td>
<td>ACACTTACATTCTGGATGCAG</td>
<td>21</td>
<td>48172-48192</td>
<td>396</td>
</tr>
<tr>
<td>gmwl-15k6 ferredoxin L</td>
<td>GCCTATATCCTGAATGAATCGA</td>
<td>22</td>
<td>47796-47817</td>
<td></td>
</tr>
<tr>
<td>gmwl-105h23 ferredoxin U</td>
<td>TAGGAGATCATGTGCTCTTAC</td>
<td>21</td>
<td>75026-75046</td>
<td>564</td>
</tr>
<tr>
<td>gmwl-105h23 ferredoxin L</td>
<td>AGCCTATTCTTGATGAAACCA</td>
<td>22</td>
<td>75569-75590</td>
<td></td>
</tr>
<tr>
<td>gmwl-15k6 germin U</td>
<td>CCAACAAATATCTTCAATAC</td>
<td>22</td>
<td>108845-108866</td>
<td>385</td>
</tr>
<tr>
<td>gmwl-15k6 germin L</td>
<td>AGTGTGGCAGATGACACCA</td>
<td>20</td>
<td>109211-109230</td>
<td></td>
</tr>
<tr>
<td>gmwl-105h23 germin U</td>
<td>ACTCAACATCTTTTGGATAC</td>
<td>22</td>
<td>7363-7384</td>
<td>355</td>
</tr>
<tr>
<td>gmwl-105h23 germin L</td>
<td>ACCAAAATCTGCTCAGACAAA</td>
<td>21</td>
<td>7029-7049</td>
<td></td>
</tr>
<tr>
<td>HHE U gmwl-15k6</td>
<td>CTTTTCTACTCGCCTCAAGT</td>
<td>20</td>
<td>28953-28972</td>
<td>532</td>
</tr>
<tr>
<td>HHE L gmwl-105h23</td>
<td>CACTTAACATCTTTTGGAGAA</td>
<td>22</td>
<td>29464-29485</td>
<td></td>
</tr>
<tr>
<td>HHE U gmwl-105h23</td>
<td>CTACCTCTCATGAATGGCATA</td>
<td>21</td>
<td>119381-119401</td>
<td>417</td>
</tr>
<tr>
<td>HHE L gmwl-105h23</td>
<td>CCTCCATTAGCGGAAGTAAAC</td>
<td>20</td>
<td>119779-119798</td>
<td></td>
</tr>
<tr>
<td>HHE U gmwl-15k6</td>
<td>CATTGATGGTATTTGAGCAT</td>
<td>23</td>
<td>106201-106223</td>
<td>639</td>
</tr>
<tr>
<td>HHE L gmwl-15k6</td>
<td>TCCCCAGCTCTGTACGATG</td>
<td>19</td>
<td>106822-106840</td>
<td></td>
</tr>
</tbody>
</table>

1 Length of primer in base
2 Location of primer within BAC sequence, relative to 5' end of
3 Size of PCR amplicon
gmw1-105h23 lipoic U
gmw1-105h23 lipoic L
gmw1-15k6 melvionate U
gmw1-15k6 melvionate L
gmw1-105h23 melvionate U
gmw1-105h23 melvionate L
gmw1-15k6 unknown110 U
gmw1-15k6 unknown110 L
gmw1-105h23 unknown110 U
gmw1-105h23 unknown110 L
gmw1-105h23 unknown130 U
gmw1-15k6 unknown130 L
gmw1-105h23 unknown130 U
gmw1-105h23 unknown130 L
gmw1-15k6 VAMP U
gmw1-15k6 VAMP L
gmw1-105h23 VAMP U
gmw1-105h23 VAMP L
gmw1-15k6 viral U
gmw1-15k6 viral L
gmw1-105h23 viral U
gmw1-105h23 viral L

