ABSTRACT

DELHEIMER, JACOB CHARLES. Agronomic and Molecular Analysis of Populations Developed from a Single Cross of Adapted x Wild Soybean. (Under the direction of Dr. Thomas E. Carter, Jr.).

An underutilized source of allelic diversity in cultivated soybean ([*Glycine max* (L.) Merr.] is the progenitor species wild soybean, *G. soja* (Sieb. And Zucc.). The underutilization of wild soybean in soybean breeding programs results from the poor agronomic performance of progeny derived from crosses of the wild to the domesticate. Typically, soybean breeders have implemented backcross breeding as a way to overcome the inherent shortcomings of the progeny derived from these crosses. However, backcrossing has typically been viewed as the most conservative approach to cultivar development, and typically little progress is made in improving quantitative traits. The focus of this research has been on the novel application of plant breeding methods to improve soybean through crosses to the wild soybean. Two replicate populations were derived from nearly 1.3 million F_3 progeny of a single cross of wild X domesticated soybean. Selection of F_3 plants was centered around an erect main stem phenotype, a rare, yet easily identifiable trait found amongst progeny. Nearly 225, F_4-derived progeny were evaluated in replicated yield trials from 2008-2010. 50 breeding lines yielded within 75% of the conventional *G. max* checks. Transgressive segregation for 100 seed weight and seed protein content was observed. Analysis of breeding lines using 558 single nucleotide polymorphism (SNP) markers showed that lines maintained from 20-48% of the *G. soja* alleles.

Patterns of diversity found within these breeding were analyzed by calculating genetic distance using simple matching coefficients (SMC) calculated from the 558 SNP

markers. The MDS plot revealed a distinct relationship between the breeding lines and the *G. max* parent. A further analysis of SMC coefficients revealed an average relationship between breeding lines of 0.63, which is greater than the expected relationship of 0.5 based on pedigree probabilities. Single marker analysis of variance and multiple linear regression revealed three genomic regions were associated with increased seed yield derived from the *G. soja* parent, as well as four genomic regions that were associated with greater 100 seed weight. Analysis of the breeding lines also revealed that a judiciously selected group of 16 lines was able to capture all of the SNP diversity found in the breeding material.

A separate study was conducted using a large population of $F_{2:3}$ family derived from a cross of domesticated x wild soybean to study the inheritance of erect plant growth in breeding populations. No F_2 plant or $F_{2:3}$ completely exhibited erect plant growth. However, 118 $F_{2:3}$ families produced at least one F_3 plant that exhibited erect plant growth. A follow study on the segregation patterns of $F_{3:4}$ families showed that one to three genes were still segregating for after selection. Based on these results, it appears that the segregation of plant type can be explained by the segregation of seven, ten, or twelve genes, and the selection within F_3 populations is a viable option for developing breeding material.

The last study focused on developing near isogenic lines (NILs) to test for any negative associations between morphological traits of *G. soja* and agronomic performance. No negative associations were detected in yield trials conducted in three locations. These results show that particular traits associated with *G.soja* such as purple flower color, may not be associated with lower seed yield or negative linkages can be broken.

Taken together, the results of these studies show that the use of large population size combined with visual selection of rare upright plants is an efficient strategy for introgressing allelic diversity from *G. soja* into soybean breeding programs. This method can offer breeders access to genetic diversity at a level previously never thought possible.

Agronomic and Molecular Analysis of Populations Developed from a Single Cross of Adapted x Wild Soybean

by
Jacob Charles Delheimer

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Crop Science

Raleigh, North Carolina

2012

APPROVED BY:

Dr. Thomas E. Carter, Jr. Committee Chair	Dr. Joseph W. Burton
Dr. Thomas G. Isleib	Dr. Jason Osborne
Dr. David Hyten	

DEDICATION

To Mike, Mary, and Mike Jr.

BIOGRAPHY

Jacob Charles Delheimer was born to Mike and Mary Delheimer in Streator, IL on October 24th, 1983. He attended St. Mary's Grade School and graduated from Streator Twp. High School in 2002. He attended the University of Illinois at Urbana-Champaign, where he received his B.S. degree in Integrative Biology in 2005. While home for the summer of 2004, he worked as a farm hand for Johnson Farms outside of Leonore, IL. It was during this time that he first developed in interest in agriculture.

After returning to school in the fall, he began working part time with the corn breeding group at the U of I. This was his first exposure to plant breeding and genetics. After graduating in December 2005, he accepted a graduate research assistantship to work for Dr. Brian Diers at the University of Illinois. The author's research was focused on comparing different sources of soybean cyst nematode resistance for their associated effects on nematode reproduction and agronomic performance. He received his M.S. in Crop Sciences from the University of Illinois in July of 2008. That August he moved to Raleigh, NC and began working for Dr. Tommy Carter as graduate research assistant in the USDA soybean breeding and genetics lab.

ACKNOWLEDGMENTS

I would like to begin by thanking Drs. Tommy Carter and Brian Diers, my current and past adviser, for their support and direction over the last few years. Their guidance and passion for soybean breeding have helped me develop as a scientist and will leave a deep impression on me as I move forward with my career. It has truly been a great experience working for both of them.

I would also like to thank my committee members: Dr. Joseph Burton, Dr. Tom Isleib, Dr. Jason Osborne, and Dr. David Hyten for all of their guidance and insight as I worked on this dissertation. I would also like to thank Dr. Lilian Miranda and Dr. Perry Cregan for their aid in the molecular portions of my project from DNA extractions to SNP analysis.

In addition, I would like to thank past and present members the USDA soybean staff: Dr. Margarita Villagarcia, Robert Masson, Earl Huie, William Novitzky, John Graber, Bobby McMillen, Dr. Phil Rzewnicki, Connie Bryant, Dr. Andrew Scaboo, Cory Callahan, Stella Kreilkamp, Sarah Albright, Andre Haden, Garret Lloyd, Johnnie Bailey, Savannah King, Katherine Novitzky, Mike Novitzky, Thomas Kevin, Amanda Noble, Jimmy Kanowski, Hailey Huie, Mackenzie Gibbons, Leah Ruff, and Rochelle Strednak. Without your help, this project would not even be close to finished.

I would also like to thank my family and friends, especially those that I have made since I have arrived in Raleigh: Charlie and Katie Zila, Jill Recker, Emma Flemmig, Dave Eickholt, Krista Sondergaard, Amy Heffelfinger, Stine Petersen, Peter Maloney, Jose Santa

Cruz, Margaret Worthington, Jessica Palermo, and Zebulon Rouse. Without your support and encouragement I would not have succeeded.

TABLE OF CONTENTS

LIST OF TABLES ix

LIST OF FIGURES xii

CHAPTER 1 A NOVEL MEGA-POPULATION APPROACH FOR BREEDING WITH WILD SOYBEAN 1

ABTRACT 1
INTRODUCTION 2
MATERIALS AND METHODS 9
DEVELOPMENT OF POPULATION JCD-1 9
DEVELOPMENT OF POPULATION JCD-2 10
BREEDING LINE EVALUATION 11
YIELD TESTS 11
PLOT TECHNIQUE 12
TRAITS EVALUATED 13
DNA ISOLATION AND EVALUATION 15
STATISTICAL ANALYSIS 15
AGRONOMIC TRAITS 15
ASSOCIATION OF SNP MARKERS WITH AGRONOMIC TRAITS 16
RESULTS 17
AGRONOMIC EVALUATION OF POPULATION JCD-1 AND JCD-2 17
DISCUSSION 17
REFERENCES 23
TABLES 30
FIGURES 47

CHAPTER 2 GENETIC DIVERSITY PATTERNS IN SINGLE CROSS *GLYCINE MAX* BY *GLYCINE SOJA* POPULATIONS. 51

ABSTRACT 51
INTRODUCTION 52
MATERIALS AND METHODS 54
DNA EXTRACTION AND SNP ANALYSIS 55
GENETIC DISTANCE AND MULTIVARIATE ANALYSIS 55
GRAPHICAL DEPICTION OF MDS RESULTS 56
SEGREGATION DISTORTION IN HIGHLY SELECTED BREEDING LINES 57
SINGLE MARKER ANOVA 58
MULTIPLE LINEAR REGRESSION 59
CAPTURING *G.SOJA* DIVERSITY IN A SUBSET OF LINES 59

RESULTS 60
GENETIC DISTANCE AND MULTIVARIATE ANALYSIS 60
SEGREGATION DISTORTION IN HIGHLY SELECTED BREEDING LINES 60
SINGLE MARKER ANALYSIS FOR KNOWN GENES 61
SINGLE MARKER ANOVA 62
CREATING A COLLECTION OF BREEDING LINES CAPTURING *G. SOJA* DIVERSITY 63
DISCUSSION 63
GENETIC DISTANCE AND MDS ANALYSIS 63
SEGREGATION DISTORTION IN HIGHLY SELECTED BREEDING LINES 65
SINGLE MARKER ANALYSIS FOR KNOW QTL: FLOWER AND PUBESENCE COLOR, AND SEED PROTEIN CONTENT 66
PUTATIVE YIELD ALLELES FROM SINGLE MARKER ANOVA AND MULTIPLE REGRESSION 67
PUTATUVE 100-SEED WEIGHT ALLELES FROM SINGLE MARKER ANOVA AND MULTIPLE REGRESSION 70
PUTATIVE SEED OIL CONTENT ALLELES FROM SINGLE MARKER ANOVA AND MULTIPLE REGRESSION 71
PUTATUVE LODGING ALLELES FROM SINGLE MARKER ANOVA AND MULTIPLE REGRESSION 71
CREATING A COLLECTION OF BREEDING LINES CAPTURING *G. SOJA* DIVERSITY 72
IMPLICATIONS TO SOYBEAN BREEDING 73
REFERENCES 75
TABLES 80
FIGURES 101

CHAPTER 3 STRATEGIES FOR EFFICIENT SELECTION OF UPRIGHT PLANTS IN A SINGLE CROSS *GLYCINE MAX* AND *GLYCINE SOJA* 116

ABSTRACT 116
INTRODUCTION 117
MATERIALS AND METHODS 120
POPULATION DEVELOPMENT 120
PLOT TECHNIQUE AND VISUAL EVALUATION OF $F_{2:3}$ FAMILIES .121
PLOT TECHNIQUE AND VISUAL EVALUATION OF $F_{3:4}$ FAMILIES .122
DATA ANALYSIS 123
RESULTS 124
OBSERVATIONAL RESULTS OF PLANT TYPE IN $F_{2:3}$ AND $F_{3:4}$ FAMILIES 124
GENETIC MODEL FITTING OF OBSERVATIONAL RESULTS 125

DISCUSSION 126
ESTIMATING THE NUMBER OF GENES CONTROLLING AN ERECT MAIN STEM 126
IMPLICATIONS TO SOYBEAN BREEDING 127
REFERENCES 130
TABLES 133
FIGURES 146

CHAPTER 4 CHARACTERIZING THE EFFECTS OF PHENOTYPIC TRAITS OF *GLYCINE SOJA* PI 366122 ON AGRONOMIC PERFORMANCE THROUGH THE USE OF NEAR ISOGENIC LINES 147

ABSTRACT 147
INTRODUCTION 148
MATERIALS AND METHODS 148
DEVELOPMENT OF NEAR ISOGENIC LINES 148
YIELD TESTING 149
FIELD PLOT TECHNIQUE, TRAITS EVALUATED, AND STATISTICAL ANALYSIS 150
RESULTS 151
DISCUSSION 151
REFERENCES 153
TABLES 155

APPENDIX 158
APPENDIX A. Means of agronomic traits for breeding lines tested in yield test TCLP 384, averaged over 5 environments from 2008 to 2010. 159

LIST OF TABLES

Table 1.1. Means of agronomic traits for breeding lines tested in yield test TCLP 381, averaged over 5 environments from 2008 to 2010. 30

Table 1.2. Means of agronomic traits for breeding lines tested in yield test TCLP 382, averaged over 5 environments from 2008 to 2010. 32

Table 1.3. Means of agronomic traits for breeding lines tested in yield test TCLP 383, averaged over 5 environments from 2008 to 2010. 34

Table 1.4. Means of agronomic traits for breeding lines tested in yield test TCLP 405A, averaged over 3 environments from 2009 to 2010. 36

Table 1.5. Means of agronomic traits for breeding lines tested in yield test TCLP 406A, averaged over 3 environments from 2009 to 2010. 38

Table 1.6. Means of agronomic traits for breeding lines tested in yield test TCLP 407A, averaged over 3 environments from 2009 to 2010. 40

Table 1.7. Means of agronomic traits for breeding lines tested in yield test TCLP 408A, averaged over 3 environments from 2009 to 2010. 42

Table 1.8. Means of agronomic traits for breeding lines tested in yield test TCLP 409A, averaged over 3 environments from 2009 to 2010. 44

Table 1.9. Means of agronomic traits for breeding lines tested in yield test TCLP 410A, averaged over 3 environments from 2009 to 2010. 46

Table 2.1 Summary of allele frequencies that are significantly skewed toward the *G. soja* allele within and across populations. All significance at $p<0.01$.....80

Table 2.2 Results of single marker ANOVA of agronomic traits for combined data of all breeding lines. 87

Table 2.3 Results of model selection and multiple linear regression of loci associated with increased percent of check yield derived from *G. soja*. R-square value of the model was 0.37. 97

Table 2.4 Results of model selection and multiple linear regression of loci associated with increased 100-seed weight derived from *G. soja*. R-square value of the model was 0.32. 97

Table 2.5 Results of model selection and multiple linear regression of loci associated with increased seed oil content derived from *G. soja*. R-square value of the model was 0.34. 97

Table 2.6 Results of model selection and multiple linear regression of loci associated with increased seed protein content derived from *G. soja*. R-square of the model was 0.43. 98

Table 2.7 Results of model selection and multiple linear regression of loci associated with lodging tolerance derived from *G. soja*. R-square of the model was 0.21. 98

Table 2.8 Summary of SNP diversity collection created from JCD-1.....98

Table 2.9 Summary of SNP diversity collection created from JCD-2.....99

Table 2.10 Summary of SNP diversity collection created from both populations. 99

Table 2.11 Summary of SNP diversity collection created by breeding lines that yield within 70% of the *G. max* checks. Collection contains at least 1 homoyzgote for the *G. soja* allele at 556 SNP loci of a total of 558.....100

Table 3.1. Theoretical number of F2 plants needed to recover, on average, one "Max-Like" plant from a wild x domesticated soybean hybridization, for various genetic models. Calculations are based on the binomial distribution and the assumption that upright growth habit is a minority phenotype conditioned by 1 to 10 segregating genes. For each gene number scenario, the effect of dominance on recovery of upright plants is examined. For example, given that one gene controls upright growth habit, one in four F_2 plants would be expected to be upright on average. If the locus exhibits dominance for upright growth habit, then 3 of 4 F_2 plants (1 homozygous plant and 2 heterozygotes on average) would be expected to be upright.....133

Table 3.2 Theoretical number of F3 plants needed to recover, on average, one "Max-Like" plant from a wild x domesticated soybean hybridization, for various genetic models. Calculations are based on the binomial distribution and the assumption that upright growth habit is a minority phenotype conditioned by 1 to 10 segregating genes. For each gene number scenario, the effect of dominance on recovery of upright plants is examined. 134

Table 3.3 Expected number of plants to be identified with an erect main stem in an F_3 population of 251,550. An estimated 251,550 F_3 plants were grown from 2009-2010, from which 296 upright plants were recovered. 135

Table 3.4 Results of Chi Square Tests for Model Fitting 2009-2010 $F_{2:3}$ observational data. 296 F_3 plants were selected from an estimated population of 251,550. 136

Table 3.5 Segregation patterns observed within $F_{2:3}$ families in 2009 and 2010. 137

Table 3.6 Segregation patterns of $F_{3:4}$ families grown in 2010. 140

Table 4.1 Agronomic Data of NIL populations grown in TCLP 436 averaged across 3 locations from 2010-2011. Lines from same F_7 row trace back to same F_6 plant. 155

Table 4.2 Agronomic Data of NIL populations grown in TCLP 437 averaged across 3 locations from 2010-2011. Lines from same F_7 row trace back to same F_6 plant. 156

Table 4.3 Mean comparisons across NIL groups of TCLP 436 for the various traits of interest. 157

Table 4.4 Mean comparisons of the leaf shape groups of TCLP 437. 157

LIST OF FIGURES

Figure 1.1 (Top to bottom, left to right) Regression percent check yield versus maturity, and percent *G. soja* alleles versus percent check yield, protein content, and oil content, for breeding lines from population JCD-1.....47

Figure 1.2 (Top to bottom, left to right) Regression of percent *G. soja* alleles versus lodging score, seed size, and maturity date for breeding lines from population JCD-1. 48

Figure 1.3. (Top to bottom, left to right) Regression percent check yield versus maturity, and percent *G. soja* alleles versus percent check yield, protein content, and oil content, for breeding lines from population JCD-2.....49

Figure 1.4 (Top to bottom, left to right) Regression of percent *G. soja* alleles versus lodging score, seed size, and maturity date for breeding lines from population JCD-2. 50

Figure 2.1 MDS plot of 120 breeding lines from population JCD-1(Blue) and JCD-2 and parental lines N7103 and *G. soja* PI 366122. 102

Figure 2.2 Graphical representation of allele frequencies of SNPs on Chromosomes 1-20. Frequencies below black line represent skewness towards N7103 allele, while frequencies above represent skewness toward *G. soja* PI 366122. 105

Figure 3.1 Distribution of the segregation patterns in $F_{3:4}$ families grown in 2010. 146

CHAPTER 1

A NOVEL MEGA-POPULATION APPROACH FOR BREEDING WITH WILD SOYBEAN

ABSTRACT

The wild soybean, *Glycine soja* (Sieb. and Zucc.) remains a relatively untapped source of diversity in soybean [*Glycine max* (L.) Merr.] breeding. The underutilization of this genetic resource arises from the strikingly poor performance of the progeny, when *G. soja* is hybridized with the cultivated form. This problem has led most researchers to employ the backcross breeding method when incorporating alleles from wild soybean into applied breeding programs. However, no commodity cultivars have wild soybean pedigree in North America to date, suggesting that backcrossing is an imperfect answer to the problem. We theorized that the agronomic barrier posed by wild soybean could be better resolved by returning to the single cross approach and applying the plant breeding adage, 'breeding is a numbers game', in a novel way, using mega-populations and intense selection. Cultivar 'N7103' x wild soybean PI 366122 were hybridized and a total of more than 5000 F_2 plants were bulk harvested. Over 1.2 million F_3 plants were grown and, approximately 375 individual F_3 plants exhibited upright growth habit similar to soybean. To our knowledge, this is the largest nursery ever devoted to a single hybridization combination in soybean. Through subsequent pedigree selection, we developed a total of 192 F_4-derived lines and evaluated their agronomic performance in three to five environments from 2008 to 2010 at Kinston and Clayton NC. A total of 558 polymorphic SNP markers were used to verify

pedigree correctness of the lines. Mean yield of the breeding lines was 70% of the commodity checks. A total of 50 F_4-derived breeding lines yielded at least 75% and 11 yielded at least 90% of commodity checks. Lodging scores for 65 lines were comparable to check cultivars. Pod dehiscence was minimal, with an average of 0.5% across environments. Several lines exhibited transgressive segregation for large seed size. Marker data indicated intense selection reduced the average percentage of wild soybean SNP alleles from the expected 50% under random selection to an actual 25% in the breeding lines, the equivalent of one backcross. Individual lines ranged from 14% to 47% percentage of wild soybean alleles. An increased percentage of wild soybean alleles in breeding lines was associated ($p<0.05$) with decreased yield performance in one population but not the other, based on genotypic means over environments. Increased lodging and decreased seed weight were associated ($p<0.05$) with increased percentage of *G. soja* alleles in both populations. We conclude that mega-population sizes coupled with intense selection in the manner described here offers an effective alternative to backcrossing for introducing diversity from wild soybean into the soybean breeding pool.

INTRODUCTION

Favorable genetic diversity for quantitative traits is important in applied crop breeding, because genetic gain from selection is directly proportional to the inherent genetic variation being selected upon within the population (Falconer and Mackay 1996). The genetic diversity in applied soybean [*Glycine max* (L.) Merr.] breeding programs in the USA has been narrowed, in comparison to the original diversity present in its progenitor, wild soybean *Glycine soja* (Sieb. And Zucc.) by three separate historical events: i. the

domestication of soybean from the wild in Asia, ii. its initial introduction into the United States *via* only a few ancestral types, and iii. a continual interbreeding of related types and associated fixation of alleles, as a by-product of successful breeding in the USA over the last 70 years. Hyten et al. (2006) indicated that the greatest contributing factor to the loss of diversity in cultivated soybean was the domestication from *G. soja*. The authors determined that domestication resulted in a 50% reduction in sequence diversity and the elimination of 81% of the rare alleles found in *G. soja*. Gizlice et al. (1994) showed that 85% of the genetic base in North American soybean cultivars can be traced back to only 16 ancestral introductions. Hyten et al. (2006) found that although North American ancestors contained nearly as much DNA sequence diversity as all available Asian soybean landraces, there was a significant loss in rare alleles among those 16 ancestors. Carter et al. (1993) calculated that modern North American cultivars now have an average relation of half sibs, when using coefficient of parentage as a measure of diversity.

