ABSTRACT

Title of Dissertation: GENETIC DIVERSITY AND LINKAGE DISEQUILIBRIUM IN WILD SOYBEAN, LANDRACES, ANCESTRAL, AND ELITE SOYBEAN POPULATIONS

David Lee Hyten, Jr., Doctor of Philosophy, 2005

Dissertation Directed By: Associate Professor Jose M. Costa, Department of Natural Resource Sciences and Landscape Architecture

Domestication, founder effects, and artificial selection can impact populations by reducing genome diversity and increasing the extent of linkage disequilibrium (LD). To understand the impact of these genetic bottlenecks and selection on sequence diversity and LD within soybean [*Glycine max* (L.) Merr.], 111 genes and three chromosomal regions located on linkage groups A2, G, and J were characterized in soybean. Four soybean populations were evaluated: 1) the wild ancestor of soybean (*G. soja*), 2) the population resulting from domestication (landraces), 3) Asian introductions from which North American cultivars were developed (ancestors), and 4) elite cultivars from the 1980's (elite). A total of 438 single nucleotide polymorphisms (SNPs) and 58 insertions-deletions were discovered within the 102 genes. Sequence diversity was lower than expected in *G. soja* with an overall theta equal to 0.00235, and was less than half that value (theta = 0.00115) in the landraces. Domestication eliminated most unique haplotypes with *G. soja* containing 240 unique haplotypes while the landraces only contained 42 unique haplotypes. The founder effect of the introduction of soybean to North America followed by intensive artificial selection, resulted in only a 30% decrease in nucleotide diversity. A total of 738 SNPs were discovered and genotyped in the four populations throughout three chromosomal regions. In *G. soja* LD did not extend past 100 kb while in the three cultivated soybean populations LD extended from 90 kb up to 600+ kb, most likely as a result of increased inbreeding and domestication. The three chromosomal regions varied in the extent of LD within the populations. *G. soja* is the greatest resource for unique alleles and may be best suited for fine mapping utilizing association analysis. The landraces do not contain much more variability than the elite cultivars but may have enough diversity to facilitate genetic improvement of elite cultivars. Finally, due to the extended levels of LD in the landraces and the elite cultivars, whole genome association analysis may be possible for the discovery of QTL.

GENETIC DIVERSITY AND LINKAGE DISEQUILIBRIUM IN WILD SOYBEAN, LANDRACES, ANCESTRAL, AND ELITE SOYBEAN POPULATIONS

By

David Lee Hyten, Jr.

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2005

Advisory Committee:
Associate Professor Jose M. Costa, Chair
Adjunct Professor Perry B. Cregan
Associate Professor Charles B. Fenster
Professor William J. Kenworthy
Professor Marla S. McIntosh

Dedication

I would like to dedicate this dissertation to my wife Aimee for giving me unlimited love, support, and encouragement, and to my family especially my parents, David Lee Hyten, Sr. and Barbara Joyce Hyten, for supporting and giving me motivation all of these years.

Acknowledgements

I wish to thank all those who helped me in completing my degree in Doctorate of Philosophy in the Natural Resources and Landscape Systems Department. I thank Dr. Perry B. Cregan for being a great research advisor. He has not only given me great advice on how to plan and conduct my research but has been a good friend that has helped me throughout my project. I would also like to thank my academic advisor and advisory committee chair professor Dr. Jose M. Costa. He has gone the extra mile in giving me good advice and helping me make it through to get my degree. I would also like to thank my other committee members, Dr. Charles B. Fenster, Dr. William J. Kenworthy, and Dr. Marla S. McIntosh for giving me great advice on my project along with being very encouraging. I thank the department of Natural Resources and Landscape Systems and The United States Department of Agriculture-Beltsville Agricultural Research Station, and the Soybean Genetics and Improvement Laboratory at the USDA-ARS. I thank all the people that have collaborated with me on this project which include Dr. Randy Nelson for selecting the *G. soja* and the landrace populations, Dr. Randy Shoemaker for donating the BAC end sequences for the BAC contig on linkage group J, and Dr. James Specht for mapping discovered SNPs for placement on the soybean composite map. I thank all those that have given me technical support which includes Dr. Lakshmi Matukumalli for setting up the software for SNP detection, Dr. Qijiang Song for helping me with analysis of my data, and Dr. Tad Sonstegard and Mrs. Tina Sphon for running my sequencing reactions at the BFGL-ANRI BARC East DNA Sequencing Facility. I thank everybody that has given me general encouragement and advice throughout this project; Dr. Ik-Young

Choi, Mr. Raymond Fernald, Mr. Edward Fickus, Dr. Nicolle Hofmann, Ms. Eun-Young Hwang, Mr. Chuck Quigley, Mrs. Rita Walker, and Mrs. Ronghui Yang. I would like to the United Soybean Board and USDA-ARS for funding my project.

Last but not least I would like to thank my family, my wife, and my friends, whose help, support, and encouragement gave me the opportunity and the motivation to achieve this project.

Table of Contents

List of Tables

List of Figures

Chapter 1: Literature Review

Introduction

Evolution is a powerful force affecting biological systems in many complex ways. The genomes within populations demonstrate the effects of evolutionary events in the form of single nucleotide differences in homologous DNA fragments plus small insertions and deletions (indels), collectively referred to as single nucleotide polymorphism (SNP) variation. Evolutionary events also have the potential to affect the extent and the pattern of linkage disequilibrium (LD) between SNPs. Linkage disequilibrium is the non-random association of alleles in a population. Few studies in plants have explored the amount of SNP variation and extent of LD before and after evolutionary events such as domestication and human selection. Recently, advances in large-scale DNA sequencing and genotyping have led to many new possibilities for studying population variation at the sequence level. Information on SNP variation and the pattern of LD will facilitate the mining of germplasm for unique alleles and quantitative trait loci (QTL) discovery and fine mapping through association analysis.

Plant geneticists have been slow to adopt SNPs and LD as tools for QTL discovery and for the study of the effects of domestication and selection on natural and artificial populations. Maize is the only plant species in which preliminary studies have gained insights into the amount of variation found before and after domestication and the creation of highly inbred lines used for modern plant breeding (TENAILLON *et al.* 2001; TENAILLON *et al.* 2004). Association analysis, which uses information about the structure and extent of LD in an organism, is emerging as an alternative method for QTL discovery and fine mapping and has been successfully applied in a few plant species (PALAISA *et al.* 2004; THORNSBERRY *et al.* 2001; TROGNITZ *et al.* 2002). However, research on the extent of LD in plant species is needed in order to develop strategies to implement association analysis. To date, most studies exploring SNP variation and LD in plants have sampled single genes or small continuous regions of DNA (HAGENBLAD and NORDBORG 2002; NORDBORG *et al.* 2002; REMINGTON *et al.* 2001; TENAILLON *et al.* 2001). With additional research, a clearer understanding of SNP variation, LD extent, and LD structure in plants could facilitate the practical application of association analysis for genetic improvement of crops.

Soybean provides an opportunity to assess the effects of domestication, founder population effects, and intensive artificial selection because of the availability of populations occurring before and after these evolutionary events. Four main populations have been formed through soybean domestication, dissemination, and crop improvement. The four soybean populations include germplasm which represent pre-domestication (*G. soja*), post-domestication landraces (landraces), Asian introductions from which North American cultivars were developed (ancestors), and North American elite cultivars (elite). The development of a model for studying SNP variation and LD in soybean could be helpful to mine germplasm for unique alleles and QTL discovery followed by fine mapping of genes responsible for the QTL through association analysis.

SNP Variation

SNPs are usually bi-allelic and can be used as molecular markers and for a number of purposes including the assessment of diversity, QTL discovery, association analysis, and marker assisted selection. SNPs have two main advantages over other molecular markers; firstly they are the most common form of genetic variation within genomes and secondly a wide array of technologies have been developed for high throughput SNP analysis (LEE *et al.* 2004). Some of the technologies have achieved multiplexing capabilities of 1536 SNPs in a single assay and over 500,000 genotypes achieved with two technicians in a three day period (www.illumina.com). Only recently have there been large scale SNP discovery efforts in a number of species to obtain an understanding of the amount of SNP variation within different genomes and the effects evolutionary events have on SNP variation (SCHMID *et al.* 2003; TENAILLON *et al.* 2004; ZHU *et al.* 2003).

One method of SNP discovery involves resequencing of gene fragments in multiple individuals. Through this method an understanding of the level of sequence variation is starting to develop. The frequency of SNPs varies greatly between species and between genes within the same species (BROWN *et al.* 2004; ZHU *et al.* 2003). Two common measurements used to quantify SNP variation are the expected heterozygosity per nucleotide site (π), (TAJIMA 1983) and nucleotide diversity (θ), which is the proportion of polymorphic sites in a sample corrected for sample size that is insensitive to the allele frequency of segregating nucleotides (WATTERSON 1975).

Several early studies explored SNP variation on a single gene level. The first estimates of nucleotide diversity in soybean came from single gene studies and found diversity ranged from θ = 0.00085 (SCALLON *et al.* 1987) to θ = 0.015 (ZHU *et al.* 1995). The 17-fold difference between these two estimates suggests that an ideal way to study SNP variation within a species is to analyze multiple genes among multiple individuals. Currently, few studies are available that give estimates of nucleotide diversity on multiple genes within a species. Some of the available studies include nucleotide diversity estimates for *Drosophila*, human, maize, loblolly pine, and soybean (BROWN *et al.* 2004; CARGILL *et al.* 1999; HALUSHKA *et al.* 1999; TENAILLON *et al.* 2001; WANG *et al.* 2004; ZHU *et al.* 2003). Estimates of nucleotide diversity in these five species indicate that nucleotide variation varies over 10-fold between species. Humans have been found to have the lowest levels of nucleotide diversity among the five species. Cargill *et al*. (1999) resequenced 196 kb of sequence in 106 genes from 57 individuals composed of European, Asian, African American, and African Pygmy descent. Nucleotide diversity ranged from θ = 0.0 to 0.0026 with an average θ = 0.00054. Halushka *et al*. (1999) found θ ranged from 0.0 to 0.0032 with an average θ = 0.00083. Their study consisted of 75 genes in 80 individuals of African, European American, and Northern European descent. A recent study in soybean has found very similar nucleotide diversity patterns. Zhu *et al*. (2003) resequenced 76 kb of genomic DNA in 116 genes and 28 non-genic regions from 25 diverse individuals. They found nucleotide diversity was only 0.00097 with a range of 0.0 to 0.00126. The range of nucleotide diversity is smaller in soybean than the previous two studies in humans.

The similar estimates of nucleotide diversity in humans and soybean are not representative of other organisms. Other species such as loblolly pine (θ = 0.0041) (BROWN *et al.* 2004), *Arabidopsis* (θ = 0.0071) (SCHMID *et al.* 2005), *Drosophila* (θ = 0.0070) (MORIYAMA and POWELL 1996), and maize (θ = 0.0096) have 4.2 to 10-fold more variation than soybean and humans. The most surprising difference comes from comparing *Arabidopsis* to soybean. An average θ value of 0.0071 for *Arabidopsis* (SCHMID *et al.* 2005) is 7.3 times greater than for soybean (ZHU *et al.* 2003). The two species are >99% self-fertilizing which is thought to decrease nucleotide variation due to background selection and genetic hitchhiking with selective sweeps (SCHMID *et al.* 2005). It is apparent that other genetic bottlenecks or other evolutionary events have decreased soybean nucleotide diversity.

Evolutionary events effect on SNP Variation

Mutation, migration, selection, and random genetic drift are the four factors that affect SNP frequencies in genomes (HALLIBURTON 2004). These factors encompass many evolutionary events including domestication, founding events, artificial selection, and natural selection (HALLIBURTON 2004). Domestication and founding events have the ability to create genetic bottlenecks greatly reducing the amount of genetic variation within a species (TANKSLEY and MCCOUCH 1997). Natural and artificial selection can also have the effect of decreasing or increasing genetic variation.

The effects of evolutionary factors on nucleotide sequence diversity in different populations have only begun to be studied in maize (TENAILLON *et al.* 2001; TENAILLON *et al.* 2004). Maize underwent a single domestication event in Mexico approximately 9000 years ago with several domesticated landraces being formed (MATSUOKA *et al.* 2002). This domestication event only reduced maize nucleotide diversity by ~20% as estimated from eight neutral genes located on chromosome one (TENAILLON *et al.* 2004). Four additional genes located on chromosome one have likely undergone selection after domestication with SNP variation reduced by >65% (TENAILLON *et al.* 2004). The later finding of a large reduction in the genes under selection is somewhat surprising. The effect of domestication on one of the four genes (*tb1*) has been extensively studied and was found to only have reduced variation in the 5' untranslated region of the gene while the exons and the upstream 5' untranslated region had similar amounts of diversity in maize and teosinte (WANG *et al.* 1999). Further work with many more loci on multiple chromosomes is needed to understand the true genome-wide effects that domestication has on a crop species.

Maize has also undergone a major genetic bottleneck common to most crop species. This genetic bottleneck was the result of the creation of elite inbred lines which served as parents for modern maize hybrid. The elite U.S. inbred lines were selected from only a few landraces which probably represent a small fraction of the genetic diversity available in the landraces (TENAILLON *et al.* 2001). The reduction in genetic diversity between 16 exotic landraces and nine U.S. elite inbred lines was 33% in 21 loci (TENAILLON *et al.* 2001). This is a slightly greater reduction than that which occurred as a result of domestication but is still not a statistically significant reduction. While these two bottlenecks of domestication and the development of inbred lines may significantly reduce variation when combined they do not significantly reduce variation when taken alone. If these genetic bottlenecks have relatively small effects on nucleotide diversity, it raises the question of the extent to which genetic bottlenecks have affected LD in populations after such events.

The Extent of LD in Humans

Most knowledge of the structure or the patterns of LD has been gained from research on the human genome. Several recent studies suggest LD structure in humans is best described using a haplotype block model. Haplotype blocks are consecutive sites in high LD flanked by blocks demonstrating historical recombination (DALY *et al.* 2001; GABRIEL *et al.* 2002). Several advances in human research have come from haplotype analysis of LD blocks. Haplotype analysis of LD blocks has helped to detect natural selection acting upon disease resistance alleles (TISHKOFF *et al.* 2001). In addition, complex migration patterns of humans have been resolved through haplotype analysis (REICH *et al.* 2001). Case-control studies have been successfully used in humans as a result of a causative mutation being in an LD block with markers in the same block (HELMS *et al.* 2003).

Daly *et al*. (2001) studied a 500 kb region responsible for a risk factor of Crohn’s disease on chromosome 5q31. A block-like pattern of LD emerged from their case-control study. The 500 kb region contained 11 blocks of sequence in which a SNP contained within each block was in high LD with every other SNP within the same block irrespective of distance. The size of the blocks ranged from 3 kb up to 92 kb containing only two to four haplotypes that could be identified with a small subset of SNPs contained within the block. Another study by Jeffreys *et al*. (2001) reported a haplotype block structure of LD within a 216 kb segment of the class II region of the major histocompatibility complex. The blocks of high LD were found to be flanked by 1 to 2 kb regions of rapidly decaying LD, which corresponded to recombinational hotspots found through sperm typing. This led to the hypothesis that most variation within the human genome may be transmitted from generation to generation as blocks broken up by recombination hot-spots at defined locations.

To test this hypothesis, studies were undertaken to examine several regions in the human genome to determine if most of the genome is maintained as relatively unchanging haplotype blocks and also how these blocks vary across different populations (GABRIEL *et al.* 2002; PATIL *et al.* 2001; REICH *et al.* 2001; SHIFMAN *et al.* 2003). Larger scale studies have confirmed that most variation between genomes is contained within haplotype blocks (GABRIEL *et al.* 2002; REICH *et al.* 2001). One study looked at LD structure in 51 autosomal regions of 250 kb each (GABRIEL *et al.* 2002). LD blocks ranged from <1 to 173 kb in different populations with over half of the genome contained in blocks of 22 kb or larger in African and African-American samples and 44 kb or larger in European and Asian samples (GABRIEL *et al.* 2002).

Controversy still remains concerning the nature of the block-like structure in humans (TISHKOFF and VERRELLI 2003; WALL and PRITCHARD 2003). While it is accepted that a block-like structure is present, earlier studies may have oversimplified the complexity of LD structure in humans (WALL and PRITCHARD 2003). These authors suggest that recombinational hotspots may not be enough to explain the creation of the block structure (WALL and PRITCHARD 2003). In addition, population demographics and gene conversion may contribute to the formation or degradation of blocks (FRISSE *et al.* 2001; STUMPF and GOLDSTEIN 2003). The International HapMap Project (http://www.hapmap.org/) aimed at the creation of a haplotype map of all the blocks within different human subpopulations may need many more markers than anticipated to detect the underlying complexity of LD structure.

LD in Plants

General aspects of LD structure in humans may not be predictive of what will be found in plants. Large-scale LD studies of a wide sampling of plants and animals are needed to increase understanding of genome structure. Most studies exploring LD in plants have come from maize, an outcrossing crop species and *Arabidopsis*, an autogamous non-crop species. Comparisons with human data and comparisons between different plant species are limited since most plant LD data are from single genes or very small continuous regions of DNA.

The first study to explore the extent of LD in maize sampled 21 loci throughout chromosome one (TENAILLON *et al.* 2001). A very diverse and small sampling of exotic landraces and U.S. inbred lines was tested. The level of LD found in these loci was surprisingly low, ranging between 100-200 bp on average for the exotic landraces. LD was greater in the U.S. inbred lines, frequently extending over 1 kb. The slower decline of LD in U.S. inbred lines was later confirmed via the assay of six genes in a maize population consisting of 102 inbred lines (REMINGTON *et al.* 2001). They found that the average LD of the six genes extended about 1.5 kb. Although average LD extended about 1.5 kb, the LD decline of individual genes varied from 200 bp to over 8,000 bp (REMINGTON *et al.* 2001). The locus with the highest level of LD was the sugary1 (*su1*) locus (WHITT *et al.* 2002). The extensive LD at the *su1* locus was attributed to selection. A third study examining LD decline in 18 maize genes using 36 diverse inbred lines reported a rapid decline of LD averaging about 500 bp (CHING *et al.* 2002). These three studies established that the extent of LD in maize, like humans, varies between regions of the genome and between different populations. It is also apparent that the creation of the U.S. inbred lines increased LD. No studies in maize have reported the presence of a block-like structure of LD analogous to that described in humans.