ATCGACGTTGATATTCTGACACT 23 10207-10229 638
TCCCCAGCCCTGTACGAAG 19 9591-9609
CTGTCTAAATTGACGTAATAG 22 114010-114031 500
CTTGAAAGGGTCCCTCCAATC 20 113531-113550
GTTCATAATTGACGCAACAG 20 1455-1474 524
ATGCCACGAGTAGCATGTG 19 1961-1979
TGCCCTGCCACTTCCAATGC 19 95948-95967 309
GTTCGAGGTGTCTTGCACAC 20 96238-96257
TCCTGCGCTGCCACTTTTAC 20 20583-20602 302
GTTCGAGGTGGCCTTGAAGC 20 20300-20319
GTTGTGCAAAGCTTATTCTTC 21 111997-112017 256
ACATCAACTCCCCCTCAAACA 20 111761-111780
CATGATAGAAACCTGGCTTG 21 76132-76153 563
GAACATCAACTCCCTTCAAAATA 22 76673-76695
TGGAAGAAGAGATGAAACTGCT 21 11667-11687 388
ACAGACTTAAACTCTCCAGTTC 22 12034-12055
GAGAGAAGATGAAACTGTTG 20 97732-97751 369
CTTAAACTTCTTATGGCTCTA 21 98081-98101
GCACAATCTTCTTGCCCATG 20 74806-74825 508
ATGATCAGCATCTAACCA 20 75295-75314
GCACAATCTTCTTGCCCTCG 20 93967-93987 288
ATCATCAGCATCTCAGTA 20 94235-94255
<p>| | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>gmwl-15k6</td>
<td>ACT U</td>
<td>GATGTAACATTGATGTTGAAT</td>
<td>22</td>
<td>77735-77756</td>
<td>508</td>
</tr>
<tr>
<td>gmwl-15k6</td>
<td>ACT L</td>
<td>AACTGGGGAATTGGTATAAG</td>
<td>20</td>
<td>77248-77267</td>
<td></td>
</tr>
<tr>
<td>gmwl-105h23</td>
<td>ACT U</td>
<td>GATGTAACATTGATGTTGAGT</td>
<td>22</td>
<td>41946-41967</td>
<td>504</td>
</tr>
<tr>
<td>gmwl-105h23</td>
<td>ACT L</td>
<td>AACTGGCGATTGGTATAGG</td>
<td>20</td>
<td>42431-42450</td>
<td></td>
</tr>
<tr>
<td>unknown 70</td>
<td>U</td>
<td>CAATTGCGTCTCACCACATC</td>
<td>19</td>
<td>59965-59983</td>
<td>393</td>
</tr>
<tr>
<td>gmwl-15k6</td>
<td>L</td>
<td>GAAAATGGGCTTGGTGGAGG</td>
<td>19</td>
<td>59590-59608</td>
<td></td>
</tr>
<tr>
<td>unknown 70</td>
<td>L</td>
<td>CTCATTTCTAGTCTCAATCGCAT</td>
<td>23</td>
<td>62396-62418</td>
<td>372</td>
</tr>
<tr>
<td>unknown 70</td>
<td></td>
<td>CCACCTAACAGCGGATTGAT</td>
<td>21</td>
<td>62748-62768</td>
<td></td>
</tr>
</tbody>
</table>
Appendix 2. Reverse-transcriptase results for homeologs between gmw1-15k6 and gmw1-105h23