Although yield improvements continue to be made in applied breeding programs in North America, there is concern that incremental advances are harder to achieve now than in the past. Intense selection over decades may have fixed major yield alleles early in breeding programs, leaving mainly alleles with smaller effects as the basis for yield advancement today. Guzman et al. (2007) identified quantitative trait loci (QTL) for seed yield from exotic sources, but the majority of the QTL had been identified in US cultivars previously, which supports this theory that yield advancement today may be based on the accumulation of 'small-effect' alleles. For this reason, public and private soybean breeders and molecular

geneticists all consider the broadening of the genetic base of applied breeding as a high priority research goal for the public sector. The United Soybean Board, a national soybean commodity group, now funds a large project specifically designed to address the perceived problem of a narrow genetic base of North American soybean breeding (Steve Muench and Randy Nelson, 2011, personal comm.). The approach in this public sector effort is to develop high yielding new breeding stock *via* hybridization with exotic soybean accessions and make these available for public and private breeding programs

Although exotic accessions of domesticated soybean are perhaps the most easily and most often used genetic source of novel genetic variation for applied soybean breeding, the wild soybean has received increased interest in recent years as a unique source of desirable traits (Hyten et al., 2006). Wild soybean is a diploid (2n=40) species that crosses freely to the domesticate and has long been thought to be its progenitor (Hymowitz 2004). The time and location of domestication has been a subject of debate, but recent archeological evidence suggests that soybean may have been domesticated in multiple regions of the Asian continent, nearly 5,500 years ago (Lee et al., 2011). More than 1,100 accessions of wild soybean are available from the USDA soybean germplasm collection as parental stock for breeding programs.

In the search for desirable traits and alleles in wild soybean, Diers et al. (1992) identified alleles from *G. soja* which increase seed protein content. Rebetzke et al. (1997) concluded that alleles from three *G. soja* plant introductions (PIs) were associated with increased palmitic and linoleic acid content of soybean seed oil content. Pantalone et al. (1997) identified novel alleles in wild soybean that control the expression of desaturase

enzymes that regulate linolenic acid production in soybean oil. The seed storage proteins found in *G. soja* influence soyfoods properties greatly, with the ratio of 11S/7S seed proteins varying from about 1.7 to 4.9 in wild soybean (Kwanyuen et al., 1997, Xu et al., 1990). This ratio can affect the nutritional desirability of the soybean protein meal. In addition to seed quality traits, *G. soja* accessions have been identified which are resistant to the soybean cyst nematode (*Heterodera glycines* Ichinohe) and salt stress (Wang et al., 2001, Luo et al., 2005).

Alleles which improve the yield potential of soybean may also reside in the wild soybean. Li et al. (2008) identified two quantitative trait loci (QTL) from *G. soja* PI 245331 (USDA-ARS, NGRP, 2012a) that imparted a combined 6.3% yield advantage across three environments, employing the Tanksley and Nelson (1996a) backrossing method. These effects were confirmed in a separate population, developed from a cross of the same parents. Concibido et al. (2003) also identified a yield enhancing QTL from *G. soja* PI 407305 (USDA-ARS, NGRP, 2012b) that conveyed a 9% yield increase over several locations. The QTL allele from this latter study was backcrossed subsequently into six additional genotypes and found to have a positive effect in two. Even though this QTL was not stable across genetic backgrounds, these studies, taken together, showed a potential for yield enhancement using wild soybean. Other attempts to identify yield enhancing alleles using this method have been unsuccessful (Perry Cregan, 2012, person comm.). However, novel genetic alleles for yield need not be universally favorable in order to be useful. In fact, Context-Specific MAS (CSM) is already showing great promise as an effective strategy for MAS of complex trait QTL that are highly context dependent (Sebastian et al, 2010; Sebastian et al, 2012)

Although an abundance of studies suggest that *G. soja* has great potential as a novel source of variation in applied breeding programs, it remains relatively unused at present. Only 2 of 651 Chinese cultivars released by 1999 had a *G. soja* parent (Cui et al., 1999). In the 1950s and 1960s, public North American soybean cultivars were released for the natto market in which the small seed size characteristic of these cultivars was obtained from wild soybean. Small-seeded cultivars AC Colombe, IA2055, IA2005, and IA2023 have approximately 12.5% or less *G. soja* pedigree in their background (Cui et al., 2004). The underutilization of *G. soja* arises from the large genotypic and phenotypic differences between cultivated and wild soybean. When F_2 plants are produced from a mating of domesticated and wild soybean, they tend to be procumbent, small seeded and prone to shattering, much like the wild parent, even when thousands of F_2 plants are generated (Weber, 1950).

When challenged with difficult-to-use breeding stock such as wild soybean, breeders often focus on qualitative traits and rely on the backcross breeding method to transfer such traits to an adapted genetic background (Fehr 1993). This method increases the frequency of desirable segregates in a breeding population with a minimum of breeding effort. The Tanksley and Nelson (1996a) approach is the perhaps the most refined backcross breeding method attempted, thus far, to move alleles from poorly adapted wild species to domesticated stock.

Less well appreciated is the idea that backcrossing can play an important role in quantitative trait breeding, where individual alleles are unlikely to be identified. Backcrossing can be an important breeding tool in a quantitative breeding framework for a

trait such as seed yield, where the donor parent is very low yielding. This is the case when wild and domesticated soybean are used as contrasting parents. Isleib (1999) concluded that the utility of the backcrossing method is a function of the number of allelic differences between the parents, the proportion of desirable alleles in the better parent, and the number of loci homozygous for desirable alleles by which a progeny must exceed the number in the better parent to be selected. If the proportion of desirable alleles in the better parent is close to 0.5, then a bi-parental cross followed by a generation of self-pollination and selection is likely a better breeding method than backcrossing. However, as the proportion of favorable alleles approaches one, backcrossing becomes the more desirable method. Previous studies indicate that this ratio of favorable yield alleles in *G.max* vs. *G. soja* may be close to one. Ertl and Fehr (1985) concluded that breeding populations containing as much as 25% *G. soja* germplasm, the equivalent of a single backcross, based on pedigree information, cannot maintain acceptable agronomic quality. Carpenter and Fehr (1986) showed that, in the absence of selection, three backcross generations (i.e. populations with only 1/16 of their pedigree from wild soybean) are needed in order for the breeder to recover a sufficient frequency of segregates that are agronomically similar to the recurrent parent in soybean.

Thus far, and for the reasons above, breeding with wild soybean has been limited to some form of backcrossing breeding methodology. However, this approach clearly limits a breeder's ability to extract quantitative variation from wild soybean for a trait such as seed yield. For example, Tanksley and Nelson's (1996a) backcrossing breeding method is not designed to identify multiple 'yield' alleles with small effects that may interact epistatically. Such non-additive effects are typical of 'yield QTL' (Sebastian et al, 2010; Sebastain et al,

2012). One other alternative to backcrossing is recombinant inbred line (RIL) development for QTL mapping. However, RIL development does not usually produce agronomic lines from a wild and domesticated soybean hybridization, even when more than 500 are developed. Thus, RIL development is not usually considered an applied breeding methodology for wild soybean.

If useful quantitative variation exists, and is to be extracted from the wild soybean, a more insightful approach than backcrossing or RIL development may be necessary. One option is to develop an extremely large bi-parental population beyond the bounds of normal breeding and expose it to intense phenotypic selection against poor agronomic appearance in early generations of inbreeding. This approach is similar to mutagenesis in that the desired progeny are expected to be very rare, but unique in that the desired progeny are produced via recombination as opposed to induced mutation. The first objective of this study was to test the hypothesis that agronomic breeding lines can be developed from single-cross populations of domesticated x wild soybean, *via* novel mega-populations and intense selection. The second was to determine the equivalency of this approach, *post hoc*, to backcrossing by characterizing the percentage of *G. soja* alleles present in breeding material, based on single nucleotide polymorphism (SNP) markers. The third objective was to determine the association between percentage of alleles from *G. soja* in a breeding line and its agronomic performance.

MATERIALS AND METHODS

DEVELOPMENT OF POPULATION JCD-1

The population JCD-1 was developed from a cross of *G. max* variety N7103 (Carter et al. 2003) and *G. soja* PI 366122 (USDA-ARS, NGRP, 2012c). N7103 was adopted as a parent because of its resistance to lodging, while PI 366122 was selected from a group of *G. soja* accessions received from Dr. Randy Nelson (2011 personal comm.), based on reduced levels of infection by *Xanthomonas campestris* pv. *Glycines* (the casual agent in bacterial pustule) and overall desirable appearance in the field in NC. The two parents were crossed in the summer of 2002. The F_1 seed were harvested in the fall and F_1 plants were grown subsequently at the USDA-ARS Tropical Agriculture Research Station (TARS) winter nursery in Isabella, Puerto Rico from December 2002 through April 2003. Approximately 2,500 F_2 plants were grown in Clayton, NC at the Central Crops Research Station over the following summer. As the F_2 plants neared maturity, late maturing F_2 plants (approximately 100) were removed. The remainder of the F_2 plants was harvested in bulk using a plot combine. The combine was cleaned thoroughly prior to harvest. Pod dehiscence was negligible. The harvest yielded approximately 900,000 F_3 seed into two large burlap bags, designated as one and two. These large bulks were sent to the University of Georgia for scarification with an alfalfa scarifier in order to ensure high germination. In 2004 at the Caswell Research Station at Kinston, NC, three ha of F_3 plants were grown in four blocks based on seed size. There was a block corresponding to the largest 15% seed from bag one, the smallest 85% from bag one, the largest 15% from bag two, and the smallest 85% from bag two. Average 100-seed weight for the largest groups was approximately 7 grams. Seed

were planted in 1 m row width and at approximately 4 seed per m of row. In the fall of 2004, approximately 300 F_3 plants were selected from across these four blocks. Plants were selected based solely on the criterion of upright growth habit and harvested individually. Approximately 95% of the plants selected came from blocks planted with the larger seed. Approximately 92 individually harvested plants were discarded because of they appeared to be contaminates in the population or had low seed set. In 2005, the remaining 208 $F_{3:4}$ progeny rows were grown in Clayton, NC and approximately 700 individual F_4 plants were selected from individual rows to create F_4–derived lines. In 2006, these $F_{4:5}$ lines were grown for seed increase and visual observation at Clayton, NC. Four hundred of the most visually desirable lines were bulk harvested. The $F_{4:6}$ lines were yield tested in a series of yield trials employing a randomized complete block design (RCBD) with three replications at the Central Crops Research Station near Clayton , NC during 2007 (data not shown). Based on yield results, 104 F_4-derived lines were selected for multi-location yield trials from 2008-2010. These lines traced back to 72 individual F_3 plants.

DEVELOPMENT OF POPULATION JCD-2

Population JCD-2 was developed from the same parental cross as population JCD-1. The initial cross between N7103 and PI 366122 was made in the summer of 2003. The F_1 plants were grown at the TARS winter nursery from December 2003 through April 2004. Approximately 2,500 F_2 plants were grown in Clayton, NC during the summer of 2004 and harvested as a bulk in the fall using a plot combine. The F_3 seed were sieved based on seed size, and the largest 10% were kept for harvest. This 10% was then broken into four groups using sieves with contrasting hole diameters. Group 1 consisted of seed that rested upon a

sieve of 0.46 cm hole diameter, group 2 consisted of seed that rested on a sieve of 0.44 cm hole diameter, group 3 consisted of seed that rested on a sieve of 0.42 cm hole diameter, and group 4 consisted of seed that rested on a sieve of 0.40 cm hole diameter. In 2005, F_3 seed were planted in 4 blocks based on seed sieving. The four blocks covered 0.6 hectare of field space. The F_3 population was planted in 1 m row width at 4 seeds per m. Approximately 275 individual F_3 plants were selected based on plant erectness. Plants that had a 100 seed weight less than five grams were discarded, as well as plants with poor seed quality. In total, 203 F_3 plants were retained. In the summer of 2006, $F_{3:4}$ progeny rows were grown in Clayton, NC. Approximately 600 F_4 plants were selected based on upright plant growth. The $F_{4:5}$ lines were grown as an unreplicated yield trial in the summer of 2008 at Clayton, NC (data not shown). Of these nearly 600 lines, approximately 200 were retained and grown in replicated yield trials at Clayton, NC in 2009, as well as at Clayton and Kinston, NC in 2010. These lines traced back to 65 individual F_3 plants.

BREEDING LINE EVALUATION

YIELD TESTS

The 102 F_4-derived breeding lines from population JCD-1 were divided into three separate yield trials and designated as TCLP 381, TCLP 382, and TCLP 383. Lines were subdivided into smaller groups based on maturity observations. The three individual tests each consisted of 40 entries, including check cultivars, and were grown in proximity to each other in six environments: Clayton, NC during 2008, Clayton and Kinston, NC during 2009 and in Kinston and Mt. Olive, NC in 2010. Test TCLP 381 included the four conventional checks: ‘Stressland’ (Cooper et al. 1999), ‘Holladay’ (Burton et al. 1996), ‘5601T’

(Pantalone et al. 2003), and 'NC-Roy' (Burton et al. 2005). TCLP 382 included two entries each of the conventional checks 5601T and NC-Roy. TCLP 383 included two entries each of the conventional checks: Holladay, 5601T, NC-Roy, and 'NC-Raleigh' (Burton et al. 2006).

The 134 F4-derived breeding lines from population JCD-2 were divided into six separate yield trials designated as TCLP 405A, TCLP 406A, TCLP 407A, TCLP 408A, TCLP 409A, and TCLP 410A and tested in proximity to each other in a total of 3 environments: Clayton, NC during 2009, and Clayton and Kinston, NC during 2010. Each test included the conventional checks 5601T, NC-Roy, N7103, and NC-Raleigh. In 2010, 44 breeding lines were dropped from population JCD-2 because of poor agronomic performance, and the populations were truncated to the final 90. Piepho and Mohring (2006) pointed out the issues that can arise when pooling test data when selections are made between years. However, due to the explorative nature of the study and because the deleted lines were assumed to be low yielding, bias effects were considered minimal.

PLOT TECHNIQUE

The USDA-ARS breeding program in Raleigh, NC managed all yield plots grown in Clayton and Kinston, NC. All plots contained three rows, 97 cm wide, which were end trimmed to a length of 4.5 m near maturity. The center row was harvested from these plots, constituting a final plot area of 4.5 m^2. Plant populations were approximately 306,000 plants ha^{-1}. Collaborators managed the plots grown at Mt. Olive, NC. Each plot consisted of four rows planted 4.88 m long spaced 76.2 cm apart in which the center two rows were harvested (7.44 m^2). Plant populations were approximately 344,000 plants ha^{-1}.

TRAITS EVALUATED

Plots were rated for height, maturity, lodging, 100-seed weight, seed yield and pod dehiscence (shattering). Height was recorded for each plot as the mean of 3 plants, and maturity was recorded as the first day on which at least 95% of the pods turned a mature color. Lodging was measured on a 1 to 5 scale, where 1 indicates all plants were erect and 5 indicates all plants were prostrate. In some environments, pod dehiscence (shattering) was detected. Visual estimates of shattering were made on the day of harvest using a 0 to 100 scale to indicate the percentage of pods affected. Lines were also evaluated for seed oil and seed protein content using two different methods. Lines with a yellow seed coat were analyzed for protein and oil content using near-infrared spectroscopy (American Association of Cereal Chemistry, 1999) at the National Center for Agricultural Utilization Research, USDA-ARS, in Peoria, IL. The seed coat color of some breeding lines was sufficiently dark to preclude analysis using near-infrared spectroscopy. For these lines, protein and oil content analysis was performed at Raleigh, NC by the USDA-ARS Soybean and Nitrogen Fixation Unit. Oil and moisture content of seed were determined by a pulsed proton NMR using a Maran pulsed NMR (Resonance Instruments, Witney, Oxfordshire, UK) by the Field Induction Decay-Spin Echo procedure of Rubel (1994). Oil content was reported on a zero moisture basis. Protein content was determined by the Dumas combustion method(American Oil Chemists' Society, 1997) using an Elementar Rapid N III Nitrogen-Analyzer (Elementar Americas, Inc., 520 Fellowship Rd.,Suite D-408, Mt. Laurel, NJ 08054 USA). Meal was oven dried overnight at 80°C. Samples (0.2g) were prepared in tin foil packets for combustion analysis. Protein was calculated from nitrogen (N) values using the following

conversion factor: Protein content=6.25x N concentration. All control cultivars were yellow seeded and, thus, were used in both analyses so that results from Peoria and Raleigh, could be pooled. A regression analysis was conducted on the mean protein content of the check cultivars for the two methods. The regression line was used to adjust the values obtained from the smaller data set generated at Raleigh to an equivalency for the seed composition values from Peoria. This method was also used to generate unbiased estimates of oil content for the trials.

The maturity dates of the breeding lines developed in both populations varied from mid maturity group (MG) V to early MG VII. Relative maturity was estimated for each breeding line on a test by test basis, by first regressing the recorded maturity dates of the commodity checks against published relative maturity estimates and then using this regression equation subsequently to estimate a relative maturity for the breeding line. To allow for yield comparisons across maturity groups, we calculated yield as a percentage of check for each breeding line (%EYC). We accomplished this by first regressing the average seed yield of the commodity checks against their published relative maturity and then using this regression equation to estimate an appropriate check yield for all relative maturity values (EYC) in the study, on a test by test basis. We then took the average seed yield of each individual breeding line, and divided it by its estimated 'relative maturity check yield' for its particular test to obtain %EYC. A *post hoc* regression of %EYC with maturity date revealed no significant ($p<0.05$) association.

DNA ISOLATION AND EVALUATION

DNA samples were collected from all breeding lines in both JCD-1 and JCD-2 as well as from the parents, N7103 and PI 366122. DNA was isolated by sampling root tips from 20 seedlings of remnant $F_{4:6}$ seed and combining them into one sample for each line. The Qiagen Plant Easy DNA Extraction Kit (Qiagen, Hilden, Germany) was used to obtain purified DNA.

1,536 SNP markers were tested on N7103, PI 36122, and the 225 breeding lines using the GoldenGate assay and analyzed on the Illumina BeadStation 500G (Illumina, San Diego, CA) as described previously (Hyten et al., 2008). The automatic allele calling for each locus is accomplished with the GenomeStudio software (Illumina, San Diego, CA). The allele calling was then verified manually, and only SNPs that were polymorphic between the parents were used for further analysis.

Marker data were used to detect outcrossing and other contamination within the populations. Because all breeding lines were derived from a single F_4 plant, the expectation based on random chance was that any line would be heterozygous at 12.5% of the loci (Fehr 1987). Any line that varied significantly from this expected ratio was dropped from the data analysis. A total of eight lines were dropped from further analysis based using this rationale.

STATISTICAL ANALYSIS

AGRONOMIC TRAITS

Seed yield and other agronomic traits were evaluated in the field with a randomized complete block experimental design. All yield trials were replicated three times in each environment. Seed yield was recorded on all three replications, as were most other

agronomic traits. Plant height, protein content, and oil content were measured on only two replications per environment for breeding lines from population JCD-1 in Clayton and Kinston, NC during 2009. Lodging was measured on only two replications for breeding lines from population JCD-1 in Clayton, NC during 2009. Only two replications of maturity date were recorded on these lines in 2008 at Clayton. No protein and oil measurements were made on lines tested at Mt. Olive in 2010. Only two replications of protein and oil content were measured on lines tested from population JCD-2 for both years of testing. Analysis of variance was performed using the GLM procedure in SAS (SAS Institute, 2007). Genotypes were treated as fixed effects, while location and replication were treated as random effects. Fischer's protected LSD was used for comparison of genotypes over locations using the genotype x location mean square as the error estimate.

ASSOCIATION OF SNP MARKERS WITH AGRONOMIC TRAITS

The 1,536 SNP markers were used to detect polymorphisms between the two parents, N7103 and PI 366122. The allelic contribution to each breeding line of each parent was calculated using the equation: (A+1/2H)/(A+H+B), where A is the number of loci homozygous for parent A, H is the number of heterozygous loci, and B is the number of loci homozygous for parent B. We then regressed the mean agronomic performance of breeding lines against the percentage of PI 366122 alleles present in the line for each population using the REG procedure in SAS.

RESULTS

AGRONOMIC EVALUATION OF POPULATIONS JCD-1 AND JCD-2

Overall agronomic performance of breeding lines was similar in the two populations. Genotype as well as genotype by environment (GxE) interaction effects were significant ($p<0.0001$) for all traits in both populations in all trials (Data not shown). The detection of significant GxE was due to minor fluctuations in rank between some of the poorer performing lines. We were able to identify a total of 25 breeding lines in each population that yielded at least 75% of the *G. max* checks (Tables 1.1-1.9). In population JCD-2, the N7103 parental check was included, allowing detection of significant ($p<0.05$) transgressive segregation for increased 100-seed weight. Over both populations, we identified 65 breeding lines that had lodging scores not significantly different from the *G. max* checks. Breeding lines tended to be higher in seed protein content and lower in seed oil content when compared to the *G. max* checks.

Regression analysis of agronomic traits against the percentage of *G. soja* alleles revealed that increased lodging and reduced 100-seed were associated ($p<0.05$) with an increased percentage of *G. soja* alleles in the breeding lines of both populations. An increased percentage of *G. soja* alleles was also associated increased seed protein content and decreased seed oil content and yield expressed as a percent of check in population JCD-1 (Fig. 1.1 and 1.2).

DISCUSSION

The most famous adage in plant breeding is, “plant breeding is a numbers game”. That is, the more breeding lines that are evaluated in a program, the more likely one is to

identify a valuable new variety among them. In the case of wild soybean, traditional breeding approaches do not appear to have played the numbers game well enough to deliver products from wild soybean to the farmer's field. This plant breeding failure stems from the fact that upright plant growth habit is controlled by many genes and that current breeding approaches do not take this fact into account sufficiently. For example, the single seed decent (SSD) breeding method (Brim, 1966), often employed in development of populations from single crosses, is a practical breeding approach only when parents are relatively equally adapted. If the number of homozygous loci required to produce an upright inbred line is ten (not unlikely in a wild x domesticated soybean cross), a breeder must grow out at least 1,000 F_4 plants in order to ensure at least one desirable recombinant plant would be identified. In public soybean breeding programs, a breeder may want to evaluate ~200 individual breeding lines after SSD in order to ensure that breeding lines will be identified which not only stand upright but also have good yield potential. Given that the SSD method involves hand picking of pods, this scenario suggests the hand picking of 200,000 pods in the F_2 and F_3 generations and then harvesting of 200,000 individual plants in the F_4 generation. This would require one worker perhaps 30 days to pick pods in each generation and six months to harvest individual plants. Given the limited resources available to most soybean breeders, this is an insurmountable task. Normal SSD populations have far less than 5,000 F_2 plants in soybean. While backcrossing fares better than SSD, under the scenario of 10 genes controlling plant uprightness, the cycles of crossing and selection process become protracted, and it is not clear that more than a fraction of the *G. soja* genome could be transferred to an adapted soybean background.