The model organism *Arabidopsis thaliana* has been studied to determine the average rate at which LD decays. Since *Arabidopsis* is an autogamous species, which is believed to be 99% selfing, extensive LD should be present (NORDBORG *et al.* 2002). Nordborg *et al*. (2002) sequenced thirteen segments in a 250 kb region surrounding the flowering time locus FRIGIDA (FRI) in *Arabidopsis*. A total of 83 SNPs in a global sample of 20 accessions were genotyped for the FRI region along with 163 genome-wide SNPs in 76 accessions. Nordborg *et al*. (2002) found LD declined within 250 kb around the FRI region. The analysis of SNPs distributed randomly throughout the genome detected no genome-wide LD except for one region on chromosome four. The region on chromosome four had several markers that were linked to give an estimate of LD decline within 50 kb. Further study of LD around the FRI locus found LD decline is not uniform and appears to show evidence of recombination within a 400 kb region (HAGENBLAD and NORDBORG 2002). The work by Nordborg *et al*. (2002) and Hagenblad and Nordborg (2002) is significant because they were the first to study LD structure and extent in an autogamous plant species.

The high level of LD at the FRI locus was expected due to the selfing nature of *Arabidopsis*. Two other studies suggest that LD decay in FRI may not be typical of the *Arabidopsis* genome. A disease resistance gene (*rsp5*) responsible for specific recognition of a *Pseudomonas syringae* strain has LD decay within a 10 kb region (TIAN *et al.* 2002). In addition, the CLAVATA2 locus which encodes a leucine-rich repeat protein regulating the development of the shoot meristem has a more rapid decline in LD within a 40 kb region with LD decaying in as little as 6 kb (SHEPARD and PURUGGANAN 2003). Investigations of the extent of LD in *Arabidopsis* suggest a variable and complex pattern of LD present even in this highly autogamous species.

Two domesticated autogamous plant species, soybean and rice (*Oryza sativa* L.), also show extensive amounts of LD. Zhu *et al.* (2003) investigated LD decay in 16 diverse soybean genotypes that were direct introductions to North America from Asia. Linkage disequilibrium was measured over a distance of 12.5 cM and was estimated to decay at distances of 2.0-2.5 cM. Genome-wide LD was determined from 54 sequenced loci and was found to be low based upon all pairwise estimates. Like soybean, rice has undergone a domestication bottleneck and exhibits the same patterns of extensive LD. Significant LD was found spanning a range of 115 kb from 114 rice accessions among 18 sequenced fragments (GARRIS *et al.* 2003).

Several important questions remain about LD structure in plants. All the studies of LD structure in plants have investigated the extent of LD in a few genic regions which may not be representative of the genomes under investigation. The main lesson learned from the studies in *Arabidopsis* and maize is that the extent of LD is not uniform across the regions examined. In maize, recombination hot spots occurring within genes with little recombination in intervening sequence between genes has been reported (DOONER *et al.* 1985). This has led to the hypothesis in maize that regions of extensive LD may occur in gene-poor regions while very limited LD is present in gene-dense regions as has been reported in a few recent studies (RAFALSKI and MORGANTE 2004; REMINGTON *et al.* 2001; TENAILLON *et al.* 2001). Further work needs be done to validate this hypothesis in maize and to determine if it is applicable to other plants.

Association analysis for QTL Discovery and Fine Mapping

Genetic association analysis for gene or QTL discovery measures correlations between genetic variants and phenotypic differences on a population basis. The most common form of genetic association analysis is a case-control study, which is the comparison of allele frequencies at marker loci between subpopulations of individuals with contrasting phenotypes. When a significant difference in allele frequencies is found through a Fisher's exact test then the marker loci are putatively associated with the genes contributing to the phenotype (RISCH and MERIKANGAS 1996). This is in contrast to linkage analysis which depends on the correlation of genetic and phenotypic variation among individuals with known familial relationships. While association analysis relies on existing populations, linkage analysis often requires the time-consuming creation of populations derived from hybridization between individuals with extreme values of the phenotype under study i.e., resistant vs. susceptible to a disease. In addition, all QTL are potentially detectable in an association analysis unlike linkage analysis where only the QTL segregating in a bi-parental cross are detectable. Another suggested advantage of genetic association analysis is better resolution of the genome position of genes or QTL than traditional genetic linkage approaches (BUCKLER and THORNSBERRY 2002). This is because genetic association analysis takes advantage of recombination occurring over long periods of time rather than over the few generations available in existing families or those created via artificial hybridization. Another advantage can be the existence of populations for which phenotypic data have already been accumulated. Determining the likely success of genetic association analysis for the discovery of genes or QTL in a population requires a full understanding of the structure and extent of LD in a species and possible factors that have shaped the pattern of LD.

One major obstacle to the use of association analysis in plants is the possibility of population structure within case-control populations. This occurs when a subpopulation is represented more within either the case or control group. A higher frequency of a marker allele can occur due to the subpopulation being over-represented rather than as a result of the presence of a QTL, leading to false positive associations. Recently, methods for detecting population structure and correcting for it have been developed (PRITCHARD *et al.* 2000). Thornsberry *et al.* (2001) were the first to apply association analysis to a plant species by mapping SNPs occurring in the *Dwarf8* locus in maize for association with flowering time. The rapid decay of LD in maize allows for fine mapping of variations in the DNA sequence associated with a quantitative trait. In the *Dwarf8* locus, nine polymorphisms in LD with each other were significantly associated with flowering time (THORNSBERRY *et al.* 2001).

Additional studies in maize and potato have applied association analysis to find associations with candidate genes for qualitative and quantitative traits (PALAISA *et al.* 2003; TROGNITZ *et al.* 2002). Association analysis found that insertions in the promoter of the *Y1* gene are responsible for increased expression resulting in an increase of carotenoid content in the endosperm of maize (PALAISA *et al.* 2003).

Trognitz *et al.* (2002) applied association analysis to test 27 plant defense genes for association with potato late blight resistance in a diploid potato population. Six of the defense genes screened were significantly associated with late blight resistance. The early work demonstrating that association analysis is feasible in some plant species suggests the possibility of using association analysis in other crop species.

Domestication, dissemination, and crop evolution of soybeans

The history of soybean makes it an ideal model crop for studying the effects of evolutionary events on SNP variation, LD structure, and for implementing association analysis. Soybean is thought to have been domesticated in northern China during or before the Shang dynasty ca. 1700-1100 BC from the annual wild soybean *Glycine soja* (Seib. et Zucc.) (HYMOWITZ 2004). After domestication, the primitive soybean was most likely disseminated throughout the rest of China and possibly Korea by the first century AD (HYMOWITZ 2004). From the first century AD to the 15th and 16th centuries, soybean spread across Asia and landraces were developed in Japan, Thailand, Malaysia, Nepal, north India, Burma, Vietnam, Indonesia, and the Philippines (HYMOWITZ 1990). The widespread distribution of soybean can be attributed to sea and land trade routes (HYMOWITZ 1990). In Asia, soybean was mostly used in food products such as miso, soy sauce, tempeh, and tofu (HYMOWITZ 1990).

Soybean was introduced to North America almost a hundred years before it was grown widely as a crop (HYMOWITZ 1990). The first documented introduction to America was by Samuel Bowen in 1765 of soybean from China that was planted by a farmer in Georgia (HYMOWITZ 1990). In 1851, soybean was introduced to Illinois and the Corn Belt where it spread throughout the U.S. where it was mostly grown as a forage crop (HYMOWITZ 1990). It wasn't until the 1920's, 155 years after its introduction to the U.S. that soybean become popular as a grain crop (HYMOWITZ 1990). Beginning in the 1920's, soybean plant introductions were brought from Asia to be grown for seed in North America (HYMOWITZ 1990). Most crop improvement through the development of new cultivars in the U.S. has occurred over the past 50-60 years and has resulted in making soybean one of the most important crops in the United States (HYMOWITZ 1990).

There are an estimated 45,000 unique soybean accessions preserved in germplasm collections in the world today (CARTER *et al.* 2004). Of the 45,000 accessions, only 80 ancestors account for 99% of the parentage of U.S. soybean cultivars (CARTER *et al.* 2004). This would not be a narrow genetic base for U.S. cultivars but dissecting this number reveals some disturbing trends. When the breeding of new cultivars first began, many of the early breeders came from or had an extensive background in plant pathology. This led them to realize that disease resistance in soybeans would be an effective method for increasing yields. Many of the cultivars used as sources of disease resistance also contained many undesirable genes. This led early breeders to breed disease resistant lines by backcrossing resistance into high yielding and agronomically superior cultivars (CARTER *et al.* 2004). While this increased the number of genotypes that form the genetic base of soybeans it did not significantly increase genetic diversity. Based upon this consideration, 16 cultivars contribute 85% of the parentage of U.S. cultivars released between 1947 and 1988 and the remaining 64 ancestors all contribute less than 1% to the genetic base of elite cultivars (GIZLICE *et al.* 1994).

While backcrossing with limited parents has kept the genetic diversity low another possible factor may have contributed to the prominence of the aforementioned 16 cultivars to the North American soybean genetic base. This has been the frequent use of one outstanding line in many breeding programs due to free exchange between breeders from 1940-1990. The first example of this was the cultivar Lincoln which was the first line released in the northern Midwest (North) that resulted from hybridization. Lincoln was released in 1944 and yielded more than all other lines of this time period (CARTER *et al.* 2004). This prompted many of the breeding programs in the North to use Lincoln as a parent. Pedigree analysis reveals that Lincoln contributes 18% of the total parentage to North American cultivars released between 1947 and 1988 (GIZLICE *et al.* 1994). Another similar bottleneck occurred in the southern U.S. (South) with the development of the cultivars Lee from the cross of S-100 with CNS (CARTER *et al.* 2004). Lee became a popular parent in the South and the two parents S-100 and CNS contributed 17% of the total parentage to U.S. cultivars released between 1947 and 1988 (GIZLICE *et al.* 1994).

Until 1977, there was a subdivision of germplasm pools among the North and the South regions due to maturity differences (CARTER *et al.* 2004). This resulted in two divergent germplasm pools with the North mostly based upon Lincoln and the South mostly based upon S-100 and CNS (CARTER *et al.* 2004). This practice changed in 1977 with the release of A3127, which quickly caused another major bottleneck of diversity that occurred in the U.S. Like Lincoln, A3127 was a cultivar that out-yielded all other cultivars available at the time and was derived from a cross of a Northern parent (Williams) and a Southern parent (Essex) (SNELLER 1994). This caused A3127 to become a bridge between Northern and Southern germplasm and was used in many breeding programs (SNELLER 1994). Subsequently, A3127 and Williams have been the most commonly used parents in the history of the Northern soybean germplasm (SNELLER 1994). The successful development of a uniquely productive line from a North x South cross demonstrated the potential value of increasing diversity to increase yield.

Association Analysis in Soybean

The USDA Soybean Germplasm collection provides a spectrum of genotypes from *G. soja* to *G. max* cultivars developed by intense artificial selection and is ideal for the creation of case-control populations for QTL fine mapping and discovery. The main source of germplasm for increasing diversity via the discovery of new QTL is from *G. soja* and landrace accessions. The Institute of Crop Germplasm Resources in Beijing, China has the largest soybean germplasm collection in the world with 26,000 unique accessions of *G. max* and 6,200 unique *G. soja* accessions (CARTER *et al.* 2004). *G. soja* has recently been found to have an outcrossing rate of 13% (FUJITA *et al.* 1997). In contrast, *G. max* is a self-pollinating species with an outcrossing rate of only 1%. This difference in outcrossing rates could lead to considerably lower LD in wild soybean compared to that of domesticated soybean. Through the study of LD, researchers will be able to take advantage of populations for QTL discovery or fine mapping possessing adequate levels of LD.

LD extent and structure in the landraces and *G. soja* is yet unknown, although some information on LD in North American ancestral soybean (ancestors) is available. Zhu *et al*. (2003) described the extent of LD among ancestors to be extensive and greater than in maize or *Arabidopsis* (ZHU *et al.* 2003). The extensive LD could make whole genome scans for QTL discovery feasible in soybean but would not allow for fine mapping of QTL with a resolution greater than 2-3 cM. The population tested by Zhu *et al*. (2003) was limited in size and the extent of LD may not be representative of other soybean germplasms or even the population from which the ancestors were selected.

Another important subpopulation of soybeans is the elite cultivars. Since the introduction of the ancestors, approximately 60 years of artificial hybridization and selection have resulted in the elite cultivars. These cycles of hybridization and genetic recombination would be anticipated to have reduced the level of LD around non-selected loci. The elite germplasm pool has also undergone intensive selection for yield. Thus, regions in the genome responsible for increased yield may exhibit increased amounts of LD when compared to the ancestors making fine mapping of genes in these areas problematic. Fine mapping of genes or whole genome scans may be possible with association analysis in elite cultivars depending on the degree to which hybridization and subsequent recombination has reduced LD.

The successful use of association analysis for fine mapping to define causative mutations in soybeans seems unlikely due to its selfing nature and the likelihood of extended amounts of LD. Currently, over 1,017 putative QTL have been identified in soybean (www.soybase.org). All soybean QTL have been mapped via traditional mapping techniques utilizing F_2 populations or other traditional mapping populations with few QTL being confirmed. Resolution of the genomic location of a QTL mapped with traditional mapping populations generally identifies a chromosomal region about 20-30 cM in size (STUBER *et al.* 1999). With a better understanding of the extent and structure of LD in soybean, whole genome association analysis may produce better resolution of QTL locations than current QTL mapping strategies.

Chapter 2: Assessment of Genetic Bottlenecks and Genetic Variation in Soybean

Introduction

The study of the reduction of genetic variation of major crops due to domestication and founding events such as introduction and artificial selection is developing an increasing urgency. The narrow genetic base of elite crop cultivars is perceived as a threat to long-term food and feed security (TANKSLEY and MCCOUCH 1997). World crop production has been punctuated by epidemics such as the coffee rust epidemic in the 1870's, wheat rust epidemics in 1916, 1935, and 1953, the Bengal rice epidemic in 1942, and the Southern leaf blight epidemic of maize in 1970 (NATIONAL RESEARCH COUNCIL. COMMITTEE ON GENETIC VULNERABILITY OF MAJOR CROPS. 1972). Given the perception that many of the world's crops are susceptible to the possibility of disease epidemics, it is remarkable that still little is known about the relative genetic vulnerability of modern cultivars, crop landraces, and wild crop progenitors, in terms of DNA sequence variation.

Soybean [*Glycine max* (L.) Merr.] is a major crop grown on 74 million hectares world-wide (WILCOX 2004). The recent introduction of soybean rust to North America (STOKSTAD 2004) raises the concern of the effects genetic bottlenecks and intensive artificial selection have on creating a monoculture crop in soybean which is more susceptible to disease epidemics. Current evidence indicates that soybean was domesticated from the annual wild relative *G. soja* most likely in China during the Shang dynasty 3,000 to 5,000 years ago (HYMOWITZ 2004). Since domestication, several landraces grown by ancient farmers and their successors are the basis of the genetic diversity available in soybean. There are an estimated 45,000 unique soybean accessions in the world with 80 ancestors accounting for 99% of the parentage of U.S. soybean cultivars (CARTER *et al.* 2004). Only 17 cultivars account for 86% of the parentage of U.S. cultivars released between 1947 and 1988 with the remaining 63 ancestors contributing less than 1% of parentage each to modern cultivars (GIZLICE *et al.* 1994). The perceived genetic vulnerability of North American soybean is based upon the small number of Asian introductions that form the genetic base of currently grown cultivars as well as the intensive selection that has occurred in soybean breeding programs (NATIONAL RESEARCH COUNCIL. COMMITTEE ON GENETIC VULNERABILITY OF MAJOR CROPS. 1972; TANKSLEY and MCCOUCH 1997).

Single nucleotide polymorphisms (SNPs) are the most common form of variation within a genome. Understanding the patterns of variation due to SNPs can aid in assessing genetic diversity in populations and help to interpret historical events that may have affected a population. Evolutionary events such as domestication, founding events, and selection can affect the amount of SNP variation within a population. Domestication is a long process in which a wild species undergoes artificial selection over hundreds of generations exerted by humans in the form of both positive and negative selection to create a cultivated crop. Founding events in crops include the use of a few individuals to introduce a crop into a new region or to create an elite inbred line population. Domestication and founding events create genetic bottlenecks which can decrease genetic diversity, change allele frequencies, and eliminate most rare alleles in the subsequent population (HALLIBURTON 2004). The magnitude of these effects will depend on the number of individuals involved, the selection pressures, and the duration of the genetic bottleneck.

Two common measurements used for SNP variation or nucleotide diversity are the expected heterozygosity per nucleotide site (π), (TAJIMA 1983) and θ, which is the proportion of polymorphic sites in a sample corrected for sample size. θ is insensitive to the allele frequency of segregating nucleotides (WATTERSON 1975). The first estimates of nucleotide diversity in soybean was based on the study of single genes and reported that diversity ranged from $\theta = 0.00085$ (SCALLON *et al.* 1987) to $\theta = 0.015$ (ZHU *et al.* 1995). The 17-fold difference between these two single gene estimates suggests that estimates of nucleotide diversity must be based upon multiple genes in order to permit a valid comparison of soybean with other species. A more recent study in soybean sampled 116 genes in 25 diverse soybean genotypes which included 12 of the 17 founding introductions to North America (ZHU *et al.* 2003). This study found soybean diversity to be 5 to 8-fold lower than reports in *Arabidopsis* (KAWABE and MIYASHITA 1999; KAWABE *et al.* 2000; KUITTINEN and AGUADE 2000; PURUGGANAN and SUDDITH 1999) and 10-fold lower than maize (TENAILLON *et al.* 2001). A more extensive appraisal of sequence variation in 12 diverse accessions of *Arabidopsis* reported the assay of 334 genes with mean θ of 0.0071 (SCHMID *et al.* 2005) which is 7.3 times greater diversity than soybean (ZHU *et al.* 2003). The two species are >99% self-fertilizing which is thought to decrease nucleotide variation due to background selection and genetic hitchhiking with selective sweeps (SCHMID *et al.* 2005). It is apparent that other factors have contributed to low soybean nucleotide diversity. A likely reason for the much lower diversity in soybean may be the domestication bottleneck that does not exist in *Arabidopsis*.

