Investigation of genome duplication by polyploidization and diploidization in Glycine max L Merr

Jennifer Marie Lee
Iowa State University

Follow this and additional works at: https://lib.dr.iastate.edu/rtd

Part of the Genetics Commons

Recommended Citation
Lee, Jennifer Marie, "Investigation of genome duplication by polyploidization and diploidization in Glycine max L Merr " (2000). Retrospective Theses and Dissertations. 12698.
https://lib.dr.iastate.edu/rtd/12698
INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

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 bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI 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.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

Bell & Howell Information and Learning
300 North Zeib Road, Ann Arbor, MI 48106-1346 USA
800-521-0600

UMI®
Investigation of genome duplication by polyploidization and diploidization in

Glycine max L. Merr.

by

Jennifer Marie Lee

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

DOCTOR OF PHILOSOPHY

Major: Genetics

Major Professor: Randy C. Shoemaker

Iowa State University

Ames, Iowa

2000
Graduate College
Iowa State University

This is to certify that the Doctoral dissertation of

Jennifer Marie Lee

has met the dissertation requirements of Iowa State University

Signature was redacted for privacy.

Major Professor

Signature was redacted for privacy.

For the Major Program

Signature was redacted for privacy.

For the Graduate College
DEDICATION

This dissertation is dedicated to my parents Tom and Francy Lee who have always encouraged and supported all of my scientific and academic pursuits.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>ABSTRACT</th>
<th>vi</th>
</tr>
</thead>
<tbody>
<tr>
<td>CHAPTER 1. GENERAL INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>Introduction</td>
<td>1</td>
</tr>
<tr>
<td>Dissertation Organization</td>
<td>3</td>
</tr>
<tr>
<td>Literature review</td>
<td>5</td>
</tr>
<tr>
<td>References</td>
<td>17</td>
</tr>
<tr>
<td>CHAPTER 2. MAPPING OF DUPLICATE GENES IN SOYBEAN</td>
<td>26</td>
</tr>
<tr>
<td>Abstract</td>
<td>26</td>
</tr>
<tr>
<td>Introduction</td>
<td>27</td>
</tr>
<tr>
<td>Materials and Methods</td>
<td>30</td>
</tr>
<tr>
<td>Results</td>
<td>35</td>
</tr>
<tr>
<td>Discussion</td>
<td>41</td>
</tr>
<tr>
<td>References</td>
<td>46</td>
</tr>
<tr>
<td>CHAPTER 3. SYNTENIC RELATIONSHIP WITHIN LEGUMES AND THEIR COMPARISON TO</td>
<td>52</td>
</tr>
<tr>
<td>ARABIDOPSIS THALIANA</td>
<td></td>
</tr>
<tr>
<td>Abstract</td>
<td>52</td>
</tr>
<tr>
<td>Introduction</td>
<td>53</td>
</tr>
<tr>
<td>Materials and Methods</td>
<td>58</td>
</tr>
<tr>
<td>Results</td>
<td>62</td>
</tr>
<tr>
<td>Discussion</td>
<td>71</td>
</tr>
<tr>
<td>Acknowledgments</td>
<td>75</td>
</tr>
</tbody>
</table>
ABSTRACT

Glycine max L. Merr. (soybean) is an important agronomic crop grown throughout the world. Investigations of the soybean genome reveal that it is an ancient polyploid. We would expect duplicate functional genes in soybean to have arisen from the polyploidization event and to be found in homoeologous regions in the soybean genome. We mapped the duplicate functional genes *Pa1* and *Pa2* and found they were not located in directly homoeologous regions as might be expected. To determine the evolutionary origins of these genes the soybean genome was examined for additional duplications by comparisons with the related legumes *Vigna radiata* and *Phaseolus vulgaris*. These comparisons disclosed two regions in *G. max* showing homology to a single region in each *V. radiata* and *P. vulgaris*.

Comparisons of TBLASTX searches using soybean RFLP sequences and *Arabidopsis thaliana* genomic sequences led to the identification of regions of conservation between *G. max* and *A. thaliana*. Detection of conserved regions also allowed us to determine that two homoeologous blocks in soybean showed homology to the same region in *A. thaliana*. The physical relationship of the duplicate genes mapped in soybean appears to be the result of two rounds of duplication; a recent duplication detected by comparison of duplicate RFLP markers in soybean and an earlier event only detected by the use of *A. thaliana* as a bridging species. These results emphasize the importance in interspecies comparisons to detect evolutionary relationships of distantly related genomic segments.
CHAPTER 1. GENERAL INTRODUCTION

Introduction

Soybean (*Glycine max* L. Merr.) is believed to have been domesticated in China between 2800 BC and 1100 BC (Hymowitz, 1970). Since then soybean and its products have gained worldwide acceptance (Hymowitz and Newell, 1981). Its immense and increasing value has made soybean a species of intense investigation. Efforts to improve the species by increasing yield or content of various proteins and oils have been quite successful. In addition to investigations in protein and oil, researchers have pursued studies which will lead to understanding the evolution and organization of the soybean genome.

Previous genomic investigations have found the soybean genome to be highly duplicated (Shoemaker *et al.* 1996). Soybean has nearly twice the number of chromosomes and double the amount of DNA as the most closely related legumes which have been studied (Lackey, 1980; Arumuganathan and Earle, 1991). Numerous homologous regions can be found in the soybean genome and are identified by duplicated RFLP markers (Shoemaker *et al.* 1996). Soybean behaves functionally as a diploid. However, numerous genetic studies led researchers to suggest soybean has undergone at least one polyploidization even in its evolutionary history (Lackey, 1980; Zhu *et al.* 1994; Shoemaker *et al.* 1996). Since this polyploidization event soybean has reverted back to a diploid state through the processes of diploidization (Lackey, 1980; Zhu *et al.* 1994; Shoemaker *et al.* 1996).
Polyploidy is extremely prevalent in the plant kingdom with an estimated 70% of plants thought to be polyploid (Lewis, 1980). Many polyploids are, as soybean, 'ancient' polyploids which now behave as diploids. Plants may be identified as ancient polyploids, or paleopolyploids, utilizing methods such as comparative mapping and identification of homologous regions similar to identification of paleoploidy in soybean. Recent evidence showing duplication in the tiny genome of Arabidopsis thaliana (Paterson et al. 1996; Terryn et al. 1999; Grant et al. 2000) may indicate that an even higher percentage of plants have gone through one or more rounds of polyploidization than previously suggested.

There is much to be learned about genome organization and evolution in soybean. This dissertation is an investigation of duplicated regions, polyploid origin and diploidization in soybean. An important step in supporting the hypothesis of polyploid origin of a species and identifying regions arising from polyploidization events is to locate duplicated functional genes. As polyploid genomes evolve many duplicated genes will diverge and develop alternate function or lose function. Some may provide an advantage by remaining duplicated and will be retained as duplicated functional genes in remnant homoeologous regions. Identification of duplicated genes and segments provides a starting block for investigation of ancient polyploidy in a species.

This investigation begins with the mapping of duplicated functional genes to determine if they arose from a polyploidization event or another form of gene duplication (Lee et al. 1999). The regions to which the duplicated genes mapped were not directly homocologous. Further investigation was required to determine the origins of these genes. The limited genomic sequence information available for soybean necessitated the development of new methods for examining the origins of duplicated segments in soybean.
An extensive amount of data is being generated from *Arabidopsis* research. We anticipated the possibility of using *Arabidopsis* as a 'bridge' species to learn more about the soybean genome. A previous study detected synteny between *Arabidopsis* and soybean (Grant et al. 2000). Genome conservation between these two species can be used to identify multiple regions in soybean that have homology to a single *Arabidopsis* segment. Soybean genomic regions detecting homology to the same region in *Arabidopsis* likely arose from a duplication event in soybean. Therefore the comparison of these two species should lead to identification of additional duplications in soybean. Duplications identified by this technique would likely be missed in studies involving only legumes because of, 1) the monomorphic nature of soybean and other legumes and, 2) the lack of genomic sequence data within legumes.

Two hypotheses are explored in this dissertation. The first is that functional duplicate genes are likely the result of a polyploidization event in the evolutionary history of the soybean genome. If true, these genes should map to homoeologous regions. Our second hypothesis is that additional duplicated segments in the soybean genome will be revealed through comparison with other legumes and the well studied plant *A. thaliana*.

**Dissertation Organization**

This dissertation is organized into four chapters. Chapter 1 (this chapter) contains a literature review of current knowledge about polyploidy and diploidization. Chapters 2 and 3
each consist of a manuscript, either published or in preparation for publication. Chapter 4 is a summary of conclusions reached during the course of this dissertation research as well as recommendation for future research in this area.

Chapter 2, entitled 'Mapping of duplicate genes in soybean' was published in the October 1999 issue of Genome (Lee et al. 1999). The authors mapped functional duplicate genes (Pal and Pa2) and investigated the linkage groups on which these genes were located in relation to each other in attempts to ascertain the type of duplication event giving rise to these genes. We were unable to place the duplicated functional genes directly into homoeologous regions. We then proposed an additional duplication event or possibly a second round of polyploidization involved in the duplication of Pal and Pa2 as an explanation of their current genomic locations.

Chapter 3. 'Syntenic relationships within legumes and their comparison to Arabidopsis thaliana' has been prepared for submission to Genetics. After the placement of duplicated genes in the soybean genome we were interested in further investigating these regions to determine if we could ascertain the method of duplication from which these regions arose. By comparing the current soybean genomic structure to that of the diploid legumes (mung bean and common bean) and to A. thaliana, we were able to define regions of synteny between these distantly related species. Additionally, we found evidence for a second duplication in the region analyzed in soybean and proposed this may be indicative of an additional round of polyploidization in soybean. Our findings are not inconsistent with previous proposals for multiple rounds of polyploidy in the soybean genome (Shoemaker et al. 1996; Lee et al. 1999).
Literature Review

Introduction

Plants appear to have promiscuous genomes that can duplicate, diverge, re-organize, and hybridize with related species. Often, these evolutionary events occur while the plant continues to function with little or no effect from the changes in its genome. Gene or genome duplications are common among plants and have been suggested to play a significant role in genome evolution (Stebbins, 1950; Ohno, 1970; Lewis, 1980). Genome duplication, or polyploidy, while prevalent in the plant kingdom is seldom encountered in other kingdoms. A phenomenon so widespread in one kingdom, while infrequent in others has been discussed in countless studies.

Polyploids have the unique feature of having their entire genome duplicated. Polyploids are also subject to unique genetic processes stemming from the interaction of two complete sets of genes. Stress from multivalent formation following polyploidization creates a tendency for many polyploid genomes to revert to a diploid state through the processes of diploidization (Stebbins, 1966; Leipoldt and Schmidtke, 1982). Polyploidization followed by diploidization is believed to have occurred in soybean. This section is a review of duplication, divergence, polyploidization, and the diploidization process.

Gene duplication

Gene duplication is prevalent in plant genomes and has long been thought to be a powerful force in genome evolution (Stephens, 1951; Ohno, 1970). It can provide
redundancy, form permanent heterozygotes, or lead to new functions or specializations of genes (Ohno, 1970; Soltis and Soltis, 1993; Hughes, 1994; Force et al. 1999). Gene duplication may occur by several mechanisms. The most common types of duplications are regional, tandem, reverse transcriptase DNA insertion from an RNA intermediate, and total genomic duplication.