<table>
<thead>
<tr>
<th>Primer</th>
<th>Warm pod tissue</th>
<th>Cool pod tissue</th>
<th>C3†</th>
<th>R3</th>
<th>FU</th>
<th>UU</th>
<th>C7</th>
<th>R8</th>
<th>FT</th>
<th>UT</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>15k6-hhe</td>
<td>6-10</td>
<td>6-10 13 17 19 26</td>
<td>6-10</td>
<td>6-10 13 17 19 26</td>
<td>C3†</td>
<td>R3</td>
<td>FU</td>
<td>UU</td>
<td>C7</td>
<td>R8</td>
<td>FT</td>
</tr>
<tr>
<td>105h23-hhe</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15k6-melvonate</td>
<td></td>
<td></td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>X</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>105h23-melvonate</td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>X</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>15k6-vamp</td>
<td></td>
<td></td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>X</td>
<td>x</td>
</tr>
<tr>
<td>105h23-vamp</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15k6-unknown130</td>
<td></td>
<td></td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>105h23-unknown130</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15k6-coatomer</td>
<td></td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>X</td>
</tr>
<tr>
<td>105h23-coatomer</td>
<td></td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>15k6-ferredoxin</td>
<td></td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>105h23-ferredoxin</td>
<td></td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>15k6-germin</td>
<td></td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>105h23-germin</td>
<td></td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>15k6-lipoic</td>
<td></td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>105h23-lipoic</td>
<td></td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>15k6-act</td>
<td></td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>105h23-act</td>
<td></td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>15k6-viral</td>
<td></td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>105h23-viral</td>
<td></td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>15k6-unknown110</td>
<td></td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>105h23-unknown110</td>
<td></td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>15K6-unknown70</td>
<td></td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>105h23-unknown70</td>
<td></td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
</tbody>
</table>

*Capitol X’s for strong amplification, small x’s for weak amplification.

†C3 - cotyledons 3 DAE; R3 - roots 3DAE; FU - furled unifoliate; UU - unfurled unifoliate; C7 - cotyledons 7 DAE; R8 - roots 8 DAE; FT - furled trifoliolate; UT - unfurled trifoliolate; F - flowers
CHAPTER 5: CONCLUSIONS

Conclusions

Studies of the soybean genome have long suggested that it is not a simple diploid and consists of many large gene families (Lee and Verma 1984; Hightower and Meagher 1985; Grandbastien et al., 1986; Shoemaker et al., 2002). These conclusions have been drawn from various sources including cytogenetic studies, genetic mapping, BAC-based hybridization projects and BAC fingerprinting (Lackey 1980; Walling et al., 2006; Shoemaker et al., 1996; Marek et al., 2001; Yan et al., 2004). All of these have tried to further dissect the structure of the soybean genome, especially the duplicated nature. The research presented in this dissertation was done in an effort to build upon these investigations and further understand both the structural and functional divergence of the paleopolyploid soybean genome.

The existence of duplicate genes in the soybean genome has long been accepted. However, previous studies had not surveyed the extent of duplicate gene conservation. Expressed sequence tag (EST) projects provided an excellent resource for identification of retained transcribed duplicate genes in soybean as well as *Medicago*, rice, maize, barley, sorghum, tomato and potato. In this dissertation we identified 1392 duplicate pairs across the eight surveyed species. Through determining mixtures of Ln-normal distributions within each species based upon similar synonymous distances between pairs evidence for major genome events, likely genome duplications, was observed in all of the species. These results show that most all plant species likely have a polyploid history and that conserved transcribed duplicate genes can be identified in each of these species. We further investigated the selection pressures acting upon these duplicates by comparing nonsynonymous distance to synonymous distances. By far, most all of the genes are under purifying or negative selection. Finally, because of the range of libraries that the EST collections were derived from, an analysis comparing expression differences of duplicate genes versus their divergence was conducted. Results of these studies showed that retained duplicate genes seem to fit all models of duplicate gene retention. There does appear to be some duplicate pairs with identical expression profiles while other pairs have completely different profiles.
that might compliment one another. Most of the duplicate pairs have uncorrelated expression, meaning that there have been changes in expression between pairs that shows neither exactly the same expression profile nor completely partitioned expression.

Based upon the identification of duplicate genes in soybean, the next logical step was to investigate the structure of potential homeologous regions anchored by duplicate genes. Two BACs were identified and fully sequenced that were anchored by N-hydroxycinnamoyl/benzoyltransferase (HCBT) genes. Analysis of these regions showed that homeologous regions in soybean have extensive sequence conservation, both in gene order and orientation, although evidence for gene loss/gain and tandem duplication was also seen. An extensive analysis of the potential gene functions found that most of the homeologs are involved in transcription/DNA binding or cellular signaling. These findings suggested that there might be preferential retention of duplicate genes as had previously been hypothesized by Blanc and Wolfe (2004).