Soybean provides a model to assess how domestication, founder population effects, and intensive artificial selection have affected genetic variability in a selfing species. This model includes the bottleneck of domestication which occurred in Asia and the recent event of crop introduction from Asia to North America, followed by intensive artificial selection. The introduction of soybean into North America provides an opportunity to understand the effects of a genetic bottleneck due to a founder event. Most bottlenecks are the result of domestication, migration, or environmental disease factors occurring in the distant past. The duration and number of individuals are mostly unknown and can only be inferred from molecular data. The introduction of soybean to North America followed by intensive artificial selection to create the currently grown elite cultivars provides a well defined founding event. The bottleneck lasted one generation followed by a rapid population expansion with artificial selection for approximately 4 to 5 generations (LUEDDERS 1977). The artificial selection primarily focused on yield improvement which is a quantitative trait with low heritability controlled by many genes.

It was my objective to assess how genetic diversity in soybean has been affected by domestication, introduction into North America, and intensive artificial selection. This will help determine how susceptible current cultivars may be to disease and how much current methods of selection increase soybean's risk of disease due to genetic uniformity. Current methods for increasing diversity and hence decreasing genetic uniformity in elite lines are to utilize the landraces as the germplasm pool to increase diversity (CARTER *et al.* 2004). I will determine if the landraces contain a significant increase in variation compared to the elite lines to be an effective method for increasing diversity or if a more diverse germplasm pool such as *G. soja* is needed to significantly increase diversity.

Materials and Methods

Plant Materials

The plant material included genotypes listed in Table 2.1. The first population consisted of 26 *G. soja* plant introductions from China, Korea, Taiwan, Russia and Japan collected from 23-50.2 degrees N, 106-140 degrees E. *G. soja* is the putative ancestor of cultivated soybean with which it generally produces completely fertile hybrids (HYMOWITZ 2004). The population of landraces consisted of 52 Asian plant introductions from China, Korea, and Japan collected from 22-50 degrees N, 104-140 degrees E. The *G. soja* and the landraces were selected to represent a range of geographic origin and various maturity groups to maximize the diversity sampled. The 17 North American ancestors are *G. max* accessions from Asia that are estimated to contribute at least 86% of the genes present in the gene pool of North American soybean cultivars (GIZLICE *et al.* 1994). The North American elite cultivars consisted of 25 North American cultivars publicly released between 1977 and 1990, selected to maximize diversity based upon coefficient of parentage by Gizlice *et al*. (1996). Pure line seeds of all genotypes were obtained from the USDA Soybean Germplasm Collection courtesy of Dr. Randall Nelson (USDA-ARS, Univ. of Illinois, Urbana, IL). DNA was extracted from bulked leaf tissue of 8-10 *G. soja* plants or 30 to 50 *G. max* plants as described by Keim *et al*. (1988).

Table 2.1. Germplasm used in this study.

Type	Strain designation	Province or State	Country	cultivar	Maturity Group
Glycine soja	PI339871A	Cheju	Korea		V
	PI366120	Akita	Japan		IV
	PI393551	Taiwan	Taiwan		X
	PI407027	Akita	Japan		V
	PI407131	Kumamoto	Japan		VI
	PI407140	Kumamoto	Japan		VII
	PI407170	Kyonggi	Korea, South		V
	PI407275	Kyonggi	Korea, South		IV
	PI407282	Cheju	Korea, South		VI
	PI407288	Jilin	China		II
	PI407301	Jiangsu	China		V
	PI447004	Jilin	China		III
	PI458536	Heilongjiang	China		0
	PI458538	Heilongjiang	China		0
	PI464935	Jiangsu	China		VI
	PI468400A	Ningxia	China		IV
	PI483464A	Ningxia	China		III
	PI483465	Shaanxi	China		V
	PI518282	Unknown	Taiwan		VI
	PI549046	Shaanxi	China		III
	PI562559	Cholla Puk	Korea, South		V
	PI562565	Cholla Puk	Korea, South		IV
	PI597459D	Shandong	China		III
	PI597461A	Shandong	China		IV
	PI326582A	Primorye	Russia		II
	PI468916	Liaoning	China		III
Landraces	PI059845	Akita	Japan	Sohgetsu	V
	PI081775	Akita	Japan		I
	PI089138	Hamgyong Puk	Korea, North	Zontanoruk-on	II
	PI097094	Hwanghae Puk	Korea, North		VII
	PI398296	Kyonggi	Korea, South		II
	PI399043	Cheju	Korea, South		III
	PI407801	Kyonggi	Korea, South		VI

Table 2.1. Cont.

Type	Strain designation	Province or State	Country	cultivar	Maturity Group
Landraces	PI407849	Cholla Puk	Korea, South		III
	PI408342	Cheju	Korea, South		VI
	PI423954	Kumamoto	Japan	Shirome	0
	PI423967	Kumamoto	Japan	Nabeshima	IX
	PI424391	Cholla Puk	Korea, South		VI
	PI567258	Jiangxi	China	He pi dou	II
	PI567293	Gansu	China	Ben di huang dou	II
	PI567298	Gansu	China	Chan yao dou	V
	PI567364	Ningxia	China	Ping luo huang da dou	II
	PI567368	Ningxia	China	Xi he huang dou	IV
	PI567395	Shaanxi	China	Lai wa dou	IV
	PI567481	Hebei	China	Bao ding huang dou	II
	PI567503	Hebei	China	Niu mao huang	IV
	PI567525	Shandong	China	Cao qing huang dou	II
	PI567700	Anhui	China	Fu yang (19)	III
	PI587552	Jiangsu	China	Nan jing da ping ding huang yi 1	VII
	PI587666	Anhui	China	Er dao zao	VI
	PI587752	Hubei	China	Xian ning dong huang dou jia	V
	PI587799	Hubei	China	Wu chang zao huang dou	VIII
	PI587906	Zhejiang	China	Huang dou	IX
	PI587946	Fujian	China	Ping nan qiu da dou	X
	PI588000	Sichuan	China	Shi yue huang	X

Table 2.1. Cont.

Type	Strain designation	Province or State	Country	cultivar	Maturity Group
Landraces	PI588047	Guangdong	China	Huang ke wu dou	IX
	PI588053A	Guangdong	China	Xiao li huang	VI
	PI594451	Sichuan	China	Liu yue bao	III
	PI594554	Jiangxi	China	Huang pi tian dou	IX
	PI594579	Hunan	China	Zhong he tian cheng dou	V
	PI594597	Hunan	China	Ning yuan ba yue huang	IX
	PI594615	Guizhou	China	Liu yue zao	IV
	PI594629	Guizhou	China	Xiao hua lian	VI
	PI594770A	Guangxi	China	Fu sui chang ping hei dou	VI
	PI594773	Guangxi	China	Fu sui qu li dou	IX
	PI594777	Yunnan	China	Liu yue huang	IV
	PI594788	Yunnan	China	Da zao dou	IX
	PI602991	Shandong	China	Niu jiao qi da hei dou	V
	PI603318	Heilongjiang	China		0
	PI603336	Heilongjiang	China		II
	PI603357	Jilin	China		I
	PI603384	Jilin	China		III
	PI603420	Nei Monggol	China		II
	PI603424A	Nei Monggol	China		0
	PI603516	Shaanxi	China		VI
	PI603596	Fujian	China		III
	PI603675	Jiangsu	China		III
	PI603756	Zhejiang	China		II
Ancestors	PI548362	Unknown	Unknown	Lincoln	III
	PI 548379	Heilongjiang	China	Mandarin	0
	PI 548445	Jiangsu	China	CNS	VII
	PI 548406	Jilin	China	Richland	II
	PI 548488	Missouri	USA	S-100	V
	PI 548477	Tennessee	USA	Ogden	VI

Table 2.1. Cont.

Type	Strain designation	Province or State	Country	cultivar	Maturity Group
Ancestors	PI 548298	Unknown	China	AK [Harrow]	III
	PI 548318	Jilin	China	Dunfield	III
	PI 548391	Liaoning	China	Mukden	II
	PI 548657	North Carolina	USA	Jackson	VII
	PI 548348	Unknown	China	Illini	III
	PI 548485	Jiangsu	China	Roanoke	VII
	PI 548311	Ontario	Canada	Capital	0
	PI 548603	Indiana	USA	Perry	IV
	PI 548382	Liaoning	China	Manitoba Brown	0
	PI 548456	Pyongyang	Korea, North	Haberlandt	VI
	FC 33243	Unknown	Unknown	Anderson	IV
Elite	Weber	Iowa	USA	Weber	I
	Burlison	Illinois	USA	Burlison	II
	Century	Indiana	USA	Century	II
	Conrad	Iowa	USA	Conrad	II
	Dassel	Minnesota	USA	Dassel	0
	Dawson	Minnesota	USA	Dawson	0
	Glenwood	Minnesota	USA	Glenwood	0
	Gordon	Georgia	USA	Gordon	VII
	Hoyt	Ohio	USA	Hoyt	II
	Hutchenson	Virginia	USA	Hutchenson	V
	Kershaw	South Carolina	USA	Kershaw	VI
	Young	North Carolina	USA	Young	VI
	Lloyd	Arkansas	USA	Lloyd	VI
	Maple Glen	Ontario (Ottawa)	Canada	Maple Glen	0
	Zane	Ohio	USA	Zane	III
	OAC Libra	Ontario (Guelph)	Canada	OAC Libra	0
	OAC Musca	Ontario (Guelph)	Canada	OAC Musca	0
	Pennyrile	Kentucky	USA	Pennyrile	IV
	Perrin	South Carolina	USA	Perrin	VIII
	Pershing	Missouri	USA	Pershing	IV
	Preston	Iowa	USA	Preston	II
	Ripley	Ohio	USA	Ripley	IV

Table 2.1. Cont.

Type	Strain designation	Province or State	Country	cultivar	Maturity Group
Elite	Sprite	Ohio	USA	Sprite	III
	Thomas	Georgia	USA	Thomas	VII
	A3127	Michigan	USA	A3127	III

PCR and Sequencing

A total of 178 sequenced genes and cDNAs were previously selected from GenBank and primers designed by Zhu *et al*. (2003). Sequence data were obtained for 111 fragments whose predicted sequence length ranged from 400 to 600 bp for the 102 genes and cDNAs listed in Table 2.2 from all or most of the 120 *G. soja* and *G. max* genotypes. PCR primers and amplification conditions were previously described by Zhu *et al*. (2003). Forward and reverse sequencing reactions were performed on an ABI 3700 or ABI 3730 using ABI Prism BigDye Terminator version 3.1 cycle sequencing (Applied Biosystems, Foster City, CA). Sequence data from each amplicon were aligned and analyzed with the standard DNA analysis software Phred/Phrap and SNP detection was carried out with PolyBayes SNP detection software (MARTH *et al.* 1999). The resulting alignments and SNP predictions were visually verified using the Consed viewer (GORDON *et al.* 1998). All SNPs were resequenced if there was any ambiguity as to which allele was present.

Sequence Analysis

Small insertions and deletions (indels) were included in all analysis as SNPs. Nucleotide diversity estimated as π (TAJIMA 1983) and θ (WATTERSON 1975) were calculated for individual gene fragments as well as across all fragments. The number

Table 2.2. GenBank accessions and gene and cDNAs description of genes and cDNAs analyzed for sequence diversity.

GenBank Accession	**Description**	**GenBank Accession**	**Description**
AB003680	A3B4 Glycinin	L20310	Nodulin (nod-20)
AB003908	Phosphoenolpyruvate carboxylase	L27265	Phosphatidylinositol 3-kinase
AB004062	A5A4B3 glycinin	L27417	GTP binding protein (STGA1)
AB007127	Acidic chitinase	L28831	Ribosomal protein S11
AB018378	Early nodulin	L29770	Phosphatidylinositol 3-kinase
AB025102	Protoporphyrinogen IX oxidase	L34842	Chloroplast phytochrome A (phyA)
AB029159	GmMYB29A1	L42814	Acetyl coA carboxylase (ACCase-A)
AB030491	Thiamin biosynthetic enzyme	M10594	Uricase I I
AB030493	Thiamin biosynthetic enzyme	M10595	Peribacteroid membrane protein
AB040040	Nonclathrin coat protein	M11317	Low MW heat shock protein
AF005030	2S albumin pre-propeptide	M13759	Alpha'-type beta conglycinin storage protein
AF007211	Peroxidase precursor (GMIPER1)	M16772	Urease
AF022462	Cytochrome P450 monooxygenase	M16884	Cytochrome oxidase subunit I
AF055369	Nitrate reductase (nr2)	M21296	Beta-tubulin (S-beta-1)
AF061564	Glyceraldehyde-3-phosphate dehydrogenase 1	M64267	Iron superoxide dismutase (FeSOD)
AF079058	Alcohol dehydrogenase Adh-1	M76980	Vegetative storage protein (vspB)
AF083880	Alternative oxidase precursor (Aox 1)	M76981	vspA
AF089850	Urate-degrading peroxidase (PP1)	M80664	Late embryogenesis abundant (LEA) protein
AF105199	Glutathione reductase (GR-5)	M94012	Maturation-associated protein (MAT9)
AF117885	Seed maturation protein PM31 (PM31)	M97285	Seed maturation protein
AF124148	Trehalase 1 GMTRE1	M98871	Chalcone synthase (chs7)
AF127110	GO8 ripening related protein	U12150	Protease inhibitor
AF128443	SNF-1-like serine/threonine protein kinase	U26457	Lipoxygenase (vlxC)
AF141602	Cystathionine-gamma-synthase precursor	U31648	Ferritin
AF162283	Acetyl-CoA carboxylase (accB-1)	U32185	Guanine nucleotide regulalory protein
AF167556	Dihydroflavonol-4-reductase DFR1	U41323	Beta-1,3-glucanase (SGN1)
AF195819	Isoflavone synthase 2 (ifs2)	U47143	Nonsymbiotic hemoglobin
AJ223037	Leginsulin	U60500	Actin (Soy57)
AJ239127	Major latex protein homolog	U63726	Gamma glutamyl hydrolase

Table 2.2. Cont.

GenBank Accession	Description	GenBank Accession	Description
AJ276407	Pre-pro-subtilisin	U66836	RecA/Rad51/DMC1-like protein
D13505	Early nodulin	U82810	Early light induced protein
D13949	Lipoxygenase -2 (lox2)	U87999	Phosphoribosylpyrophosphate amidotransferase
D16107	Basic 7s globulin	X05024	Nodulin 22
D16248	Ubiquitin	X07675	NADH dehydrogenase and rps7
D26092	Ubiquitin	X16875	Ngm-75
D31700	Cysteine proteinase inhibitor	X52863	Glycinin
D50866	Beta-amylase	X60043	Stress-induced gene (SAM22)
D64115	Cysteine proteinase inhibitor	X63198	Low MW heat shock protein
D78510	Beta-glucan-elicitor receptor	X63565	Seed maturation polypeptide
E00532	Heat-shock protein	X67304	Lipoxygenase 1
E01433	Leghemoglobine c3	X68702	Alternative oxidase
E03629	Lipoxygenase	X68707	Proteinase inhibitor D-II
E13668	DNA-binding protein	X69639	Auxin down regulated gene (ADR6)
J01297	Actin 3 (Sac 3)	X71083	Coproporphyrinogen oxidase
J02746	Proline-rich protein	X78547	Epoxide hydrolase
K00821	Lectin (Le1)	X78548	Epoxide hydrolase
L00921	Maturation protein (MAT 1)	Z11980	Cytochrome oxidase subunit 2
L01433	Calmodulin (SCaM-4)	Z32795	Cysteine endopeptidase
L01447	G-box binding factor (GBF1)	Z46951	Heat shock transcription factor 29
L10292	Ascorbate peroxidase	Z46953	Heat shock transcription factor 34
L19359	Calmodulin (ScaM-5)	Z46954	Heat shock transcription factor 33

of synonymous and nonsynonymous sites were measured using DnaSP sequence polymorphism software version 3.5 (ROZAS and ROZAS 1999). Tajima's *D* was calculated without an outgroup as described by Tajima (1989). Haplotype diversity was calculated as described by Weir (1996) as $1 - \sum P^2_{ij}$, where $\sum P_{ij}$ is the frequency of the *j*th allele for *i*th locus summed across all alleles in the locus. A neighbor joining cluster was created using SNPs from the 111 fragments in the four soybean populations.

Results

Soybean Diversity

Sequence data were obtained for 111 fragments selected from 102 random genes in the four soybean populations. The amount of aligned sequence in the 120 soybean genotypes included 11 kilobases (kb) of 5' and 3'-untranslated region (UTR) sequence, 18 kb of intron sequence, 2 kb of perigenic genomic sequence, and 22 kb of coding sequence totaling 53 kb (Table 2.3). There were a total of 438 SNPs and 58 indels identified with an average of a SNP or indel every 106.9 bp. Indels and single DNA base differences are collectively referred to as SNPs throughout all subsequent analyses. There were a total of 84 nonsynonymous SNPs (those which alter the encoded amino acid) and 59 synonymous SNPs (nucleotide changes that do not alter the encoded amino acid) out of the 438 SNPs. A specific combination of linked SNPs within a contiguous segment of DNA is defined as a haplotype. Haplotype diversity provides another measure of genetic diversity. Average haplotype diversity for the four populations was 0.40. A total of 140 haplotypes was common to all four populations. *G. soja* contained 240 unique haplotypes that were not found in any of the three other populations (Figure 2.1).