Regional duplication results from a segment of DNA being duplicated anywhere within a genome. The size of the duplicated segment can vary from a few hundred bases to an entire chromosome (Li and Graur, 1991). Tandem duplication is a specialized form of regional duplication. The DNA segment is always duplicated adjacent to the original copy on a chromosome. Tandem duplication is frequently the result of unequal crossing over or slip strand mispairing and results in the same sequence repeated in a head to tail arrangement at a single location in a genome (Ohno, 1970; Elder and Turner, 1995). Tandem duplication may lead to arrays of similar sequence such as found with rDNA sequences. It is also possible for genomic processes involved in tandem duplication to result in deletion of some sequences in an array. Duplication by reverse transcriptase DNA insertion from an RNA intermediate (dispersed duplication) most frequently occurs from an mRNA transcript. Only rarely would such a DNA duplication result in a functional gene since the promoter is not duplicated. Occasionally insertion may occur near an active promoter, however, the chances are low that the promoter will be compatible with the gene. Finally, total genomic duplication, or polyploidy, results in duplication of every sequence present in a genome. Polyploidy frequently occurs by the fusion of unreduced gametes (Dawson, 1962; Harlan and DeWet, 1975).
The mechanism responsible for a duplication event can frequently be discerned from the current organization and sequence of the duplicate segments. Duplication by polyploidy is the primary method of gene duplication addressed in this dissertation. It is necessary to understand the other methods of duplication in order to determine if current genomic organization of duplicate segments are from polyploidy or another form of gene duplication.

**Fate of duplicate genes**

Following gene duplication several fates may befall the duplicated sequences. Genetic redundancy creates a large pool of genomic information which can be altered, frequently with little consequence to the organism (Pickett and Meeks-Wagner, 1995). After duplication both copies may retain their original function (Ferris and Whitt, 1977; Lee *et al.* 1999). Two mechanisms are proposed to maintain duplicated functional genes. One is 'positive selection' which creates functional independence of the duplicate genes (Pickett and Meeks-Wagner, 1995). The other is 'negative purifying selection' which eliminates individuals with mutations in either gene, thereby maintaining the duplicate copies (Ohno, 1970; Pickett and Meek-Wagner, 1995).

Other studies have demonstrated the acquisition of new functions, or divergence and specialization of function following gene duplication. Walsh (1995) calculated that for large populations duplicated genes are likely to develop a new function. Force *et al.* (1999) demonstrated that duplication of the *Enl* gene in zebrafish resulted in specialization of each gene copy. Zebrafish is believed to be an ancient polyploid and to have undergone at least one round of polyploidization (Amores *et al.* 1998). Based on current genomic location, *Enl* was likely duplicated during the polyploidization event and the specialization of the *Enl*
genes occurred during diploidization. Enl is usually expressed in both the head and pectoral appendages in embryos (Force et al. 1999). The duplicate Enl genes in zebrafish are now named engI and engIb (Force et al. 1999). The duplicated genes now have different expression patterns with engI being expressed only in the pectoral appendages and engIb expressed only in the head in zebrafish embryos (Force et al. 1999).

Finally, upon duplication one of the duplicate genes may be lost or silenced. Haldane (1933) suggested that mutation would likely inactivate one copy of a duplicated pair of genes. Ferris and Whitt (1977) presented evidence for extensive gene silencing following polyploidization. Many studies have located pseudogenes, inactive copies of functional genes in a wide range of species as further evidence for duplicate gene silencing (Lee et al. 1983; Drouin and Dover, 1987; Heschl and Billie, 1989; Currie and Sullivan, 1994; Gottlieb and Ford, 1997; Milligan et al. 1998; Nakayama et al. 1999).

The fates of duplicated genes may be the same regardless of how the duplication arose. However, divergence of duplicate genes has been found to be most prevalent when duplication is from polyploidization (Ferris and Whitt, 1977).

**Polyploidy**

The term polyploidy was coined by Winkler in 1916 (Winkler, 1916; Grant, 1971) to describe organisms with more genetic material and a greater number of chromosomes than its ancestor (Dawson, 1962). In 1926, Kihara and Ono identified polyploids as being either autopolyploids or allopolyploids (Harlan and DeWet, 1975). Autopolyploids are generally described as individuals which have been derived from the duplication of the genome of a single species (Dawson, 1962; Ohno, 1970; Harlan and DeWet, 1975). Allopolyploids are a
result of the hybridization of the genomes of two different species (Dawson, 1962; Ohno, 1970; Harlan and DeWet, 1975). Distinguishing between autopolyploidy and allopolyploidy can be difficult if the concept of a species is not well defined.

A plethora of information about polyploids has been generated since the phenomenon was first characterized. Polyploidization has been suggested to be the cause for some of the major steps in evolution (Ohno, 1970). Stebbins (1971) suggested that "chromosome doubling by itself is not a help but a hindrance to the evolutionary success of higher plants". If this were true polyploidy would not be so widespread in the plant kingdom (Lewis, 1980; Masterson, 1994).

Polyploidy is relatively rare in other kingdoms. With a few exceptions, vertebrates can tolerate little redundancy in their genomes before the effect becomes lethal (Ohno, 1970). In most cases, duplication of even a single chromosome can be deleterious or lethal (Thompson et al., 1991). The paucity of polyploids among vertebrates may be due to their dependence on chromosome-specific sex determination (Ohno, 1970) and the detrimental effects of polyploidy in the early developmental stages of animals (Leipoldt and Engel, 1983). The second generation of polyploid organisms with chromosome-specific sex determination would be either a population of sterile individuals or a generation of a single sex (Ohno, 1970). Exceptions have been found among fish and amphibians where some true polyploid species exist. In some fish and amphibians sex determination is dependent on temperature and not chromosome ratios, perhaps alleviating problems described above (Ferris and Whitt, 1977; Leipoldt and Engel, 1983). Although there are few polyploids among non-plant kingdoms, investigation into vertebrate genomes has revealed the
possibility some vertebrates may be paleopolyploids (Spring, 1997; Amores et al. 1998; Postlethwait et al. 1998).

The lack of polyploids in other kingdoms prompts the question, why is polyploidy so prevalent in the plant kingdom? In some genera there may be advantages obtained from polyploidy. Bringhurst and Voth (1983) studied cultivated octaploid strawberries (Fragaria ×ananassa Duch). When octaploid strawberry cultivars were compared with a diploid cultivar (F. vesca) the octaploid species was found to be superior on all levels compared, including cytological and morphological evaluations. The octaploids had a high degree of genetic diversity both within and between populations whereas the diploids were found to be almost uniform. When morphological features and resistance to strawberry root diseases were compared, again the octaploid strawberries were highly polymorphic while the diploids were generally uniform (Bringhurst and Voth, 1983). Bringhurst and Voth (1983) suggested polyploidy is important in strawberries because only the octaploids have evolved the necessary genes and diversity required to accumulate the advantages observed among the large-fruited strawberry cultivars. While the diploid species are viable, they are not as agronomically important as their polyploid relatives due to the superior performance and fruit size of the octaploids. A similar relationship between genetic variation and ploidy level was also detected in Claytonia (Lewis and Semple, 1977) and Draba (Brochmann and Elven, 1992) species.

Some polyploids appear to tolerate extreme climates better than their diploid counterparts (Stebbins, 1971). Brochmann and Elven measured ranges of the Draba species and noted that the polyploids occurred in a wider range of habitats. Stebbins (1971) had also detected an association of ploidy level with geographical location. By making the
assumption that species with high basic chromosome number were originally polyploid. Stebbins gathered information relating ploidy to climate. Between 85% and 95% of species from tropical rain forests were determined to be polyploids. The lowest percentage of polyploids were found in warm temperate and subtropical regions. The percentage of polyploids increased as one moved toward cooler climates as well as more tropical climates (Stebbins, 1971). Therefore, tolerance of extreme climates may be associated with polyploidy.

Multiple origins of a single polyploid species has been shown to occur from the same diploid ancestors (Doyle et al. 1990; Brochmann et al. 1992; Soltis and Soltis, 1993; Jiang and Gill, 1994; Kollipara et al. 1994;). Doyle et al. (1990) investigated chloroplast DNA of *Glycine tabacina*. Based on plastome types they determined there must have been at least six independent origins of the wild perennial soybean *G. tabacina* involving six different genome types as the female progenitor (Doyle et al. 1990). Isozyme electrophoresis banding patterns allowed researchers to demonstrate there must have been multiple origins of the polyploid genomes of *Draba cacuminum* (Brochmann et al. 1992) and *Glycine tomentella* (Kollipara et al. 1994). Jiang and Gill (1994) used N-banding chromosome staining to detect species specific translocation events in *Triticum timopheevii* and *T. turgidum*. Species-specific events in *Triticum* support the diphylectic origin of these polyploids over the alternate monophyletic origin (Jiang and Gill, 1994). The demonstration of recurrent formation of polyploids in several species may indicate that multiple origins are the rule rather than the exception in polyploid formation and account for the genetic diversity observed in many polyploid species (Soltis and Soltis, 1995; Bringshurst and Voth, 1983).
Diploidization

High ploidy levels, above the octoploid level, create stress on plants due to the difficulty in maintaining multivalent association during meiosis (Stebbins, 1966). This may limit viability (Leipoldt and Engel, 1983). The stress in polyploid genomes may be alleviated through reversion to a diploid state, a common occurrence among polyploids (Leipoldt and Schmidtke, 1982; Stebbins, 1966). Diploidization can take place on either the chromosomal or genic level (Leipoldt and Schmidtke, 1982). At the chromosome level alterations such as inversions, translocations, and deletions may cause a preferential formation of bivalents over quadrivalents during meiosis (Ohno, 1970). At the genic level divergence and/or silencing of duplicated genes contributes to diploidization, along with genes which suppress synapses of homoeologous regions of the genome (Avivi, 1976).

Diploidization was first detected in *Nicotiana tobacum*, an amphidiploid from South America (Clausen, 1941). Clausen produced an artificial amphidiploid equivalent to *N. tobacum*. By comparing chromosome behavior and effects of polyploidization in both the natural and synthetic amphidiploids, Clausen was able to conclude that alterations have occurred in *N. tobacum* which transformed it to appear more like a diploid (Clausen, 1941). Clausen observed that *N. tobacum* was less duplicated than its amphidiploid counterpart and had widespread genomic deficiencies. However, there did not appear to be any elimination of genetic material in *N. tobacum*. As a result of the diploidization of *N. tobacum*’s genome, Clausen (1941) concluded that the two original sub-genomes are no longer self-sufficient and it would be impossible to separate the two.

The work by Clausen was followed by more in-depth studies on genomic events following polyploidization. In 1951 Gilles and Randolph published results from a 10-year
study on autotetraploid maize. They observed a decrease in the number of quadrivalents and an increase in the number of bivalents formed (Gilles and Randolph, 1951). This was the first study directly showing increased bivalent formation over quadrivalents following polyploidization. Over a decade later Shaver (1963) determined that macro-rearrangement of chromosomes in maize greatly increased preferential bivalent formation. Byrne and Jelenkovic (1976) detected the presence of bivalents with few multivalents forming during meiosis in the octoploid *Fragaria* concluding diploidization also occurs in species having been through multiple rounds of polyploidy. These studies provided evidence for diploidization at the chromosome level.