If one is to make desirable selections from among progeny of wild soybean, then one must find a way to make the numbers game more manageable. A key factor in the numbers game for the plant breeder in domesticated x wild populations is phenotyping. In the case of domesticated x wild soybean hybridization, most progeny are visually and agronomically undesirable at every stage of inbreeding, resembling the wild soybean much more than the domesticate, *G. max*. Thus, visual selection for agronomic acceptability can be made with great effect and low cost on a grand scale. Rare desirable plants simply stand out in stark contrast to the other viny, prostrate progeny. This contrast is easily visible at a glance, and identifiable to even those with little knowledge of plant breeding. It is this key factor that we capitalized upon to overcome the deficiencies of backcrossing and SSD to recover adapted plants from a *G. max* x *G. soja* cross. By resorting to F_3 mega-populations (populations in excess of 1 million individuals), we were able to recover close to 600 upright plants which eventually produced adapted inbred breeding lines. To our knowledge, this is the largest nursery ever devoted to a single hybridization combination in soybean. Larger seed size is also relatively easy to select upon and appeared to be helpful as a selection tool in the F_3 generation to enrich the breeding population with upright plants. In the development of population JCD-1, we sieved the two bags of bulk harvested seed, and then planted four blocks corresponding to large and small seed bulks. Ninety-five percent of the F_3 plants selected came from blocks planted with the largest 15% of the seed. In the development of population JCD-2 we planted out only the largest 10% of seed, and achieved similar results. We did not directly quantify the relationship between seed size and plant erectness, but the repeated success of developing agronomically acceptable breeding lines from the large

seeded progeny and transgressive segregation for large seed size in the progeny for the two separate populations points to a positive relationship between these traits.

Clearly, the single cross approach, as we practiced it, was successful in developing adapted progeny from the domesticated x wild soybean hybridization. In total we identified 50 F4-derived breeding lines, 25 in each population, that yield at least 75% of conventional check varieties. These lines can be traced back to 35 individual F_3 plants. Looking at these results further, we were able to identify 11 breeding lines that yield at least 90% of the check varieties. We also identified breeding lines that had significantly greater protein contents, less oil content, and greater lodging than the maturity checks. However, NC-Roy, a commonly grown group 6 conventional variety, is lodging prone, so this effect is not of great concern. None of the lines from either population had yields that were significantly ($p<0.05$ value) higher than N7103 or any of the *G. max* checks tested. Accurate measures of seed yield of *G. soja* PIs are difficult to obtain due to the amount of labor required to manually plant and harvest yield plots, so this information was not available when parental selection was made and the study was initiated.

Tanksley et al. (1996b) used the advanced backcross method to increase fruit size in tomato with alleles from a wild relative. Our method employed here was able to achieve a similar result in regards to seed size. We were able to identify transgressive segregants for larger seed size in population JCD-2, based on the direct comparison to the *G. max* parent N7103 with the LSD values listed. N7103 was not included in the yield tests of lines from JCD-1, but appeared to have 100-seed weights beyond those measured on N7103.

Carpenter and Fehr (1986) stated that at least three backcrosses would be necessary to recover a large number of agronomically acceptable progeny, in the absence of selection during population development. Our data show that when simple phenotypic selection is used, no backcrossing would be necessary in order to recover acceptable agronomic progeny. It should be noted that their definition of agronomically acceptable differed from the one employed in this study. They took individual ratings on lodging, agronomic score, vining score, shattering, and leaf abscission. Our selection method was not as strict, as F_3 plants were only selected based solely upon upright plant growth. As selection criteria are relaxed, the number of allelic differences between parents, as referenced in Isleib (1999), decreases. This allowed for us to select a sufficient number of F_3 plants for evaluation in subsequent tests. Our selection method resulted in 13 of 225 total breeding lines that contained only 25% or less *G. soja* germplasm. If we had selected on more traits other than plant erectness, this number may have increased significantly, as the selection would have been for even more "Max-like" characteristics as outlined in our study.

Ertl and Fehr (1985) reported that soybean populations cannot contain 25% *G. soja* germplasm, the equivalent of one backcross, and maintain their desirable agronomic characteristics. The regression analysis of population JCD-1 appears to follow a similar trend to the one they reported (Fig 1.1-1.2), but within this population we identified breeding lines that yield at least 80% of the conventional checks numerically, but still maintain as much as 33% *G. soja* germplasm based on SNP marker data. Surprisingly, we were not able to detect a significant association between the percentage of *G. soja* alleles present and percent of the check yield in Population JCD-2 (Fig. 1.3). We were able to identify a breeding line, NMS5-

253-1-531, that yielded at least 86% of the conventional checks numerically, but still maintained 40% *G. soja* alleles based on molecular marker data. This finding is a direct contradiction of the findings of Ertl and Fehr (1985), and a direct validation of the methods outlined as an effective way of introgressing diversity into agronomically acceptable soybean germplasm.

The progress that can be made toward introducing novel, allelic diversity from a wild relative into a cultivated crop species has been greatly hindered by possible genetic linkages or physiological relationships that can exist between desirable and undesirable traits of the wild relative. Typically, breeders have used backcrossing as a way to overcome these difficulties, which remains as the most efficient method for addressing these issues when faced with the introgression of few major qualitative genes. As molecular marker technology has advanced, breeders have been able to successfully select against large segments of the wild relative genome, in order to increase the frequency of agronomically acceptable progeny, in hopes of identify alleles that can affect traits that are quantitative in nature. However, early generation selection based on molecular markers, which itself can be a costly and time consuming approach, can result in the loss of novel genetic combinations that affect the trait of interest. It could also lead to loss of beneficial epistatic interactions between *G. max* and *G. soja* alleles that could affect the trait of interest. Our data show that it is possible to increase allelic diversity in cultivated soybean through single crosses to the wild soybean. Thus, the ability to produce genetically diverse breeding material while maintaining the agronomic quality of domesticated soybean can be readily achieved without the process of several generations of backcrossing. We have also created diverse material that can be used

as an agronomically reliable source of diversity as future breeding needs arise, such as the identification of new disease pathogens. As interest in the *G. soja* breeding pool increases, we propose this method as a way to create novel variation for quantitative traits of interest. As more data on the variation of quanitative traits in the *G. soja* germplasm collection becomes available, this method can serve as a way for breeders to easily and efficiently increase the amount of diversity readily available within their programs.

REFERENCES

American Association of Cereal Chemistry. 1999. Near infrared methods: Guidelines for model development and maintenance. 39-00. AACC, St. Paul, MN.

Official Methods and Recommended Practices of the American Oil Chemists' Society, 5th edn., American Oil Chemists' Society, Champaign, 1997, Protein Content by Combustion Method Ba 4e-93 (03).

Brim, C.A. 1966. A modified pedigree method of selection in soybean. Crop Sci. 6: 220

Burton, J.W., T.E. Carter, and D.T. Bowman 2005. Registration of 'NC-Roy' soybean. 2005. Crop Sci. 45: 2654.

Burton, J.W., T.E. Carter, and E.B. Huie. Registration of 'Holladay' soybean. 1996. Crop Sci. 36: 467.

Burton, J.W., T.E. Carter, M.O. Fountain, and D.T. Bowman. 2006. Registration of 'NC-Raleigh' soybean. Crop Sci. 46: 2710-2711.

Carpenter, J. and W.R. Fehr. 1986. Genetic variability for desirable agronomic traits in populations containing *Glycine soja* germplasm. Crop Sci. 26: 681-686.

Carter, T.E., J.W. Burton, M.R. Villagarcia, Z. Cui, X. Zhou, M.O. Fountain, D.T. Bowman, and A.S. Niewoehner. 2003. Registration of 'N7103' soybean. Crop Sci. 43: 1128.

Carter, T.E., Z. Gizlice, and J.W. Burton. 1993. Coefficient-of-parentage and genetic-proximity estimates for 258 North American cultivars released by public agencies during 1945 and 1988. USDA Tech. Bull. 1814. U.S. Gov. Print. Office, Washington, DC.

Concibido, V.C., B.L. Vallee, P. Mclaird, N. Pineda, J. Meyer, L. Hummel, J. Yang, K. Wu, and X. Delannay. 2003. Introgression of a quantitative trait locus for yield from *Glycine soja* into commercial soybean cultivars. Theor. Appl. Genet. 106: 575-582.

Cooper, R.L., R.J. Martin, S.K. St. Martin, A. Calip-DuBois, R.J. Fioritto, and A.F. Schmitthenner. 1999. Registration of 'Stressland' soybean. Crop Sci. 39: 590-591.

Cui, Z., A.T. James, S. Miyazaki, R.F. Wilson, and T.E. Carter, Jr. 2004. Breeding specialty soybeans for traditional and new soyfoods, In *Soybeans as Funcitonal Foods and Ingredients*. AOCS Press, Champaign, IL, p.p. 264-322.

Cui, Z., T.E. Carter, Jr., J. Gai, J. Qui, and R.L. Nelson. 1999. Origin, description, and pedigree of Chinese soybean cultivars released from 1923 to 1995. United States Department of Agriculture. Technical Bulletin 1871.

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

Ertl, D.S., and W.R. Fehr. 1985. Agronomic performance of soybean genotypes from *Glycine max* X *Glycine soja* crosses. Crop Sci 25: 589-592.

Falconer, D. S., and T. F. C. Mackay, 1996. Introduction to quantitative genetics. 4th Ed. Longmans Green, Harlow, Essex, UK

Fehr, W.R. 1987. Principles of Cultivar Development. Vol. 1. Theory and technique. Macmillan, New York.

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.

Guzman, P. S., B. W. Diers, D. J. Neece, S. K. St. Martin, A. R. LeRoy, C. R. Grau, T. J. Hughes, and R. L. Nelson. 2007. QTL associated with yield in three backcross-derived populations of soybean. Crop Sci. 47: 111-122.

Hymowitz, T. 2004. Speciation and cytogenetics. p. 97-136. *In* H.R. Boerma and J.E. Specht (3rd Ed.) Soybeans: Improvement, production, and uses. American Society of Agronomy, Madison, WI.

Hyten, D., Q. Song, I. Choi, M. Yoon, J. Specht, L. Matukumalli, R. Nelson, R. Shoemaker, N. Young, and P. Cregan. 2008. High-throughput genotyping with the GoldenGate assay in the complex genome of soybean. Theor Appl Genet 116:945-952.

Hyten, D.L., Q. Song, Y. Zhu, I. Choi, R.L. Nelson, J. M. Costa, J. E. Specht, R. C. Shoemaker, and P.B. Cregan. 2006. Impacts of genetic bottlenecks on soybean genome diversity. PNAS. Vol. 103. 45: 16666-16671.

Isleib, T. G. 1999. Recovery of superior homozygous progeny from biparental crosses and backcrosses. Crop Sci. 39: 558-563.

Kwanyuen, P. V.R. Pantalone, J.W. Burton, and R.F. Wilson. 1997. A new approach to genetic alteration of soybean protein composition and quality. JAOCS. 74: 983-987.

Li, D., T.W. Pfeiffer, and P.L. Cornelius. 2008. Soybean QTL for yield and yield components associated with *Glycine soja* alleles. Crop Sci. 48: 571-581.

Luo, Q., B. Yu, Y. Liu. 2005. Differential sensitivity to chloride and sodium ions in seedlings of *Glycine max* and *G. soja* under NaCl stress. Journal of Plant Phys. 162: 1003-1012.

Pantalone, V.R., F. L. Allen, and D. Landau-Ellis. 2003. Registration of '5601T' soybean. Crop Sci. 43: 1123-1124.

Pantalone, V.R., G.J. Rebetzke, R.F. Wilson, and J.W. Burton. 1997. Genetic regulation of linolenic acid concentration in wild soybean *Glycine soja* accessions. J. Am. Oil. Chem. Soc. Vol. 74 No. 2: 159-163.

Piepho, H.P. and J. Mohring. 2006. Selection in trials: Is it ignorable? Crop Sci. 46: 192-201.

Rebetzke, G.J., V. R. Pantalone, J. W. Burton, T. E. Carter, Jr., and R. F. Wilson. 1997. Genotypic variation for fatty acid content in selected *Glycine max* x *Glycine Soja* populations. Crop Sci. 37: 1636-1640.

Rubel, G. 1994. Simultaneous determination of oil and water contents in different oilseeds by pulsed nuclear magnetic resonance. J. Am. Oil Chemists Soc. 71(9): 1057-1062.

SAS Institute. 2007. The SAS system for Microsoft Windows. Release 9.1.3. SAS Inst., Cary, NC.

Sebastian, S. A., L.G. Streit, P.A. Stephens, J.A. Thompson, B.R. Hedges, M.A. Fabrizius, J.F. Soper, D.H. Schmidt, R.L. Kallem, M.A. Hinds, L. Feng, and J.A. Hoeck. 2010. Context-specific MAS for improved grain yield in elite soybean populations. Crop Science 50: 1196-1206.

Sebastian, S. A., L. Feng, and Les C. Kuhlman. 2012. Accelerated Yield Technology™: A Platform for Marker Assisted Selection of Simple and Complex Traits. In: Designing Soybeans for 21st Century Markets, ed. Richard F. Wilson, AOCS press, pp. 297-305.

Tanksley, S.D. and J.C. Nelson. 1996a. Advanced backcross QTL analysis: a method for the simultaneous discovery and transfer of valuable QTLs from unadapted germplasm into elite breeding lines. Theor Appl Genet 92: 191-203.

Tanksley, S.D., S. Grandillo, T.M. Fulton, D. Zamir, Y. Eshed, V. Petiard, J. Lopez, and T. Beck-Bunn. 1996b. Advanced backcross QTL analysis in a cross between an elite processing line of tomato and its wild relative *L. pimpinellifolium*. Theor. Appl. Genet. 92, 213–224.

USDA-ARS National Genetic Resources Program (USDA-ARS NGRP). 2012a. *Glycine soja* Sieb. And Zucc. PI 245331. Germplasm Resources Information Network (GRIN). [Online database] National Germplasm Resources Laboratory, Beltsville, MD. Available at http://www.arsgrin.gov/cgi-bin/npgs/acc/display.pl?1189261 (verified 09 Feb. 2012).

USDA-ARS National Genetic Resources Program (USDA-ARS NGRP). 2012b. *Glycine soja* Sieb. And Zucc. PI 407305. Germplasm Resources Information Network (GRIN). [Online database] National Germplasm Resources Laboratory, Beltsville, MD. Available at http://www.arsgrin.gov/cgi-bin/npgs/acc/display.pl?1306963 (verified 09 Feb. 2012).

USDA-ARS National Genetic Resources Program (USDA-ARS NGRP). 2012c. *Glycine soja* Sieb. And Zucc. PI 366122. Germplasm Resources Information Network (GRIN). [Online database] National Germplasm Resources Laboratory, Beltsville, MD. Available at http://www.arsgrin.gov/cgi-bin/npgs/acc/display.pl?1274061 (verified 09 Feb. 2012).

Weber, C.R. 1950. Inheritance and interrelation of some agronomic and chemical characters in an interspecific cross in soybeans, *Glycine max* x *Glycine ussuriensis*. Res. Bull. 374. Agricultural Experiment Station, Iowa State College. Ames, IA.

Wang, D., P.R. Arelli, R.C. Shoemaker, B.W. Diers. 2001. Loci underlying resistance to Race 3 of soybean cyst nematode in *Glycine* soja plant introduction 468916. Theor. Appl. Genet. 103: 561-566.

Xu, B., S.H. Zou, B.C. Zhuang, Z.P. Lin, and Y.J. Zhao. 1990. Study on seed storage component 11S/7S of wild soybean (*G. soja*) [In Chinese, with English abstract], Acta Agronomica Sinica 16: 235-24.

TABLES

Table 1.1. Means of agronomic traits for breeding lines tested in yield test TCLP 381, averaged over 5 environments from 2008 to 2010

Genotype†	Yield	Protein	Oil	Maturity	Est. MG‡	Lodging§	100-Seed Weight	Yield as % Check¶	% *G. soja* SNP alleles#
	kg ha^{-1}	g kg^{-1}	g kg^{-1}	Oct. 1 = 1			g		
Stressland	1841	418	196	5	4.5	1.9	14.9	-	-
Holladay	3003	393	207	14	5.3	1.4	15.1	-	-
5601T	2580	411	200	17	5.6	1.5	14.2	-	-
NC-Roy	2762	411	193	25	6.9	2.0	13.0	-	-
NMS4-6-10	2416	403	191	12	5.2	1.8	13.0	99	26
NMS4-31-11	1547	436	169	13	5.4	3.0	8.0	62	31
NMS4-30-302	1678	433	174	14	5.4	2.4	7.4	67	34
NMS4-19-243	2213	433	181	14	5.5	3.0	9.0	88	29
NMS4-104-527	1918	431	187	15	5.5	2.9	8.2	76	28
NMS4-4-132	2397	414	180	17	5.8	2.4	14.3	92	32
NMS4-19-245	2111	424	180	17	5.8	2.9	9.0	81	29
NMS4-38-315	1865	409	177	17	5.8	2.2	8.8	71	38
NMS4-138-682	1345	437	157	17	5.9	3.5	6.2	51	30
NMS4-37-308-1	1803	443	163	19	6.1	2.4	7.6	67	41
NMS4-37-305	1881	407	190	19	6.1	3.0	8.9	70	37
NMS4-37-307-1	1705	417	180	19	6.1	1.5	8.7	63	36
NMS4-3-117	1842	436	173	20	6.2	3.4	9.1	67	33
NMS4-67-462-1	2174	425	173	21	6.3	2.5	8.6	79	34
$LSD_{0.05}$	356	12	9	3		0.5	0.8	-	-

† NMS4 corresponds to lines in population JCD-1. The first number after the hyphen corresponds to F_3 plant.
‡ Estimated from regression of maturity date of conventional check against published maturity group.
§ 1 indicates all plants erect; 5 indicates all plants prostrate.
¶ Based on comparison of yield mean vs. regression of check means.
 # Based on 558 polymorphic SNP markers.

Table 1.1 cont.

Genotype†	Yield	Protein	Oil	Maturity	Est. MG‡	Lodging§	100-Seed Weight	Yield as % Check¶	% *G. soja* SNP alleles#
	kg ha^{-1}	g kg^{-1}	g kg^{-1}	Oct. 1 = 1			g		
NMS4-50-361	1509	423	157	21	6.3	3.8	6.2	54	33
NMS4-137-677	1527	447	154	22	6.4	4.0	6.0	55	32
NMS4-12-193-2	2030	424	166	22	6.4	2.9	6.4	72	31
NMS4-7-140	1547	439	173	22	6.5	3.3	8.4	55	35
NMS4-10-184	2137	428	179	22	6.5	3.0	9.9	76	30
NMS4-11-188	2087	421	175	22	6.5	2.9	6.6	74	34
NMS4-12-193	2016	428	172	23	6.5	2.8	7.1	71	31
NMS4-9-181	1789	458	164	23	6.6	3.1	7.9	63	35
NMS4-140-689	1936	420	177	24	6.7	3.1	7.6	67	34
NMS4-12-192	2118	426	178	25	6.8	3.0	7.6	73	34
NMS4-12-200	2131	416	188	25	6.8	3.1	8.2	73	33
NMS4-11-191-1	1934	430	163	25	6.8	3.0	7.3	66	38
NMS4-8-179	2026	433	174	26	6.9	2.4	8.2	69	29
NMS4-1-45	2684	408	201	26	6.9	2.1	10.3	91	20
NMS4-12-193-1	1991	427	173	26	6.9	2.8	7.1	67	31
NMS4-24-288	2119	419	184	26	6.9	2.9	8.0	72	30
NMS4-1-46	2714	406	204	26	7.0	2.3	10.6	92	20
NMS4-1-66	2700	411	200	27	7.0	2.1	10.4	91	19
NMS4-12-195-2	1849	433	170	28	7.2	2.3	9.2	61	34
NMS4-12-195-1	2161	433	179	29	7.3	2.5	8.0	71	33
$LSD_{0.05}$	356	12	9	3		0.5	0.8		

† NMS4 corresponds to lines in population JCD-1. The first number after the hyphen corresponds to F_3 plant.
‡ Estimated from regression of maturity date of conventional check against published maturity group.
§ 1 indicates all plants erect; 5 indicates all plants prostrate.
¶ Based on comparison of yield mean vs. regression of check means.
 # Based on 558 polymorphic SNP markers.

Table 1.2. Means of agronomic traits for breeding lines tested in yield test TCLP 382, averaged over 5 environments from 2008 to 2010.