Domestication Bottleneck

The landraces retained 49% (θ) and 66% (π) of the overall diversity found in *G. soja* (Table 2.4). The smallest reduction occurred in nonsynonymous sites with the landraces retaining 77% (π) of the diversity present in *G. soja*. The ratio of $\theta_{\text{synonymous}}/\theta_{\text{nonsynonymous}}$ was 2.6 and 1.6 in *G. soja* and the landraces, respectively

Table 2.3. GenBank accessions, amount of aligned sequenced in base-pairs (bp), number of single nucleotide polymorphisms (SNPs) and indels (I/D) discovered in coding and non-coding regions within the four soybean populations. Untranslated region (UTR) sequence includes 5' and 3' UTR.

	Coding region		Non-coding regions						Total	
			UTR		Intron		Others			
GenBank Accession	bp	SNPs and (I/D)	bp	SNPs and (I/D)	bp	SNPs and (I/D)	bp	SNPs and (I/D)	bp	SNPs and (I/D)
AB003680	12	0	0	0	462	7	0	0	474	7
AB003908	99	1	154	2	283	1(1)	0	0	536	4(1)
AB004062	423	4(2)	0	0	69	0	0	0	492	4(2)
AB007127	0	0	0	0	502	7	0	0	502	7
AB018378	85	0	27	0	379	0	0	0	491	0
AB025102	115	0	78	0	264	3	0	0	457	3
AB029159	339	1	75	0	0	0	449	3(1)	863	4(1)
AB030491	0	0	0	0	364	1(1)	0	0	364	1(1)
AB030493	213	0	158	1	287	3(1)	0	0	658	4(1)
AB040040	60	1	25	1	323	3	0	0	408	5
AF005030	246	1	101	1	0	0	0	0	347	2
AF007211	211	0	93	0	0	0	0	0	304	0
AF022462	138	2	117	2(1)	0	0	0	0	255	4(1)
AF055369	234	1	0	0	0	0	0	0	234	1
AF061564	113	1	0	0	198	0	0	0	311	1
AF079058	252	1	0	0	306	0	0	0	558	1
AF083880	159	2	0	0	333	7	0	0	492	9
AF089850	324	3	0	0	0	0	0	0	324	3
AF105199	138	4	0	0	187	2	0	0	325	6
AF117885	228	0	172	6	0	0	0	0	400	6
AF124148	0	0	232	2	0	0	0	0	232	2
AF127110	0	0	0	0	201	1	0	0	201	1
AF128443	0	0	237	0	0	0	0	0	236	0
AF141602	249	1	20	0	320	1(1)	0	0	589	2(1)
AF162283	228	1	0	0	416	9	0	0	644	10
AF167556	135	3(3)	46	0	0	0	0	0	181	3(3)
AF195819	315	1	0	0	136	0	0	0	451	1
AJ223037	150	6	0	0	45	0(1)	0	0	195	6(1)
AJ239127	207	0	31	0	0	0	288	1	526	1
AJ276407	272	4	175	6	94	0	0	0	541	10
D13505	0	0	267	0	457	5	0	0	724	5
D13949	339	1	166	1	0	0	0	0	505	2
D16107	159	1	99	0	0	0	87	1	345	2
D16248	234	2	233	2	0	0	0	0	467	4
D16248	198	0	176	2(2)	0	0	0	0	374	2(2)
D26092	472	2	76	2	0	0	0	0	548	4
D26092	0	0	453	6(3)	0	0	69	2	522	8(3)
D31700	106	1	308	5	0	0	0	0	414	6
D50866	363	1	43	0	101	3	0	0	507	4
D64115	135	1	294	5	72	0	0	0	501	6

Table 2.3. Cont.

	Coding region		Non-coding regions						Total	
			UTR		Intron		Others			
GenBank Accession	bp	SNPs and (I/D)	bp	SNPs and (I/D)	bp	SNPs and (I/D)	bp	SNPs and (I/D)	bp	SNPs and (I/D)
D78510	405	2	87	2	0	0	0	0	492	4
E00532	153	0	269	1(1)	0	0	0	0	422	1(1)
E00532	60	0	363	5	0	0	0	0	428	5
E01433	141	1	0	0	293	4	0	0	434	5
E03629	453	0	0	0	3	0	0	0	456	0
E03629	465	1	0	0	74	1	0	0	539	2
E13668	11	0	299	2	0	0	0	0	310	2
J01297	276	0	0	0	0	0	0	0	276	0
J02746	0	0	0	0	0	0	475	5	475	5
K00821	174	1	30	1	0	0	174	2	378	4
L00921	351	3	145	5	0	0	0	0	496	8
L01433	0	0	433	3	0	0	0	0	433	3
L01447	243	0	123	1	626	5	0	0	992	6
L10292	135	1	56	0	244	2(1)	0	0	435	3(1)
L19359	102	0	307	1(1)	0	0	0	0	409	1(1)
L20310	330	6	54	0(3)	0	0	91	0(4)	475	6(7)
L27265	165	0	365	1(1)	401	0	0	0	931	1(1)
L27417	399	0	97	2	617	3(1)	0	0	1113	5(1)
L28831	51	0	84	1	195	2(2)	102	1	432	4(2)
L29770	81	0	0	0	509	5(1)	0	0	590	5(1)
L34842	0	0	516	4(1)	0	0	0	0	516	4(1)
L42814	63	0	0	0	388	2	0	0	451	2
M10594	42	0	0	0	378	3(1)	0	0	420	3(1)
M10595	166	5	0	0	253	4	0	0	419	9
M11317	357	7	96	3	0	0	18	1	471	11
M13759	270	0(1)	0	0	0	0	177	4	447	4(1)
M16772	90	0	0	0	344	2	0	0	434	2
M16884	841	0	0	0	0	0	0	0	841	0
M16884	0	0	434	0	0	0	0	0	434	0
M21296	292	0	0	0	329	1	0	0	621	1
M64267	126	0	181	3(1)	362	3(1)	0	0	669	6(2)
M76980	288	0	0	0	230	2	0	0	518	2
M76980	317	0	0	0	234	0	0	0	551	0
M76981	120	0	0	0	257	3	0	0	377	3
M80664	0	0	0	0	509	2	0	0	509	2
M94012	531	10	0	0	0	0	0	0	531	10
M97285	120	0	0	0	246	6	0	0	366	6
M98871	195	2	0	0	226	4(1)	0	0	421	6(1)
U12150	151	0	175	0	0	0	0	0	326	0
U26457	551	1	0	0	231	1	0	0	782	2
U31648	0	0	0	0	486	7	0	0	486	7
U32185	129	0	145	0(1)	397	4	0	0	671	4(1)
U41323	93	2	45	1	315	5(1)	0	0	453	8(1)

Table 2.3. Cont.

	Coding region		Non-coding regions						Total	
			UTR		Intron		Others			
GenBank Accession	bp	SNPs and (I/D)	bp	SNPs and (I/D)	bp	SNPs and (I/D)	bp	SNPs and (I/D)	bp	SNPs and (I/D)
U47143	198	5	16	0	239	3	0	0	453	8
U60500	328	4	0	0	143	4	0	0	471	8
U63726	171	0	48	0	216	3(1)	0	0	435	3(1)
U66836	265	0	0	0	646	3	0	0	911	3
U82810	330	0	18	1	0	0	0	0	348	1
U87999	39	2	0	0	362	11	92	0	493	13
X05024	473	2	43	0	0	0	0	0	516	2
X05024	73	0	0	0	349	2(1)	0	0	422	2(1)
X07675	331	0	348	0	0	0	0	0	679	0
X16875	0	0	189	1	0	0	114	2	303	3
X52863	288	3	0	0	335	2	0	0	623	5
X60043	351	6	121	5(4)	172	2(1)	0	0	644	13(5)
X63198	118	0	353	4	0	0	0	0	471	4
X63198	333	2(1)	0	0	0	0	85	1	418	3(1)
X63565	336	1	136	2(2)	0	0	0	0	472	3(2)
X67304	135	1	177	1	0	0	0	0	312	2
X68702	228	1	196	1	142	1	0	0	566	3
X68707	94	1	355	6	0	0	0	0	449	7
X69639	306	7	166	1	0	0	0	0	472	8
X71083	313	1	0	0	298	1(1)	0	0	611	2(1)
X71083	0	0	0	0	320	2	0	0	320	2
X78547	135	0	196	1	0	0	0	0	331	1
X78548	261	4	0	0	219	3(3)	0	0	480	7(3)
Z11980	282	2	0	0	254	5	0	0	536	7
Z32795	129	0	141	2(1)	238	1	0	0	508	3(1)
Z46951	303	2	139	2(2)	18	0	0	0	460	4(2)
Z46953	369	3	71	0	0	0	0	0	440	3
Z46954	0	0	261	6	0	0	0	0	261	6

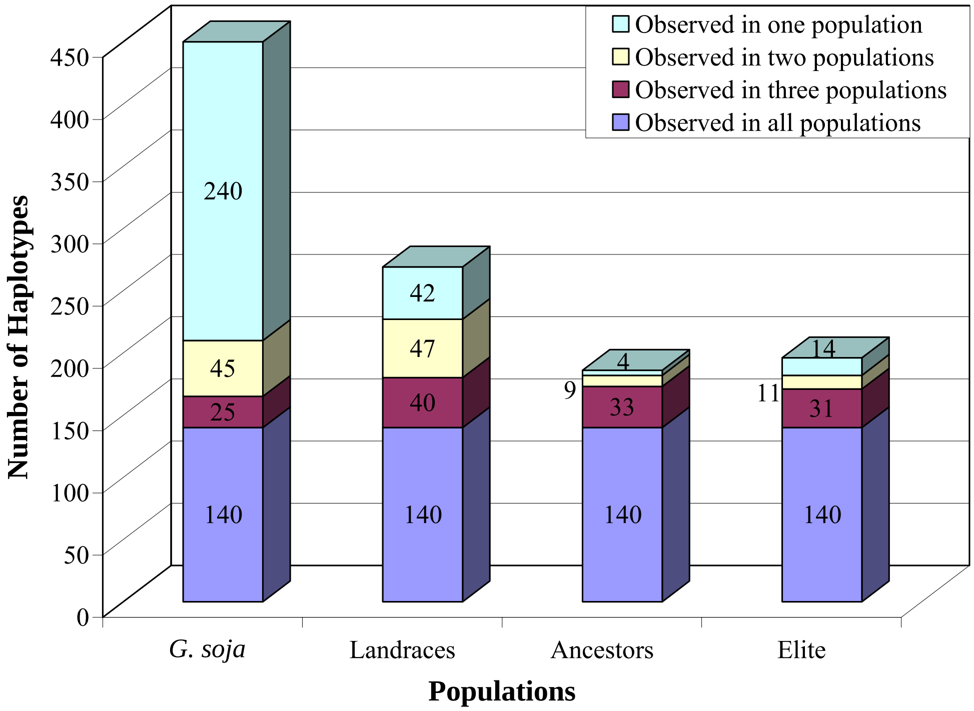

Figure 2.1. Comparison of the number of unique and shared haplotypes among the four soybean populations.

Table 2.4. Nucleotide diversity per base pair (bp) x10^3 in coding and non-coding regions within the four soybean populations. Untranslated region (UTR) sequence includes 5’ and 3’ UTR.

	Coding sequence diversity						Noncoding sequence diversity							
	Synonymous		Nonsynonymous		Total Coding		UTR		Intron		Total Noncoding		Total	
Population	π	θ	π	θ	π	θ	π	θ	π	θ	π	θ	π	θ
G. soja	4.73a*	3.15a	0.96a	1.20a	1.05a	1.63a	3.18a	3.24a	2.34a	2.65a	2.76a	3.06a	2.17a	2.35a
landraces	1.84b	1.18b	0.74ab	0.72b	0.70b	0.81b	2.02b	1.43b	1.55b	1.35b	1.77b	1.36b	1.43b	1.15b
N.A. ancestors	1.21b	1.29b	0.56b	0.58b	0.60b	0.73b	1.28b	1.07b	1.14c	1.07bc	1.36b	1.16bc	1.14c	1.00bc
N.A. Elites	1.22b	0.77b	0.60b	0.54b	0.61b	0.59b	1.10b	0.86b	0.96c	0.76c	1.22c	0.92c	1.11c	0.83c
Mean Square of Error	50.84	7.89	0.56	0.67	0.25	1.08	7.85	4.33	0.98	1.05	2.58	1.89	0.78	0.60

*Same letter within column is not significantly different based on Duncan's multiple range test (p>0.05) suggesting that the *G. soja* population has been more effective in purging deleterious alleles. Average haplotype diversity was significantly different between *G. soja* and the landraces (P<0.0001): *G. soja* = 0.51 versus landraces = 0.32. These values were consistent with nucleotide diversity and indicated that the landraces retained 63% or more than half of the haplotype diversity of *G. soja*. On an individual gene basis there was a weak positive relationship between haplotype diversity in *G. soja* and the landraces. The regression of the individual gene haplotype diversity of the landraces on the haplotype diversity of *G. soja* yielded a positive relationship (Y = 0.521x + 0.071). The relationship had a low R^2 value of 0.23 between the two populations suggesting that allele frequencies have changed as a result of domestication (Figure 2.2).

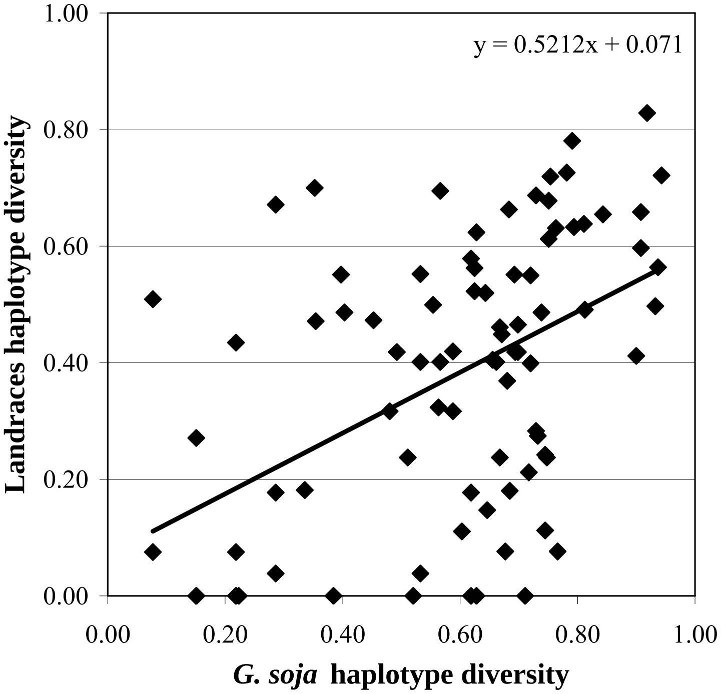

Figure 2.2. Relationship of haplotype diversity of 102 individual genes between the *G. soja* population and the landraces population.

Tajima's *D* statistic is often used to detect genetic bottlenecks since it compares π and θ to determine the presence of deviation from neutral variation (TAJIMA 1983). Very recent bottlenecks will produce positive values for Tajima's *D* because most rare alleles will be lost (SIMONSEN *et al.* 1995; TENAILLON *et al.* 2001). Conversely, a recent population expansion will produce negative Tajima's *D* values since most variation will not have had time to increase in frequency leading to an excess of rare alleles (TENAILLON *et al.* 2001). The domestication bottleneck has occurred recently and would therefore be expected to produce mostly positive Tajima's *D* values in the resulting landraces. The value of *D* was positive in 46 genes (6 significantly different from 0, $p<0.05$) while 29 genes had a negative *D* value (1 significantly different from 0, $p<0.05$). These data suggest a recent bottleneck and that the values are moving towards neutral *D* values as new mutations arise. In the case of *G. soja,* 63 genes had a negative *D* value (3 significantly different from 0, $p<0.05$) and 32 genes had a positive *D* value (1 significantly different from 0, $p<0.05$).

A comparison of *G. soja* versus the three *G. max* populations indicated that diversity has been retained in many genes but that several gene fragments have moved towards fixation. Eighteen gene fragments contained SNPs segregating in *G. soja* but were monomorphic in the three *G. max* populations. Six of the fragments that were monomorphic in *G. max* had a π value of 0.0028-0.0071 in *G. soja*, which is two to six-fold more diverse than the overall nucleotide diversity in *G. max*. In addition, the number of unique haplotypes was greater in *G. soja* than in the landraces despite the fact that there were twice as many landraces as there were *G. soja* genotypes. *G. soja* contained 246 unique haplotypes while the landraces had 65 unique haplotypes with 204 haplotypes shared between the two populations (Figure 2.3). Of the 246 unique haplotypes in *G. soja*, 198 occur at a frequency <10% suggesting many rare alleles were lost during domestication.

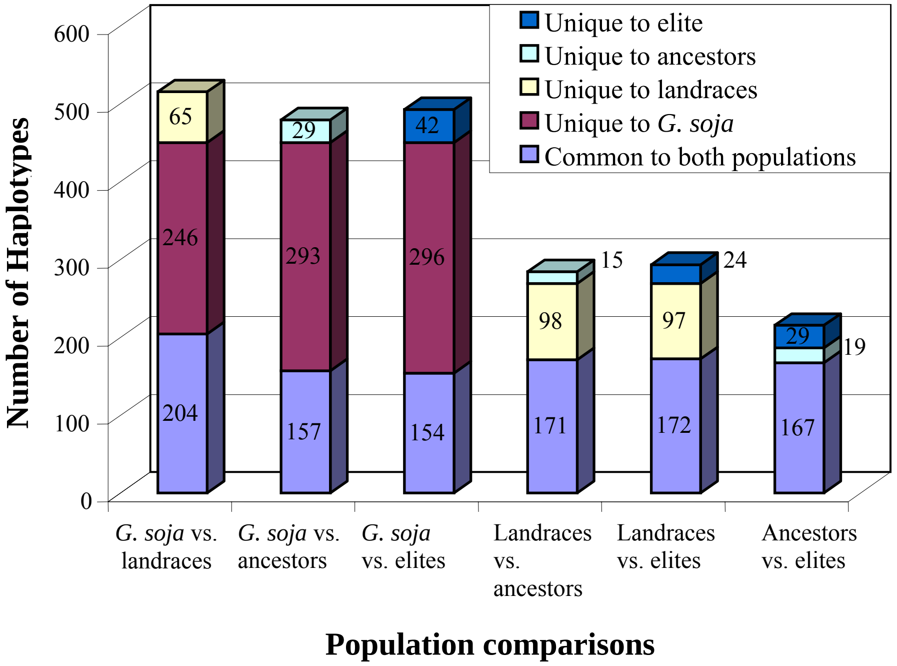

Figure 2.3. Paired comparisons of the number of unique haplotypes to each population and the number of shared haplotypes.