Riley and Chapman (1958) provided evidence for a gene involved in diploidization. The presence of this gene caused a reduction in the number of quadrivalents, trivalents and even bivalents at meiosis (Riley and Chapman, 1958). This gene was called *Phi* for 'pairing homologous' (Gill et al. 1993). Cells which express *Phi* showed no quadrivalent formation, very little trivalent formation, and averaged 1.38 bivalents per cell (Riley and Chapman, 1958). Cells deficient in *Phi* had some quadrivalents and trivalents and averaged 4.16 bivalents per cell (Riley and Chapman, 1958). Feldman (1966) noted that four doses of *Phi* slightly reduced chiasma frequency between homologues and six doses of *Phi* causes considerable asynapsis in cells. Genes, such as *Phi*, contribute to diploidization by inhibiting multivalent formation through the reduction of chiasmas between homoeologues (Riley and Chapman, 1958).

The next step in investigating diploidization might be to ask how long might it take for diploidization to occur? Studies investigating processes contributing to diploidization following polyploidization have found that some genomic changes occur rapidly following
polyploidization events (Song et al. 1995; Feldman et al. 1997; Humphreys et al. 1998). A study in Brassica demonstrated accelerated changes in polyploids compared to diploids (Song et al. 1995). Song et al. (1995) detected extensive changes occurring within a few generations in genomes of the synthetic brassica polyploids. Shortly after this discovery Feldman et al. (1997) found similar results in wheat. They determined that low copy sequences were being eliminated from both hexaploid wheat and newly synthesized amphidiploid wheat (Feldman et al. 1997). Humphreys et al. (1998) found that dramatic genomic changes occurred in newly synthesized pentaploids. Following hybridization of Lolium multijlonm (2n=4x=28) with Festuca araundinacea (2n=6x=42) many chromosomes were eliminated during meiosis (Humphreys et al. 1998). Progeny were produced with significantly less than the expected 2n=35 chromosomes (Humphreys et al. 1998).

The knowledge that diploidization occurs in genomes makes it possible to investigate presumed diploid species and identify remnants of polyploidization events. Helentjaris et al. (1988) identified duplicate genes and RFLP loci in maize and found a distribution indicative of a diploidized polyploid. The maize genome was shown to consist of intricate patterns of interspersed duplicate segments between several linkage groups (Helentjaris et al. 1988). A pattern that would not be expected to occur by random chance. Evidence of ancient polyploidy in plants has been detected in other investigations of maize (Gaut and Doebley. 1997), cotton (Reinisch et al. 1994; Brubaker et al. 1999; Muravenko et al. 1998). Sorghum (Gómez et al. 1998), Helianthus (Sossey-Alaouni et al. 1998) and Brassica (Lagercrantz. 1998) indicating paleopolyploid is widespread in the plant kingdom. Other investigations allowed Shoemaker et al. (1996) to present molecular evidence for the polyploid origin of
the soybean, and Spring (1997) to furnish evidence that humans may be ancient polyploids. Duplication in Arabidopsis has been detected and is speculated to have occurred from a polyploidization event (Paterson et al. 1996; Terryn et al. 1999; Grant et al. 2000). It has been suggested that perhaps most or even all eukaryotes are diploidized polyploids (Leipoldt and Schmidtke. 1982).

Species with extremely high numbers of chromosomes (>100) have been proposed to be ancient polyploids (e.g. Chrysanthemum lacustre with 2n=198 may be an ancient 22-ploid) (Grant, 1971). Lack of species with true polyploids above an octoploid level has led researchers to speculate that polyploids above the octoploid level are not stable, do not survive, or are quickly subjected to diploidization (Stebbins, 1966). Stebbins (1966) proposed that plants undergo repeated rounds of polyploidization followed by diploidization to reach the high ploidy levels. It is only through cycles of returning a polyploid to a diploid state that polyploidization can continue and organisms with very high chromosome numbers are created and remain viable. Without diploidization, polyploidization could only occur a few times before it would become an evolutionary dead end. When the two interact they can achieve an effective evolutionary collaboration.

Paleopolyploidy in the soybean

Soybean has been suggested to be a diploidized polyploid (Lackey, 1980; Lee and Verma, 1984; Shoemaker et al. 1996). Multiple lines of evidence ranging from, cytological to molecular, have been collected supporting polyploidy in soybean.

Soybean is a member of the family Fabaceae, subfamily Phaseoleae (Lackey, 1980). Most other members of Phaseoleae have a base chromosome number of x=11 while
soybeans have x=20 (Lackey, 1980). *Glycine* was suggested to have been derived from a diploid ancestor (x=11) which underwent an aneuploid loss to x=10, then subsequent polyploidization and diploidization to yield the genome size of x=20 seen today (Lackey, 1980). Although, there is no evidence favoring this hypothesis over polyploidization followed by aneuploid loss.

Lee and Verma (1984) generated data from leghemoglobin genes supporting the diploidized tetraploid origin of the soybean genome. They studied leghemoglobin genes in *Phaseolus vulgaris* (common bean), *G. max* (soybean), and *G. soja* (wild soybean). Four leghemoglobin genes were found at one locus in common bean. Tandem duplications of a single gene occurred prior to divergence of common bean and soybean. Soybean then underwent tetraploidization to create two loci. A large deletion then occurred at one of these duplicate loci. The result was one functional locus and one locus which contained a pseudogene in both *G. max* and *G. soja*.

Zhu *et al.* (1994) were able to use the average number of fragments per RFLP probe to predict the ploidy level of soybean compared to its diploid relative, common bean. They calculated that 1.5 fragments per probe are expected for an inbred diploid and 3.0 fragments per probe expected for an inbred tetraploid. Soybean was found to have an average of 2.3 fragments per probe and common bean an average of 1.4 fragments per probe (Zhu *et al.* 1994). Therefore, common bean is supported as being an inbred diploid, however soybean is somewhere between a diploid and tetraploid. Zhu *et al.* (1994) predicated that the outcome is what would be expected of a tetraploid having gone through diploidization and having lost duplicate sequences during the process.
Numerous duplicate regions have been identified in the soybean genome confirming that a polyploidization event has occurred in its evolutionary history (Shoemaker et al. 1996). We have now been able to map duplicate loci to homologous regions in soybean (Lee et al. 1999). Further investigation of the soybean genome indicates additional duplications have occurred and may be indicative of a second round of polyploidization in soybean. Analysis of duplicate regions in soybean has allowed us to decipher some of the events which have occurred in the evolution of its genome.

References


Langercrantz, U. 1998. Comparative mapping between *Arabidopsis thaliana* and *Brassica nigra* indicates that *Brassica* genomes have evolved through extensive genome replication accompanied by chromosome fusions and frequent recombinations. Genetics. 150:1217-1228.


Postlethwait, J. H., Y.-L. Yan, M. A. Gates, S. Horne, A. Amores, A. Brownlie, A.

Donovan, E. S. Egan, A. Force, Z. Gong, C. Goutel, A. Fritz, R. Kelsh, E. Knapik, E.

Liao, B. Paw, D. Ransom, A. Singer, M. Thomson, T. S. Abduljabbar, P. Yelick, D.


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


CHAPTER 2. MAPPING OF DUPLICATE GENES IN SOYBEAN

A paper published in Genome

Jennifer M. Lee, Arla L. Bush, James E. Specht, and Randy C. Shoemaker

Abstract

Appressed pubescence genes in soybean cause hairs on the upper surface of leaves to lie flat while pubescence remains erect elsewhere on the plant. For decades this trait was believed to be controlled in soybean by duplicated single genes, *Pa1* and *Pa2*. However, reports in the literature conflicted as to which phenotype was dominant or recessive. Two populations were developed, each approximately one hundred individuals, and each segregating for one of the appressed pubescence genes. A combination of SSRs and RFLPs were used in each of these populations to map the independent genes. Two-point analysis

---

1 Reprinted with permission of Genome, 1999 42, 829-836

2 Department of Agronomy, Iowa State University, Ames, IA 50011; A.B. current address: Identity Genetics, 2308 Sixth St. East P.O. Box 877, Brookings, SD 57006-2404

3 Department of Agronomy, University of Nebraska, Lincoln, NE

4 USDA-ARS, Corn Insect and Crop Genetics Research Group, Iowa State University, Ames, Iowa 50011; Author for correspondence
weakly linked \textit{Pa1} and \textit{Pa2} to separate linkage groups. Lack of strong linkage suggested the trait may not be controlled by single genes. When QTL analysis was performed one major locus and several minor loci were detected in each population. We report the mapping of the genes controlling appressed pubescence in soybean and their placement in homologous regions. Although oppressed pubescence was originally reported to be single duplicate genes, we report that it is actually a more complex phenotype with major duplicated genes and minor modifying genes. These results offer interesting implications regarding the evolution of duplicate genetic factors and the definition of qualitative traits.

\textbf{Introduction}

One approach to studying genome evolution is to observe duplicated genes either within or between species. Polyploid organisms are useful for studying evolution at the molecular level because all genes in their genome have been duplicated. After polyploidization, most duplicated regions begin to diverge from one another at both the sequence and chromosomal levels (Pickett and Meeks-Wagner 1995). Frequently, polyploid subgenomes will diverge so extensively that a polyploid becomes effectively diploid, a process termed diploidization by Stebbins in 1966, although the concept of diploidization was envisioned as early as 1941 (Clausen 1941). By analyzing duplicated regions and duplicated genes it may be possible to determine the origin of these segments and gain an understanding of the mechanisms of the process of diploidization.
As polyploids undergo diploidization, duplicated genes evolve resulting in one of several fates. Both may retain their original functions, they may diverge to have different functions, or one may become silenced or lost while the other remains functional (Pickett and Meeks-Wagner 1995). Duplicated genes that retain their original function are of special interest because placement of them on linkage maps may provide evidence for evolutionary relationships among linkage groups.

The cultivated soybean, *Glycine max* (L.) Merr. is agronomically important to much of the world (Hymowitz and Newell 1981). Although the belief that all species within *Glycine* are polyploid or paleopolyploid (Lackey 1980; Doyle 1991) is widely accepted, the details of the origin of this genus remains a mystery. The original hybridization event from which the polyploid progenitor of *Glycine* was created occurred so long ago that its evolutionary history is now obscured. By studying duplicated genes or regions within the soybean genome it may be possible to describe some of the events that occurred during polyploidization and subsequent diploidization.

Appressed pubescence is a phenotype previously shown to be controlled by duplicated genes (Bernard 1973). It was also among the first pubescence types described in *Glycine* (Woodhouse and Taylor 1913). This mutation causes the pubescence to be appressed (flattened) to the upper surface of the leaf (Figure 1) while the hairs remain erect elsewhere on the plant. Although there have been no reported studies to determine what, if any, advantage the trait confers to a plant, appressed pubescence is common in wild varieties of soybean, suggesting it may confer some selective advantage to the plant (Bernard 1973).
Figure 1. Extreme forms of (a) appressed and (b) erect phenotypes.
It has been shown that cultivars with appressed pubescence have, on average, a higher yield than varieties with erect pubescence (Specht et al. 1985)

In this paper we report results from the mapping of genes controlling appressed pubescence in soybean, which suggest that these genes lie in homologous regions of the soybean genome. Although appressed pubescence has previously been reported to be qualitatively controlled by single duplicate genes, we report that it is actually a more complex phenotype with major duplicated genes and minor modifying genes. The placement of the genes on the soybean linkage map provides some clues as to the evolutionary origins of these genomic regions.