Genotype†	Yield	Protein	Oil	Maturity	Est. MG‡	Lodging§	100-Seed Weight	Yield as % Check¶	% *G. soja* SNP alleles#
	kg ha^{-1}	g kg^{-1}	g kg^{-1}	Oct. 1 = 1			g		
5601T††	2241	413	195	17	5.6	1.5	14.9	-	-
NC-Roy	2725	412	193	26	6.9	2.0	13.3	-	-
NMS4-16-221	1653	442	145	20	6.0	3.2	6.6	69	33
NMS4-37-309	1505	422	181	20	6.0	3.7	8.0	63	35
NMS4-42-19	1854	403	188	20	6.1	3.0	9.0	77	33
NMS4-34-14	1518	428	177	20	6.1	3.7	8.0	63	35
NMS4-38-316	1922	413	177	20	6.1	2.3	8.8	79	34
NMS4-55-403	1767	427	171	20	6.1	2.9	8.2	73	38
NMS4-15-213	1418	435	168	21	6.2	3.6	9.0	57	34
NMS4-20-273	1828	442	164	21	6.2	3.7	7.8	74	35
NMS4-67-462	1875	434	166	22	6.3	2.7	8.8	75	34
NMS4-50-361-1	1505	406	183	22	6.4	3.7	6.6	59	32
NMS4-58-407	1653	406	185	23	6.4	2.6	8.0	65	33
NMS4-52-390	1942	438	162	23	6.5	2.8	7.7	76	32
NMS4-43-326	1807	416	181	23	6.5	3.5	7.6	70	38
NMS4-87-485	1713	414	175	23	6.5	1.9	7.3	66	35
NMS4-22-280	1975	420	188	23	6.5	2.7	9.3	76	35
NMS4-58-410	1680	407	184	24	6.6	2.8	8.2	64	32
$LSD_{0.05}$	269	11	7	3	-	0.5	0.7	-	-

† NMS4 corresponds to lines in population JCD-1. The first number after the hyphen corresponds to F_3 plant.
‡ Estimated from regression of maturity date of conventional check against published maturity group.
§ 1 indicates all plants erect; 5 indicates all plants prostrate.
¶ Based on comparison of yield mean vs. regression of check means.
 # Based on 558 polymorphic SNP markers.
†† Represents the mean of two separate entries of 5601T and NC-Roy respectively.

Table 1.2 cont.

Genotype†	Yield	Protein	Oil	Maturity	Est. MG‡	Lodging§	100-Seed Weight	Yield as % Check¶	% *G. soja* SNP alleles#
	kg ha^{-1}	g kg^{-1}	g kg^{-1}	Oct. 1 = 1			g		
NMS4-24-290	1707	433	175	24	6.6	3.0	8.0	65	31
NMS4-64-441	1928	436	164	24	6.6	3.1	6.4	74	34
NMS4-53-397	1498	432	148	24	6.6	3.0	6.8	57	38
NMS4-47-345	1774	426	155	24	6.6	3.6	7.2	68	39
NMS4-65-454	1848	429	163	24	6.6	3.1	7.8	71	30
NMS4-44-330	1801	419	182	24	6.6	3.1	8.3	68	37
NMS4-18-232	1686	438	166	24	6.7	3.1	9.9	64	35
NMS4-80-481	1720	430	164	25	6.7	3.3	6.8	65	33
NMS4-78-478	1485	430	162	25	6.7	3.2	7.6	56	33
NMS4-13-206	1344	437	163	25	6.8	3.4	9.3	50	31
NMS4-52-371	1841	437	165	26	6.9	3.2	7.2	68	33
NMS4-21-276	1975	435	172	27	7.1	3.0	7.6	71	35
NMS4-14-209	1431	427	174	28	7.1	3.3	7.2	51	43
NMS4-62-433	1895	420	173	28	7.2	2.6	8.7	67	34
NMS4-81-482	1888	407	187	29	7.3	3.3	8.0	66	33
NMS4-77-476-1	1922	442	175	29	7.3	2.7	9.2	67	32
NMS4-77-476	1875	434	171	29	7.3	3.2	7.9	65	33
NMS4-87-484	1868	434	172	30	7.4	3.1	7.6	64	30
$LSD_{0.05}$	269	11	7	3	-	0.5	0.7	-	-

† NMS4 corresponds to lines in population JCD-1. The first number after the hyphen corresponds to F_3 plant.
‡ Estimated from regression of maturity date of conventional check against published maturity group.
§ 1 indicates all plants erect; 5 indicates all plants prostrate.
¶ Based on comparison of yield mean vs. regression of check means.
 # Based on 558 polymorphic SNP markers.
†† Represents the mean of two separate entries of 5601T and NC-Roy respectively.

Table 1.3. Means of agronomic traits for breeding lines tested in yield test TCLP 383, averaged over 5 environments from 2008 to 2010

Genotype†	Yield	Protein	Oil	Maturity	Est. MG‡	Lodging§	100-Seed Weight	Yield as % Check¶	% *G. soja* SNP alleles#
	kg ha^{-1}	g kg^{-1}	g kg^{-1}	Oct. 1 = 1			g		
Holladay††	2815	381	209	16	5.3	1.4	14.6	-	-
5601T	2416	404	200	15	5.6	1.5	13.9	-	-
NC-Roy	2768	407	197	25	6.9	1.9	12.9	-	-
NC-Raleigh	2735	386	207	31	7.5	2.2	13.7	-	-
NMS4-97-502	1928	421	186	21	6.2	3.3	8.3	72	33
NMS4-110-559	1660	434	162	22	6.4	3.7	7.4	62	38
NMS4-152-704	1525	421	184	23	6.5	3.7	7.0	57	35
NMS4-111-569	1915	441	168	24	6.6	2.6	7.8	71	38
NMS4-103-520	1848	429	168	24	6.7	3.2	8.1	68	33
NMS4-146-693	1868	415	180	24	6.7	3.5	7.4	69	34
NMS4-123-651	2083	425	181	25	6.8	2.9	7.4	77	35
NMS4-124-653	1848	430	173	25	6.8	3.1	7.0	68	33
NMS4-101-515	1888	416	179	25	6.8	3.3	6.9	70	33
NMS4-97-504-1	1660	433	177	25	6.8	2.5	7.2	61	35
NMS4-127-657	2016	415	186.5	26	6.9	3.5	7.2	74	33
NMS4-144-653-1	1861	436	168	26	6.9	2.9	7.0	69	33
NMS4-99-505	1566	440	178	26	7.0	3.2	7.0	58	30
$LSD_{0.05}$	309	13	8	3	-	0.4	0.8	-	-

† NMS4 corresponds to lines in population JCD-1. The first number after the hyphen corresponds to F_3 plant.
‡ Estimated from regression of maturity date of conventional check against published maturity group.
§ 1 indicates all plants erect; 5 indicates all plants prostrate.
¶ Based on comparison of yield mean vs. regression of check means.
 # Based on 558 polymorphic SNP markers.
†† Represents the mean of two separate entries of Holladay, 5601T, NC-Roy, and NC-Raleigh respectively

Table 1.3 cont.

Genotype†	Yield	Protein	Oil	Maturity	Est. MG‡	Lodging§	100-Seed Weight	Yield as % Check¶	% *G. soja* SNP alleles#
	kg ha^{-1}	g kg^{-1}	g kg^{-1}	Oct. 1 = 1			g		
NMS4-102-517	1398	434	136	26	7.0	3.8	6.4	51	37
NMS4-108-550	1868	427	160	27	7.0	3.1	8.0	69	35
NMS4-112-594	1901	432	178	27	7.0	3.2	7.0	70	32
NMS4-175-709	2614	403	203	27	7.0	2.0	10.5	96	20
NMS4-152-703	1848	431	168	27	7.1	2.6	7.2	68	36
NMS4-115-600	1895	428	175	28	7.1	3.0	7.1	70	39
NMS4-1-83	2714	399	203	28	7.2	1.9	10.8	100	20
NMS4-1-65	2694	401	203	28	7.2	1.9	10.7	99	20
NMS4-1-77	2661	398	207	28	7.2	2.1	10.7	97	21
NMS4-79-20	1485	433	175	28	7.2	2.9	7.4	54	40
NMS4-169-706	2177	412	189	28	7.2	2.2	8.3	80	28
NMS4-1-73	2667	402	202	28	7.3	2.0	10.9	98	21
NMS4-44-329	2271	426	185	28	7.3	2.7	10.3	83	37
NMS4-8-146	2090	405	188	29	7.3	2.2	8.7	76	25
NMS4-48-352	1730	423	175	29	7.4	3.6	8.7	63	31
NMS4-107-543	1801	409	184	30	7.5	3.6	6.8	66	35
NMS4-8-178	2083	423	187	30	7.5	2.3	8.2	76	29
NMS4-51-365	1848	433	175	31	7.6	2.9	7.6	67	35
$LSD_{0.05}$	309	13	8	3	-	0.4	0.8	-	-

† NMS4 corresponds to lines in population JCD-1. The first number after the hyphen corresponds to F_3 plant.
‡ Estimated from regression of maturity date of conventional check against published maturity group.
§ 1 indicates all plants erect; 5 indicates all plants prostrate.
¶ Based on comparison of yield mean vs. regression of check means.
 # Based on 558 polymorphic SNP markers.
†† Represents the mean of two separate entries of Holladay, 5601T, NC-Roy, and NC-Raleigh respectively

Table 1.4. Means of agronomic traits for breeding lines tested in yield test TCLP 405A, averaged over 3 environments from 2009 to 2010.

Genotype†	Yield	Protein	Oil	Maturity	Est. MG‡	Lodging§	100-Seed Weight	Yield as % Check¶	% *G. soja* SNP alleles#
	kg ha^{-1}	g kg^{-1}	g kg^{-1}	Oct. 1 = 1			g		
5601T††	2822	413	198	14	5.6	1.8	14.1	-	-
NC-Roy	2787	411	193	24	6.9	2.2	12.5	-	-
N7103	2451	404	193	28	7.5	1.6	7.6	-	-
NC-Raleigh	2894	391	209	30	7.5	2.2	14.3	-	-
NMS5-156-3-341	1857	464	142	16	5.7	2.4	7.3	66	36
NMS5-145-2-319	1502	455	162	16	5.8	3.2	7.1	53	40
NMS5-145-1-318	1464	435	175	18	6.0	3.2	7.8	51	37
NMS5-252-3-536	2076	422	172	19	6.2	2.6	8.4	73	35
NMS5-269-3-573	1806	441	156	19	6.3	2.8	5.1	63	33
NMS5-274-3-580	2067	457	155	20	6.4	2.6	6.1	72	20
NMS5-269-2-572	1893	445	157	22	6.6	2.9	5.7	66	33
$LSD_{0.05}$	323	11	6	4	-	0.5	0.8	-	-

† NMS5 corresponds to lines in population JCD-2. The first number after the hyphen corresponds to F_3 plant.
‡ Estimated from regression of maturity date of conventional check against published maturity group.
§ 1 indicates all plants erect; 5 indicates all plants prostrate.
¶ Based on comparison of yield mean vs. regression of check means.
 # Based on 558 polymorphic SNP markers.
†† Represents the mean of three separate entries of 5601T, NC-Roy, and N7103. NC-Raleigh was duplicated twice.

Table 1.4 cont.

Genotype†	Yield	Protein	Oil	Maturity	Est. MG‡	Lodging§	100-Seed Weight	Yield as % Check¶	% *G. soja* SNP alleles#
	kg ha^{-1}	g kg^{-1}	g kg^{-1}	Oct. 1 = 1			g		
NMS5-214-1-464	1834	423	172	23	6.7	3.3	8.4	64	43
NMS5-267-2-563	1277	440	155	24	6.9	2.9	5.5	44	32
NMS5-217-2-478	1836	444	156	24	6.9	2.6	7.1	64	35
NMS5-216-5-476	1684	454	152	25	7.0	2.6	7.5	58	35
NMS5-96-2-182	1657	466	151	27	7.3	3.1	10.2	57	37
NMS5-283-1-602	1318	469	144	31	7.7	2.9	5.7	45	22
$LSD_{0.05}$	323	11	6	4	-	0.5	0.8	-	-

† NMS5 corresponds to lines in population JCD-2. The first number after the hyphen corresponds to F_3 plant.
‡ Estimated from regression of maturity date of conventional check against published maturity group.
§ 1 indicates all plants erect; 5 indicates all plants prostrate.
¶ Based on comparison of yield mean vs. regression of check means.
 # Based on 558 polymorphic SNP markers.
†† Represents the mean of three separate entries of 5601T, NC-Roy, and N7103. NC-Raleigh was duplicated twice.

Table 1.5. Means of agronomic traits for breeding lines tested in yield test TCLP 406A, averaged over 3 environments from 2009 to 2010.

Genotype†	Yield	Protein	Oil	Maturity	Est. MG‡	Lodging§	100-Seed Weight	Yield as % Check¶	% *G. soja* SNP alleles#
	kg ha^{-1}	g kg^{-1}	g kg^{-1}	Oct. 1 = 1			g		
5601T††	2364	408	201	14	5.6	1.7	13.7	-	-
NC-Roy	2517	406	197	24	6.9	2.2	12.2	-	-
N7103	2394	398	195	29	7.5	1.5	7.1	-	-
NC-Raleigh	2516	386	210	31	7.5	2.2	13.9	-	-
NMS5-48-2-75	1694	440	159	19	6.2	3.3	9.6	70	37
NMS5-48-3-76	1856	420	176	20	6.4	3.5	9.8	77	37
NMS5-48-6-79	1878	434	155	20	6.4	3.0	10.0	77	38
NMS5-208-4-448	1770	426	180	21	6.5	2.7	6.2	73	44
NMS5-85-1-146	1695	431	179	21	6.5	3.1	8.7	70	38
NMS5-255-6-548	1680	404	189	22	6.6	3.2	5.4	69	38
NMS5-266-2-559	1514	419	145	22	6.6	3.1	5.9	62	33
NMS5-282-2-600	1781	428	167	23	6.7	1.9	4.8	73	18
$LSD_{0.05}$	316	13	9	3	-	0.4	1.0	-	-

† NMS5 corresponds to lines in population JCD-2. The first number after the hyphen corresponds to F_3 plant.
‡ Estimated from regression of maturity date of conventional check against published maturity group.
§ 1 indicates all plants erect; 5 indicates all plants prostrate.
¶ Based on comparison of yield mean vs. regression of check means.
 # Based on 558 polymorphic SNP markers.
†† Represents the mean of two seperate entries of 5601T and NC-Raleigh. NC-Roy and N7103 were replicated three times respectively.

Table 1.5 cont.

Genotype†	Yield	Protein	Oil	Maturity	Est. MG‡	Lodging§	100-Seed Weight	Yield as % Check¶	% *G. soja* SNP alleles#
	kg ha^{-1}	g kg^{-1}	g kg^{-1}	Oct. 1 = 1			g		
NMS5-241-1-522	1684	411	162	24	6.8	2.9	6.3	69	43
NMS5-255-5-547	2059	402	190	24	6.8	2.6	6.6	84	36
NMS5-246-2-528	1520	407	182	24	6.9	2.7	5.7	62	29
NMS5-101-5-206	1926	426	177	24	6.9	2.6	8.8	78	37
NMS5-188-1-406	2045	425	178	24	6.9	2.7	9.8	83	38
NMS5-282-1-599	1630	418	171	26	7.1	2.4	4.9	66	18
NMS5-253-1-537	2155	413	181	28	7.3	2.5	8.5	86	40
$LSD_{0.05}$	316	13	9	3	-	0.4	1.0	-	-

† NMS5 corresponds to lines in population JCD-2. The first number after the hyphen corresponds to F_3 plant.
‡ Estimated from regression of maturity date of conventional check against published maturity group.
§ 1 indicates all plants erect; 5 indicates all plants prostrate.
¶ Based on comparison of yield mean vs. regression of check means.
 # Based on 558 polymorphic SNP markers.
†† Represents the mean of two seperate entries of 5601T and NC-Raleigh. NC-Roy and N7103 were replicated three times respectively.

Table 1.6. Means of agronomic traits for breeding lines tested in yield test TCLP 407A, averaged over 3 environments from 2009 to 2010.

Genotype†	Yield	Protein	Oil	Maturity	Est. MG‡	Lodging§	100-Seed Weight	Yield as % Check¶	% *G. soja* SNP alleles#
	kg ha^{-1}	g kg^{-1}	g kg^{-1}	Oct. 1 = 1			g		
5601T††	2570	411	198	15	5.6	1.7	14.4	-	-
NC-Roy	2773	417	191	25	6.9	2.1	12.6	-	-
N7103	2638	407	191	30	7.5	1.5	7.7	-	-
NC-Raleigh	2822	390	209	31	7.5	2.2	14.1	-	-
NMS5-144-3-317	1615	435	153	17	5.9	3.6	6.9	62	45
NMS5-20-7-32	1959	419	181	20	6.2	3.2	9.0	73	39
NMS5-98-1-190	2157	405	188	20	6.3	3.0	9.7	81	37
NMS5-90-2-161	1996	420	163	21	6.3	2.3	9.0	74	36
NMS5-109-1-229	2135	442	161	22	6.4	2.6	9.2	79	31
NMS5-75-5-134	1650	414	187	22	6.5	3.3	10.5	61	32
NMS5-179-1-391	1877	419	174	23	6.6	2.2	9.6	69	35
NMS5-85-2-147	2330	456	152	23	6.6	2.9	8.5	86	31
NMS5-100-3-199	1822	418	178	23	6.7	2.8	8.7	67	33
$LSD_{0.05}$	363	10	8	3	-	0.5	0.8	-	-

† NMS5 corresponds to lines in population JCD-2. The first number after the hyphen corresponds to F_3 plant.
‡ Estimated from regression of maturity date of conventional check against published maturity group.
§ 1 indicates all plants erect; 5 indicates all plants prostrate.
¶ Based on comparison of yield mean vs. regression of check means.
 # Based on 558 polymorphic SNP markers.
†† Represents the mean of two separate entries of 5601T, N7103, and NC-Raleigh. NC-Roy was entered three separate times.

Table 1.6 cont.

Genotype†	Yield	Protein	Oil	Maturity	Est. MG‡	Lodging§	100-Seed Weight	Yield as % Check¶	% *G. soja* SNP alleles#
	kg ha^{-1}	g kg^{-1}	g kg^{-1}	Oct. 1 = 1			g		
NMS5-62-9-116	2065	447	154	24	6.8	2.7	8.2	75	30
NMS5-121-2-257	2097	423	185	24	6.8	2.9	7.9	77	30
NMS5-98-4-193	2028	410	178	24	6.8	3.2	10.5	74	35
NMS5-109-2-230	2193	443	161	25	6.8	2.8	9.9	80	30
NMS5-115-3-246	1876	428	164	25	6.9	2.6	8.7	68	38
NMS5-68-4-122	1939	415	187	26	7.0	2.2	9.6	70	38
NMS5-225-6-496	1696	417	166	26	7.0	3.2	7.6	61	38
NMS5-51-1-80	1840	436	150	27	7.1	2.9	9.2	66	47
NMS5-139-1-304	1518	430	151	27	7.2	2.7	8.0	54	34
NMS5-139-2-305	1648	444	144	28	7.3	2.4	7.6	59	33
NMS5-68-4-122	1939	415	187	26	7.0	2.2	9.6	70	38
$LSD_{0.05}$	363	10	8	3	-	0.5	0.8	-	-

† NMS5 corresponds to lines in population JCD-2. The first number after the hyphen corresponds to F_3 plant.
‡ Estimated from regression of maturity date of conventional check against published maturity group.
§ 1 indicates all plants erect; 5 indicates all plants prostrate.
¶ Based on comparison of yield mean vs. regression of check means.
 # Based on 558 polymorphic SNP markers.
†† Represents the mean of two separate entries of 5601T, N7103, and NC-Raleigh. NC-Roy was entered three separate times.

Table 1.7. Means of agronomic traits for breeding lines tested in yield test TCLP 408A, averaged over 3 environments from 2009 to 2010.

Genotype†	Yield	Protein	Oil	Maturity	Est. MG‡	Lodging§	100-Seed Weight	Yield as % Check¶	% *G. soja* SNP alleles#
	kg ha^{-1}	g kg^{-1}	g kg^{-1}	Oct. 1 = 1			g		
5601T††	2450	414	199	14	5.6	1.8	13.6	-	-
NC-Roy	2585	419	189	25	6.9	1.9	12.5	-	-
N7103	2553	408	189	29	7.5	1.4	7.7	-	-
NC-Raleigh	2679	386	207	30	7.5	2.0	13.9	-	-
NMS5-11-1-4	2048	424	187	27	7.2	2.1	10.0	78	40
NMS5-11-4-7	1760	431	183	25	6.9	2.1	9.3	68	40
NMS5-12-1-9	1975	405	175	26	7.1	2.4	10.5	75	38
NMS5-17-3-22	2070	432	180	27	7.1	2.7	10.5	79	34
NMS5-20-5-30	1961	408	177	22	6.6	3.1	8.6	77	31
NMS5-32-2-50	1896	410	187	23	6.7	2.5	11.4	74	31
NMS5-32-5-53	1711	424	185	22	6.6	2.6	11.0	67	30
NMS5-37-10-70	2104	423	171	28	7.3	2.1	10.0	80	32
NMS5-70-1-124	1863	439	183	28	7.3	2.6	8.2	70	28
$LSD_{0.05}$	417	15	9	3	-	0.4	0.7	-	-

† NMS5 corresponds to lines in population JCD-2. The first number after the hyphen corresponds to F_3 plant.
‡ Estimated from regression of maturity date of conventional check against published maturity group.
§ 1 indicates all plants erect; 5 indicates all plants prostrate.
¶ Based on comparison of yield mean vs. regression of check means.
 # Based on 558 polymorphic SNP markers.
†† Represents the mean of two separate entries of 5601T, N7103, and NC-Raleigh. NC-Roy was entered three separate times.

Table 1.7 cont.

Genotype†	Yield	Protein	Oil	Maturity	Est. MG‡	Lodging§	100-Seed Weight	Yield as % Check¶	% *G. soja* SNP alleles#
	kg ha^{-1}	g kg^{-1}	g kg^{-1}	Oct. 1 = 1			g		
NMS5-75-2-131	1662	417	184	26	7.1	2.5	10.5	63	31
NMS5-88-2-154	1875	448	168	26	7.1	2.7	9.6	72	40
NMS5-100-5-201	1834	420	174	24	6.8	2.2	9.7	71	33
NMS5-109-3-231	2067	450	157	24	6.8	2.6	9.1	80	28
NMS5-172-3-374	1238	446	128	25	6.9	3.2	6.7	48	37
NMS5-208-2-446	1691	411	141	27	7.2	3.3	5.3	64	47
NMS5-268-3-570	1819	435	160	28	7.3	2.4	6.1	69	45
NMS5-99-2-196	1703	408	172	19	6.3	2.3	7.5	68	42
$LSD_{0.05}$	417	15	9	3	-	0.4	0.7	-	-

† NMS5 corresponds to lines in population JCD-2. The first number after the hyphen corresponds to F_3 plant.
‡ Estimated from regression of maturity date of conventional check against published maturity group.
§ 1 indicates all plants erect; 5 indicates all plants prostrate.
¶ Based on comparison of yield mean vs. regression of check means.
 # Based on 558 polymorphic SNP markers.
†† Represents the mean of two separate entries of 5601T, N7103, and NC-Raleigh. NC-Roy was entered three separate times.