Demographics of Domestication and Population Structure

A neighbor joining tree of the four soybean populations was created based on the SNPs in the 102 genes (Figure 2.4). Three of the four populations grouped together in the tree. The *G. soja* population formed a distinctive clade with only one branch joining it to the landraces suggesting a single domestication event. Due to the size of the data set bootstrapping was only performed on the polymorphism data which leads to low bootstrapping values. *G. soja* as a distinctive clade was present 67% of the time after 5000 bootstraps were performed. The area of domestication can also be inferred from the geographical origin of the closest *G. soja* genotypes that

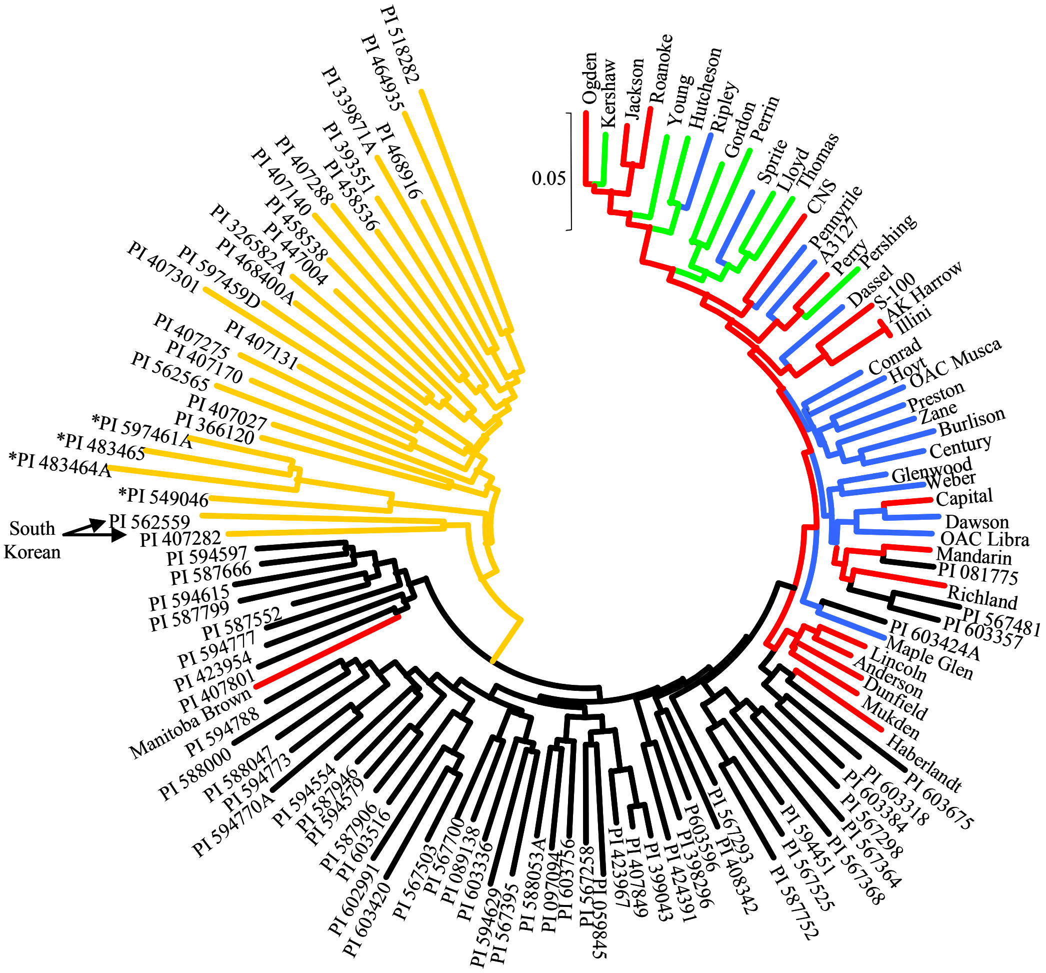

Figure 2.4. Cladogram created by the neighbor joining method of the Northern 🟦 and Southern 🟩 elite cultivars, ancestors 🟥, landraces ⬛, and soybean wild ancestor *G. soja* 🟨 populations based on 102 gene sequences. The arrows indicate two *G. soja* lines collected from South Korea and the asterisks indicate *G. soja* lines collected along the Yellow River in China.

are most closely related to *G. max*. The *G. soja* accessions closest to cultivated soybean were collected from South Korea. The next clade of *G. soja* accessions closest to cultivated soybean were accessions collected from along the Yellow River (Figure 2.4). Only four landraces were grouped with the ancestors and elites while one ancestor (Manitoba Brown) was quite distinct from the other ancestors. There was little population structure between the ancestors and the elites. Within the two populations there is a divide between cultivars from the Northern maturity groups (maturity groups 00-IV) and the Southern maturity groups (V-VIII). The exceptions are the elite cultivars Pennyrile, A3127, Sprite, and Ripley. These four cultivars are unique because they are the only elite cultivars included in the study that are derived from crosses between parents from the Northern and Southern maturity groups.

Introduction of Soybean to N. America

The ancestors was the only population containing genotypes not selected for maximum diversity since this population was defined based upon individual contribution to the North American soybean germplasm pool (GIZLICE *et al.* 1994). The estimates for π and θ in the ancestors were not significantly different from the landraces except for overall π and intron sequence π (Table 2.4). Overall, π and θ retained 80% and 87% of the nucleotide diversity found in the landraces. The ratio of $\theta_{synonymous}/\theta_{nonsynonymous}$ in the ancestors was 2.2 which is intermediate between the values for the *G. soja* and the landraces. Average haplotype diversity was 0.30 in the ancestors which is 94% of, and not significantly different ($p>0.55$) from the haplotype diversity in the landraces. The regression of the individual-gene haplotype diversities of the ancestors on the landraces ($y = 0.857x + 0.022$, $R^2 = 0.56$) indicated a strong positive relationship between the haplotype diversities of the genes in the two populations (Figure 2.5). The estimates of Tajima *D* in the ancestors indicated that 42 genes had a positive *D* value (3 significantly different from 0, p<0.05) while 21 genes had a negative *D* values (none significantly different from 0, p<0.05). A total of 39 genes were monomorphic in the ancestors while 14 of the 39 were polymorphic in the landraces.

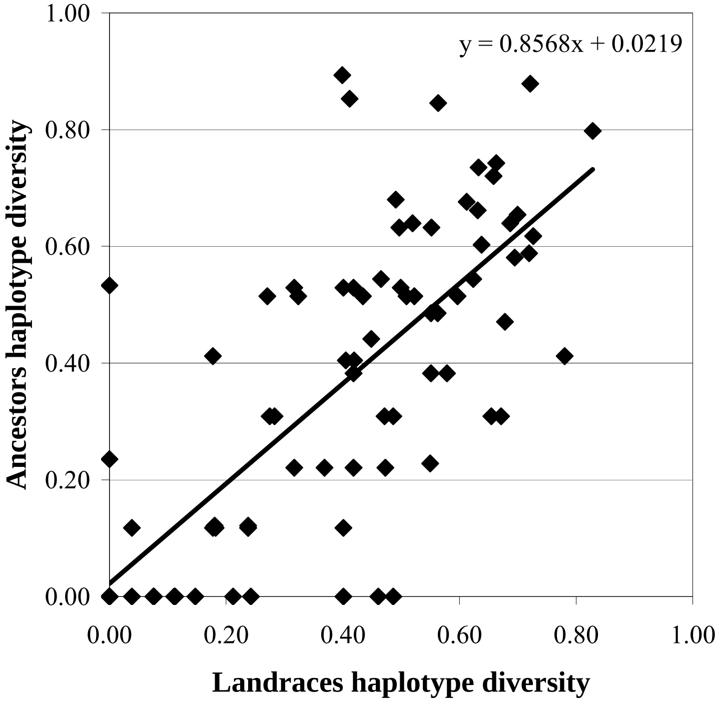

Figure 2.5. Relationship of haplotype diversity of 102 individual genes between the landrace population and the ancestor population.

It is common for low frequency SNPs to become fixed during a genetic bottleneck. However, one gene (L20310) became fixed in the ancestors despite the fact that it contained a high level of diversity (π = 0.0031) in the landraces compared to the mean of the ancestors (π = 0.0011). Three of the 14 genes that were fixed in the ancestors were polymorphic in the elite genotypes suggesting that alleles had been introduced from a source other than these 17 ancestors or that mutations had occurred since introduction. Several haplotypes were not found in the ancestors that were unique to the landraces. A total of 171 haplotypes were shared between the two populations. The landraces contained 98 unique haplotypes not found in the ancestors while the ancestors contained 15 unique haplotypes not found in the landraces (Figure 2.3).

Creation of the elite cultivars

The short duration of the introduction bottleneck followed by rapid population expansion and intense artificial selection in the expanding population resulted in little change in the nucleotide diversity of the elite cultivars, retaining 83% (θ) and 97% (π) of the total nucleotide diversity of the ancestors from which they were derived (Table 2.4). In relation to the landraces, the elite cultivars had significantly ($p<0.05$) reduced nucleotide diversity retaining 72% (θ) and 78% (π) of the total nucleotide diversity (Table 2.4). In addition, the introduction bottleneck combined with selective breeding did not significantly reduce the haplotype diversity between the elite cultivars (0.28) and the N.A. ancestors (0.30) or the Asian landraces (0.32). The haplotype diversity on an individual gene basis had a strong positive relationship between the elite cultivars and the landraces ($R^2 = 0.63$) as well between the elites and the ancestors ($R^2 = 0.78$) (Figure 2.6a and Figure 2.6b, respectively). The ratio of $\theta_{synonymous}/\theta_{nonsynonymous}$ in the elite cultivars was 1.4 which was comparable to the landraces. Four genes were monomorphic in only the elite population but contained SNPs in the other three populations. The gene AB007127 was monomorphic in the

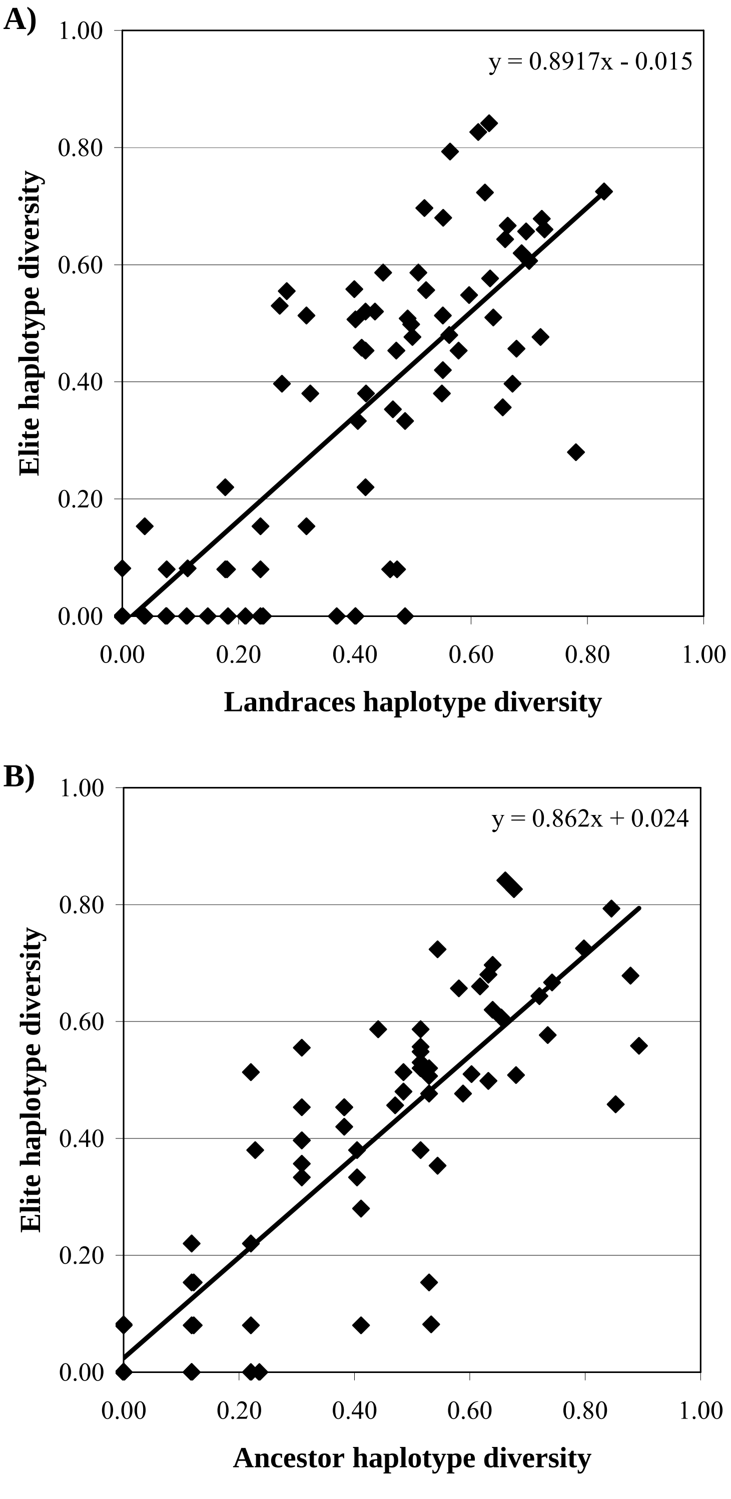

Figure 2.6. Relationship of haplotype diversity of 102 individual genes between A) landraces population and elite population, and B) ancestor population and elite population.

elite cultivars but had a π = 0.003 and θ = 0.004 in the ancestors which is three to five times greater than the average nucleotide diversity of the elite cultivars. Additionally, 36 genes were monomorphic in the elites that were also monomorphic in one or all of the other three populations.

Discussion

The goals of my study were to evaluate the amount of diversity found in soybean and its wild ancestor *G. soja* and to assess the importance of past events such as domestication which could impact the amount of variation present in soybean. Due to the wide differences of nucleotide diversity found between genes in previous studies I decided to estimate sequence variation via the sequence analysis of a large sampling of genes. I first explored the effects of domestication by comparing the diversity found in *G. soja* and the landraces. Maize is another crop that has undergone a domestication bottleneck approximately 9000 years ago (MATSUOKA *et al.* 2002). One study recently sampled nucleotide diversity in maize from multiple genes to characterize the effects of domestication. Tenaillon *et al*. (2004) sequenced 12 genes in maize elite inbred lines, maize exotic landraces, and the wild ancestor of maize (*Zea mays* ssp. *parviglumis*) and found that maize retained 80% of nucleotide diversity of its wild ancestor (TENAILLON *et al.* 2004). My data indicate that the amount of variation retained through the domestication bottleneck in soybean was 50% suggesting a more severe bottleneck in the case of soybean. This is also emphasized by the loss of many rare alleles and unique haplotypes through domestication. While six genes had significant Tajima *D* values in the landraces, 18 gene fragments were fixed during domestication.

Inbreeding is often cited as a cause for reduced variation due to background selection and genetic hitchhiking. *Arabidopsis* is a non-domesticated species with

>99% self-fertilization and its SNP variation is 1.4-fold lower than the outcrossing species maize (SCHMID *et al.* 2005; TENAILLON *et al.* 2001). *G. soja* was expected to harbour more SNP variation than *Arabidopsis* since it is also a non-domesticated species but with a greater outcrossing rate of 13% (FUJITA *et al.* 1997). However, *G. soja* contained 3-fold less diversity than *Arabidopsis* and approximately 4-fold less diversity than maize. This suggests that some factor such as an environmental stress resulting in a reduction in population size has acted upon *G. soja* to reduce its overall diversity. It has been suggested that the wild ancestors for most crops including soybean contain a reservoir of unique alleles for crop improvement (TANKSLEY and MCCOUCH 1997). While the wild ancestor of soybean is a reservoir of unique alleles and diversity, the reservoir is much smaller than expected and may pose future problems when additional diversity is needed for crop improvement and to forestall future disease epidemics.

My results indicate that *G. max* arose from a single domestication event based upon the single branch connecting *G. soja* to *G. max* in the cladogram (Figure 2.4). The area of domestication is less clear since my results suggest domestication occurred in South Korea because two *G. soja* genotypes originating from South Korea were the closest relatives to *G. max*. The eastern half of northern China rather than Korea is the region where soybean is thought to have been domesticated (HYMOWITZ 2004). However, the *G. soja* clade that is the next most closely related to *G. max* is composed of four accessions collected along the Yellow River. The Yellow River is the region that is generally accepted as the region where domestication is most likely to have occurred (HYMOWITZ 2004). This discrepancy relative to the exact geographic origin of soybean domestication warrants further research with a larger sampling of genotypes from throughout Asia.