**Material and Methods**

*Plant materials*

Two F2.3 populations were developed (94-1059 and 94-1060), each of which segregates for one of the appressed pubescence genes. Each of these populations is a 1993 mating of a female 'Clark' (palpa2) near-isogenic line (NIL) with two male 'HaroSoy' NILs; 1059 (palPa2) and 1060 (Pa1pa2) (Bernard 1973). The F1 plant was grown in the 1993-94 winter greenhouse at University of Nebraska-Lincoln. Population 94-1059 contains 113 F2 individuals and segregates for Pa2 in a genetic background of palpa1; 94-1060 contains 102 F2 individuals and segregates for Pa1 in a genetic background of pa2pa2.

Seed for parental lines were grown in the USDA greenhouse at Iowa State University during the spring of 1995. Leaves were collected from seedlings and used for DNA
extraction. F2 plants were scored for appressed pubescence \( (palpa2pa2) \) (Figure 1A) versus erect pubescence \( (Pa_+) \) (Figure 1B). Leaves from plants in the F2.3 generation identified as having completely appressed pubescence were collected from the University of Nebraska-Lincoln campus nursery field during the summer of 1995. Seed for the remainder of the population was grown in the USDA greenhouse during the winter of 1996 at Iowa State University. Heterozygous and homozygous erect F2 plants were identified by examining these F2.4 lines. F2.4 lines exhibiting both erect and appressed pubescence orientations were classified as heterozygotes. F2.4 lines from which only completely erect pubescence plants were seen were classified as homozygous erect. Several leaves from each of about ten plants were collected to use for DNA extraction. All lines which were not classified as completely erect pubescence phenotype were also grown in the USDA field at Iowa State University during the summer of 1997 to confirm phenotype. Seed from F2.4 plants with appressed pubescence were grown in the USDA greenhouse at Iowa State University during the fall of 1997 to confirm phenotype.

Prior to 1997 the pubescence orientation was scored qualitatively as either appressed or erect. The plants grown in 1997 to confirm phenotype were observed express a range of variation between the two extreme phenotypes, in both populations, of completely appressed or completely erect pubescence. Therefore, plants also were scored for degrees of appressed pubescence to permit QTL analysis. F2.4 lines were scored as: 0 - all F4 plants show nearly to completely erect pubescence, 1 - F4 plants segregate for erect and appressed pubescence, and 2 - all F4 plants show nearly to completely appressed pubescence.
**DNA extraction**

DNA was extracted from dried leaves using 1% CTAB as described in Keim *et al.* (1990). DNA was then quantified on a fluorometer. Concentration and purity of DNA samples were checked by running 1 µg of undigested DNA and 1 µg of *HindIII* digested DNA on 1% agarose gels for three hours at 80 to 100 volts. DNA that was not of sufficient concentration or high enough quality was re-extracted.

**DNA markers**

Of the 217 RFLP (designated with a starting letter of A, B, K, L, R, or T in figures) probes and 351 SSR (designated with a starting letter of S in figures) primers tested, 85 RFLP probes (39%) and 93 SSR primers (26%) were found to detect polymorphisms between the 'Clark' and 'Harosoy' parental lines. Probes were selected based on known linkage maps in soybean such that markers were not more than 40 cM apart (Shoemaker and Specht 1995). This marker density should be adequate to detect linkage anywhere within the genome (Tanskley 1993). *Phaseolus vulgarus* (kindly provided by E. Vallejos, University of Florida) and *G. max* genomic probes were used to identify RFLPs. SSR primers were kindly provided by P. Cregan, USDA-ARS (Cregan *et al.* 1998).

**Southern blotting and hybridization**

Eight µg DNA was digested with each of the five enzymes: *DraI, EcoRI, EcoRV, HindIII*, or *TaqI* for four hours at the recommended temperature and reaction conditions. One tenth the reaction volume of a solution of 25% ficoll, 10% 1.0M Tris, and 0.25% Bromophenol Blue was added as the loading buffer. Gels were electrophoresed at 20-30
volts for 16-24 hours on 6mm 1% agarose gels containing ethidium bromide. *HindIII* digested lambda DNA was used as a molecular size marker. Gels were photographed with an AlphalImager 2000.

Southern hybridizations (Southern 1975) were performed similarly to that described in Keim *et al.* (1990) with the following changes. Blots were prehybridized overnight at 65°C. Probes were labeled as described except that uncut lambda was added to the labeling solution instead of being labeled separately. This reaction was incubated at room temperature for four hours then denatured for five minutes at 95°C and added to the prehybridization solution. Blots were hybridized overnight at 65°C. The blots were rinsed in a 2X SSC, 0.4% SDS solution, then washed 20 minutes in the same solution for a low stringency wash at 62.5°C. Two high stringency washes then were done in 0.6X SSC, 0.12% SDS for 20 minutes each at 62.5°C. If blots still had significant background radiation after exposure to X-ray film, they were washed for an additional 15 minutes in 0.1X SSC, 0.1% SDS at 63°C.

*Simple sequence repeats*

The SSR reaction consisted of 60 ng DNA, 2 mM MgCl$_2$, 0.2 mM of each dNTP, 1X reaction buffer, 0.5 U Taq polymerase, and 0.15 μM primers. Cycling was performed in a MJ Research, Inc. PTC-100 Programmable Thermal Controller. Forty-five cycles of 30 seconds at 94°C, 30 seconds at 47°C, 30 seconds at 6°C were done followed by a 4°C hold. After adding 5μl of 5X ficoll loading dye (12.5% ficoll 400, 0.25% Bromophenol Blue) to each reaction the full reaction volume was loaded in a 4mm thick 2.5% Amresco SFR agarose
gels containing ethidium bromide and electrophoresed for two to four hours at 150 volts. A 20 base pair ladder was used as a marker. The gels were visualized on a ultraviolet light box and photographed with an Alphalmager 2000. Patterns on gels were scored as male parental type, heterozygous, or female parental type.

Data analysis

Marker segregation data was entered into Map Manager 2.6.5 (Manly 1993). Data then was exported as an intercross to a MapMaker 2.0 file (Lander et al. 1987) and revised in Text-Edit Plus. This data was opened in MapMaker 2.0 (Lander et al. 1987) and two-point analysis was performed using Mapmaker’s ‘group 3.0 30’ command, which assembled the markers into their respective linkage groups based on a minimum LOD score of 3.0 and a maximum distance of 30cM. Markers were ordered within the linkage groups by use of the ‘best order’ command. Phenotypes scored as qualitative traits were placed on the genomic map with the ‘try’ command.

QTL analysis was performed on the phenotypes scored by the 0,1,2 rating system with MapMaker/QTL (Lander and Botstein 1989). A threshold LOD cutoff of 2.0 was used and linkage groups and map distances were assigned as determined in MapMaker 2.0 (Lander et al. 1987).
Results

Populations

Appressed pubescence is a characteristic in soybean which causes the hairs to be appressed to the upper surface of the leaf while they remain erect elsewhere on the plant (Figure 1) (Karasawa 1936). For decades appressed pubescence was believed to be controlled by duplicate qualitative genes (\(P_{al}\) and \(P_{a2}\)). However, the literature is conflicting as to which phenotype was dominant or recessive (Bernard 1973; Ting 1946; Karasawa 1936). In this study two populations were developed that each segregated for only one of the appressed pubescence genes. The appressed phenotype was originally scored as recessive and the erect pubescence form as dominant (Bernard 1973). The erect pubescence F2 plants were further subdivided into homozygous dominant and heterozygous classes based on segregation data of F2.3 families. When F2.4 lines were grown to confirm the phenotype of the F2 plants, visual inspection indicated that only a few plants had completely appressed or completely erect pubescence but that there was actually a range of variation between the two extreme phenotypes. Degrees of appression were difficult to score, so phenotypes were divided into the three categories expected for monogenic inheritance: lines in which all plants were mostly or completely erect, putatively homozygous dominant (\(P_{al}P_{al}\)); lines in which all plants where mostly or completely appressed putatively homozygous recessive (\(pa1pa1\)); and lines in which both erect and semi-appressed forms were observed, putatively heterozygous (\(Pa1pa1\)); if plants were in a \(pa2pa2\) background. In a \(palpa1\) background the genotypes would be \(Pa2Pa2\) for putatively
homozygous dominant, \(pa2pa2\) for putatively homozygous recessive, and \(Pa2pa2\) for putatively heterozygous.

**Data analysis**

A combination of RFLP and SSR markers were tested with a total of 568 potential markers screened. Of these, 175 (85 RFLPs and 90 SSRs) were found polymorphic in 94-1059 and 168 (80 RFLPs and 88 SSRs) in 94-1060. All polymorphic markers were mapped. Two-point analysis in MapMaker 2.0 (Lander et al. 1987) revealed \(Pa2\) was linked to markers on linkage group F. The greatest LOD score (5.7) involved \(Pa2\) with Sat\_120. \(Pal\) was linked to markers on linkage group B1/S. The greatest LOD score (4.2) involved \(Pal\) with A381\_RI (Figure 2). When attempts were made to place the genes at discrete positions on their respective linkage groups it could not be done without generating gaps of at least 20cM which were presumed to be artifactual. Qualitative genes should be able to be mapped to a single point on a linkage group without significantly disrupting the relative positions of other markers on the group. The inability to precisely map the loci and our observation of a range of variation of the appressed verse erect pubescence phenotypes suggested that the appressed pubescence genes may be genes whose phenotypes are not qualitatively distinct.

Segregation data from the appressed pubescence populations fit more than one mendelian ratio. For the 94-1059 population neither a 3:1 nor a 9:7 erect:appressed pubescence segregation ratios could be rejected (83 erect pubescent plants, 30 appressed pubescent plants). In population 94-1060 neither 3:1 nor 13:3 erect:appressed pubescence segregation ratios could be rejected (82 erect pubescent plants, 20 appressed pubescent
Figure 2. Homoeologous relationships between linkage groups B1/S, H, and F. Positions of Pa1 and Pa2 major loci are indicated by the hashed bars with the arrow pointing to the most likely placement. Histone H3 genes are indicated in bold illustrating that other functional duplicate genes lie in these regions.

plants). Data from previous studies showed similar results (Bernard 1973, Karasawa 1936). We hypothesized that these aberrant segregation ratios may be an effect of qualitatively classifying an inherent quantitative trait.

The segregation data then was converted to a quantitative form which could be used for QTL analysis and analyzed in MapMaker/QTL (Lander and Botstein 1989). For a
population of approximately 100 individuals a maximum distance between markers of 35cM is suggested to detect a QTL of moderate strength in an interval (Tanksley 1993). Our average interval is 17.5cM. Although a few markers were more than 35cM apart (as estimated by anchored loci from previous mapping studies) these were infrequent.

QTL analysis detected multiple loci controlling pubescence morphology. A single major locus and several minor loci were detected in each population (Figure 3). The major \( Pa1 \) and \( Pa2 \) QTLs had mapped to positions identical to those detected using two-point analysis (\( Pa2 \) on LG-F; \( Pa1 \) on LG-B1/S). Of the minor QTL, one was in common between the two populations (on LG-D1b/W) and potentially represented the same locus, while others were unique to each population (Figure 3).