Table 1.8. Means of agronomic traits for breeding lines tested in yield test TCLP 409A, averaged over 3 environments from 2009 to 2010.

Genotype†	Yield	Protein	Oil	Maturity	Est. MG‡	Lodging§	100-Seed Weight	Yield as % Check¶	% *G. soja* SNP alleles#
	kg ha^{-1}	g kg^{-1}	g kg^{-1}	Oct. 1 = 1			g		
5601T††	2217	421	196	13	5.6	1.6	13.4	-	-
NC-Roy	2467	416	190	25	6.9	1.8	12.4	-	-
N7103	2084	419	186	30	7.5	1.4	7.9	-	-
NC-Raleigh	2669	390	206	30	7.5	1.9	13.9	-	-
NMS5-60-2-106	1541	435	165	22	6.6	2.2	9.2	63	28
NMS5-11-3-6	1638	418	186	22	6.6	2.3	9.6	67	37
NMS5-55-2-91	1530	438	156	22	6.6	2.8	9.6	63	31
NMS5-249-4-531	1479	440	151	22	6.7	2.3	6.7	60	34
NMS5-20-6-31	1627	427	168	23	6.7	2.7	9.2	66	37
NMS5-111-6-240	2136	431	177	24	6.8	2.4	9.2	86	32
NMS5-55-7-96	1608	453	149	24	6.9	2.4	9.3	64	33
NMS5-130-2-275	1709	423	159	25	6.9	2.7	7.9	68	36
NMS5-101-4-205	1637	442	171	25	7.0	2.1	9.3	65	35
NMS5-90-6-165	1828	426	159	25	7.0	2.7	8.9	72	36
$LSD_{0.05}$	477	13	12	3	-	0.5	1.0	-	-

† NMS5 corresponds to lines in population JCD-2. The first number after the hyphen corresponds to F_3 plant.
‡ Estimated from regression of maturity date of conventional check against published maturity group.
§ 1 indicates all plants erect; 5 indicates all plants prostrate.
¶ Based on comparison of yield mean vs. regression of check means.
 # Based on 558 polymorphic SNP markers.
†† Represents the mean of two duplicate entries of 5601T, NC-Roy, N7103, and NC-Raleigh

Table 1.8 cont.

Genotype†	Yield	Protein	Oil	Maturity	Est. MG‡	Lodging§	100-Seed Weight	Yield as % Check¶	% *G. soja* SNP alleles#
	kg ha^{-1}	g kg^{-1}	g kg^{-1}	Oct. 1 = 1			g		
NMS5-80-5-145	1651	453	169	27	7.2	3.3	10.0	64	38
NMS5-102-1-207	1936	444	137	27	7.2	3.1	8.4	75	34
NMS5-101-3-204	1675	446	166	27	7.2	2.3	8.2	65	37
NMS5-98-3-192	2131	420	174	27	7.2	2.4	10.8	83	36
NMS5-300-2-634	1768	406	177	28	7.3	2.6	4.7	68	26
NMS5-30-3-47	1899	446	162	29	7.4	2.2	11.7	72	37
NMS5-130-1-274	1538	427	155	29	7.5	3.3	7.7	58	32
$LSD_{0.05}$	477	13	12	3	-	0.5	1.0	-	-

† NMS5 corresponds to lines in population JCD-2. The first number after the hyphen corresponds to F_3 plant.
‡ Estimated from regression of maturity date of conventional check against published maturity group.
§ 1 indicates all plants erect; 5 indicates all plants prostrate.
¶ Based on comparison of yield mean vs. regression of check means.
 # Based on 558 polymorphic SNP markers.
†† Represents the mean of two duplicate entries of 5601T, NC-Roy, N7103, and NC-Raleigh.

Table 1.9. Means of agronomic traits for breeding lines tested in yield test TCLP 410A, averaged over 3 environments from 2009 to 2010.

Genotype†	Yield	Protein	Oil	Maturity	Est. MG‡	Lodging§	100-Seed Weight	Yield as % Check¶	% *G. soja* SNP alleles#
	kg ha^{-1}	g kg^{-1}	g kg^{-1}	Oct. 1 = 1			g		
5601T††	2356	409	202	14	5.6	1.7	13.5	.	.
NC-Roy	2725	408	197	24	6.9	2.0	12.3	.	.
N7103	2391	408	193	29	7.5	1.4	7.5	.	.
NC-Raleigh	2693	387	209	30	7.5	1.9	13.7	.	.
NMS5-128-2-269	1772	428	167	16	5.8	3.1	7.4	73	40
NMS5-108-2-226	1752	428	179	17	5.9	2.7	9.2	72	37
NMS5-294-6-629	1684	424	164	18	6.1	2.3	5.1	68	33
NMS5-112-1-241	2284	400	190	19	6.2	1.8	7.6	91	28
NMS5-112-2-242	2137	403	192	19	6.2	2.1	8.5	85	28
NMS5-294-5-628	1854	423	168	19	6.2	2.6	5.5	74	35
NMS5-115-4-247	1887	447	158	21	6.5	1.8	8.5	74	36
NMS5-112-4-243	2243	410	182	21	6.5	1.9	7.1	87	25
NMS5-294-4-627	1742	423	159	22	6.6	2.6	5.3	67	33
NMS5-276-3-586	1972	417	186	23	6.6	2.1	5.2	76	14
$LSD_{0.05}$	477	13	12	3	-	0.5	1.0	-	-

† NMS5 corresponds to lines in population JCD-2. The first number after the hyphen corresponds to F_3 plant.
‡ Estimated from regression of maturity date of conventional check against published maturity group.
§ 1 indicates all plants erect; 5 indicates all plants prostrate.
¶ Based on comparison of yield mean vs. regression of check means.
 # Based on 558 polymorphic SNP markers.
†† Represents the mean of two duplicate entries of 5601T, NC-Roy, and N7103. NC-Raleigh was entered only once.

FIGURES

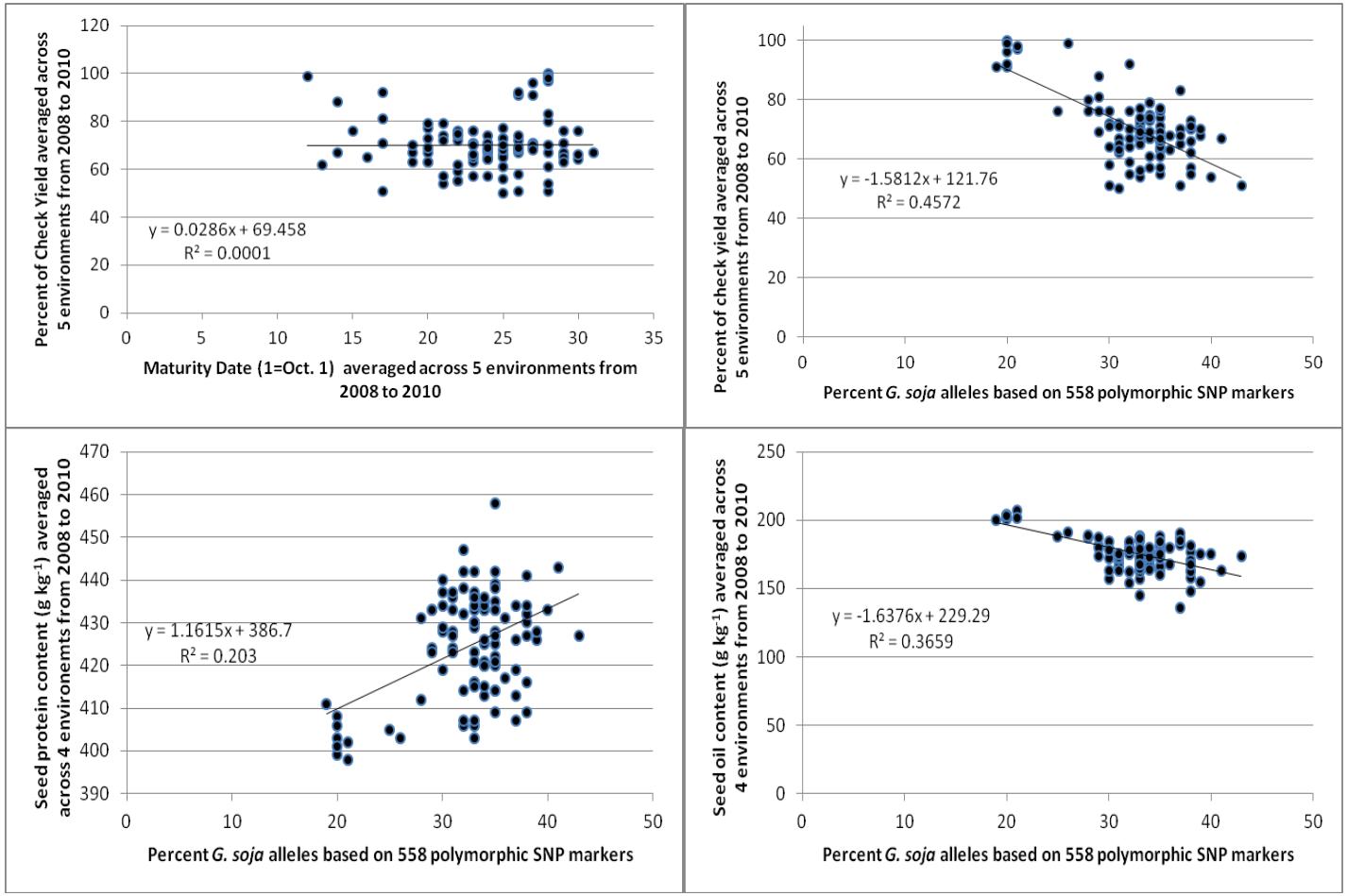


Fig. 1.1 (Top to bottom, left to right) Regression percent check yield versus maturity, and percent *G. soja* alleles versus percent check yield, protein content, and oil content, for breeding lines from population JCD-1.

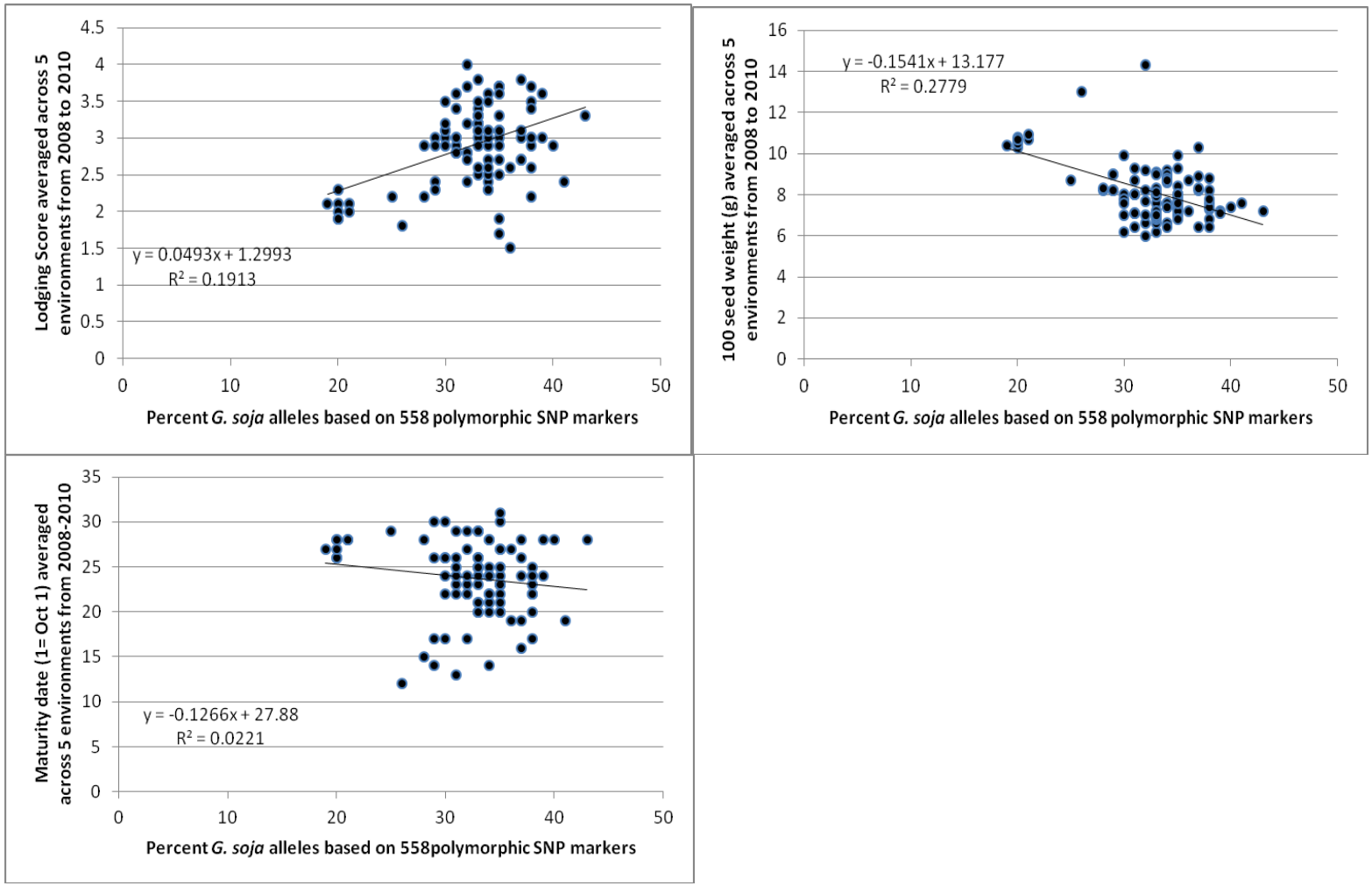


Fig. 1.2 (Top to bottom, left to right) Regression of percent *G. soja* alleles versus lodging score, seed size, and maturity date for breeding lines from population JCD-1.

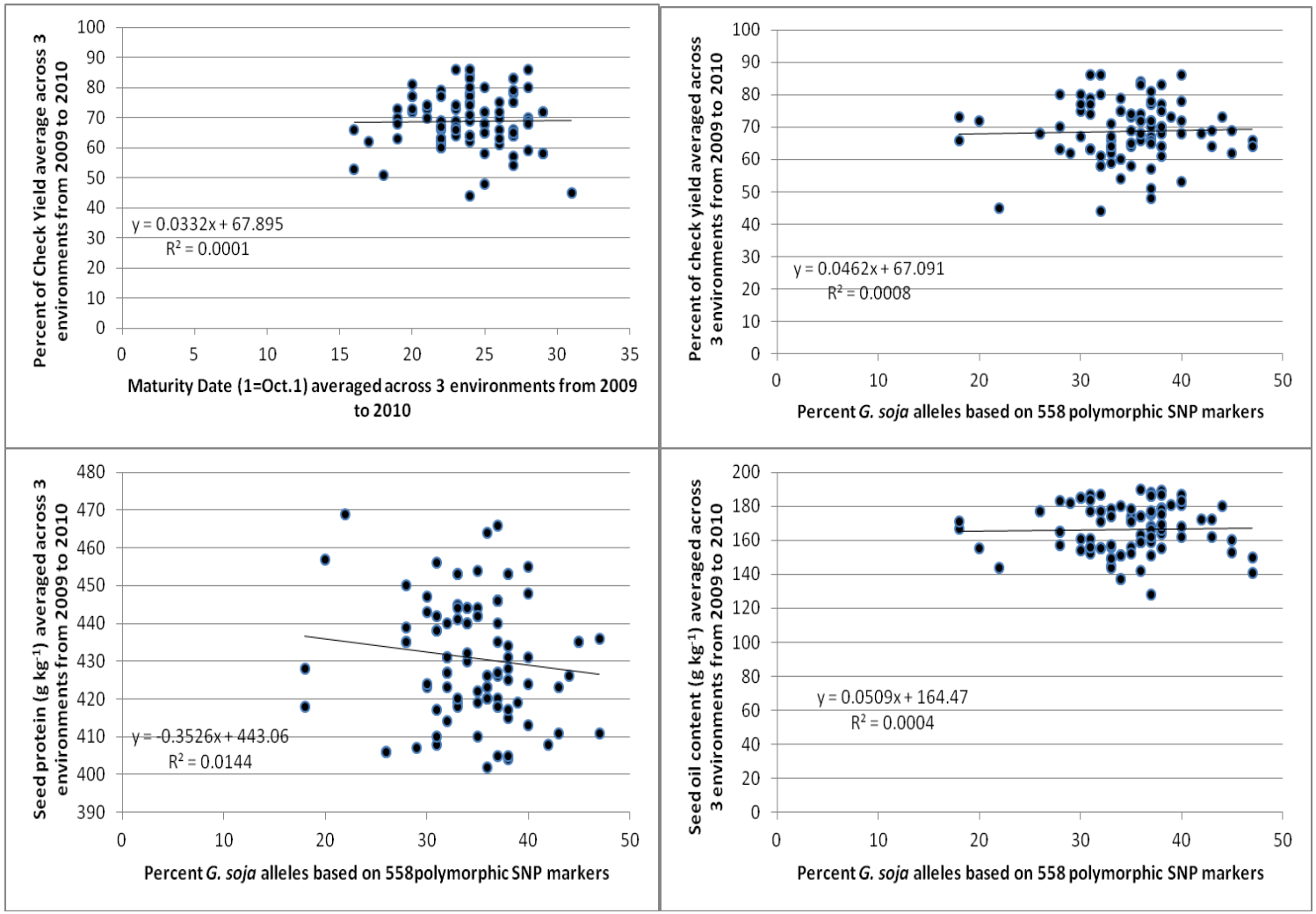


Fig 1.3. (Top to bottom, left to right) Regression percent check yield versus maturity, and percent *G. soja* alleles versus percent check yield, protein content, and oil content, for breeding lines from population JCD-2.

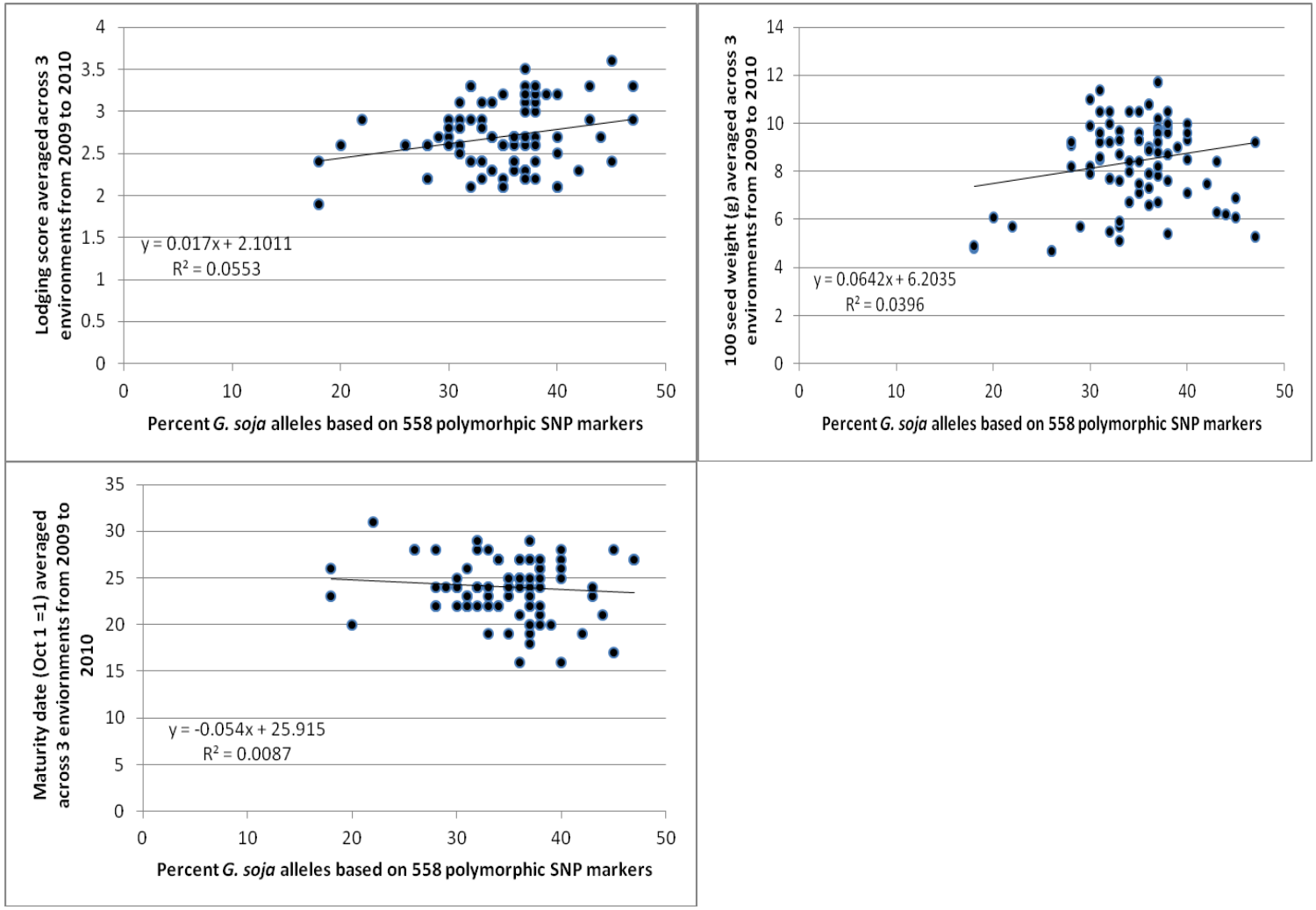


Fig. 1.4 (Top to bottom, left to right) Regression of percent *G. soja* alleles versus lodging score, seed size, and maturity date for breeding lines from population JCD-2.