It is widely accepted that the introduction bottleneck greatly reduced diversity in soybean (NATIONAL RESEARCH COUNCIL. COMMITTEE ON GENETIC VULNERABILITY OF MAJOR CROPS. 1972; TANKSLEY and MCCOUCH 1997). Many hybridizations between Asian introductions were made during the early years of North American soybean breeding but only a small number produced progeny that became cultivars and parents in subsequent cycles of improvement. Despite the use of only a few plant introductions combined with intensive artificial selection there has been only a minimal reduction in sequence diversity with the elite cultivars retaining 83% (θ) and 97% (π) of the diversity present in the ancestors and 72% of the diversity found in the landraces. This is surprisingly close to the amount of diversity maize elite inbred lines have retained from exotic landraces. The mean diversity retained across 21 loci in the maize inbred lines is 77% of the diversity found in maize landraces (TENAILLON *et al.* 2001). My data indicate that the introduction bottleneck minimally reduced diversity because it was of short duration and was followed by a rapid population expansion. In addition, the ancestors appear to be an adequate sampling of the landraces based upon the data indicating that the ancestors retained 87% of the landraces diversity. Another possible reason for the minimal reduction in diversity due to introduction and intensive selection is that until 1977 there was a separation between the breeding programs in the two major regions in North America which correspond to the Northern and the Southern germplasm pools (CARTER *et al.* 2004). This separation is apparent in Figure 2.4. Due to maturity and other differences, hybridizations were made only between genotypes within the two separate germplasm pools. This practical constraint may have helped to maintain the overall diversity in the elite genotypes. After 1977, some hybridizations occurred between genotypes from the Northern and Southern germplasm pools creating cultivars such as A3127, Ripley, Sprite, and Pennyrile. Interestingly, these northern maturity germplasm lines grouped with the southern germplasm (Figure 2.4) which suggests that genes for early maturity from the northern parent were selected while keeping most of the southern parental background.

Overall, the combined effects of the domestication and introduction bottlenecks combined with artificial selection during improvement has reduced sequence diversity in soybean by 35, 51, and by 56% as measured by θ, π, and haplotype diversity, respectively. In addition, the comparison of unique haplotypes between *G. soja,* the landraces, and the elite cultivars reveals that the landraces have 4 times the number of unique haplotypes than the elites while *G. soja* has 7 times the number of unique haplotypes when compared to the elite cultivars. One explanation for the discrepancy between π vs. θ and haplotype diversity vs. unique haplotypes is that *G. soja* contains many infrequent SNPs which inflate θ and creates many unique haplotypes in *G. soja* while half of the common variation has been retained in the elite cultivars through the bottlenecks.

Many see the landraces as a large source of diversity (CARTER *et al.* 2004). My study has shown that in relation to elite cultivars, the landraces contain only some additional diversity while the wild germplasm of *G. soja* is the greatest resource of diversity in the form of rare and unique alleles. Most efforts to increase diversity have come through the integration of landraces in the hopes of capturing alleles to widen the genetic base *per se* and thereby reduce genetic vulnerability. It is unlikely that such efforts will protect against any specific disease such as soybean rust. In soybean, *G. soja* is likely to be the best source of genetic resistance to new diseases and searches for resistance must be done on a case by case basis. The widely held assumption that intensive crop breeding using small numbers of introductions from the center of soybean origin has resulted in genetic vulnerability (NATIONAL RESEARCH COUNCIL. COMMITTEE ON GENETIC VULNERABILITY OF MAJOR CROPS. 1972; TANKSLEY and MCCOUCH 1997) does not appear valid. Rather, my data suggest that there is an unusually low level of genetic variability in the wild progenitor of soybean, *G. soja*, and that this low level of variability is reflected in cultivated soybean.

Chapter 3: Different Patterns of LD around Three Disease Resistance Loci in Four Soybean Populations

Introduction

Linkage disequilibrium (LD) is the non-random association of alleles and has become a subject of renewed interest since the development and availability of large scale sequencing and genotyping technology. Population geneticists have been interested in the study of LD because it is affected by many factors that occur as populations evolve. Factors such as domestication, population subdivision, founding events, and selection are likely to increase LD throughout the genome or around selected loci (RAFALSKI and MORGANTE 2004). LD is decreased in a population through recombination until equilibrium is restored between loci. The study of LD also has a practical application since it is the basis of conventional genetic mapping and QTL discovery as well as genetic association analysis for the discovery and fine mapping of quantitative trait loci (QTL) in natural populations (THORNSBERRY *et al.* 2001; WILSON *et al.* 2004). Genetic association analysis for gene or QTL discovery measures correlations between genetic variants and phenotypic differences on a population basis and is dependent on the level of LD present within the population for the detection of significant associations (FLINT-GARCIA *et al.* 2003).

LD has been found to have structure in humans which is best described using a haplotype block model. Haplotype blocks are consecutive sites in high LD flanked by blocks demonstrating historical recombination (DALY *et al.* 2001; GABRIEL *et al.* 2002). This type of structure can obscure the true meaning of a significant association if the structure of LD is unknown since a significant association of a SNP with a trait can place the gene or QTL anywhere within the haplotype block. A few questions have been answered about the extent of LD in plants. The most extensive studies have been in maize, an outcrossing species, and *Arabidopsis*, a selfing species (CHING *et al.* 2002; NORDBORG *et al.* 2002; TENAILLON *et al.* 2001). The studies exploring LD in plants have sampled mostly single genes or a single continuous region of DNA. Additionally, studies have assayed different populations thus making it difficult to compare studies to obtain a full understanding of LD and how evolutionary events affect it.

Maize is the best characterized species for the variability of LD present across the genome. An early study of maize sampled loci throughout chromosome 1 and found LD only extended 100-200 bp on average for exotic landraces while in U.S. inbred lines LD extended more than 1 kb of genomic sequence (TENAILLON *et al.* 2001). The most extensive LD in maize has been found in regions known to have undergone historical selection resulting in LD extending up to 600 kb downstream of the *Y1* gene on chromosome 6 which is responsible for endosperm color (PALAISA *et al.* 2004). The model plant *Arabidopsis thaliana* has also been studied to determine the average rate of LD decay. Since *Arabidopsis* is an autogamous species, which is believed to be 99% selfing, it was expected to have extensive LD. Single nucleotide polymorphisms (SNPs) in thirteen short segments of the FRI (flowering time) locus were analyzed in 20 diverse accessions and LD was found to extend over 250 kb corresponding to roughly 1 cM (NORDBORG *et al.* 2002). The extent of LD in other genomic regions of *Arabidopsis* has been characterized and determined to contain considerably less extensive LD than the FRI locus. The disease resistance gene *rsp5* chromosomal region has LD decaying within 10 kb (TIAN *et al.* 2002). LD decays within 6 kb in a 40 kb region containing the CLAVATA2 locus (SHEPARD and PURUGGANAN 2003). Both of these regions were sampled with different germplasm than was used to assess the FRI region so the lesser LD could be due to the different populations or to different evolutionary factors acting on the different regions.

The differences between the extent of LD in maize and *Arabidopsis* have mostly been explained through their method of reproduction. Outcrossing species such as maize are expected to harbor lesser amounts of LD due to a higher rate of effective recombination leading to decay in LD. A selfing species like *Arabidopsis* is expected to contain 50-fold more extensive LD due to the reduction in the rate of effective recombination (FLINT-GARCIA *et al.* 2003). This assertion has been challenged by a recent report from wild barley (*Hordeum vulgare* ssp. *spontaneum*), a self-fertilizing species, which has been found to harbor levels of LD comparable to maize rather than *Arabidopsis* (MORRELL *et al.* 2005).

Soybean is a major crop plant grown worldwide on 74 million hectares (WILCOX 2004) and is a species in which there is the potential to apply genetic association analysis for QTL discovery and fine mapping. Soybean was domesticated approximately 3000 to 5000 years ago from the wild species *G. soja* (Seib. et Zucc.) (HYMOWITZ 2004). While cultivated soybean is widely known as a self-fertilizing species with outcrossing rates of <1%, the wild progenitor *G. soja* has been reported to have an outcrossing rate as high as 13% (FUJITA *et al.* 1997). The greater amount of outcrossing in *G. soja* increases the effective recombination rate leading to a prediction of an 11-fold lower extent of LD in *G. soja* as compared to *G. max* (FLINT-GARCIA *et al.* 2003). The largest resource of soybean germplasm is the landraces created after domestication. U.S. plant explorers collected this germplasm beginning in the early 20th century. Selections from the landraces became the first cultivars grown by North American farmers. This was followed by breeding programs based upon hybridization and selection resulting in the release of improved cultivars beginning in 1947. Gizlice *et al.* (1994) analyzed the pedigrees of 257 publicly developed cultivars released between 1947 and 1988 and determined that over 86% of the parentage could be traced to only 17 ancestors selected from the landraces. Thus, the current North American soybean germplasm pool as defined by Gizlice *et al.* (1994) is the result of several cycles of selection and effective recombination among a relatively small number of selections from the landraces. Therefore, *G. soja*, the landraces, the ancestors and currently grown elite soybean cultivars are four distinct soybean populations that provide the opportunity to study the effects of evolutionary factors such as domestication, founding events, and selection on LD extent and structure. In addition, the four populations have potential for germplasm mining through association analysis and thus it is of great interest to determine the level of LD in each.

My objective was to examine the extent and structure of LD present in these four soybean populations to estimate the effects that domestication, selection, and the autogamous nature of soybean have had on LD. In addition, three chromosomal regions were selected to determine if different regions of the genome have differing amounts of LD. This will help to determine the optimal strategies for implementing association analysis in soybean through the selection of the optimum dispersion of markers needed and the best populations for whole genome association analysis or fine mapping of genes responsible for traits.

Materials and Methods

Plant Materials

The plant materials included genotypes from four soybean populations listed in Table 2.1. The first population consisted of 26 *G. soja* plant introductions from China, Korea, Taiwan, Russia and Japan collected from 23-50.2 degrees N, 106-140 degrees E. *G. soja* is the putative ancestor of *G. max* with which it generally produces completely fertile hybrids (HYMOWITZ 2004). The population of landraces consisted of 52 Asian plant introductions from China, Korea, and Japan collected from 22-50 degrees N, 104-140 degrees E. The *G. soja* and the landraces were selected to represent a range of geographic origin and various maturity groups to maximize the diversity sampled. The 17 ancestors were selected from *G. max* accessions from Asia and are estimated to contribute at least 86% of the genes present in the gene pool of North American soybean cultivars (GIZLICE *et al.* 1994). The population of elite cultivars consisted of 25 North American cultivars publicly released between 1977 and 1990, selected to maximize diversity based upon coefficient of parentage estimations by Gizlice *et al*. (1996). Seeds of all genotypes were obtained from the USDA Soybean Germplasm Collection courtesy of Dr. Randall Nelson (USDA-ARS, Univ. of Illinois, Urbana, IL). DNA was extracted from bulked leaf tissue of 8-10 *G. soja* plants or 30 to 50 *G. max* plants as described by Keim *et al*. (1988).

Source of Genomic Sequences

Only three regions of contiguous sequence over 300 kb in length are currently available in soybean. Two genomic regions have been deposited in GenBank under accessions AX196295, AX196296, AX196297, and AX197417 (www.ncbi.nlm.nih.gov). The program bl2seq (www.ncbi.nlm.nih.gov) was used for all comparisons of sequences from GenBank. GenBank accessions AX196295 and AX196296 completely align with a sequence length of 336 kb and were considered one sequence. AX196295 was placed on the genetic map by aligning it to GenBank accession # BH126500 which is the microsatellite marker BARC-Satt309 mapping to soybean linkage group (LG) G (SONG *et al.* 2004). BARC-Satt309 is tightly linked to the soybean disease resistance gene for soybean cyst nematode (*rhg1*) (CREGAN *et al.* 1999). The sequence region of AX196295 will be referred to as chromosomal region-G (CR-G). GenBank accessions AX196297 and AX197417 have a 50 kb overlap to form a complete sequence with a 513 kb total length. AX196297 was placed on the genetic map by aligning it to GenBank accession BH126793 which is the microsatellite marker BARC-Satt632 mapping to soybean linkage group A2 (SONG *et al.* 2004). BARC-Satt632 is tightly linked to the soybean disease resistance gene for soybean cyst nematode (*Rhg4*) (CREGAN *et al.* 1999). The sequence region of AX196297 and AX197417 will be referred to as CR-A2. The third chromosomal region studied is a BAC contig constructed by Graham *et al*. (2000) located on LG J and has an estimated physical distance length of 574 kb. The BAC ends of all contigs have been sequenced and the sequence data were provided by Randy Shoemaker

(USDA-ARS, Univ. of Iowa, Ames, IA). This BAC contig region will be referred to as CR-J.

SNP Discovery and Genotyping

Primers were designed throughout the three chromosomal regions with Array Designer 2.0 (Premier Biosoft International, Palo Alto, CA). Primers were initially screened in genomic DNA from the soybean accessions Archer, Minsoy, Noir 1, Evans, Peking, and PI 209332 whose sequence analysis are reported to discover 93% of the common SNPs (frequency >0.10) in a diverse germplasm sample (ZHU *et al.* 2003). PCR primers and amplification conditions were previously described by Zhu *et al.* (2003). Forward and reverse sequencing reactions were performed on an ABI 3700 or ABI 3730 using ABI Prism BigDye Terminator version 3.1 cycle sequencing (Applied Biosystems, Foster City, CA). Evenly distributed fragments containing one or more SNPs in the six genotypes were selected throughout the three chromosomal regions for genotyping in each of the individuals in the four populations listed in Table 2.1. The genotyping was done via resequencing.

Sequence Analyses

Sequence data from each amplicon were aligned and analyzed with the standard DNA analysis software Phred/Phrap and SNP detection was carried out with PolyBayes SNP detection software (MARTH *et al.* 1999). The resulting alignments and SNP predictions were visually verified using the Consed viewer (GORDON *et al.* 1998). SNPs were resequenced if there was any ambiguity as to which allele was present. The pairwise estimates D' and r^2 along with haplotype blocks were calculated using SNPs with a frequency >10% using the software package Haploview (BARRETT *et al.* 2005).

Results

SNP Discovery and Coverage

The amplicons ranged from 500 to 800 bp and were tested in the six genotypes, Archer, Minsoy, Noir 1, Evans, Peking, and PI 209332. A total of 309 PCR primer pairs were tested from the three chromosomal regions with 54% giving a robust sequence tagged site (STS) (Table 3.1). Overall, 558 SNPs were discovered in 73% of the STSs in the six genotypes. The remaining 27% of the STSs were monomorphic. Seventy-five polymorphic STS were selected to give maximum coverage across the three chromosomal regions with an average of one STS every 13.5 kb on CR-A2, 12.4 kb on CR-G, and 57.4 kb on CR-J. The lack of coverage on CR-J was due to a lack of complete sequence data available for this region and because many of the STSs from this region came from BAC ends that were clustered rather than spread across the 574 kb contig.

Variability of LD within the Genome and Populations

Few summary statistics are available for characterizing LD across large chromosomal regions in large datasets (GAUT and LONG 2003). The most common methods are to calculate the pairwise comparison D′ and r^2 between all physically linked polymorphic marker combinations and plot these values against distance or in a matrix form (GAUT and LONG 2003). Figure 3.1 shows the matrix of D′ values along the three chromosomal regions in the four populations. Haplotype blocks are

Table 3.1. Summary of fragments tested and single nucleotide polymorphisms (SNPs) found in the three regions on soybean linkage group A2, G, and J.

Chromosomal Region	Type of Sequence	Size of region analyzed	No. of Primers Pairs	No. of Sequence Tag Sites	No. of SNPs found in Six Genotypes	Fragments sequenced in all populations	Total SNPs genotyped in 120 individuals
CR-A2	BAC contig	513	181	106	226	38	323
CR-G	BAC contig	336	91	38	169	27	291
CR-J	BAC end Sequence	574	37	23	163	10	124
Total			309	167	558	75	738

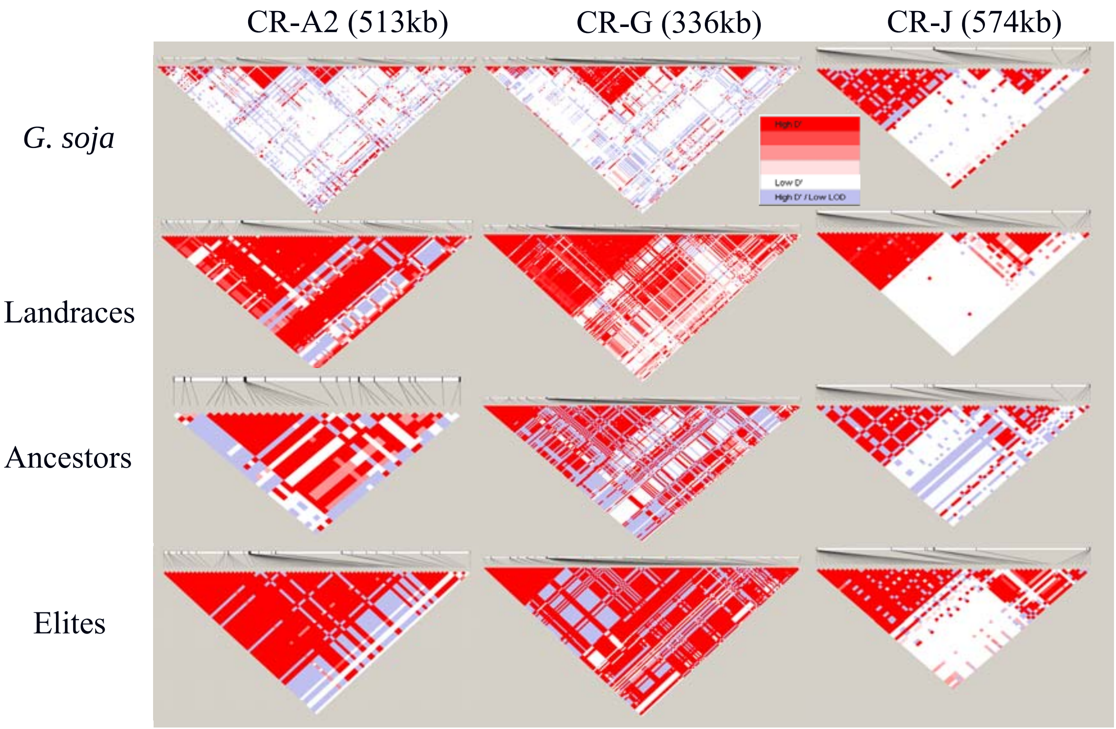

Figure 3.1. The D´ plots of the three genomic regions in the four soybean populations.