Map positions

Map locations of duplicate loci can provide insight into genome evolution. The results of this appressed pubescence gene study offers some interesting implications as to the history of the soybean genome. A close evaluation of duplicated molecular markers and genome organization revealed that the major QTLs of these populations mapped to homologous regions within the genome (Figure 2). Linkage groups B1/S and H are highly homoeologous with seven markers in common between them, including functional histone H3 loci. The probability that seven duplicated markers would coincidentally be syntenic on linkage groups of equal size in a map with 20 linkage groups is equal to \((\frac{1}{20})^7\), or \(1.56 \times 10^{-8}\) (Reinisch et al. 1994). LG-F also shows homology to LG-H and LG-B1/S. There are five markers in common between LG-F and LG-H (a likelihood of \(6.25 \times 10^{-6}\), most of
Figure 3. QTL mapping results for populations 94-1059 and 94-1060. The vertical lines represent linkage groups, the numbers to the left of them markers on that linkage group, the linkage group designation is given in bold above each group. The horizontal lines represent an association of a locus involved with appressed pubescence in that interval. The longer the line the more strongly associated that interval is with a QTL. There were three additional linkage groups in population 94-1060 which detected association with a QTL however these three were all very intense peaks in large gaps (greater than 50cM). In our experience we have noticed that this type of result frequently indicates a type I error and there is likely not a QTL in those intervals.
which were located on LG-H in other mapping populations and are detected on LG-F only in the appressed pubescence populations. These relationships are strong evidence of homoeology between these three linkage groups and suggest clues to their possible origins.

**Discussion**

Duplicate genes have been advantageous for studying genome evolution in polyploids (Hughes and Hughes 1993; Huang et al. 1992). The appressed pubescence genes mapped in this study are no exception. Although these genes long have been thought to be qualitative genes, the phenotypic data collected here suggests a more quantitative nature. Of equal interest to the discovery of the quantitative nature of these genes is the placement of the major QTLs controlling this trait in homoeologous regions of molecular maps in each population (Figure 2).

The Bl/S and F linkage groups to which Pal and Pa2 map do not share homology based on molecular markers used in this and other mapping studies. However, both linkage groups share homology to a third linkage group, H (Figure 2). Which two linkage groups are the homoeologues that arose from the most recent polyploidization event and which are paralogues that are derived by other means of duplication is not entirely clear. A comparative examination of these three linkage groups suggests that LG-B1/S has homology almost exclusively with the top of LG-H, and LG-F has homology almost exclusively with the bottom of LG-H (Figure 2). There are single exceptions for both sets of homoeologous regions. These observations may indicate some possible evolutionary origins of these
regions. The linkage groups analyzed may have arisen from a chromosome breakage or fusion event. LG-B1/S and LG-F at one time may have been a single linkage group whose homoeologue was LG-H. At some point during soybean evolution LG-B1/S and LG-F broke, creating two chromosomes that are homoeologous to the upper and lower portions, respectively, of the single linkage group of H. Alternatively, LG-H at one point may have been two linkage groups (LG-Htop and LG-Hbottom) that were joined. LG-Htop would have been homoeologous to LG-B1/S and LG-Hbottom homoeologous to LG-F. These four linkage groups may or may not have been paralogous from a more ancient genome duplication. Then LG-Htop and LG-Hbottom fused creating one linkage group. The duplicate marker which appear to map in opposite directions as the rest of the markers (Figure 2) may reflect additional rearrangements that have likely occurred in the soybean genome since its polyploidization event as the genome has undergone diploidization. It is also possible these markers represent loci that once occurred in four places in the soybean genome (two homoeologous regions) that underwent alternative silencing, so the two detectable loci do not fall in homoeologous regions. Based on previous data, it is likely that soybean has undergone anueploid loss during its genomic evolution (Lackey 1980). The end result of chromosome fusion would be equivalent to anueploid loss, which may make chromosome fusion a more likely alternative.

There has been some speculation that soybean is a paleohexaploid (Shoemaker et al. 1996). If this were the case it would not be surprising to see three homoeologues in the soybean genome, two being derived from the tetraploid and one from the diploid to which it hybridized. Which two linkage groups would be expected to be most closely related in such a
case would depend on whether the tetraploid was an auto- or an allopolyploid and whether
the diploid was more closely related to one or the other of the tetraploid subgenomes then
they were to each other.

That presumed monogenic traits were found to be quantitative is in alignment with
our growing understanding of quantitative genetics. Individual cells require complex
interactions of DNA, RNA and proteins. In multicellular organisms this complexity
increases dramatically. No single gene or protein functions entirely independently
(Robertson 1989). All genes interact with other genes or gene products at some level in a cell.
In light of this it is difficult to imagine how one might map a single gene as a distinct entity
since, by nature, it must be interacting with other genes, either directly or indirectly. A
gradient of interaction is likely present when all genes in a genome are considered, which is
usually ignored or not detected (Robertson 1989). This gradient would range from simple
traits that are controlled almost entirely by a single gene and have only minimal interaction
with other genes, to extreme quantitative traits in which numerous loci interact intimately to
create a phenotype.

Quantitative traits show a range of variation between two extreme phenotypes. In
previous literature, the appressed pubescence genes were classified in qualitative terms
(Bernard 1973; Ting 1946; Karasawa 1936). This may have led to two types of
misclassifications. Plants that showed completely erect pubescence may have been scored as
the erect phenotype, and plants not showing completely erect pubescence were grouped
together as appressed pubescence phenotype. This would result in appressed pubescence
appearing to be the dominant phenotype (Karasawa 1936; Ting 1946). Alternately, if plants
with completely appressed pubescence were identified as the appressed pubescence phenotype, and plants not having completely appressed pubescence were classified as the erect pubescence phenotype, then erect would appear to be the dominant phenotype (Bernard 1973). Phenotypic dominance (or recessiveness) is determined by the researchers' choice of which parental phenotype is easier to distinguish for the heterozygous class.

In QTL studies, a major locus detected in one population may be detectable in a second population (Lin et al. 1997; Bush and Wise 1996). In the appressed pubescence populations the major QTLs segregating in each population is purposefully different. These populations were developed to segregate at just one of the major loci and selected to be homozygous recessive at the other major locus. We expect that in a population in which a locus is selected to be homozygous for all individuals, that locus may not be detected in a QTL analysis.

The lack of minor QTLs in common between populations could be explained if the major QTL that codes for the protein which plays a predominant role in the mechanism that appresses pubescence interacts with the proteins from different minor loci. Then only the major locus in each population and the minor loci which modify or regulate it in each population would be detected in a QTL study, these would not necessarily be detected as a minor QTL in other populations. The differences we see in the collection of minor QTL in the two appressed pubescence populations also may be explained by the evolutionary history of these genes. If they arose from an ancient polyploidization event we would expect them to lie in homoeologous regions of the soybean genome. As the genome became diploidized, the major QTL would have diverged, each of them co-evolving with their own
compliment of modifying genes. On occasion the same minor loci were retained in both populations, as on linkage group D1b/W (Figure 3), but more frequently different loci were maintained in each population (Figure 3). We can not speculate how these multiple loci interact without further analysis.

Following genome duplication, many duplicate sequences will diverge from one another (Shoemaker et al. 1996; Reinisch et al. 1994; Helentjaris et al. 1988). In the absence of any selective advantage for retaining two genes with the same function, one copy should be lost or its function altered (Pickett and Meeks-Wagner 1995). Maintenance of duplicated sequences in homoeologous regions has been shown previously in polyploids and paleopolyploids (Van Deynze et al. 1995; Nelson et al. 1995; Helentjaris et al. 1988; Berhan et al. 1993; Lin et al. 1995; Hughes 1994; Begun 1997; Lee and Verma 1984). Persistence of redundant copies of functional genes suggests the involvement of selective pressures in their evolution (Pickett and Meeks-Wagner 1995). The observation that major QTLs controlling appressed pubescence map to homoeologous regions suggests that the function of these duplicated genes have been maintained following polyploidization. Alternatively, it is possible that there just has not been sufficient time for one of the redundant copies to be lost.

There has been some suggestion that pubescence orientation may be involved in insect resistance (Bernard 1973). The map location of Pa2 lends support for the involvement of the appressed pubescence genes in resistance to pathogens. Disease resistance genes have previously been shown to be clustered in soybean (Polzin et al. 1994, Tamulonis et al. 1997a and 1997b) as well as in other species (Holub 1997). The incidence of chromosomal
clustering of genes possessing related functions has been discussed in Chen and Shoemaker (1998). The major QTL for Pa2 maps on LG-F to a cluster of at least eight known QTLs linked to resistance (Tamulonis et al. 1997a and 1997b) thus strengthening the possibility of its involvement in resistance to pathogens.

The degree of homoeology versus paralogy of the Pal and Pa2 genes is still not entirely clear. There are multiple possibilities for placement of these genes to homologous regions. Additional investigation of linkage groups B1/S, F, and H may further elucidate the true origins of these regions. The compilation of molecular relationships of these regions has contributed to the understanding of the evolution of the soybean genome. It is apparent that the evolutionary history of the soybean genome is painstakingly complex and great care will have to be taken to determine the true origins of its duplicated regions.

References


Chen Z. and Shoemaker, R.C. 1998. Four genes affecting seed traits in soybeans map to linkage group F. Journal of Heredity, 89:000-000


Keim, P., Diers, B., Olson, T.C., and Shoemaker, R.C. 1990. RFLP mapping in soybean: association between marker loci and variation in quantitative traits. Genetics, 126:735-742.


CHAPTER 3. SYNTENIC RELATIONSHIPS WITHIN LEGUMES AND THEIR COMPARISON TO ARABIDOPSIS THALIANA

A paper submitted to Genetics
Jennifer M. Lee¹, David Grant¹,², C. Eduardo Vallejos¹, and Randy C. Shoemaker¹,²

Abstract

Analysis of molecular linkage groups within the soybean (Glycine max L. Merr.) genome reveals many homoeologous regions, reflecting the ancient polyploidy of the soybean. The fragmented arrangement of the duplicated regions suggests that extensive rearrangements as well as additional duplications have occurred since the initial polyploidization event. In this study we used comparisons between homoeologous regions in soybean and the homologous regions in the related diploids Phaseolus vulgaris and Vigna radiata to elucidate the evolutionary history of the three legume genomes. Our results show

¹Department of Agronomy, Iowa State University, Ames, IA 50011 USA
²USDA-ARS Corn Insect and Crop Genetics Research Group, Iowa State University, Ames IA 50011 USA
³Horticultural Science Department, University of Florida, Gainesville, FL 32611 USA
that there is not only conservation of large regions of the genomes but that these conserved linkage blocks are also represented twice in the soybean genome. To gain a better understanding of the process of genome evolution in dicots, molecular comparisons have been extended to another well-studied species, Arabidopsis thaliana. Interestingly, the conserved regions we identified in the legume species are also relatively conserved in Arabidopsis. Our results suggest that there is conservation of blocks of DNA between species as distantly related as legumes and brassicas, representing 90 MY of divergence. We also present evidence for additional duplication of large regions in soybean, which may be indicative of a second, earlier polyploidization event.

Introduction

The dynamic nature of plant genomes has made them an area of intense investigation. Research in plant genetics has revealed a great deal about genome duplications and polyploidization. Although a low percentage of stable polyplolds exist in nature, numerous polyploid events are suggested to have occurred during evolution of the plant kingdom (Lewis, 1980). There appears to be a natural tendency for polyploid genomes to revert to a diploid state following polyploidization, a phenomenon termed diploidization (Stebbins, 1966). Diploidization occurs at both the chromosomal and genic levels. At the chromosomal level processes such as inversions and translocations allow preferential formation of bivalents over quadrivalents during meiosis. Genes, such as Ph1 in wheat, have
been identified which prevent pairing of homoeologues during meiosis, which in turn contributes to diploidization (Riley and Chapman, 1958; Gill et al. 1993).