Similarly, four BACs anchored by \(\omega-6\) fatty acid desaturase genes, FAD2, were sequenced to completion. These regions provided another glimpse at the structure of the soybean genome. Two BACs seemed to support the findings from the HCBT genes in that homeologous regions of the soybean genome are highly conserved. In fact, these two FAD2 BACs showed even more genomic conservation than the HCBT BACs, with the exception of a large inverted region. The other FAD2 regions provide a very different picture as to soybean genome structure, they have only the FAD2 genes in common between all the FAD2 BACs. One regions contains numerous retrotransposon insertions and has two to three times lower gene density than the other regions. These results taken together suggest that the soybean genome is a mosaic. Some regions will have more of a polyploid nature with highly conserved duplicate sequences while other regions will seem diploid in nature with only a single conserved duplicate gene, or none at all.

An analysis of the functional divergence across a varied to developmental stages and tissues via reverse-transcriptase PCR of duplicate genes was performed for all six of the BACs. Homeologs from the HCBT BACs showed that expression differences between retained duplicates are less than expected with most genes having identical profiles as well as many of the genes showing transcript accumulation in all of the tissues studied. Between
duplicate bHLH-like encoding genes one copy showed transcript accumulation in unfurled unifoliolates relative to the other copy. Similarly, four of the HCBT genes seemed to show some evidence for changes in expression with two copies appearing to be expression only in roots. Similar results were seen between homeologs from the highly similar FAD2 BACs. In light of the agronomic impact of the FAD2 genes, semi-quantitative RT-PCR was performed. The desire was to screen the FAD2 genes for temperature dependent transcript accumulation in developing pods. Results showed that one copy, FAD2-2C, might be a candidate for the gene responsible for increases in ω-6 fatty acid desaturase conversion of oleic acid to linoleic acid in developing pods grown under cool conditions.

**Future research**

The research presented here provides a resource for continuing studies on the structure of the soybean genome. Recently, soybean has been chosen for whole-genome shotgun sequencing. In light of these findings, portions of the soybean genome will likely be shown to exist in duplicate. It will be intriguing to see from the duplicate regions sequenced in this analysis to what extent homeology will be conserved and at what point along a chromosome it will begin to degenerate. Identification and characterization of further duplicate regions will provide an excellent look at the process of diploidization in a paleopolyploid. In addition, with the sequencing of the model legumes, *Medicago truncatula* and *Lotus japonicus*, this will allow for syntenic studies between potential duplicate regions in three legumes.

Each of the identified duplicate genes from the EST study provides a point to continue studying changes in expression between duplicate genes in a polyploid. Further, these duplicate genes can be used to continue targeted sequencing of duplicate genome regions as has been done in this dissertation. These duplicate genes would also be an excellent resource to create macroarrays looking specifically at expression differences between duplicate genes in a variety of tissues.
Selective sequencing of duplicate regions identified a number of interesting genes. HCBT has a role in phytoalexin biosynthesis and signaling during disease response. Each of these gene copies could be surveyed to see what pathogen might elicit expression of each copy. Similarly, the FAD2 genes warrant further exploration especially in light of their agronomic importance. The ferredoxin gene that was shown to make within an iron efficiency QTL is another gene of interest. This gene might be a candidate for resistance to iron deficiency chlorosis in soybean.

References


ACKNOWLEDGMENTS

I would like to thank Dr. Randy Shoemaker for his constant encouragement, support, guidance and advice through my career. Without his support I would not be where I am today. I would also like to thank each of my committee members, Dr. Volker Brendel, Dr. Patrick Schnable, Dr. E. Charles Brummer, Dr. Jonathan Wendel, and Dr. Phillip Dixon for their guidance and for constantly challenging me. I am also grateful for the various opportunities presented to me in graduate school as well as the collaborators that I have been honored to work with. Finally, I want to thank Shannon Schluetter my husband who has been by my side through all of this and my son Harmon for simply being who he is.