consecutive sites in high LD flanked by blocks demonstrating historical recombination (Daly *et al.* 2001; Gabriel *et al.* 2002). Haplotype blocks are present in *G. soja* covering a small fraction of the three chromosomal regions. Using a common method to define haplotype blocks (Gabriel *et al.* 2002), CR-A2 contained 13 blocks covering 46 kb of the 513 kb fragment and nine of the 13 blocks were <1 kb. CR-G contained 13 blocks with 11 of the 13 blocks <1 kb and the other two blocks only covered 8 kb of the 336 kb fragment. CR-J contained three blocks all with a size <1 kb throughout the 574 kb fragment. The lack of LD in *G. soja* across the three chromosomal regions is also apparent when r^2 is plotted against distance (Figure 3.2). A cutoff of r^2=0.1 is often used to determine when LD has

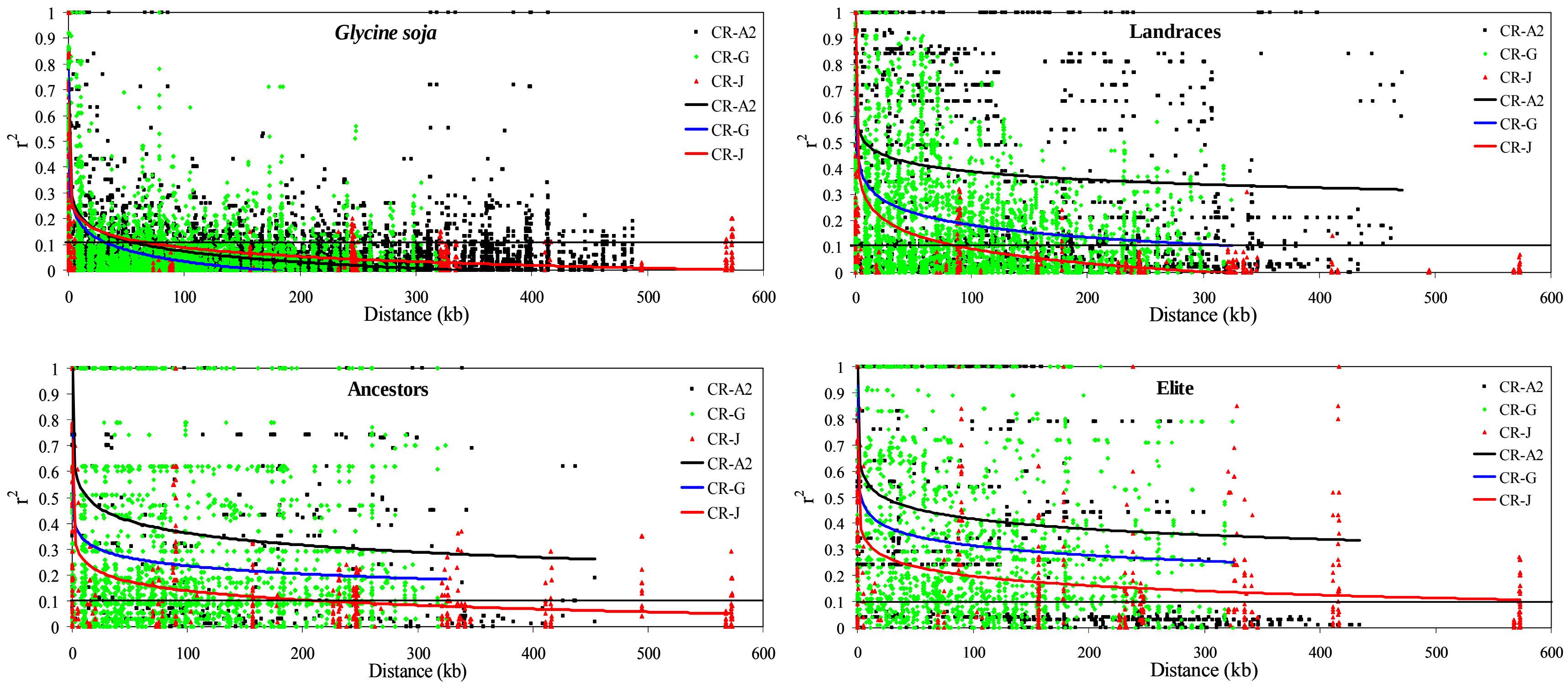

Figure 3.2. Linkage disequilibrium plots of r^2 vs. distance for the three chromosomal regions CR-A2, CR-G, and CR-J.

sufficiently decayed to a point that it is no longer useful for association analysis (KRUGLYAK 1999; PRITCHARD and PRZEWORSKI 2001). This is because very large case and control sample sizes are needed to detect a significant association when r^2 is less than 0.1 making an association analysis impractical (KRUGLYAK 1999; PRITCHARD and PRZEWORSKI 2001). To determine the average decay I determined in all cases a model with a logarithmic trend line was significant ($p<0.001$) and had the highest R^2 value from other models with a linear, power, polynomial with up to two orders, and exponential trend line. The average decay of LD reached an $r^2 = 0.1$ in *G. soja* between 36 kb and 77 kb (Figure 3.2).

The landrace population had variable amounts of LD throughout the three chromosomal regions. Only CR-J demonstrated any type of haplotype block structure with one block <1 kb present and another possible block being degraded through historical recombination (Figure 3.1). CR-G in the landraces showed evidence of gradual LD decay through half of the region but distinct haplotype blocks of high LD flanked by blocks of low LD did not appear to be present. CR-A2 had extensive LD throughout the whole region with no evidence of historical recombination.

Figure 3.2 showed the variability of LD extent between the three fragments in the landraces. In CR-J LD decayed to $r^2 = 0.1$ at a distance of approximately 90 kb and in CR-G LD decayed to $r^2 = 0.1$ around 340 kb while LD did not significantly decay throughout CR-A2 (Figure 3.2). The same pattern of LD found in the landraces was present in the ancestors and the elites for the three chromosomal regions. The only fragment that decayed to an $r^2 = 0.1$ in the ancestors and the elites was CR-J which reached this level at 212 kb in the ancestors and 654 kb in the elite cultivars (Figure 3.2).

Pairwise Measurement Comparisons

The CR-G region was selected to explore the measurements of r^2 and D′ in more detail between populations since its LD decay was intermediate to that observed in CR-A2 and CR-J. The r^2 and D′ values between pairs of loci in one population were plotted against the corresponding r^2 and D′ value for the same pair of loci in a second population (Figure 3.3). A comparison of the r^2 and D′ values plotted for all four populations on CR-G showed wide variability in the LD measurements between the four populations (Figure 3.3). The scatter plots for D′ were extremely variable between the four populations with no distinct patterns present in Figure 3.3. This was also reflected in the low correlations observed for D′ (Table 3.2) that would not be useful in comparing LD between populations. The scatter plots of r^2 for the three *G. soja* comparisons all show a clustering of points toward lower r^2 values ranging from 0 to 0.3 in *G. soja* and ranging from 0 to 1.0 in the other three populations. The correlations of the r^2 between populations were relatively high with the ancestors and the elites having the highest correlation of r = 0.78 and *G. soja* and elites having the lowest r = 0.52 (Table 3.2).

Discussion

This study has characterized LD in wild soybean and domesticated soybean throughout multiple chromosomal regions. It is apparent that effective recombination is not a good predictor of the amounts of LD in a population. Based on an

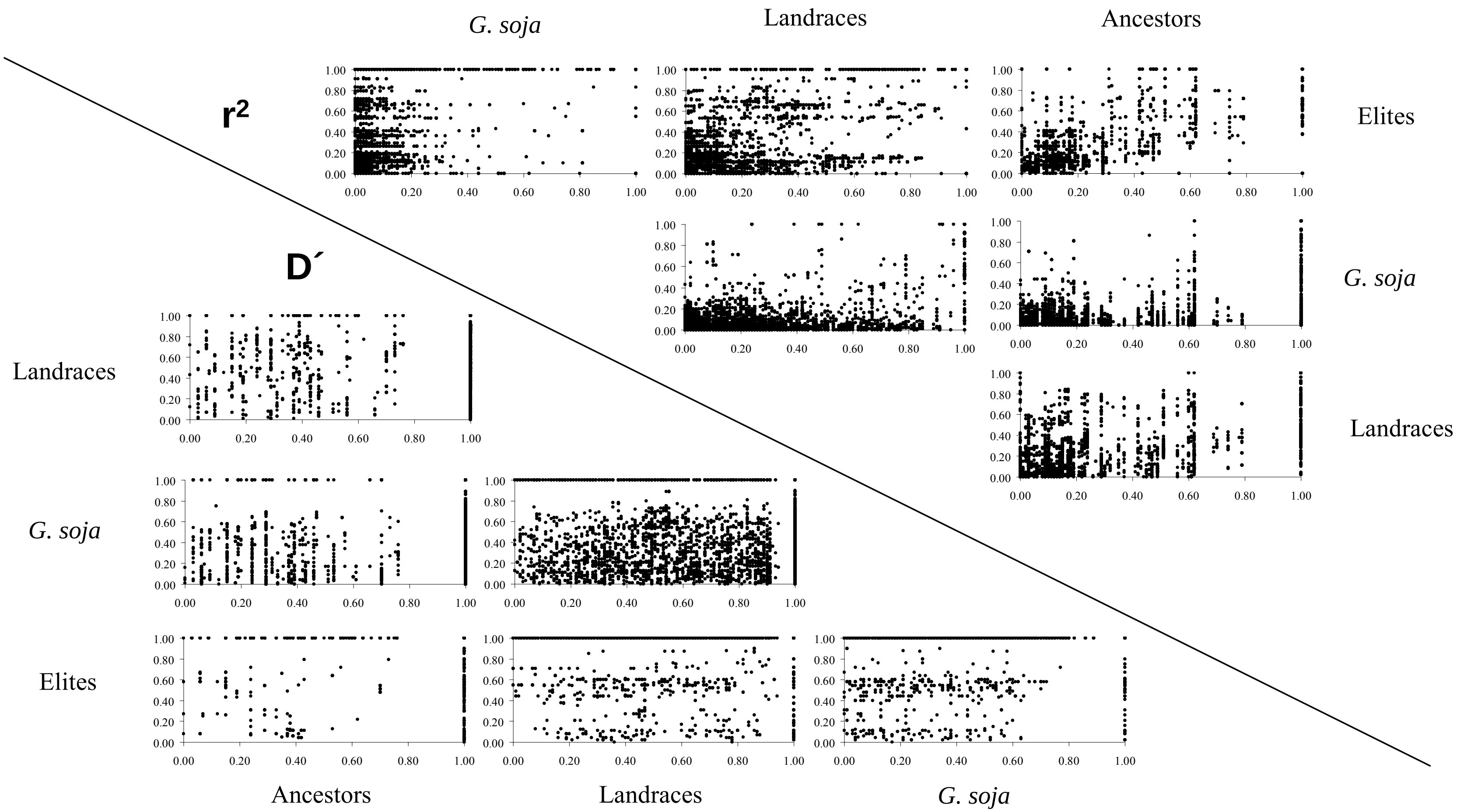

Figure 3.3. Comparisons of r^2 and D´ values between populations. The scatter plots show LD values for all marker pairs located on CR-G with allele frequencies >10% shared between populations.

Table 3.2. Correlations between the pairwise measures r^2 and D' for CR-G in the four populations. The upper right quadrant is r^2 values and the lower left quadrant is D' values.

Population	*G. soja*	Landraces	Ancestors	Elites
G. soja		0.63	0.63	0.53
Landraces	0.19		0.65	0.52
Ancestors	0.24	0.38		0.78
Elites	0.10	0.22	0.23	

outcrossing rate of 13%, *G. soja* is expected to have 4-fold greater LD than maize which is an outcrossing species, and 11-fold less LD than *Arabidopsis*, wild barley, and domesticated soybean which all have an outcrossing rate of <1%. Only CR-G in the landraces had 9-fold greater LD than *G. soja* which is close to the 11-fold predicted amounts of LD. Although effective recombination predicts this amount of LD, the effect of the domestication genetic bottleneck was predicted to also increase LD. CR-A2 is closer to the prediction of extensive LD attributable to both decreased effective recombination and domestication with LD extending throughout the entire CR-A2 fragment of 513 kb. In CR-J, LD decays similarly between *G. soja* and the landraces which is contrary to expectation. This region must contain either recombinational hotspots or was a region in which there was selection for recombinants either during or after domestication. Other species also contradict the hypothesis of effective recombination as predictor of LD. Based on effective recombination, *G. soja* is expected to harbor 4-fold more LD than maize. Instead, the exotic maize landraces have LD decay within 100-200 bp (TENAILLON *et al.* 2001) which is over 300-fold less LD than was found in wild soybean. The inbred lines of maize contain 36-fold more LD than wild soybean (TENAILLON *et al.* 2001). *Arabidopsis* has 3 to 7-fold more LD around the FRI locus than the mean of the three regions analyzed in *G. soja* (NORDBORG *et al.* 2002). In contrast, other genomic regions, such as the *rsp5* and CLAVATA2 regions characterized in different *Arabidopsis* populations are reported to have 5 to 8-fold less LD when compared to *G. soja* (SHEPARD and PURUGGANAN 2003; TIAN *et al.* 2002). Wild barley is reported to have 99% inbreeding which would help predict 11-fold more extensive LD in wild barley than *G. soja*. Instead, wild barley has LD levels that are equal to maize (MORRELL *et al.* 2005), and considerably less than the extent of LD present in *G. soja*.

While effective recombination rate is the most likely factor contributing to the LD decay, many factors such as domestication, selection, founding events, population subdivision, and population stratification can all contribute to increased LD (FLINT-GARCIA *et al.* 2003). Little is known about how these different factors affect the whole genome. The landraces resulted from domestication and it was expected that domestication would increase LD throughout the entire genome due to the genetic bottleneck of domestication. Furthermore, loci directly associated with domestication would contain larger regions of extensive LD due to selective sweeps. An example is the level of LD on CR-A2 which is more extensive than CR-G while CR-G in the landraces has more extensive LD than CR-G in *G. soja*. CR-A2 contains the *I* locus which is a region of chalcone synthase gene duplication affecting hilum color and seed-coat color (PALMER *et al.* 2004) and has been a trait that was likely under selection for during domestication. Through a sampling of landraces that contain the different alleles of the *I* locus it should be possible to determine if CR-A2 is a region that was affected by a selective sweep. The suggestion that domestication should have an impact on LD is refuted by data from CR-J where LD decay is almost identical in *G. soja* and the landraces. LD of domesticated soybean with an outcrossing rate of only 1% has had only 3000-5000 years to decay to the levels of the wild ancestor but yet this appears to have occurred in the case of CR-J. Before general conclusions about LD decay in soybean can be reached, more regions in the genome need to be characterized to see if other regions have extensive LD decay similar to CR-J or if LD across other genomic regions are more similar to that observed in CR-A2 or CR-G.

The ancestors are the founding population for the elites. Both populations have evidence of increased LD on CR-G and CR-J in relation to the landraces. A previous study characterized LD in 16 direct introductions to North America which had 12 accessions in common with the current study and found LD was extensive over a 50 kb region and dissipated at 2-3 cM (ZHU *et al.* 2003). This agrees with my study where LD did not dissipate below the r^2= 0.1 threshold in the 17 ancestors on CR-G. Another study to assay LD distal to CR-G is needed to assess the extent of LD in the ancestors.

The increased LD in the elite population versus the landraces and the ancestors could be due to multiple factors. Selection occurring on or near the three chromosomal regions could be responsible for the increased LD. Another factor may be population subdivision. Up until 1977 there was a subdivision of germplasm pools between two regions in North America due to maturity differences which correspond to the North and the South (CARTER *et al.* 2004). Most breeding programs in the Northern U.S. (North) only used northern genotypes as parents and likewise most programs in the Southern U.S. (South) only used later maturing lines from the South in their crosses. This resulted in two divergent germplasm pools with the North mostly based upon the cultivar Lincoln and the South mostly based upon the cultivars S-100 and CNS (CARTER *et al.* 2004). Similarly, an increase in LD due to the creation of elite inbred lines has also been shown in an elite maize population (TENAILLON *et al.* 2001).

The variable patterns of LD among fragments and populations indicate that more LD data are needed for association analysis to be efficient. This suggests that a population will need to be picked to create an LD map to help guide future researchers as they perform association analysis in soybean. To determine the effectiveness of creating an LD map in one population as a predictor of LD for other populations, the correlative properties between populations was explored for the common measurements of LD: r^2 and D′. D′ does not correlate well between populations while r^2 was a better predictor between populations. This is in agreement with a similar study between different human populations (EVANS and CARDON 2005). The overall best population to predict LD in other populations was the ancestors since its r^2 values had the highest correlations with the other three populations. This may not be an ideal population to use since the population size is limited and this population would not be used directly for association analysis. The next best population to use as a predictor of LD in the other populations was the landraces. The landraces performed similarly to the ancestors in predicting LD in *G. soja* and the elite population. Additionally, the landraces also are an ideal population to use for whole genome association analysis. There are an estimated 45,000 unique landraces preserved in germplasm collections around the world that have been characterized for many traits (CARTER *et al.* 2004). Several case-control populations can be created to perform whole genome scans for any trait once the level of LD across the entire genome is characterized. Besides new QTL discovery there are currently, over 1,017 putative QTL identified in soybean (www.soybase.org). All soybean QTL have been mapped via traditional mapping techniques but few of these QTL have been confirmed. Resolution of the genomic location of a QTL mapped with traditional mapping populations generally identifies a chromosomal region about 20-30 cM in size (STUBER *et al.* 1999). Given the extent of LD found in the landraces it may be possible to fine map QTL with a resolution of 1-3 cM using association analysis. Once a QTL is fine mapped in this fashion a case-control population can be created for the trait in *G. soja* to permit further fine mapping down to a resolution of 30-80 kb.