In many lineages, remnants of duplicate regions from ancient polyploid events are still detectable (Helentjaris et al. 1988; Whitkus et al. 1992; Shoemaker et al. 1996; Gaut and Doebley, 1997; Amores et al. 1998; Gómez et al. 1998; Sossey-Alaouni et al. 1998; Lagercrantz, 1998; Brubaker et al. 1999). Duplications of large blocks of genomes have been observed in plants which behave as diploids. Helentjaris et al. (1988) showed intricate patterns of interspersed duplicated segments between several linkage groups in the maize genome.

Certain complements of genes and/or DNA sequences remain together after many evolutionary events. Paterson et al. (1996) demonstrated that segments of DNA ≤3 cM remain co-linear in species separated by 130-200 million years of divergence. These authors proposed that this might be the result of DNA being reshuffled in small conserved segments over long evolutionary periods. Evidence for rearrangement of genomes by movement of blocks of DNA has been shown in several lineages, including grasses (Moore et al. 1995) and brassicas (Kowalski et al. 1994; Lagercrantz, 1998). The reason for DNA rearrangement in conserved blocks is unclear. However, conserved blocks may represent a genetic organization which confers a fitness advantage (Paterson et al. 1996).

Soybean has been suggested to be an ancient polyploid based on chromosome numbers (Lackey, 1980; Bruneau et al. 1994), genome size (Arumuganathan and Earle, 1991), duplication of classical genes and molecular markers (Zhu et al. 1994; Shoemaker et al. 1996; Lee et al. 1999), and identification homoeologous regions within soybean (Shoemaker et al. 1996). If soybean were the product of polyploidization we would expect
duplicate functional genes to map to homoeologous regions in soybean. Duplicated functional genes for appressed pubescence (Pa1 and Pa2) have previously been mapped in soybean (Lee et al. 1999). The Pa1 and Pa2 genes may confer some selective advantage to soybean and have been suggested to be involved in pest resistance (Bernard, 1973) and are associated with increased yields (Specht et al. 1985). Pa1 and Pa2 did not map to directly homoeologous regions as might be expected (Lee et al. 1999). Pa1 mapped to LG-B1 and Pa2 to LG-F. Although not directly homoeologous, LG-B1 and LG-F are each homoeologous to a third linkage group, LG-H (Figure 1A). These linkage groups were investigated further in an attempt to determine the relationship of regions containing duplicated functional genes.

Limited sequence information in soybean restricts direct genomic comparisons between soybean and other species. It therefore is necessary to develop alternative methods for genomic study. The Arabidopsis Genome Project has assembled extensive information on a plant genome and has made complex analyses of this genome possible. The *Arabidopsis thaliana* project will provide a better understanding of the composition of a plant genome as well as information on location and function of genes and will provide tools for more elaborate analysis of other plant genomes. *A. thaliana* may be used as a ‘bridging’ species to complete detailed analyses of the soybean genome.

This paper presents findings from an analysis of regions in soybean identified by duplicate functional genes and compares these regions to genomes of other species. We found two pairs of homoeologous regions within the soybean genome that are also syntenic to each other as evidenced by common markers and their relationship with *A. thaliana*. This
Figure 1. Homologous and homocologous relationships within legumes. (A)

Homocologous relationships within *G. max*. (B) The homologous relationships between *G. max* and *P. vulgaris* and (C) *G. max* and *V. radiata* of one homologous region conserved in legume evolution (the burgundy region). Blocks in figure which are not shown to have probes in common with another linkage group are homologous or homocologous to a linkage group or region of a linkage group not depicted. LG-D1a/Q and LG-F are shown in the inverted orientation as found in soybase (http://genome.cornell.edu/cgi-bin/WebAcc/webace?db=soybase).
A  Glycine max-Glycine max homoeologous relationships

B  Glycine max - Phaseolus vulgaris homologous relationships

C  Glycine max - Vigna radiata homologous relationships
supports the hypothesis of multiple polyploidization events in soybean previously suggested (Shoemaker et al. 1996; Lee et al. 1999).

**Materials and Methods**

**Homologous/ Homoeologous Identification within Legumes:** Homologous and homoeologous regions within legumes were identified by comparing molecular maps derived with common markers. As a result of the polyploidization event(s) during soybean evolution its' genome is highly duplicated and many RFLP probes map to multiple locations within the genome. Using the equation \( P = \left(1/A\right)^{B-1} \), where \( A \) is the number of linkage groups in the organism and \( B \) is the number of markers in common between two linkage groups (Reinisch et al. 1994; Shoemaker et al. 1996), we were able to determine if two linkage groups have a significant number of markers in common (homoeologous) or if their association may occur by chance.

Soybean linkage groups in this study were selected based on the results of a duplicate gene mapping study (Lee et al. 1999). Markers which had previously been mapped to regions of interest in soybean but not yet mapped in *Phaseolus vulgaris* (common bean) were sent to the University of Florida for mapping. Additional mapping was done in *P. vulgaris* as described in Vallejos et al. (2000). *Vigna radiata* (mung bean), a member of the same subtribe (Phaseolinae) as *P. vulgaris* was also included in this study.
**Sequencing:** Clones were purified using the Wizard miniprep kit (Promega, Madison, Wisconsin) and sent to the Iowa State University DNA Synthesis and Sequencing Facility (Ames, Iowa) for sequencing. Sequence was generated from both ends of each genomic clone using Universal Forward 21M13 and Reverse M13-USB primers on an ABI 377 automated sequencer. The sequence was edited in Sequencher™ 3.1 where the 5' vector sequence was trimmed off and the 3' end was edited for length. The ‘assemble automatically’ command set to default values was used on both end sequences for each RFLP probe to ascertain whether we had sequenced completely through the clone. Clones with overlapping end sequence were used as a single sequence for the remainder of the analysis.

**Homology searches:** The edited sequences were used for homology searches using BLAST programs (Altschul *et al*. 1990) at NCBI and AtDB. BLASTX was used to search the nonredundant database to determine whether the legume sequences corresponded to known sequences (http://www.ncbi.nlm.nih.gov/blast/blast.cgi). TBLASTX was used to find homologies between *A. thaliana* and legume sequences (http://arabidopsis.org/blast/). In all cases only returns with fewer than 15 hits to *A. thaliana* and E values less than $10^{-5}$ were accepted as significant.

**Legume/Arabidopsis comparison:** Homologous legume linkage blocks were compared to the *A. thaliana* genome. To identify regions of significant synteny a list of the *A. thaliana* BAC and P1 clones in their order of location on the five *A. thaliana* chromosomes (as found on http://arabidopsis.org/ in June 1999) was created. Clone identifiers of markers from the four homologous legume blocks were placed next to *A. thaliana* clone names for each RFLP clone sequence for which a TBLASTX search to *A. thaliana*. 
*Arabidopsis thaliana* detected homology. The data was analyzed for locations containing a high number of legume clones in a region of the *A. thaliana* genome that may indicate a conserved block. We ran simulations (see below) to determine whether observed relationships could occur by chance or if the pattern seen on the *A. thaliana* chromosomes was significant and may be indicative of syntenic between the legume and brassica families.

**Detection of additional duplication in soybean:** To determine if more than one duplication event occurred in the evolution of the soybean genome *A. thaliana* was used as a bridging species to compare two sets of conserved homoeologous blocks from soybean. RFLP probe sequences from Set 1 (Linkage Block B1 and its homoeologue on the top of Linkage Block H; Figure 1), and sequences from Set 2 (Linkage Block F and its homoeologue on the bottom of Linkage Block H; Figure 1) were used in this comparison. Pairs of legume markers, e.g. a marker from Set 1 and a marker from Set 2, for which TBLASTX results suggested amino acid similarity to regions of the same *A. thaliana* BAC were used to search for indirect syntenic relationships between these two sets of soybean homoeologous groups. Simulations were conducted to determine the probability that each pair of sequences from soybean would hit a single *A. thaliana* BAC by chance. Low probabilities suggest that the two legume sequences arose from genomic regions sharing a common lineage.

**Simulations:** Two simulations were performed. The first tested significance of syntenic relationships detected between legumes and *A. thaliana* to determine if the observed results were likely to occur by chance. The second simulation determined the probability that two sequences would randomly detect the same BAC in *A. thaliana* with a
TBLASTX search. To develop simulation models the strategy of Grant et al. (2000) was followed.

To determine if the relationship between legumes and *A. thaliana* was significant the *A. thaliana* genome of 586cM (http://arabidopsis.org/chromosomes/) was divided into 24 bins of 25cM each. Bins of this size were selected because the regions of homology detected in *A. thaliana* were approximately 25cM. The number of hits to *A. thaliana* detected in a TBLASTX search with each legume marker was then placed randomly into the 24 *A. thaliana* bins. This was repeated 20,000 times to determine how many times in 20,000 a pattern similar to that observed for legume marker distribution on the *A. thaliana* genome was seen (Grant et al. 2000). To ensure a conservative test, marker order was not considered.

Simulations to determine the probability two legume sequences would have a TBLASTX hit to the same BAC in *A. thaliana* by chance were accomplished by creating 1409 bins for the *A. thaliana* genome with each bin representing a single BAC or P1 clone. The number of homologous sequences in *A. thaliana* for each legume probe sequence was determined (x for sequence 1 and y for sequence 2) then randomly placed into the 1409 *A. thaliana* bins. This was repeated 10,000 times to determine how many times in 10,000 two markers with x and y number of homologous sequences to the total *A. thaliana* genome would detect the same *A. thaliana* clone.
Results

Soybean homoeologous relationships: Multiple homoeologous regions have been identified throughout the soybean genome (Shoemaker et al. 1996). In this study we focused on only a subset of these and have more precisely defined the borders of duplicated regions (Figure 1A). Linkage groups investigated here were chosen based on the map locations of duplicate genes Pa1 and Pa2 (Lee et al. 1999). LG-F and LG-B1 each were found to contain a major QTL for appressed pubescence. Although LG-F and LG-B1 are not directly homoeologous, both are homoeologous to a third linkage group (LG-H) (Figure 1A). If Pa1 and Pa2 were duplicated as a result of polyploidization we would expect them to map to homoeologous regions. These three linkage groups were investigated in attempts to identify the relationship of the regions surrounding the duplicated genes, Pa1 and Pa2.

For each linkage group, the relationships within G. max were analyzed. Six duplicate loci were detected between LG-F and LG-H (Figure 1A). The probability this would occur by chance is $3.13 \times 10^{-7}$. Seven loci are in common between LG-H and LG-B1 (Figure 1A) which would have a probability of $1.56 \times 10^{-8}$ of occurring by chance. For LG-B1 and LG-D1a/Q there are three common markers which has a probability of $2.5 \times 10^{-3}$ to occur by chance. We detected few instances of altered locus order between homoeologous groups. Most rearrangements observed were over short map distances and their map order discrepancy may be attributed to the poor resolution of map data at small intervals. Two loci (A162 and A069) stood out as being distant from their expected locations (Figure 1A). It is possible that these are due to chromosomal rearrangements or a duplication and insertion
event. However, we will discuss an alternative hypothesis for the unexpected locations of these markers.