CHAPTER 2

GENETIC DIVERSITY PATTERNS IN SINGLE CROSS *GLYCINE MAX* BY *GLYCINE SOJA* POPULATIONS

ABSTRACT

One relatively unused source of genetic diveristy is the wild soybean, *Glycine soja* (Subb. And Zucc.). Its chief limitation in soybean breeding is the poor agronomic performance of progeny, when used as parent. Recently, this barrier has been overcome to produce a large number of agronomic breeding lines from the F_1 of domesticated x wild soybean hybridization, using mega-populations plus intense visual selection. The objective of the present study was to shed light on the underlying basis for this accomplishment. Two replicate populations (120 breeding lines in total) were assayed for 558 polymorphic SNP markers and segregation, simple matching coefficients (SMC) were calculated, and the resulting matrix was subjected to multidimensional scaling analysis (MDS). Segregation distortion combined with single marker analysis of variance and multiple regression were used to identify chromosomal regions and alleles from wild soybean that were associated with agronomic traits. The MDS analysis revealed that selection has placed these lines closer to the domesticated soybean [*Glycine max* (L.) Merr.] parent, while creating lines with an average SMC relationship of 0.63, as opposed to the typical relationship of 0.5, based on pedigree probabilities. The single marker analysis of variance and stepwise regression identified three genomic regions in which the *G. soja* SNP allele was associated with higher yields as percent of check, as well as four genomic regions associated with larger 100-seed weight. One of these regions had a significant effect on both seed yield and 100 seed weight,

but the other four regions were idependent of each other. The effects of the genomic regions associated with seed yield ranged from a six to seven percent increase, while the regions associated with 100-seed weight averaged a one g increase. We also identified 16 breeding lines that, in the aggregate, contained at least one homozygous *G. soja* allele at all 558 polymorphic SNP loci. It appears that this novel mega-population method may be critically important in soybean breeding as a means of overcoming the genetic bottleneck created by the domestication of soybean, offering breeders access to diversity to a degree never thought possible.

INTRODUCTION

Genetic diversity is an important aspect of plant improvement and can be quantified through various methods. In soybean, Carter et al. (1993) used pedigree information to show that only 14 ancestral lines contribute 70% of the ancestry for 258 cultivars released from 1945 to 1988. Gizlice et al. (1996) used coefficient of parentage estimates combined with multidimensional scaling (MDS) to characterize diversity patterns in these 258, revealing nine major clusters of soybean cultivars which developed through decades of applied breeding. Employing DNA marker technology, Thompson et al. (1998) used 281 random amplified polymorphic DNA (RAPD) markers in combination with MDS to characterize diversity found in 18 soybean ancestral lines as well as 17 plant introductions (PIs). Brown-Guedira et al., (2000) further evaluated these 18 ancestral types in relation to 87 PIs that were used in the USDA germplasm enhancement program at Urbana, IL, using a combination of simple sequence repeat (SSR) and RAPD markers and MDS analysis. They were able to identify 11 distinct clusters in the exotic materials.

More recently, Hyten et al. (2008) demonstrated the use of high throughput single nucleotide polymorphisms (SNP) in genotyping soybean and created a universal soybean linkage panel which employs 1536 SNP markers (USLP 1.0) (Hyten et al., 2010). This SNP high throughput assay allows for excellent genome coverage, optimizes minor allele frequencies for both adapted and exotic germplasm, and allows for easy, automated scoring. This assay, along with one containing 50000 SNP markers, is now widely used for QTL discovery and mapping in soybean populations and for quantifying patterns of genetic diversity in soybean.

One potentially important source of allelic diversity for soybean breeders is the progenitor species, *G.soja*. Hyten et al. (2006) demonstrated that the greatest loss of diversity in the US soybean germplasm pool occurred during the domestication of soybean from its wild progenitor. Currently, wild soybean does not appear in the pedigree of North American commodity-type cultivars, indicating the difficulty in using this species in applied breeding. The problem is that the wild soybean is extremely procumbent or lodging prone, as are the vast majority of progeny derived from it. To overcome these issues, breeders have historically implemented backcrossing procedures in order to recover a sufficient number of agronomically acceptable segregants for applied selection. Delheimer (2012) recently demonstrated that it is possible to develop agronomically acceptable soybean breeding lines from a single cross of *G. max* and *G. soja* accessions N7103 (Carter et al., 2003) x PI 366122 (USDA-GRIN Database), when extremely large population sizes and intense selection are employed. They identified several breeding lines that yield 90% of the *G. max* checks, and still retain 25% *G. soja* parentage based on analysis of 558 polymorphic SNPs. However, the

study did not quantify diversity present within those lines. The 192 agronomic breeding lines developed by Delheimer (2012) from the *G. max* x *G. soja* single cross offer a unique perspective from which to dissect the extensive introgression of *G. soja* diversity into a *G. max* background. Because this set of advanced breeding lines was developed from independent replicate breeding populations, they offer a unique opportunity to validate the efficacy of introducing genetic diversity into cultivated soybean using the single cross, mega-population method outlined by Delheimer (2012). The objectives of this study were to (i) characterize the diversity patterns present within the two *G. max* X *G.soja* populations studied by Delheimer (2012) using 558 polymorphic SNP markers and MDS analysis, to (ii) identify possible regions of useful diversity derived from *G. soja*, and to (iii) create a collection of breeding lines that capture the complete array diversity identified *via* SNP marker analysis.

MATERIALS AND METHODS

The genotypes used in this study were the breeding lines developed from a single cross of *G.max* by *G. soja* evaluated by Delheimer (2012). A preliminary analysis of all 193 breeding lines in their study, revealed an average simple matching coefficient (SMC) of 0.95 between lines developed from the same F_3 plant, as well as a few individual breeding lines developed from different F_3 plants. Interpretive bias that can be introduced from making comparisons based on essentially duplicate breeding lines, and, thus, a single breeding line from each F_3 family was selected to represent that family in the overall analysis. The

truncated dataset consisted of 120 individuals (61 from population JCD-1, 59 from population JCD-2) from the 192 studied by Delheimer (2012) and was used for all analyses reported here.

DNA EXTRACTION AND SNP ANALYSIS

DNA samples were obtained from all breeding lines in both populations as well as from the parents, N7103 and PI 366122. DNA was isolated by taking the root tip from 20 seedlings and combining them into one sample for each line. The Qiagen Plant Easy DNA Extraction Kit (Qiagen, Hilden, Germany) was used to obtain purified DNA. The 1,536 SNP markers were tested on N7103, PI 36122, and the 192 breeding lines using the GoldenGate assay and analyzed on the Illumina BeadStation 500G (Illumina, San Diego, CA) as described previously (Hyten et al., 2008). The automatic allele calling for each locus is accomplished with the GenomeStudio software (Illumina, San Diego, CA). The allele calling is further checked manually, and only SNPs that were polymorphic between the parents were used in subsequent analysis.

GENETIC DISTANCE AND MULTIVARIATE ANALYSIS

Simple matching coefficients (SMC) (Sokal and Michener 1958) were obtained for each breeding line, based on polymorphic markers, using the DISTANCE procedure in SAS (SAS Institute, 2007), employing the MATCH method. A simple matching coefficient is defined as the total number of matching allele calls divided by the sum of the number of matches and mismatches. Heterozygous loci were eliminated from the SMC calculation as they lead to underestimation of simple matching coefficient relationship. The output matrix from the DISTANCE procedure was then exported to the MDS procedure of SAS. The

ABSOLUTE option was specified to preserve the distance scale between 0 and 1 for ease of interpretation and graphing. The MDS procedure creates an initial multidimensional configuration for which the eigen value of each dimension is equal to the sum of variance which that dimension explains. After the initial step, an iterative method is employed to find a final set of coordinates in Euclidean space whose interpoint distances best match the distances of the original data matrix on average, based on a user specified number of dimensions. Separate analyses must be run for each number of dimensions in order to determine which number of dimensions best fits the model. Analyses were run on the data until the distance correlation between the fitted data and the original data reached 0.95, which occurred with a model of 9 dimensions, indicating that 95% of the variation in the 120 x 120 matrix of matching coefficients could be captured in the 22 dimensions.

GRAPHICAL DEPICTION OF MDS RESULTS

Difficulty in interpretation can arise in graphically representing 9 dimensions, so the MDS procedure was also employed using only 3 dimensions and the results plotted. This model had a distance correlation of 0.78, meaning that 78% of the variation in the original matrix was captured in these three dimensions. An initial three dimensional graph placed N7103 within a large cluster of breeding lines and PI 366122 outside this cluster, but it did not accurately portray the fact that the parents are, by definition and by empirical calculation of matching coefficients, the maximally diverse genotypes in this study. To display this important fact graphically in plots derived from MDS, we weighted the parents more heavily than the progeny by replicating parental SNP allelic data 50 times within the original data file. This weighting procedure produced figures that were easily interpretable (Fig 2.1). To

test the validity of this weighting approach (i.e. that the resulting graph was not an artifact of methodology), 20 random pairs of breeding lines were chosen and treated in sequential analyses in the same manner as the parents (i.e. no replication vs. 50 replications) in MDS analyses. These weightings of random progeny, unlike the weighting of the parents, had no appreciable effect on the resulting graphs. Thus the final graph used in this study included the weighting of the parents as given above. This weighted version had a distance correlation of 0.97, meaning that 97% of the original variation was captured.

SEGREGATION DISTORTION IN HIGHLY SELECTED BREEDING LINES

The frequency of wild vs. domesticated alleles was determined at each SNP locus as an average over the 120 lines in this study. Segregation distortion was defined as deviation from the 0.5 frequency that would be expected as a result of random segregation. Segregation distortion at each SNP locus was tested for statistical significance using a normal approximation of the binomial distribution (Snedecor and Cochran, 1989). Results were presented graphically in Excel for the two individual populations in and the overall mean, with all SNP loci positioned using map positions reported in USLP 1.0 (Hyten et al., 2010). Accurate map positioning was not possible using the data directly from the present study, because of intense selection during the inbreeding process.

Multiple comparisons in a single dataset lead to risks of multiple Type I errors. To declare significant segregation distortion, all 558 individual tests were compared at an alpha level of 0.01. The probability that at least one Type I error occurring among these 558 tests would be 0.99 (Lander and Bostein 1989). However, this probability decreases to 0.42 with

an individual alpha level of 0.001. The majority of the significant differences from the test of segregation distortions were still significant at this alpha level.

SINGLE MARKER ANOVA

Each of the 558 SNP markers was subjected to single factor analysis of variance with the GLM procedure in SAS. The analysis was looking for differences between the genotypic classes based on phenotypic data reported by Delheimer (2012). The traits evaluated were: lodging, 100-seed weight, relative maturity, seed oil content, seed protein content, seed yield, flower color, and pubescence color. Agronomic data were collected on Population JCD-1 over five environments from 2008-2010, while data were collected on population JCD-2 over three environments from 2009-2010 (Delheimer 2012). Comparison of common check cultivars used in both populations revealed only minor overall differences in check means between the two populations and contrasting maturity groups for all traits other than seed yield (Delheimer 2012). Higher seed yields were seen for later maturing breeding lines. To minimize maturity bias, seed yield was re-calculated as a percent of an appropriate check for each breeding line, and a single combined dataset was created which included all breeding lines from both populations. Maturity effects were addressed in the following way. Maturity dates of all checks were regressed against published relative maturity, and the regression curve was used to estimate the relative maturities of all breeding lines. This allowed for easy calculation of yield as a percentage of *G. max* check cultivar performance, regardless of the maturity of the breeding line. No extrapolation was made beyond the range of the check cultivars. Significant differences in percent of check were declared at an alpha level of 0.05

due to the low heritability of the trait. All other traits were compared to an alpha level of 0.01.

MULTIPLE LINEAR REGRESSION

SNP markers identified with significant positive associations derived from the *G. soja* parent were further analyzed with multiple linear regression in order to prioritize further analysis of genomic regions. A correlation analysis was done on allele calls of all markers declared significant for a trait of interest. Forward, backward, and stepwise regression analysis was performed on uncorrelated markers to determine which model explains the most variation in the trait of interest (Rawling et al., 1998). These tests were conducted using the REG procedure in SAS. A significance level of 0.40 was set as the criteria for a SNP to enter the model, while a significance level of 0.15 was used for a SNP to remain in the model.

CAPTURING *G. SOJA* DIVERSITY IN A SUBSET OF LINES

The SNP assay revealed a total of 558 polymorphisms between N7103 and *G. soja* PI 366122. It would be advantageous, in terms of storage and logisitical concerns in further research, to capture all of this variation in a small of a number of breeding lines for applied breeding efforts. Therefore, the allelic data were manipulated manually in EXCEL order to identify the approximate smallest number of breeding lines necessary to capture at least one homozygous *G. soja* allele at all polymorphic loci. Two scenarios were investigated, one in which we examined the amount of diversity present when percent of check yield considerations are taken into account, and the other in which the minimal number of lines are determined, using only breeding lines that yield within 70 percent of the *G. max* checks.

RESULTS

GENETIC DISTANCE AND MULTIVARIATE ANALYSIS

The MDS plot of the truncated dataset revealed two similar but distinct clustering patterns for populations JCD-1 and JCD-2 (Figure 2.1). The average simple matching coefficient between all pairs of lines within population JCD-1 was 0.67, while the average SMC within population JCD-2 was 0.63. The average SMC between the populations was 0.60. The SMC for lines within population JCD-1 ranged from 0.48 to 0.93 and had a variance of 0.008, while the range within population JCD-2 was 0.45 to 0.93 and had a variance of 0.005. The between population SMC ranged from 0.43 to 0.75 and had a variance of 0.002.

SEGREGATION DISTORTION IN HIGHLY SELECTED BREEDING LINES

Of the 558 polymorphic SNP markers, 405 in Population JCD-1 and 397 in Population JCD-2 were significantly different ($p<0.01$) from the 0.5 frequency expected in the absence of selection. While most distortion favored the domesticated soybean alleles, 71 in Population JCD-1, and 46 in Population JCD-2 favored the *G. soja* allele significantly ($p<0.01$) at an average *G. soja* allele frequency of 0.64 (Table 2.1). In the analysis of the combined data across populations, the trend for selection to favor a large number *G. soja* alleles was even more apparent. Of the 444 loci with significant ($p<0.01$) deviations from the expected 0.5 frequencies (Fig. 2.2), 58 were skewed toward *G. soja* allele (Table 2.1). A Chi Square test was used to compare differences in allele frequencies between populations for each marker. Of the 558 SNPs, 276 showed a significant difference between the populations. Of these 276 differences, 174 were due to differences in the magnitude of distortion but not

direction, and 102 were due to the *G. soja* allele being favored in one population, and the *G. max* allele in the other. The differences were detected at allele frequency difference between the populations of approximately 0.24.

SINGLE MARKER ANALYSIS FOR KNOWN GENES

To test the validity of using single marker analysis on the quantitative data collected, we converted qualitative data collected on flower and pubescence color to numeric variables. We then conducted comparison of the homozygous genotypic classes (wild soybean vs. domesticate) at all marker loci for these traits. The analysis revealed a significant differences ($p<0.01$) in the homozygous genotypic classes at 44 loci for flower color and 78 loci for pubescence color. A close inspection of these markers associated with flower color reveal 5 loci that have p-values that are less than 1×10^{-16}. These 5 loci, BARC-050787-09886, BARC-029581-06217, BARC-014411-01355, BARC-016891-02361, and BARC-041237-07944 all map to chromosome 13 (LG F) from 21.0 cM to 40.0 cM. The marker BARC-041237-07944, which maps to 21.0 cM on chromosome 13 (LG F) had a t-value of 20.6 for the comparison of genotypic means, as compared to the other markers which had t-values less than or equal to 6.5. The major flower color gene, W1, is located on chromosome 13 at position 28.9 (Cregan et al., 1999). A close inspection of results for pubescence color revealed 10 loci with significant p-values less than or equal to 1×10^{-16}. Seven of these loci map to a region of chromosome 6 (LG C2) spanning from position 90.9 to 108.6. One marker in this region, BARC-066175-19800 (position 100.9 on chromosome 6) had a raw t-value of 39.9. The major gene controlling pubescence color in soybean, *T*, is at position 112.5 cM on chromosome 6 (Cregan et al., 1999).

SINGLE MARKER ANOVA

Initially, each replicate population was subjected to individual analysis. Single marker analysis of variance was performed for six traits on the 558 SNP loci, for a total of 3,348 distinct analyses per population. In population JCD-1, we identified 220 SNP loci in which there was a significant ($p<0.05$) difference in phenotype in the marker genotype classes. Of these 220 differences, 103 were associated with differences in seed yield expressed as a percent of check, 33 were associated with differences in lodging score, 22 were associated with differences in relative maturity, 32 were associated with differences in seed oil content, 14 were associated with differences in seed protein content, and 16 were due to differences in 100-seed weight. In population JCD-2, we identified 172 significant ($p<0.05$) differences between the homozygous genotypic classes. Of the 172 differences, 51 were associated with percent of check yield, nine were associated with lodging score, 22 were associated with relative maturity, 32 were associated with differences in seed oil content, 14 were associated with seed protein content, and 44 were associated with differences in 100 seed weight. In the combined analysis over populations, we observed 335 SNP marker loci with significant ($p<0.05$) differences between the homoyzgous genotypic classes (Table 2.2). Of these 335 differences, 98 were associated with differences in the percent of check yield, 72 were associated with differences in lodging score, 47 were associated with differences in seed oil content, 34 were were associated with differences in seed protein content, 21 were associated with differences in relative maturity, and 63 were associated with differences in seed size. Of these 335 analyses, 83 were due to positive effects on traits derived from the *G. soja* parent.

CREATING A COLLECTION OF BREEDING LINES CAPTURING *G. SOJA* DIVERSITY

Considering the entire collection of breeding lines reported by Delheimer (2012), population JCD-1 contained 10 breeding lines that yielded at least 90% of the conventional *G. max* checks and these carried a total of 261 distinct *G. soja* SNPs alleles in homozygous form (Table 2.8). Including all breeding lines that yielded at least 80% of the checks (an additional 4 breeding lines), the number of recovered G. soja alleles increased to 362. We were not able to recover all 558 polymorphisms within population JCD-1, because four loci were fixed for the *G. max* allele, and three loci contained SNP alleles only in heterozygous form. In population JCD-2, only one breeding line yielded within 90% of the *G. max* checks. This line was homozygous for the *G. soja* allele at 114 individual loci (Table 2.9). An additional 12 breeding lines yielded within 80% of the checks, and these 13 total breeding lines carried a total of 472 *G. soja* alleles in homozygous form . Overall, considering both replicate populations, we observed that all polymorphisms could be captured in 168 breeding lines that yielded at least 60% of the checks (Table 2.10). Manual manipulation of the dataset revealed that a collection of only 16 judiciously selected breeding lines that yield within 70% of the *G. max* would contain at least one homozygote for the *G. soja* allele for 556 of the 558 SNPs (Table 2.11).

DISCUSSION

GENETIC DISTANCE AND MDS ANALYSIS

The initial objective of this study was to assess the patterns of genetic diversity found within the single cross populations of *G.max* X *G.soja* that were described by Delheimer

(2012). The graphical results from MDS analysis revealed a strong genetic relationship between populations JCD-1 and JCD-2 (Fig. 2.1), and that these populations clustered near the *G. max* parent, N7103. The closeness of the two populations is theorized to result from the intense selection pressure toward the *G. max* phenotype while the more subtle distinction between the two is likely the result of a founder effect in that the populations arise from different F2 plants. The analysis of SMC also revealed effects of selection on the populations *versus* genetic drift that may have resulted from founder effects. Average SMC for the breeding lines, based on a large number of SNPs, as was the case here, should be near 0.5 in the absence of selection, as would be the case for coefficient of parentage (Malecot 1948), assuming that the lines trace to different F2 plants. In this study, F2 plants were bulk harvested within each population so that no tracking of individual F2 plants was possible and, thus, selection and genetic drift effects could not be dissected. However, the two independent populations shared no common F2 plants. Thus, a SMC value above 0.5 between populations should represent selection not drift effects. The actual mean SMC was 0.62 across all breeding lines, 0.67 and 0.63 within JCD-1 and JCD-2, respectively, and 0.60 between populations. A mean SMC value above 0.5 between populations, plus the similarity of SMC values within and between populations strongly indicates that intense selection rather than genetic drift or founder effects is a likely explanation of the frequency toward the domesticated parent. Algebraic analysis of finite numbers of F2 plants and effects on mean coefficient of parentage in F4 lines supports this notion, because the number of F2 plants within a population, must be very small (i.e. less than 5) to have a large effect on SMC for the population sizes employed here (data not shown).

SEGREGATION DISTORTION IN HIGHLY SELECTED BREEDING LINES

Our second objective was to look at patterns of selection throughout the populations, and to determine if there are any associations between agronomic performance and the SNPs under study. The allelic frequencies at any loci derived from the mating of two unrelated individuals should show a 1:1 ratio in the absence of selection or genetic drift (Falconer and Mackay 1996). In an analysis of the combined dataset over populations, 446 SNPs varied significantly ($p<0.01$) from this ratio. This could be expected in any wide cross undergoing intense selection to recover the *G. max* phenotype. While most deflections favored the *G. max* alleles, a surprising number 58 SNPs significantly ($p<0.01$) favored the *G. soja* allele and, thus, were favored by conscious or unconscious selection. Comparing the allele frequency data presented in table 2.1 to the results of the single marker ANOVA data presented in table 2.2, 19 of the 58 SNP markers showing a shift in frequency toward the wild allele also showed a difference phenotypic mean for the *G. soja* genotypic classes. In some of these 19 cases, the phenotype associated with the wild type allele might be expected, based on the wild soybean phenotype, although the underlying basis for selection is not clear. For example, a region on chromosome 17 (LG D2) from map position 59.9 to 77.4, appears to be associated with lower 100 seed weight derived from the *G. soja* alleles, as is a region on chromosome 18 (LG G) from map position 40.4 to 41.2, which is associated with greater lodging scores derived from the *G. soja* alleles. In contrast, a region on chromosome 7 (LG M) from map position 62.2 to 65.4, provides an example where the exact opposite phenomena is observed. Lines that are homozygous for the *G. soja* allele have higher 100-seed weights. Delheimer (2012) reported observing transgressive segregants for 100-seed

weight in the breeding lines. Based on our analysis, this region could be a factor in explaining those results.