Bibliography

BARRETT, J. C., B. FRY, J. MALLER and M. J. DALY, 2005 Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics **21:** 263-265.

BROWN, G. R., G. P. GILL, R. J. KUNTZ, C. H. LANGLEY and D. B. NEALE, 2004 Nucleotide diversity and linkage disequilibrium in loblolly pine. Proc. Natl. Acad. Sci. U.S.A. **101:** 15255-15260.

BUCKLER, E. S., IV, and J. M. THORNSBERRY, 2002 Plant molecular diversity and applications to genomics. Curr. Opin. Plant Biol. **5:** 107-111.

CARGILL, M., D. ALTSHULER, J. IRELAND, P. SKLAR, K. ARDLIE *et al.*, 1999 Characterization of single-nucleotide polymorphisms in coding regions of human genes. Nat. Genet. **22:** 231-238.

CARTER, T. E., R. NELSON, C. H. SNELLER and Z. CUI, 2004 Genetic Diversity in Soybean, pp. 303-416 in *Soybeans: improvement, production, and uses*, edited by H. R. BOERMA and J. E. SPECHT. American Society of Agronomy, Crop Science Society of America, Soil Science Society of America, Madison, Wis.

CHING, A., K. S. CALDWELL, M. JUNG, M. DOLAN, O. S. SMITH *et al.*, 2002 SNP frequency, haplotype structure and linkage disequilibrium in elite maize inbred lines. BMC Genet. **3:** 19.

CREGAN, P. B., J. MUDGE, E. W. FICKUS, L. F. MAREK, D. DANESH *et al.*, 1999 Targeted isolation of simple sequence repeat markers through the use of bacterial artificial chromosomes. Theor. Appl. Genet. **98:** 919-928.

DALY, M. J., J. D. RIOUX, S. F. SCHAFFNER, T. J. HUDSON and E. S. LANDER, 2001 High-resolution haplotype structure in the human genome. Nat Genet **29:** 229-232.

DOONER, H. K., E. WECK, S. ADAMS, E. RALSTON, M. FAVREAU *et al.*, 1985 A molecular genetic analysis of insertions in the bronze locus in maize. Mol. Gen. Genet. **200 (2):** 240-246.

EVANS, D. M., and L. R. CARDON, 2005 A Comparison of Linkage Disequilibrium Patterns and Estimated Population Recombination Rates across Multiple Populations. Am. J. Hum. Genet. **76:** 681-687.

FLINT-GARCIA, S. A., J. M. THORNSBERRY and E. S. BUCKLER, IV, 2003 Structure of linkage disequilibrium in plants. Annu. Rev. Plant Biol. **54:** 357-374.

FRISSE, L., R. R. HUDSON, A. BARTOSZEWICZ, J. D. WALL, J. DONFACK *et al.*, 2001 Gene conversion and different population histories may explain the contrast between polymorphism and linkage disequilibrium levels. Am. J. Hum. Genet. **69:** 831-843.

FUJITA, R., M. OHARA, K. OKAZAKI and Y. SHIMAMOTO, 1997 The extent of natural cross-pollination in wild soybean (Glycine soja). J. Hered. **88:** 124-128.

GABRIEL, S. B., S. F. SCHAFFNER, H. NGUYEN, J. M. MOORE, J. ROY *et al.*, 2002 The structure of haplotype blocks in the human genome. Science **296:** 2225-2229.

GARRIS, A. J., S. R. MCCOUCH and S. KRESOVICH, 2003 Population structure and its effect on haplotype diversity and linkage disequilibrium surrounding the xa5 locus of rice (Oryza sativa L.). Genetics **165:** 759-769.

GAUT, B. S., and A. D. LONG, 2003 The lowdown on linkage disequilibrium. Plant Cell **15:** 1502-1506.

GIZLICE, Z., T. E. CARTER, JR. and J. W. BURTON, 1994 Genetic base for North American public soybean cultivars released between 1947 and 1988. Crop Sci. **34:** 1143-1151.

GIZLICE, Z., T. E. CARTER, JR., T. M. GERIG and J. W. BURTON, 1996 Genetic diversity patterns in North American public soybean cultivars based on coefficient of parentage. Crop Sci. **36:** 753-765.

GORDON, D., C. ABAJIAN and P. GREEN, 1998 Consed: a graphical tool for sequence finishing. Genome Res. **8:** 195-202.

GRAHAM, M. A., L. F. MAREK, D. LOHNES, P. CREGAN and R. C. SHOEMAKER, 2000 Expression and genome organization of resistance gene analogs in soybean. Genome **43:** 86-93.

HAGENBLAD, J., and M. NORDBORG, 2002 Sequence variation and haplotype structure surrounding the flowering time locus FRI in Arabidopsis thaliana. Genetics **161:** 289-298.

HALLIBURTON, R., 2004 *Introduction to population genetics*. Pearson/Prentice Hall, Upper Saddle River, NJ.

HALUSHKA, M. K., J. B. FAN, K. BENTLEY, L. HSIE, N. SHEN *et al.*, 1999 Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. Nat. Genet. **22:** 239-247.

HELMS, C., L. CAO, J. G. KRUEGER, E. M. WIJSMAN, F. CHAMIAN *et al.*, 2003 A putative RUNX1 binding site variant between SLC9A3R1 and NAT9 is associated with susceptibility to psoriasis. Nat. Genet. **35:** 349-356.

HYMOWITZ, T., 1990 Soybeans: The success story, pp. 159-163 in *Advances in new crops: proceedings of the First National Symposium NEW CROPS, Research, Development, Economics, Indianapolis, Indiana, October 23-26, 1988*, edited by J. JANICK and J. E. SIMON. Timber Press, Portland, Or.

HYMOWITZ, T., 2004 Speciation and Cytogenetics, pp. 97-136 in *Soybeans: improvement, production, and uses*, edited by H. R. BOERMA and J. E. SPECHT. American Society of Agronomy, Crop Science Society of America, Soil Science Society of America, Madison, Wis.

JEFFREYS, A. J., L. KAUPPI and R. NEUMANN, 2001 Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. Nat. Genet. **29:** 217-222.

KAWABE, A., and N. T. MIYASHITA, 1999 DNA variation in the basic chitinase locus (ChiB) region of the wild plant Arabidopsis thaliana. Genetics **153:** 1445-1453.

KAWABE, A., K. YAMANE and N. T. MIYASHITA, 2000 DNA polymorphism at the cytosolic phosphoglucose isomerase (PgiC) locus of the wild plant Arabidopsis thaliana. Genetics **156:** 1339-1347.

KEIM, P., T. C. OLSON and R. C. SHOEMAKER, 1988 A rapid protocol for isolating soybean DNA. Soybean Genet. Newsl. **15:** 150-152.

Kruglyak, L., 1999 Prospects for whole-genome linkage disequilibrium mapping of common disease genes. Nat. Genet. **22:** 139-144.

Kuittinen, H., and M. Aguade, 2000 Nucleotide variation at the CHALCONE ISOMERASE locus in Arabidopsis thaliana. Genetics **155:** 863-872.

Lee, S. H., D. R. Walker, P. B. Cregan and H. R. Boerma, 2004 Comparison of four flow cytometric SNP detection assays and their use in plant improvement. Theor. Appl. Genet. **110:** 167-174.

Luedders, V. D., 1977 Genetic improvement in yield of soybeans. Crop Sci. **17:** 971-972.

Marth, G. T., I. Korf, M. D. Yandell, R. T. Yeh, Z. Gu *et al.*, 1999 A general approach to single-nucleotide polymorphism discovery. Nat. Genet. **23:** 452-456.

Matsuoka, Y., Y. Vigouroux, M. M. Goodman, G. J. Sanchez, E. Buckler *et al.*, 2002 A single domestication for maize shown by multilocus microsatellite genotyping. Proc. Natl. Acad. Sci. U.S.A. **99:** 6080-6084.

Moriyama, E. N., and J. R. Powell, 1996 Intraspecific nuclear DNA variation in Drosophila. Mol. Biol. Evol. **13:** 261-277.

Morrell, P. L., D. M. Toleno, K. E. Lundy and M. T. Clegg, 2005 Low levels of linkage disequilibrium in wild barley (Hordeum vulgare ssp. spontaneum) despite high rates of self-fertilization. Proc. Natl. Acad. Sci. U.S.A.

National Research Council. Committee on Genetic Vulnerability of Major Crops., 1972 *Genetic vulnerability of major crops*. National Academy of Sciences, Washington.

Nordborg, M., J. O. Borevitz, J. Bergelson, C. C. Berry, J. Chory *et al.*, 2002 The extent of linkage disequilibrium in Arabidopsis thaliana. Nat. Genet. **30:** 190-193.

Palaisa, K., M. Morgante, S. Tingey and A. Rafalski, 2004 Long-range patterns of diversity and linkage disequilibrium surrounding the maize Y1 gene are indicative of an asymmetric selective sweep. Proc. Natl. Acad. Sci. U.S.A. **101:** 9885-9890.

PALAISA, K. A., M. MORGANTE, M. WILLIAMS and A. RAFALSKI, 2003 Contrasting effects of selection on sequence diversity and linkage disequilibrium at two phytoene synthase loci. Plant Cell **15:** 1795-1806.

PALMER, R. G., T. W. PFEIFFER, G. R. BUSS and T. C. KILEN, 2004 Qualitative Genetics, pp. 137-233 in *Soybeans: improvement, production, and uses*, edited by H. R. BOERMA and J. E. SPECHT. American Society of Agronomy, Crop Science Society of America, Soil Science Society of America, Madison, Wis.

PATIL, N., A. J. BERNO, D. A. HINDS, W. A. BARRETT, J. M. DOSHI *et al.*, 2001 Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. Science **294:** 1719-1723.

PRITCHARD, J. K., and M. PRZEWORSKI, 2001 Linkage disequilibrium in humans: models and data. Am. J. Hum. Genet. **69:** 1-14.

PRITCHARD, J. K., M. STEPHENS, N. A. ROSENBERG and P. DONNELLY, 2000 Association mapping in structured populations. Am. J. Hum. Genet. **67:** 170-181.

PURUGGANAN, M. D., and J. I. SUDDITH, 1999 Molecular population genetics of floral homeotic loci. Departures from the equilibrium-neutral model at the APETALA3 and PISTILLATA genes of Arabidopsis thaliana. Genetics **151:** 839-848.

RAFALSKI, A., and M. MORGANTE, 2004 Corn and humans: recombination and linkage disequilibrium in two genomes of similar size. Trends Genet. **20:** 103-111.

REICH, D. E., M. CARGILL, S. BOLK, J. IRELAND, P. C. SABETI *et al.*, 2001 Linkage disequilibrium in the human genome. Nature **411:** 199-204.

REMINGTON, D. L., J. M. THORNSBERRY, Y. MATSUOKA, L. M. WILSON, S. R. WHITT *et al.*, 2001 Structure of linkage disequilibrium and phenotypic associations in the maize genome. Proc. Natl. Acad. Sci. U.S.A. **98:** 11479-11484.

RISCH, N., and K. MERIKANGAS, 1996 The future of genetic studies of complex human diseases. Science **273:** 1516-1517.

ROZAS, J., and R. ROZAS, 1999 DnaSP version 3: an integrated program for molecular population genetics and molecular evolution analysis. Bioinformatics **15:** 174-175.

SCALLON, B. J., C. D. DICKINSON and N. C. NIELSEN, 1987 Characterization of a null-allele for the *Gy_4* glycinin gene from soybean. Mol. Gen. Genet. **208:** 107-113.

SCHMID, K. J., S. RAMOS-ONSINS, H. RINGYS-BECKSTEIN, B. WEISSHAAR and T. MITCHELL-OLDS, 2005 A multilocus sequence survey in Arabidopsis thaliana reveals a genome-wide departure from the standard neutral model of DNA sequence polymorphism. Genetics.

SCHMID, K. J., T. R. SORENSEN, R. STRACKE, O. TORJEK, T. ALTMANN *et al.*, 2003 Large-scale identification and analysis of genome-wide single-nucleotide polymorphisms for mapping in Arabidopsis thaliana. Genome Res. **13:** 1250-1257.

SHEPARD, K. A., and M. D. PURUGGANAN, 2003 Molecular population genetics of the Arabidopsis CLAVATA2 region. The genomic scale of variation and selection in a selfing species. Genetics **163:** 1083-1095.

SHIFMAN, S., J. KUYPERS, M. KOKORIS, B. YAKIR and A. DARVASI, 2003 Linkage disequilibrium patterns of the human genome across populations. Hum. Mol. Genet. **12:** 771-776.

SIMONSEN, K. L., G. A. CHURCHILL and C. F. AQUADRO, 1995 Properties of statistical tests of neutrality for DNA polymorphism data. Genetics **141:** 413-429.

SNELLER, C. H., 1994 Pedigree analysis of elite soybean lines. Crop Sci. **34:** 1515-1522.

SONG, Q. J., L. F. MAREK, R. C. SHOEMAKER, K. G. LARK, V. C. CONCIBIDO *et al.*, 2004 A new integrated genetic linkage map of the soybean. Theor. Appl. Genet. **109:** 122-128.

STOKSTAD, E., 2004 Agriculture. Plant pathologists gear up for battle with dread fungus. Science **306:** 1672-1673.

STUBER, C. W., M. POLACCO and M. L. SENIOR, 1999 Synergy of empirical breeding, marker-assisted selection, and genomics to increase crop yield potential. Crop Sci. **39:** 1571-1583.

STUMPF, M. P., and D. B. GOLDSTEIN, 2003 Demography, recombination hotspot intensity, and the block structure of linkage disequilibrium. Curr. Biol. **13:** 1-8.

TAJIMA, F., 1983 Evolutionary relationship of DNA sequences in finite populations. Genetics **105:** 437-460.

TAJIMA, F., 1989 Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics **123:** 585-595.

TANKSLEY, S. D., and S. R. MCCOUCH, 1997 Seed banks and molecular maps: unlocking genetic potential from the wild. Science **277:** 1063-1066.

TENAILLON, M. I., M. C. SAWKINS, A. D. LONG, R. L. GAUT, J. F. DOEBLEY *et al.*, 2001 Patterns of DNA sequence polymorphism along chromosome 1 of maize (Zea mays ssp. mays L.). Proc. Natl. Acad. Sci. U.S.A. **98:** 9161-9166.

TENAILLON, M. I., J. U'REN, O. TENAILLON and B. S. GAUT, 2004 Selection versus demography: a multilocus investigation of the domestication process in maize. Mol. Biol. Evol. **21:** 1214-1225.

THORNSBERRY, J. M., M. M. GOODMAN, J. DOEBLEY, S. KRESOVICH, D. NIELSEN *et al.*, 2001 Dwarf8 polymorphisms associate with variation in flowering time. Nat. Genet. **28:** 286-289.

TIAN, D., H. ARAKI, E. STAHL, J. BERGELSON and M. KREITMAN, 2002 Signature of balancing selection in Arabidopsis. Proc. Natl. Acad. Sci. U.S.A. **99:** 11525-11530.

TISHKOFF, S. A., R. VARKONYI, N. CAHINHINAN, S. ABBES, G. ARGYROPOULOS *et al.*, 2001 Haplotype diversity and linkage disequilibrium at human G6PD: recent origin of alleles that confer malarial resistance. Science **293:** 455-462.

TISHKOFF, S. A., and B. C. VERRELLI, 2003 Role of evolutionary history on haplotype block structure in the human genome: implications for disease mapping. Curr. Opin. Genet. Dev. **13:** 569-575.

TROGNITZ, F., P. MANOSALVA, R. GYSIN, D. NINIO-LIU, R. SIMON *et al.*, 2002 Plant defense genes associated with quantitative resistance to potato late blight in Solanum phureja x dihaploid S. tuberosum hybrids. Mol. Plant Microbe Interact **15:** 587-597.

WALL, J. D., and J. K. PRITCHARD, 2003 Haplotype blocks and linkage disequilibrium in the human genome. Nat Rev Genet **4:** 587-597.

WANG, R. L., A. STEC, J. HEY, L. LUKENS and J. DOEBLEY, 1999 The limits of selection during maize domestication. Nature **398:** 236-239.

WANG, W., K. THORNTON, J. J. EMERSON and M. LONG, 2004 Nucleotide variation and recombination along the fourth chromosome in Drosophila simulans. Genetics **166:** 1783-1794.

WATTERSON, G. A., 1975 On the number of segregating sites in genetical models without recombination. Theor. Popul. Biol. **7:** 256-276.

WEIR, B. S., 1996 *Genetic data analysis II: methods for discrete population genetic data*. Sinauer Associates, Sunderland, Mass.

WHITT, S. R., L. M. WILSON, M. I. TENAILLON, B. S. GAUT and E. S. BUCKLER, IV, 2002 Genetic diversity and selection in the maize starch pathway. Proc. Natl. Acad. Sci. U.S.A. **99:** 12959-12962.

WILCOX, J. R., 2004 World distribution and trade of soybean, pp. 1-14 in *Soybeans: improvement, production, and uses*, edited by H. R. BOERMA and J. E. SPECHT. American Society of Agronomy, Crop Science Society of America, Soil Science Society of America, Madison, Wis.

WILSON, L. M., S. R. WHITT, A. M. IBANEZ, T. R. ROCHEFORD, M. M. GOODMAN *et al.*, 2004 Dissection of maize kernel composition and starch production by candidate gene association. Plant Cell **16:** 2719-2733.

ZHU, T., L. SHI, J. J. DOYLE and P. KEIM, 1995 A single nuclear locus phylogeny of soybean based on DNA sequence. Theor. Appl. Genet. **90:** 991-999.

ZHU, Y. L., Q. J. SONG, D. L. HYTEN, C. P. VAN TASSELL, L. K. MATUKUMALLI *et al.*, 2003 Single-nucleotide polymorphisms in soybean. Genetics **163:** 1123-1134.