**Legume homologous relationships:** Homologous regions for the linkage groups studied here were more fully characterized between legume species than previously reported (Boutin *et al.* 1995). *P. vulgaris* and *V. radiata* are closely related to each other and as a result a large portion of their genomic maps are almost completely co-linear. For instance, six common markers are detected between *Phaseolus* LG-Fl-J and *Vigna* LG-Minn-5. At the same time map comparisons showed these linkage groups not to have common markers with any other linkage group in either species (data not shown).

Map comparisons between soybean (subtribe Glycineae) and *P. vulgaris* and *V. radiata* revealed fewer co-linear regions. However, significant similarities were still found. We detected two regions in *G. max* showing homology to a single region in each of the diploid legumes, *P. vulgaris* and *V. radiata* (Figure 1B-C). There appears to be complete marker order conservation between LG-H and LG-Fl-J. However, there is one possible marker (*Bng067*) rearrangement between LG-B1 and LG-Fl-J. This apparent rearrangement is detected within a very short map distance and could be explained by mapping error. Few deviations from marker order conservation were detected between LG-H and LG-Minn-5. These rearrangements could represent duplication of smaller segments. Similar observations are seen for linkage groups not presented in this manuscript.

Disjunct regions on LG-D1a/Q show homology to the same conserved block in *P. vulgaris* and *V. radiata* (Figure 2). These regions are separated on LG-D1a/Q by another homoeologous conserved block (to LG-N) in soybean. This probably occurred by either an
Glycine max D1a/Q homologous relationships with Phaseolus vulgaris and Vigna radiata

Figure 2. Displaced homoeologous segment on linkage group D1a/Q in soybean as revealed through comparisons with P. vulgaris and V. radiata.

inversion event within the linkage group or insertion of the region homoeologous to LG-N after the divergence of Glycineae and Phaseolineae.

The homologous blocks defined among the legumes appear to be larger than predicted by comparing any two linkage groups. Using mapping data we were able to compare all legume homologous and homoeologous linkage groups concurrently. The block
shown on any single linkage group represents a composite of the conserved region as detected from the other three homologous or homoeologous regions. Thus, the borders of a conserved block are generally not defined by a single related pair of linkage groups.

**Legume-Arabidopsis relationships:** Synteny between Arabidopsis and soybean has been previously detected (Grant et al. 2000). In order to assess the extent of conservation between legumes and *A. thaliana*, DNA sequences from RFLP markers defining a set of legume homoeologous/homologous regions were compared to the *A. thaliana* sequence database. We found evidence for conservation of the same blocks of DNA in *A. thaliana* as were detected in the legume lineage. Within a 25 cM region near the bottom of *A. thaliana* chromosome V (arabV) 13 TBLASTX results were returned from sequences of probes mapping to homologous blocks in legumes and having E ≤ 10\(^{-5}\) (Figure 3A). As evidence of genome duplication in *A. thaliana* we also observed 16 TBLASTX returns with E ≤ 10\(^{-5}\) within a 25 cM region near the bottom of *A. thaliana* chromosome II (arabII) which was detected with sequences from probes mapping to the same homologous legume blocks (Figure 3B). Markers mapping to more than one of the four legume blocks were counted only once in these totals.

In order to determine if the co-occurrence of markers was due to evolutionary relationship of these regions in legumes and *A. thaliana* we used simulations to determine how often the pattern we found might occur by chance. In 20,000 simulations the number and pattern of TBLASTX hits we observed in these regions was never observed. Therefore, the probability that our observations occurred by chance is < 5x10\(^{-5}\). Although these RFLP probe sequences show homology to regions other than arabII and arabV, no other *A.
Figure 3. Syntetic relationship between four homologous regions of the three legume species and two regions in *A. thaliana* demonstrating duplication in *A. thaliana*. 
*thaliana* regions were found to have a significant number of homologies to sequences from these legume homologous blocks.

**Additional duplication in soybean:** Two soybean RFLP markers (*A069* and *A162*) and a pair of duplicated genes (*Pa1*, *Pa2*) mapped to unexpected locations based upon homoeologous relationships within soybean (Figure 1A). Interestingly, *Pa1* and one map location each of *A069* and *A162*, mapped to one pair of homoeologous regions while *Pa2* and a second map location of *A069* and *A162* mapped to another pair of homoeologous regions (Figure 4A; burgundy colored region and blue colored region, respectively). These locations suggested a possible evolutionary relationship between these two pairs of homoeologous regions. An earlier duplication in the region investigated might explain why the functional duplicated genes (Lee *et al.* 1999; *Pa1* and *Pa2*) were not located in directly homoeologous regions.

To investigate the possibility of an evolutionary relationship between the two sets of homoeologous regions we used TBLASTX to identify the genomic positions of all pairs of legume RFLP probe sequences with significant ($E < 10^{-4}$) homologies to sequences found on a single *A. thaliana* BAC or PI clone. In 75% of the cases (33 of 44) whenever two RFLP probe sequences are associated with the same *A. thaliana* BAC or PI clone, the RFLP markers are linked or map to known homoeologous regions within the soybean genome. This frequency is likely an underestimate since banding patterns of most RFLP probes indicate their sequence is present in more locations in soybean than have been mapped. Thus, some homoeologous regions have not yet been identified.

Several instances were observed in which sequences of markers from both sets of homoeologous regions were homologous to the same *A. thaliana* BACs (Table 1). We then
Figure 4. Paleoeeoctoploid model of genome evolution leading to the current organization of the soybean genome. This model invokes two rounds of polyploidization each followed by diploidization. (A) The current genome organization in soybean showing homoeologous regions. The block on LG-F includes a dotted box representing soybean sequences not detected in the blue homoeologous regions from comparisons within legumes, but inferred through comparisons to *A. thaliana*. (B) Ancestral regions of homoeologous groups identified in (A). The ancestral regions represent a composite of the markers that were likely present in the ancestor. Some markers mapping to one region in soybean but not in the homoeologous group may be present but not mapped for lack of polymorphisms. It is possible the DNA sequences have been eliminated or recently attained in one of the homoeologous regions during diploidization. The ancestral blue and burgundy regions were compared for similarity and found to be homoeologous. They have three markers in common (*Pa*, *A069*, and *A162*) and at least three sequences detected with a TBLASTX search to *A. thaliana* are descended from a common region. Additional segmental duplications of sequences occurred independently in each region following duplication. The common markers are connected by solid lines; the RFLPs shown to be related through a TBLASTX search are connected by dashed lines. One pair of RFLPs are connected by a thicker dashed line, this pair was found concurrently on three separate BACs in *A. thaliana*. (C) The ancestral regions are likely descended from a common ancestor. Purple is used to signify it as an ancestor of the blue and burgundy regions. The duplicate loci in A and B are only present once in C for clarity. *Pa* is used to identify the ancestral gene of *Pa1* and *Pa2*. 
Table 1. Pairs of RFLP probes with hits to a single *A. thaliana* BAC.

<table>
<thead>
<tr>
<th>Burgundy ancestral RFLP marker</th>
<th>Blue ancestral RFLP marker</th>
<th><em>A. thaliana</em> BAC</th>
<th>Probability of occurrence by chance</th>
</tr>
</thead>
<tbody>
<tr>
<td>A131 (3)</td>
<td>A245 (1)</td>
<td>MVE11</td>
<td>0.0019</td>
</tr>
<tr>
<td>A131 (3)</td>
<td>Bng100 (5)</td>
<td>T30D6</td>
<td>0.0114</td>
</tr>
<tr>
<td>Bng067 (1)</td>
<td>K102 (1)</td>
<td>MYA6</td>
<td>0.0001</td>
</tr>
<tr>
<td>H3-28a (13)</td>
<td>K102 (1)</td>
<td>MYA6</td>
<td>0.0078</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A132 (5)</td>
<td>B212 (15)</td>
<td>T8P21</td>
<td>0.0000</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Burgundy and Blue genomic regions are as depicted in Figures 1 and 4.*

*Numbers in parentheses indicate the number of individual homologous sequences identified in *A. thaliana* using TBLASTX queries and the soybean RFLP marker sequence as the query sequence.

*A132 and B212 hit the same *A. thaliana* BAC three separate times. In this case probability was determined for three occurrences of two sequences showing homology to the a single *A. thaliana* BAC.*

Ran simulations to determine the probability that TBLASTX results would detect two RFLP probe sequences in a single *A. thaliana* BAC or P1 clone by chance. The probability ranged from 0.0000 to 0.0114 in 10,000 simulations (Table 1). Therefore, the likelihood of multiple homologous relationships occurring by chance is near zero. These results strongly support...
our hypothesis of an evolutionary relationship between the two sets of homoeologous segments (Figure 4) and argues for a duplication event prior to the recent polyploidization in soybean.

Discussion

Our investigation began when it was determined that duplicated functional genes in soybean (Pa1 and Pa2) did not map to directly homoeologous regions as expected if they were duplicated as a result of polyploidization (Lee et al. 1999). To determine the origins of Pa1 and Pa2 we analyzed the genomic regions surrounding these genes and compared these regions to homologues in other species.

As expected, homoeologous blocks within soybean are easily identifiable as are significant homologies with other legumes (Figure 1). Homologous/homoeologous blocks detected between soybean and other legumes are larger than those previously reported (Boutin et al. 1995) and frequently encompass large segments of linkage groups. We confirm the polyploid origin of soybean by showing that duplicated blocks in G. max are homologous to a single block in each of the diploid legumes, P. vulgaris and V. radiata. We extended our study to determine whether conserved blocks detected in legumes have remained intact from a time prior to divergence of legumes and brassicas; approximately 90 MY (Gandolfo et al. 1998). Comparisons of soybean genomic clone sequence data to the A. thaliana genome database provided evidence of significant genomic conservation between
legumes and *A. thaliana*. The use of BLAST searches to detect homology between legumes and *A. thaliana* also identified a duplication between arabII and arabV (Figure 3). Grant *et al.* (2000) reported evidence for genomic duplication in *A. thaliana* involving other chromosomal regions. Taken together, these results suggest polyploidization occurred during the evolution of the *A. thaliana* genome.

Results from our study suggest that genome evolution occurs by reshuffling of conserved blocks of DNA in the legume family and we present evidence that these blocks have remained intact since before the split of the legume and brassica lineages. Reorganization of blocks of DNA in genomes has been demonstrated in other species (Moore *et al.* 1995; Kowalski *et al.* 1994; Lagercrantz, 1998). The occurrence of this phenomenon in several lineages supports there being a selective advantage in the retention of certain cassettes of genes or DNA sequences in a genome.

Extensive diploidization at the chromosomal level following polyploidization is reflected in the fragmented structure of the duplicated regions within soybean. Genomic change occurring rapidly following polyploidization events has been reported in brassicas and wheat (Song *et al.* 1995; Feldman *et al.* 1997). Extensive changes may have also occurred in soybean. It is also possible that the genome duplication in soybean was the result of an allopolyploidization event bringing together two distinct genomes which had diverged prior to joining in a single nucleus, although there is no current evidence favoring allopolyploidy over autopolyplody in soybean.

It is possible that homologous regions in soybean are not as fragmented as current genetic maps suggest. The soybean genome is relatively monomorphic, making mapping of multiple locations of RFLP probes difficult and prevents detection of duplicated loci and
homoeologous regions. RFLP markers present in one region may also be in the duplicated homoeologue but have no known map location there due to lack of polymorphisms. As more sensitive mapping techniques become available it should be possible to identify additional homoeologous regions in soybean.