SINGLE MARKER ANALYSIS FOR KNOWN QTL: FLOWER AND PUBESCENCE COLOR, AND SEED PROTEIN CONTENT

Single marker ANOVA of flower and pubescence color revealed two markers that are almost completely predictive. (BARC-041237-07944 maps to 21.0 cM on chromosome 13 (LG F) is within 10 cM of the major flower color gene, *W1*. BARC-066175-19800 maps to position 100.9 cM on chromosome 6 (LG C2), within 10 cM of the location of the major pubescence color gene, *T*. We also identified 16 SNP loci that were associated with elevated seed protein content derived from the *G. soja* parent. Five of these markers are located on chromosome 20, with four of them occurring in a five centimorgan region (LG I, USLP 1.0 map position 25 to 30) Several research groups have identified a QTL affecting protein and oil content on this chromosome (Diers et al., 1992; Mansur et al., 1996; Csanádi et al., 2001; Chung et al., 2003). We also identified 8 SNP loci that were associated with elevated seed protein on chromosome 15, with six of these loci occurring within a 10 centimorgan region (LG E, USLP 1.0 map position 77 to 87). A major seed protein QTL has been reported on chromosome 15 previously (Fasoula et al., 2004). However, this region is nearly 30 cM away from the region reported here. Follow up analysis would be required to determine the relationship between these findings. An additional *G. soja* putative 'high protein allele' on chromosome 13(LG F) (BARC-055229-13122) is within 10 centimorgans of a seed protein QTL discovered by Hyten et al. (2004). Multiple linear regression of these three genomic

regions suggests that all three regions are significant (p-value <0.05) and explain 42% of the variation in seed protein content (Table 2.6)

These results indicate that single marker ANOVA, despite the highly selected nature of the population may, have utility in assessing specific effects of *G. soja* alleles on agronomic traits.

PUTATIVE YIELD ALLELES FROM SINGLE MARKER ANOVA AND MULTIPLE REGRESSION

The single marker analysis identified 23 SNPs in which breeding lines homozygous for the *G. soja* allele were higher yielding than those homozygous for the *G. max* allele. The positive *G. soja* effects ranged from four to eight percentage units greater for yield expressed as a percent of checks. (Table 2.2). Of these 23 loci, a few were detected in the same 20 cM region, suggesting a total of 13 independent regions in which higher yields corresponded to the *G. soja* allele. A subsequent analysis of the 23 loci was conducted to determine any correlation between markers, and the resulting dataset was used for regression analysis. Multiple linear regression (forward, backward, and stepwise selection) all indicated that a model including three genomic regions was a best fit. A model including BARC-035255-07160, BARC-017645-02642, BARC-063365-18347, and BARC-031311-07043 had R-square value of 0.37 (Table 2.3). The p-values associated with all of these loci were less than 0.05 in this model, with the exception of marker BARC-063365-18347. This could be due to its close proximity to the marker BARC-035255-07160. These two markers were not correlated, but are positioned within 10 cM of each other on chromosome 10 (LG O). The lack of correlation of the loci in the selected lines plus the independent effects revealed

though regression suggest that these regions maybe more additive than epistatic in their effects.

The putative *G. soja* yield allele for SNP marker BARC-031311-07043 (5% increase, 4.1 cM on chromosome 5 (LG A1) was also significantly ($p < 0.01$) associated with higher seed oil content. The positive relationship between seed oil and yield has been well documented (Brim and Burton 1979, Wilcox and Guodong 1997, Wilcox 1998). No previous studies have reported *G. soja* as a source of higher oil content. The range in seed oil content of *G. soja* accessions typically ranges from 10 to 15 percent, while *G. max* accessions typically range from 12 to 25 percent. No selection for seed composition occurred in this study.

A region on chromosome 10 (LG O, BARC-035255-07160 and BARC-063365-18347 which map to postions 41.8 and 51.0 respectively) was associated with increased 100-seed weights (average increase of 1 g) , lower lodging scores, and a 7% increase in yield for the breeding lines homozygous for the *G. soja* allele. The relationship between seed weight and yield has generally been thought to be linear, (Burton 1987). Higher yields can lead to greater lodging in G. max cultivars when yield levels are high, suggesting that the relations between the two traits could be a function of physics. The varied association of putative yield alleles with other agronomic traits suggest that the underlying basis for yield response may have multiple underlying physiological bases. Development of near isogenic line and QTL populations should help clarify the effects of these alleles.

We were able to detect one chromosomal region where we see a positive effect on percent of check yield coming from the *G. soja* parent that is not associated with any other

phenotypic traits. BARC-017645-02642, which maps to 63.8 cM on chromosome 4 (LG C1) respectively, resulted in a 7 percent increase in percent of check yield for lines homozygous for the *G. soja* allele. Previous QTL studies have placed a seed yield QTL in a very similar position (Orf et al., 1999; Yuan et al., 2002; Smalley et al., 2004; Guzman et al., 2007), in a 5 cM region between positions 75 and 80 cM. All of these studies were conducted using both exotic and elite *G. max* germplasm,. Allelism tests would be required to test the novelty of this finding.

Previous studies searching for seed yield QTL in *G. soja* have produced positive results. Li et al. (2008) was able to identify two seed yield QTL from *G. soja* PI 245331, while Concibido et al. (2003) identified one QTL from *G. soja* PI 407305. Of the positive associations we identified, surprisingly none of them were placed in regions identified in these previous studies. The populations used in those studies were developed using the advanced backcross method first described by Tanksley and Nelson (1996). Populations JCD-1 and JCD-2 were developed from visual selection on large F_3 populations. Delheimer (2012) suggested this method as a way to maintain greater variation from the wild soybean that may be lost in backcrossing schemes that intentionally, but randomly remove a majority of the wild relative genome. Although we are able to make only tentative conclusions in the yield analyses reported here, these materials are being used to develop validation tests.

We were also identified 61 single SNPs that in which the *G. max* alleles were associated with significantly higher yield. Based on the mapping positions, there may be 28 separate chromosomal regions. These regions also occur in areas where there is greater lodging associated with the *G. soja* alleles. The *G. soja* exhibits a viny growth habit, and

many believe that during the domestication of *G. max*, upright growth (lodging tolerance) may have been one of the very first traits selected upon. These regions also tend to be associated with larger seed size and higher seed oil contents derived from the *G. max* parent. Further studies of *G. max* X *G. soja* populations may clarify the association of these genomic regions with domestication and their interaction with putative *G. soja* yield alleles.

PUTATIVE 100-SEED WEIGHT ALLELES FROM SINGLE MARKER ANOVA AND MULTIPLE REGRESSION,

Delheimer (2012) identified transgressive segregants for 100-seed weight based on their analysis of agronomic data of single cross *G. max* X *G. soja* breeding lines. Therefore, it is not surprising that we were able to identify a total of 20 SNP loci in which the breeding lines that are homozygous for the *G. soja* alleles had significantly ($p<0.01$) larger 100 seed weights. Of these 20 loci, it appears that we have identified 14 independent regions with significant effects on 100 seed weight derived from the *G. soja* parent. The effects of these regions ranged from 0.9 to 1.4 g difference in 100 seed weights between the two genotypic classes. BARC-040459-07745, which is positioned on chromosome 5 (LG A1) at 36.1 cM, is close to a position where a seed weight QTL was previously reported (Orf et al., 1999). BARC-023732-03474 and BARC-054099-12340 are both located on chromosome 16 (LG J) at positions 11.1 and 25.8 cM respectively. A previous study on seed weight placed a QTL very close to these positions as well (Mian et al., 1996). The markers BARC-28853-06030 and BARC-900926-00961 map to chromosomes 8 and 13 (LGs A2 and F) respectively. Correlations of the allele calls on these regions do not reveal a significant relationship, so it appears that they segregate independently. Stepwise regression suggests a model in which

four markers explain the greatest variation in the model. These markers included BARC-029185-06108, BARC-055141-13084, BARC-028853-06030, and BARC-062193-17703 (Table 2.4). No previous studies have reported significant associations with seed weight in the regions where these markers are located. It should be noted that the previous studies focused on alleles derived from *G. max* parents, so allelism tests would be required to test the possible novelty of these associations.

PUTATIVE SEED OIL CONTENT ALLELES FROM SINGLE MARKER ANOVA AND MULTIPLE REGRESSION

A total of 12 SNPs showed a significant ($p<0.01$) difference between genotypic classes, in which lines homozygous for the *G. soja* allele had a higher mean seed oil content. Two of these markers, BARC-23411-05376 and BARC-031311-07043, both map to 4 cM on chromosome 5 (LG A1). Another marker, BARC-064455-18689, maps to 72 cM on chromosome 16 (LG J), while BARC-061049-17016 maps to position 77.4 on chromosome 17 (LG D2). The other markers are located on three distinct regions of chromosome 8 (LG A2). Regression analysis of these separate regions reveals that the markers BARC-061049-17016, BARC-031311-07043, BARC-022387-04319, and BARC-050171-09440 have an R-square of 0.34 (Table 2.5). It appears that these three regions have a significant effect on seed oil content, and further analysis of these regions should be able to quantify their effects.

PUTATIVE LODGING ALLELES FROM SINGLE MARKER ANOVA AND MULTIPLE REGRESSION

In total we identified, 57 loci that were associated with differences in lodging (Table 2.2). Of these 57, 45 were the result of the lines that were homozygous for the *G. soja* allele

having a greater lodging score. This is to be expected from a cross of any crop species to a wild relative, as lodging would be a deleterious trait that would result in the progeny. However, the remainder of these differences could be attributed to lodging tolerance being derived from the *G. soja* parent. Multiple linear regression identified a model that included four of these loci (Table 2.7). BARC-059885-16192 which maps to position 51.4 on chromosome 10 (LG O), BARC-055907-13843 which maps to position 61.7 on chromosome 12 (LG H), BARC-050205-09457 which maps to position 76.9 on chromosome 11 (LG B1), and BARC-050595-09732 which maps to position 111.9 on chromosome 11 (LG B1) were all included as significant effects in the model. These markers do not appear to be near to any reported QTL. Lodging is a complex, multigenic trait as seen in the number of positive associations we detected. Further analysis of populations developed from crosses of *G.max* X *G. soja* should be able to further quantify these effects, and to possibly detect any relationships between this trait and possible domestication of soybean.

CREATING A COLLECTION OF BREEDING LINES CAPTURING *G.SOJA* DIVERSITY

We were able to identify a set of breeding lines that maintained good agronomic performance. Table 2.6 illustrates that we can identify at least one homozygous genotype at 287 loci and still have yield within 90 percent of the *G. max* checks. If we lower this restriction to 80 percent, we are able to identify 518 polymorphisms with 27 breeding lines. If particular alleles were to be shown beneficial in soybean breeding, this collection could serve as an agronomically acceptable platform of *G. soja* diversity for future studies. If yield restrictions are limited to breeding lines that yield within 70% of the *G. max* checks, we were able to identify 556 of the 558 polymorphisms in only 16 breeding lines. Further genetic

analysis, such as sequencing breeding lines, would be necessary to empirically state the exact amount of the *G. soja* genome that is contained within this collection. Also, some of these lines may contain the centromeric regions of chromosomes, which would include regions of heterochromatic DNA that would increase the amount of introgression from *G. soja* that has occurred.

IMPLICATIONS TO SOYBEAN BREEDING

Overall, these results show that the use of mega-populations and visual selection offers a viable alternative to the backcross methods typically employed by plant breeders working to improve crop species with wild relatives. We identified several genomic regions in *G. soja* that may be associated with agronomic improvement of soybean. Further studies are needed to validate and better understand the effects of the SNP alleles mentioned above. However, it is clear that this breeding material, by virtue of its uniqueness and diversity, offers a starting point for a greater understanding of genetic diversity present within *G. soja*. One of the main arguments that has been proposed against using the advanced backcross method proposed by Tanksley and Nelson (1996), has been that introgressing small sections of the genome may prevent the discovery of useful multigeneic, epistatic interactions (Stuber et al., 1999) in that stringent selection against the wild species genome can be counterproductive. In maize, an allele associated with a 10.5 bushel per acre increase (Furbeck 1993) may have gone undetected if the population had been developed in the way described by Tanksley and Nelson (1996). The proper synergy of molecular marker technology and plant breeding has long been debated. The method described by Tanksley and Nelson (1996) was proposed for the detection of QTL from wild relatives, while

simultaneously developing improved elite varieties. The method described by Delheimer (2012) allows for large amounts of diversity to be introgressed into easily manageable germplasm. Our assessment from marker analysis of this population is that that 99% percent of the *G. soja* SNP alleles can be captured in as few as 16 breeding lines, while still maintaining yields as 70% of *G. max* checks. Although these lines will not be commercially released cultivars, their use as parental stock may lead to genetic gains in quantitative traits. For example, the lines that were identified by Delheimer (2012) that yield within 90% of the check, all contain the putative yield alleles that we have identified.

The initial analysis conducted by Delheimer (2012) showed that the use of mega-populations and visual selection can be implemented to create diverse, agronomic breeding lines. It did offer a glimpse into the useful diversity that could be created with this method as they were able to report on positive transgressive segregation for 100 seed weight, derived from the *G. soja* parent. The results here begin to provide an underlying basis for that phenotypic result. Five genomic regions associated with greater yield, as well as five genomic regions associated with greater 100 seed weight show that this method can provide useful diversity on a large scale. The single genomic region associated with lower seed oil content derived from *G. soja* is a result that has never before been reported. These results, combined with the results of Delheimer (2012), shows that the use of mega-populations and visual selection are critical in the continued mining of germplasm and the breaking of the genetic bottleneck created by the domestication of *G. max.*

REFERENCES

Brim, C.A., and J.W. Burton. 1979. Recurrent selection in soybeans. II. Selection for increased percent protein in seeds. Crop Sci. 19: 494-498.

Brown-Guedira, G.L., J.A. Thompson, R.L. Nelson, and M.L. Warburton. 2000. Evaluation of genetic diversity of soybean introductions and North American ancestors using RAPD and SSR markers. Crop Sci. 40: 815-823.

Burton, J.W. 1987. Quantitative genetics: Results relevant to soybean breeding. In Wilcox, J.R (ed.) Soybeans: Improvement, production and uses. 2nd ed. Agron. Monogr. 16. ASA, CSSA, and SSSA, Madison, WI.

Csanádi, G., J. Vollmann, G. Stift, and T. Lelly. 2001. Seed quality QTLs identified in a molecular map of early maturing soybean. Theor. Appl. Genet. 103:912–919.

Carpenter, J. and W.R. Fehr. 1986. Genetic variability for desirable agronomic traits in populations containing *Glycine soja* germplasm. Crop Sci. 26: 681-686.

Carter, T.E., J.W. Burton, M.R. Villagarcia, Z. Cui, X. Zhou, M.O. Fountain, D.T. Bowman, and A.S. Niewoehner. 2003. Registration of 'N7103' soybean. Crop Sci. 43: 1128.

Carter, T.E., Z. Gizlice, and J.W. Burton. 1993. Coefficient-of-parentage and genetic-proximity estimates for 258 North American cultivars released by public agencies during 1945 and 1988. USDA Tech. Bull. 1814. U.S. Gov. Print. Office, Washington, DC.

Chung, J. H. L. Babka, G. L. Graef, P. E. Staswick, D. J. Lee, P. B. Cregan, R. C. Shoemaker, and J. E. Specht. 2003. Crop Sci. 43: 1053-1067.

Concibido, V.C., B.L. Vallee, P. Mclaird, N. Pineda, J. Meyer, L. Hummel, J. Yang, K. Wu, and X. Delannay. 2003. Introgression of a quantitative trait locus for yield from *Glycine soja* into commercial soybean cultivars. Theor. Appl. Genet. 106: 575-582.

Delheimer, J.C. 2012. Agronomic performance of breeding lines developed from a single cross of *Glycine max* X *Glycine soja*. Dissertation, Chapter 1.

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

Ertl, D.S., and W.R. Fehr. 1985. Agronomic performance of soybean genotypes from *Glycine max* X *Glycine soja* crosses. Crop Sci 25: 589-592.

Falconer, D. S., and T. F. C. Mackay, 1996. Introduction to quantitative genetics. 4th Ed. Longmans Green, Harlow, Essex, UK.

Fasoula, V.A., D.K. Harris, and H.R. Boerma. 2004. Validation and Designation of Quantitative Trait Loci for Seed Protein, Seed Oil, and Seed Weight from Two Soybean Populations. Crop Sci. 44: 1218-1225.

Furbeck, S.M. 1993. The development and evaluation of molecular-marker derived near isogenic lines to study quantitative traits in maize. Ph.D. thesis, North Carolina State University, Raleigh (Diss. Abstr. DA9330287).

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.

Guzman, P. S., B. W. Diers, D. J. Neece, S. K. St. Martin, A. R. LeRoy, C. R. Grau, T. J. Hughes, and R. L. Nelson. 2007. QTL associated with yield in three backcross-derived populations of soybean. Crop Sci. 47: 111-122.

Hyten, D., I. Choi, Q. Song, J.E. Specht, T.E. Carter, Jr., R.C. Shoemaker, E. Hwang, L.K. Matukumalli, and P.B. Cregan. 2010. A high density integrated genetic linkage map of soybean and the development of a 1536 universal soy linkage panel for quantitative trait locus mapping. Crop Sci 50: 960-968.

Hyten, D., Q. Song, I. Choi, M. Yoon, J. Specht, L. Matukumalli, R. Nelson, R. Shoemaker, N. Young, and P. Cregan. 2008. High-throughput genotyping with the GoldenGate assay in the complex genome of soybean. Theor Appl Genet 116:945-952.

Hyten, D.L., Q. Song, Y. Zhu, I. Choi, R.L. Nelson, J. M. Costa, J. E. Specht, R. C. Shoemaker, and P.B. Cregan. 2006. Impacts of genetic bottlenecks on soybean genome diversity. PNAS. Vol. 103. 45: 16666-16671.

Hyten, D.L., V.R. Pantalone, C.E. Sams, A.M. Saxton, D. Landau-Ellis, T.R. Stefaniak, and M. E. Schmidt. 2004. Seed quality QTL in a prominent soybean population. Theor. Appl. Genet. 109: 552-561.

Li, D., T.W. Pfeiffer, and P.L. Cornelius. 2008. Soybean QTL for yield and yield components associated with *Glycine soja* alleles. Crop Sci. 48: 571-581.

Malecot, G. 1948. Les mathematiques de l'he´re´ dite´ . Masson, Paris. English translation. The mathematics of heredity. 1969. W.H. Freeman and Co., San Francisco, CA.

Mansur, L.M., J.H. Orf, K. Chase, T. Jarvik, P.B. Cregan, and K.G. Lark. 1996. Genetic mapping of agronomic traits using recombinant inbred lines of soybean. Crop Sci. 36:1327–1336.

Mian, R., M.A. Bailey, J.P. Tamulonis, E.R. Shipe, T.E. Carter, Jr., W.A. Parrott, D.A. Ashley, R.S. Hussey, and H.R. Boerma. 1996. Molecular markers associated with seed weight in two soybean populations. Theor. Appl. Genet. 93: 1011-1016.

Orf, J.H., K. Chase, F. R. Adler, L. M. Mansur, and K. G. Lark. 1999. Genetics of Soybean Agronomic Traits: II. Interactions between Yield Quantitative Trait Loci in Soybean. Crop Sci 39: 1652-1657.

Rawlings, J.O., S.G. Pantula, and D.A. Dickey. 1998. Applied regression analysis: a research tool. 2nd Ed. Springer-Verlag, New York, NY.

Smalley, M.D., W.R. Fehr, S.R. Cianzio, F. Han, S.A. Sebastian, and L.G. Streit. 2004. Quantitative trait loci for soybean seed yield in elite and plant introduction germplasm. Crop Sci. 44:436–442.

Snedecor and Cochran. 1989. Statistical methods. 8th Ed. Iowa State University Press, Ames, Iowa.

Sokal, R.R., and C.D. Michener. 1958. A statistical method for evaluating systematic relationships. Univ. Kans. Sci. Bull. 38:1409–1438.

Stuber, C.W., M. Polacco, and M.L. Senior. 1999. Synergy of empirical breeding, marker-assisted selection, and genomics to increase crop yield protection. Crop Sci. 39: 1571-1583.

Tanksley, S.D. and J.C. Nelson. 1996a. Advanced backcross QTL analysis: a method for the simultaneous discovery and transfer of valuable QTLs from unadapted germplasm into elite breeding lines. Theor Appl Genet 92: 191-203.

Thompson, J.A., R.L. Nelson, and L.O. Vodkin. 1998. Identification of diverse soybean germplasm using RAPD markers. Crop Sci 38: 1348-1355.

USDA-ARS National Genetic Resources Program (USDA-ARS NGRP). 2012c. *Glycine soja* Sieb. And Zucc. PI 366122. Germplasm Resources Information Network (GRIN). [Online database] National Germplasm Resources Laboratory, Beltsville, MD. Available at http://www.arsgrin.gov/cgi-bin/npgs/acc/display.pl?1274061 (verified 09 Feb. 2012).

Wilcox, J.R. 1998. Increasing seed protein in soybean with eight cycles of recurrent selection. Crop Sci. 38:1536–1540.