Comparison of soybean to *V. radiata* and *P. vulgaris* reflects the evolutionary relationship of the three legumes. Rearrangements in chromosomes and gene order would be expected since the genomes/subgenomes have been diverging since their last common ancestor. However, in one case it appears that a linkage group in soybean (LG-H) has more homology to the diploids than its homoeologue (LG-B1) (Figure 1B-C). However, no other pairs of homoeologs in *G. max* show a similar relationship to the diploids (data not shown). Therefore, this observation may be due to insufficient mapping data for Linkage Group B1.

Two RFLP markers (*A069* and *A162*) and one gene (*Pa*) were identified in common between the ancestral groups (shown by blocked yellow loci in Figure 4). These loci mapped to locations other than what would be predicted from a single polyploidization event in the soybean genome although their locations may be explained by a second round of duplication followed by diploidization events (Figure 4). If each homoeologous region in soybean contains a subset of gene/markers present in the ancestral region prior to polyploidization we can predict the ancestral DNA segment for each of the homoeologous regions (Figure 4B).

Analysis of the ancestral regions indicates additional duplications likely occurred in one or more species or conserved blocks during evolution of the region in Figure 4. This is inferred from the sequence of one marker from one homoeologue (e.g. *K102*; top of blue region; Figure 4B) and sequences of two markers from the other homoeologue (e.g. *Bng067* and *H3_28*; top of burgundy region; Figure 4B) showing homology to a single *A. thaliana*
BAC (Table 1). Therefore, either K102 has been duplicated in *A. thaliana* or a segmental duplication encompassing *Bng067* and *113_28* was duplicated in the ancestor of the burgundy colored region. A similar relationship is also detected with marker *A131* from the burgundy region and *Bng190* and *A242* from the blue region (Figure 4B; Table 1). The segment encompassing *A132* and *B212* appears to have been duplicated in *A. thaliana*. We can then conservatively estimate three segments in common between the burgundy and blue colored ancestral regions as revealed through the use of *A. thaliana* as a bridging species.

If we employ the equation $P = \left(\frac{1}{A}\right)^{(B-1)}$ and assume 10 chromosomes and six sequences in common between the ancestral regions (three using the bridging species + three from direct comparison of the regions; Figure 4B) there is a probability of only $10^{-5}$ that we would see such similarity by chance. This strongly suggests these regions are descended from a common ancestor. Thus, soybean may not simply be a paleotetraploid, but may be a paleooctoploid.

If two rounds of polyploidy occurred in the soybean genome then four homoeologous locations would be predicted in the current genome. *Pa1* and *Pa2* may be from an earlier duplication as shown in Figure 4. We would then predict that there were at one time four *Pa* genes. The two currently identified by phenotype plus two in the directly homoeologous regions. *Pa* related genes not currently identified have been either deleted, silenced, or are present but have not been detected in any mapping populations.

In soybean it was possible to identify regions from the most recent polyploidization event by investigating the soybean genome alone (Shoemaker *et al.* 1996). However, evidence for an earlier duplication was not detected until we included *A. thaliana* as a
bridging species. Based on our hypothesis, a duplication or polyploidization event in \textit{P. vulgaris} and \textit{V. radiata} would be as ancient and difficult to detect as the earlier duplication event in \textit{G. max}. If similar studies were carried out involving \textit{P. vulgaris} and \textit{V. radiata} it may also be possible to reveal an ancient duplication equivalent to the earlier event we detected in soybean.

Our detection of ancient duplication and possibly polyploidization events by using \textit{A. thaliana} as a bridging species demonstrates how this can be a powerful technique for the investigation of plant genomes for evidence of paleopolyploidy and/or for identification of multiple rounds of paleopolyploidy not only within the legume family but also in other lineages.

\textbf{Acknowledgments}

This is Journal Paper J\________ of the Iowa Agriculture and Home Economics Experiment Station, Ames, Iowa (Project 3236), and from the U.S. Department of Agriculture (USDA), Agricultural Research Service, Corn Insect and Crop Genetics Research Unit. The mention of a trademark or proprietary product does not constitute a guarantee or warranty of the product by Iowa State University or the USDA and does not imply its approval to the exclusion of other products that may also be suitable.
References


Lagercrantz, U., 1998 Comparative mapping between *Arabidopsis thaliana* and *Brassica nigra* indicates that Brassica genomes have evolved through extensive genome
replication accompanied by chromosome fusions and frequent rearrangements.

Genetics 150: 1217-1228.


Moore, G., T. Foote, T. Helentjaris, K. Devos, N. Kurata, et al., 1995  Was there a single
ancestral cereal chromosome?  Trends In Genet. 11: 81-82.

unified genetic map of higher plants, transcending the monocot-dicot divergence.
Nature Genetics 14: 380-382.

RFLP map of cotton.  Gossypium hirsutum X Gossypium barbadense:  chromosome

Riley, R. and V. Chapman, 1958  Genetic control of the cytologically diploid behaviour of

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

Song, K., P. Lu, K. Tang and T. C. Osborn, 1995  Rapid genome change in synthetic
Sci. USA 92: 7719-7723.


CHAPTER FOUR: CONCLUSIONS

Conclusions

Throughout the past century numerous studies have investigated various aspects of soybean biology. Projects have ranged from classical genetic studies to improve yield to basic research investigating the organization and evolution of the genome. The research presented in this dissertation was conducted in an effort to advance the current knowledge of the soybean genome and contribute to the understanding of ancient polyploid genomes. The primary foci of this study were polyploidy as a mechanism of gene duplication and identification of genomic changes based on current genomic organization.

This dissertation began with the mapping of duplicated functional genes ($Pal$ and $Pa2$), the first duplicate genes mapped in soybean which could be attributed to a phenotype. $Pal$ and $Pa2$ were proposed to have arisen through polyploidization and would be expected to map to homoeologous regions. $Pal$ and $Pa2$ did not map to directly homoeologous regions which prompted a more extensive investigation of the regions surrounding these genes. We found that each of the linkage groups to which $Pal$ and $Pa2$ mapped (LG-B1 and LG-F respectively), although not directly homoeologous, showed homoeology to linkage group LG-H.

Identification of duplicate regions in soybean has been previously described (Shoemaker et al., 1996) as well as the conservation of genomic blocks between soybean and other legumes (Boutin et al., 1995). In this study, concurrent comparison of genetic maps of
Glycine max, Phaseolus vulgaris, and Vigna radiata allowed us to extend the boundaries of the blocks conserved between these three species. We demonstrated that two regions in G. max showed homology to a single region in P. vulgaris and V. radiata. We also discovered that conserved blocks in soybean have been reshuffled since the most recent polyploidization event, similar to results observed in grasses (Moore et al., 1995; Helentjaris et al., 1988) and brassicas (Kowalski et al., 1994; Lagercrantz, 1998).

Genomic conservation between soybean and Arabidopsis thaliana has been demonstrated previously (Grant et al., 2000). We compared soybean with A. thaliana to investigate the origins of the duplicate genes, Pal and Pa2. We found evidence that the blocks of DNA conserved among legumes appeared to also be conserved within the brassica family. The retention of this cassette of DNA sequences over 90 MY suggests that it may confers a selective advantage. Although random stochastic process could result in the retention of large regions of DNA. However, if other conserved regions are also detected the probability these regions are conserved for a biological reason would be favor of retention through random processes.

In addition we found evidence that two pairs of homoeologous regions in soybean are related through an earlier duplication event. This additional duplication in soybean may be indicative of a second round of polyploidization. Because the identifiable sequences in each block were a subset of sequences present in the ancestral region, we predicted the complement of sequences present in a segment prior to polyploidization. The ancestral regions of the current homoeologous pairs were then compared for similarities. There was a significant amount of synteny between the predicted ancestral regions of LG-B1 and LG-H (top) with LG-H (bottom) and LG-F, as presented in chapter 3. These regions are likely the
result of a duplication event prior to the most recent polyploidization in soybean. The hypothesis that soybean may have undergone more than one round of polyploidization has been previously proposed (Shoemaker et al., 1996; Lee et al., 1999). Evidence of related pairs of homoeologous ancestral regions support this hypothesis.

*Pa*1 and *Pa*2 may be the result of two duplication events in soybean. We would then predict that four copies of appressed pubescence genes are present in the soybean genome. However, only two locations are known. This may be a result of the undetected loci being lost or silenced during diploidization or a lack of populations segregating for two of the *Pa* genes.

The utilization of *A. thaliana* as a bridging species provided evidence for two rounds of genome duplication in soybean. However, more regions of the soybean genome should be investigated to support or reject this hypothesis. Other hypotheses may also explain the duplication in soybean and should be considered.

**Future Research**

The research presented here provides a springboard for numerous projects involving soybean and the inclusion of additional species. Additional research may provide evidence for the discrimination of polyploidization verses segmental duplication in soybean and *A. thaliana*. It may also be possible to determine if duplication occurred once before the divergence of soybean and *A. thaliana* or if distinct events occurred in each lineage.
Detailed investigation of the remainder of the soybean genome may reveal evidence of additional duplications in the genome. If two rounds of duplications are detected in other regions of the soybean genome polyploidization may be the cause of the earlier duplication. However, if the additional duplication identified between the homoeologous pairs LG-B1/LG-H and LG-H/LG-F is the exception, segmental duplication of this region would be favored. Other legumes should also be investigated for genome duplications to determine if the additional duplication detected in soybean occurred prior to the divergence of Glycineae and Phaseolinoae or subsequent to this division.

The genome sequencing project of *A. thaliana* may make it possible to distinguish if polyploidization or segmental duplication occurred in *A. thaliana*. If the majority of sequences in the *A. thaliana* genome are duplicated we can predict they arose from a polyploidization event. If duplicated sequences are the exception, then the duplication detected in *A. thaliana* is more likely from a segmental event. The difficulty will be to distinguish the amount of duplication necessary to favor polyploidy over segmental duplication. How much divergence may be accounted for from diploidization? If only a limited amount of sequences are determined to be duplicated in *A. thaliana* should polyploidy followed by extensive diploidization be eliminated as a possibility? Comparisons of legumes and brassicas may also permit researchers to determine if duplication occurred before the divergence of these two families or individually in each lineage.
References


Shoemaker, R. C., K. Polzin, J. Labate, J. Specht, E. C. Brummer, T. Olson, N. Young, V.


ACKNOWLEDGEMENTS

There are many people I would like to thank for their encouragement, support, or guidance during my graduate career. I would begin by thanking my major professor Dr. Randy Shoemaker. Although my graduate career was not always easy, we have become friends. To I am grateful to the rest of my POS committee members, Dr. Jonathan Wendel, Dr. James Colbert, Dr. Reid Palmer, and Dr. John Imsande, for encouragement and guidance as I complete my graduate research and for ensuring I have obtained a quality education. I am appreciative to the members of my lab, our collaborators and the Interdepartmental Genetics Program, especially our outstanding secretary Linda Wild, for assistance throughout my graduate work. I would last like to thank all my friends and family for both encouragement and support during my time in graduate school. Most of all my parents, Tom and Francy Lee. My sisters, Wendy, Cyndi and of course my twin sister Sandy, my five adorable nephews, Brian, Lance, Daniel, Josh, and Jacob and my dear friends James Rudd, Stefan and Andrea Kaskle, Shannon and Brad Zingula, Larissa and Eric Mottl, and Xiangqin Cui.