Wilcox, J.R., and Z. Guodong. 1997. Relationships between seed yield and seed protein in determinate and indeterminate soybean populations. Crop Sci. 37:361–364.

Yuan, J., V.N. Njiti, K. Meksem, M.J. Iqbal, K. Triwitayakorn, M.A. Kassem, G.T. Davis, M.E. Schmidt, and D.A. Lightfoot. 2002. Quantitative trait loci in two soybean recombinant inbred line populations segregating for yield and disease resistance. Crop Sci. 42:271–277

TABLES

Table 2.1 Summary of allele frequencies that are significantly skewed toward the *G. soja* allele within and across populations. All significance at p<0.01.

SNP	Population	Chromosome	Linkage Group	Position in cM	*G. soja* allele frequency
BARC-047699-10383	1	1	D1A	17.3	0.89
BARC-047699-10383	Total	1	D1A	17.3	0.76
BARC-024477-04900	1	1	D1A	24.8	0.91
BARC-024477-04900	Total	1	D1A	24.8	0.74
BARC-011057-00831	1	1	D1A	39.9	0.62
BARC-057545-14810	1	2	D1B	46.9	0.63
BARC-047945-10443	1	2	D1B	71.7	0.87
BARC-047945-10443	2	2	D1B	71.7	0.67
BARC-047945-10443	Total	2	D1B	71.7	0.77
BARC-064823-18809	1	2	D1B	82.2	0.62
BARC-030479-06875	1	2	D1B	103.6	0.79
BARC-030479-06875	Total	2	D1B	103.6	0.65
BARC-021647-04164	1	2	D1B	103.9	0.75
BARC-021647-04164	Total	2	D1B	103.9	0.63
BARC-021465-04122	1	3	N	47.8	0.66
BARC-048987-10780	1	5	A1	0.0	0.86
BARC-048987-10780	Total	5	A1	0.0	0.71
BARC-040651-07808	1	5	A1	2.5	0.69
BARC-023411-05376	1	5	A1	4.1	0.86
BARC-031311-07043	1	5	A1	4.1	0.87
BARC-014883-01912	1	5	A1	24.1	0.67
BARC-014883-01912	2	5	A1	24.1	0.66
BARC-014883-01912	Total	5	A1	24.1	0.66
BARC-041257-07953	1	5	A1	78.8	0.75
BARC-041257-07953	Total	5	A1	78.8	0.66
BARC-042045-08161	2	6	C2	23.8	0.70
BARC-056271-14211	1	6	C2	41.5	0.62
BARC-056271-14211	2	6	C2	41.5	0.91

Table 2.1 cont.

SNP	Population	Chromosome	Linkage Group	Position in cM	*G. soja* allele frequency
BARC-056271-14211	Total	6	C2	41.5	0.75
BARC-063259-18282	2	6	C2	66.1	0.68
BARC-038885-07387	2	6	C2	114.1	0.80
BARC-044841-08824	1	7	M	45.1	0.64
BARC-054331-12480	1	7	M	48.4	0.75
BARC-054331-12480	Total	7	M	48.4	0.68
BARC-060941-16977	1	7	M	57.3	0.68
BARC-060941-16977	2	7	M	57.3	0.72
BARC-060941-16977	Total	7	M	57.3	0.70
BARC-062457-17799	1	7	M	62.2	0.64
BARC-062457-17799	2	7	M	62.2	0.83
BARC-062457-17799	Total	7	M	62.2	0.74
BARC-019987-03748	1	7	M	64.7	0.63
BARC-019987-03748	2	7	M	64.7	0.72
BARC-019987-03748	Total	7	M	64.7	0.68
BARC-062193-17703	2	7	M	65.8	0.72
BARC-062193-17703	Total	7	M	65.8	0.60
BARC-020517-04647	2	7	M	66.4	0.83
BARC-020517-04647	Total	7	M	66.4	0.73
BARC-907715-01049	Total	7	M	73.4	0.59
BARC-045199-08906	1	8	A2	31.2	0.70
BARC-054355-12497	2	8	A2	133.9	0.69
BARC-054355-12497	Total	8	A2	133.9	0.65
BARC-056517-14441	1	8	A2	135.4	0.64
BARC-056517-14441	2	8	A2	135.4	0.75
BARC-056517-14441	Total	8	A2	135.4	0.69
BARC-054887-12192	1	8	A2	136.3	0.72
BARC-054887-12192	2	8	A2	136.3	0.66

Table 2.1 cont.

SNP	Population	Chromosome	Linkage Group	Position in cM	*G. soja* allele frequency
BARC-054887-12192	Total	8	A2	136.3	0.69
BARC-053255-11775	1	8	A2	143.0	0.68
BARC-053255-11775	2	8	A2	143.0	0.63
BARC-053255-11775	Total	8	A2	143.0	0.65
BARC-022031-04262	1	9	K	5.3	0.75
BARC-022031-04262	2	9	K	5.3	0.65
BARC-022031-04262	Total	9	K	5.3	0.70
BARC-055141-13084	1	9	K	6.9	0.92
BARC-055141-13084	2	9	K	6.9	0.68
BARC-055141-13084	Total	9	K	6.9	0.80
BARC-059521-15876	1	9	K	9.3	0.75
BARC-059521-15876	Total	9	K	9.3	0.65
BARC-051595-11169	1	9	K	10.1	0.80
BARC-051595-11169	Total	9	K	10.1	0.68
BARC-055301-13192	1	9	K	23.3	0.76
BARC-055301-13192	Total	9	K	23.3	0.59
BARC-061977-17602	2	9	K	40.1	0.67
BARC-061977-17602	Total	9	K	40.1	0.62
BARC-056323-14257	2	9	K	41.7	0.67
BARC-056323-14257	Total	9	K	41.7	0.62
BARC-052125-11359	2	9	K	81.3	0.71
BARC-052125-11359	Total	9	K	81.3	0.63
BARC-051769-11236	1	9	K	93.8	0.76
BARC-051769-11236	2	9	K	93.8	0.67
BARC-051769-11236	Total	9	K	93.8	0.71
BARC-035255-07160	2	10	O	41.8	0.65
BARC-025873-05130	1	11	B1	45.4	0.92
BARC-025873-05130	Total	11	B1	45.4	0.68

Table 2.1 cont.

SNP	Population	Chromosome	Linkage Group	Position in cM	*G. soja* allele frequency
BARC-042299-08241	2	11	B1	92.5	0.67
BARC-044181-08640	2	12	H	7.2	0.65
BARC-055731-13669	1	12	H	31.7	0.69
BARC-050429-09617	2	12	H	56.1	0.75
BARC-050429-09617	Total	12	H	56.1	0.63
BARC-050787-09886	1	13	F	33.1	0.64
BARC-018551-02971	1	13	F	34.1	0.64
BARC-024663-05516	1	13	F	34.8	0.64
BARC-061813-17543	1	13	F	34.8	0.65
BARC-061813-17543	Total	13	F	34.8	0.60
BARC-050657-09804	1	13	F	45.0	0.71
BARC-014285-01304	1	14	B2	4.9	0.69
BARC-014285-01304	2	14	B2	4.9	0.69
BARC-014285-01304	Total	14	B2	4.9	0.69
BARC-061279-17151	Total	14	B2	63.5	0.60
BARC-901431-00997	2	14	B2	64.2	0.67
BARC-031269-07029	1	15	E	0.0	0.98
BARC-031269-07029	Total	15	E	0.0	0.78
BARC-018923-03037	1	15	E	12.0	0.71
BARC-018923-03037	Total	15	E	12.0	0.63
BARC-030171-06819	1	15	E	27.2	0.71
BARC-030171-06819	2	15	E	27.2	0.66
BARC-030171-06819	Total	15	E	27.2	0.68
BARC-027786-06670	1	15	E	33.2	0.67
BARC-025663-04988	1	15	E	47.9	0.66
BARC-025663-04988	2	15	E	47.9	0.82
BARC-025663-04988	Total	15	E	47.9	0.74
BARC-018461-02916	1	15	E	61.4	0.82

Table 2.1 cont.

SNP	Population	Chromosome	Linkage Group	Position in cM	*G. soja* allele frequency
BARC-018461-02916	2	15	E	61.4	0.68
BARC-018461-02916	Total	15	E	61.4	0.75
BARC-058675-17461	2	15	E	68.1	0.66
BARC-058689-17465	2	15	E	68.1	0.67
BARC-055571-13451	2	15	E	81.6	0.80
BARC-055571-13451	Total	15	E	81.6	0.63
BARC-057281-14664	2	15	E	93.7	0.80
BARC-054099-12340	1	16	J	25.8	0.68
BARC-054099-12340	Total	16	J	25.8	0.64
BARC-050447-09631	1	16	J	42.4	0.63
BARC-050447-09631	2	16	J	42.4	0.64
BARC-050447-09631	Total	16	J	42.4	0.63
BARC-048299-10543	1	16	J	43.0	0.63
BARC-048299-10543	2	16	J	43.0	0.64
BARC-048299-10543	Total	16	J	43.0	0.63
BARC-064455-18689	1	16	J	71.6	0.62
BARC-044031-08587	1	16	J	81.4	0.73
BARC-044031-08587	2	16	J	81.4	0.88
BARC-044031-08587	Total	16	J	81.4	0.80
BARC-019215-03395	2	16	J	86.8	0.71
BARC-048135-10500	2	16	J	86.8	0.79
BARC-048135-10500	Total	16	J	86.8	0.69
BARC-031515-07105	2	16	J	88.5	0.72
BARC-020357-04569	1	17	D2	5.1	0.80
BARC-065705-19668	1	17	D2	16.4	0.75
BARC-013637-01186	1	17	D2	59.9	0.83
BARC-013637-01186	Total	17	D2	59.9	0.64
BARC-013969-01290	1	17	D2	62.8	0.82

Table 2.1 cont.

SNP	Population	Chromosome	Linkage Group	Position in cM	*G. soja* allele frequency
BARC-013969-01290	Total	17	D2	62.8	0.64
BARC-047685-10379	1	17	D2	64.7	0.81
BARC-047685-10379	Total	17	D2	64.7	0.62
BARC-047829-10399	2	17	D2	72.2	0.74
BARC-047829-10399	Total	17	D2	72.2	0.67
BARC-059719-16041	2	17	D2	75.1	0.64
BARC-061049-17016	1	17	D2	77.4	0.77
BARC-061049-17016	2	17	D2	77.4	0.63
BARC-061049-17016	Total	17	D2	77.4	0.70
BARC-043197-08552	1	18	G	0.9	0.70
BARC-043197-08552	2	18	G	0.9	0.65
BARC-043197-08552	Total	18	G	0.9	0.67
BARC-G01477-00243	1	18	G	10.0	0.70
BARC-047665-10370	1	18	G	16.0	0.73
BARC-047665-10370	Total	18	G	16.0	0.65
BARC-901121-00988	1	18	G	40.4	0.88
BARC-901121-00988	Total	18	G	40.4	0.64
BARC-063581-18909	1	18	G	41.2	0.89
BARC-063581-18909	Total	18	G	41.2	0.67
BARC-051587-11167	1	18	G	44.1	0.71
BARC-051587-11167	Total	18	G	44.1	0.64
BARC-016859-02354	1	18	G	45.0	0.70
BARC-054849-12183	1	18	G	47.5	0.71
BARC-064283-18606	1	18	G	48.2	0.71
BARC-059485-15839	1	18	G	48.9	0.69
BARC-063705-18440	1	18	G	54.4	0.68
BARC-047504-12947	1	18	G	55.6	0.90
BARC-047504-12947	Total	18	G	55.6	0.74

Table 2.1 cont.

SNP	Population	Chromosome	Linkage Group	Position in cM	*G. soja* allele frequency
BARC-024489-04936	2	18	G	70.6	0.64
BARC-028567-05954	1	19	L	54.9	0.73
BARC-028567-05954	Total	19	L	54.9	0.65
BARC-064609-18739	1	19	L	86.3	0.70
BARC-039753-07565	1	20	I	64.0	0.64
BARC-039753-07565	2	20	I	64.0	0.67
BARC-039753-07565	Total	20	I	64.0	0.65
BARC-024489-04936	2	18	G	70.6	0.64
BARC-028567-05954	1	19	L	54.9	0.73
BARC-028567-05954	Total	19	L	54.9	0.65
BARC-064609-18739	1	19	L	86.3	0.70
BARC-039753-07565	1	20	I	64.0	0.64
BARC-039753-07565	2	20	I	64.0	0.67
BARC-039753-07565	Total	20	I	64.0	0.65

Table 2.2 Results of single marker ANOVA of agronomic traits for combined data of all breeding lines.

SNP	Chr†	LG‡	Pos§	Trait¶	Number of Individuals Homozygous for *G. max* allele			Number of Individuals Homozygous for *G. soja* allele			Mean of Individuals Homozygous for *G. max* allele			Mean of Individuals Homozygous for *G. soja* allele			Prob (F)		
					1	2	T	1	2	T	1	2	T	1	2	T	1	2	T
BARC-060833-16926	1	D1A	37.1	MG	40	46	86	15	10	25	6.7	6.8	6.8	6.2	6.8	6.4	0.0016	0.8078	0.0057
BARC-011057-00831	1	D1A	39.9	LOD	23	33	56	38	13	51	2.9	2.5	2.7	3.1	2.9	3.0	0.1874	0.0186	0.0009
BARC-061099-17047	1	D1A	41.6	LOD	26	32	58	32	12	44	2.9	2.6	2.7	3.1	2.9	3.0	0.2931	0.0527	0.0028
BARC-050267-09542	1	D1A	41.7	LOD	40	39	79	12	10	22	2.9	2.6	2.7	3.2	2.9	3.1	0.0583	0.0067	0.0009
BARC-064293-18611	1	D1A	42.6	LOD	23	43	66	36	7	43	3.0	2.6	2.7	3.1	2.9	3.0	0.4684	0.2222	0.0054
BARC-065083-19095	1	D1A	45.6	MG	40	41	81	15	10	25	6.7	6.8	6.8	6.2	6.7	6.4	0.0024	0.879	0.0092
BARC-024909-10358	1	D1A	54.9	%CY	29	39	68	25	13	38	71	70	71	64	65	64	0.0382	0.1832	0.0344
BARC-024909-10358	1	D1A	54.9	LOD	29	39	68	25	13	38	2.9	2.6	2.7	3.2	2.7	3.0	0.0778	0.0292	0.0037
BARC-060037-16311	1	D1A	58.8	%CY	46	41	87	9	10	19	70	71	70	59	65	62	0.0096	0.2065	0.0569
BARC-060037-16311	1	D1A	58.8	LOD	46	41	87	9	10	19	2.9	2.6	2.8	3.5	2.8	3.1	0.0049	0.046	0.003
BARC-060037-16311	1	D1A	58.8	OIL	46	41	87	9	10	19	174	166	170	160	159	160	0.0008	0.2584	0.0006
BARC-018755-03002	1	D1A	59.8	%CY	47	41	88	7	7	14	70	71	70	59	65	62	0.0374	0.1871	0.0302
BARC-039805-07589	1	D1A	83.8	%CY	57	44	101	4	10	14	68	71	70	62	65	64	0.2536	0.0407	0.0458
BARC-039805-07589	1	D1A	83.8	OIL	57	44	101	4	10	14	173	168	171	168	158	161	0.4689	0.059	0.0077
BARC-048593-10672	2	D1B	0.9	SDWT	24	41	65	19	12	31	8.4	8.5	8.5	7.3	7.5	7.4	0.0361	0.0004	0.0021
BARC-027478-06590	2	D1B	19.8	%CY	26	45	71	18	7	25	70	70	70	64	64	64	0.112	0.1501	0.0311
BARC-027478-06590	2	D1B	19.8	SDWT	26	45	71	18	7	25	8.6	8.4	8.5	7.3	7.2	7.3	0.0127	0.0122	0.0008
BARC-016063-02049	2	D1B	28.4	SDWT	20	40	60	29	11	40	8.9	8.4	8.5	7.6	7.7	7.6	0.0079	0.0632	0.0017
BARC-007889-00156	2	D1B	36.1	%CY	27	42	69	23	7	30	71	69	70	62	66	63	0.0047	0.646	0.0033
BARC-007889-00156	2	D1B	36.1	LOD	27	42	69	23	7	30	2.9	2.7	2.8	3.2	2.7	3.1	0.0167	0.6507	0.0053
BARC-028393-05861	2	D1B	40.3	%CY	28	42	70	24	7	31	72	69	70	61	69	63	<0.0001	0.7487	0.0166
BARC-028393-05861	2	D1B	40.3	LOD	28	42	70	24	7	31	2.9	2.7	2.8	3.2	2.7	3.1	0.0177	0.7794	0.0014
BARC-014233-03138	2	D1B	82.2	%CY	30	48	78	19	6	25	71	71	71	63	60	63	0.0201	0.0232	0.0007
BARC-014233-03138	2	D1B	82.2	SDWT	30	48	78	19	6	25	8.4	8.4	8.4	7.8	6.2	7.5	0.1334	0.0037	0.0036

† Chromosome number.
‡ Linkage Group.
§ Position of SNP on chromosome in centimorgans, based on USLP 1.0 map positions.
¶ Traits: LOD=Lodging score, %CY=Yield as Percent of Check, SDWT=100 seed weight in grams, OIL=Seed oil content, PRO=Seed protein content, MG=Relative Maturity.

Table 2.2 cont.

SNP	Chr †	LG‡	Pos§	Trait¶	Number of Individuals Homozygous for *G. max* allele			Number of Individuals Homozygous for *G. soja* allele			Mean of Individuals Homozygous for *G. max* allele			Mean of Individuals Homozygous for *G. soja* allele			Prob (F)		
					1	2	T	1	2	T	1	2	T	1	2	T	1	2	T
BARC-030479-06875	2	D1B	103.6	SDWT	11	26	37	45	26	71	8.7	8.8	8.8	8.0	7.7	7.9	0.2747	0.0024	0.0007
BARC-021647-04164	2	D1B	103.9	SDWT	13	27	40	43	26	69	8.9	8.7	8.7	7.8	7.7	7.8	0.0514	0.016	0.001
BARC-045053-08869	3	N	27.5	%CY	31	27	58	21	2	23	71	71	71	64	74	65	0.0386	0.6166	0.004
BARC-010837-00763	3	N	30.5	%CY	32	44	76	22	9	31	71	70	71	64	69	66	0.0468	0.6699	0.0576
BARC-021465-04122	3	N	47.8	%CY	19	39	58	37	10	47	72	70	71	65	68	66	0.0269	0.8593	<0.0001
BARC-021465-04122	3	N	47.8	LOD	19	39	58	37	10	47	2.9	2.6	2.7	3.1	2.8	3.0	0.1615	0.575	0.0009
BARC-029185-06108	3	N	60.7	SDWT	52	27	79	8	25	33	7.9	7.2	7.7	8.8	8.7	8.7	0.2351	0.0037	0.005
BARC-028539-05944	3	N	74.2	LOD	31	40	71	25	7	32	2.9	2.6	2.8	3.1	2.9	3.1	0.0812	0.1746	0.0041
BARC-048557-10665	3	N	79.1	LOD	31	43	74	24	6	30	2.9	2.6	2.7	3.1	2.9	3.1	0.0174	0.0674	0.0002
BARC-014699-01621	3	N	85.2	LOD	24	38	62	32	11	43	3.0	2.6	2.7	3.1	2.8	3.0	0.5973	0.1185	0.0049
BARC-028321-05831	4	C1	4.9	%CY	48	37	85	9	3	12	70	72	71	57	69	60	0.0003	0.8336	0.0004
BARC-057913-15004	4	C1	10.0	%CY	47	46	93	9	6	15	71	71	71	57	67	61	0.0007	0.1264	0.031
BARC-057913-15004	4	C1	10.0	OIL	47	46	93	9	6	15	175	168	172	162	160	161	0.0018	0.3726	0.0058
BARC-022353-04316	4	C1	16.7	%CY	46	37	83	14	11	25	71	71	71	60	66	63	0.0005	0.2208	0.0059
BARC-022353-04316	4	C1	16.7	LOD	46	37	83	14	11	25	2.9	2.7	2.8	3.4	2.9	3.1	0.0009	0.2583	0.0012
BARC-022353-04316	4	C1	16.7	OIL	46	37	83	14	11	25	175	168	172	163	161	162	0.0007	0.4446	0.0061
BARC-031733-07217	4	C1	19.5	%CY	52	33	85	5	13	18	69	71	70	56	67	64	0.0127	0.2646	0.0448
BARC-031733-07217	4	C1	19.5	OIL	52	33	85	5	13	18	175	169	173	153	159	157	<0.0001	0.1263	<0.0001
BARC-021219-04011	4	C1	20.3	%CY	52	34	86	5	14	19	69	71	70	56	67	64	0.0127	0.3307	0.0166
BARC-021219-04011	4	C1	20.3	OIL	52	34	86	5	14	19	175	169	173	153	161	159	<0.0001	0.2304	0.0002
BARC-019015-03051	4	C1	28.0	OIL	61	39	100	-	11	11	172	169	171	-	154	154	-	0.0116	0.0006
BARC-017645-02642	4	C1	63.8	%CY	39	40	79	13	12	25	64	69	67	77	72	74	0.0002	0.7293	0.0019
BARC-044523-08716	4	C1	63.9	%CY	38	40	78	14	12	26	64	69	67	77	72	74	0.0002	0.7383	0.0072
BARC-044373-08692	4	C1	80.8	%CY	54	45	99	4	7	11	67	69	68	78	73	75	0.0664	0.3012	0.0437

† Chromosome number.
‡ Linkage Group.
§ Position of SNP on chromosome in centimorgans, based on USLP 1.0 map positions.
¶ Traits: LOD=Lodging score, %CY=Yield as Percent of Check, SDWT=100 seed weight in grams, OIL=Seed oil content, PRO=Seed protein content, MG=Relative Maturity.

Table 2.2 cont.

SNP	Chr †	LG‡	Pos§	Trait¶	Number of Individuals Homozygous for *G. max* allele			Number of Individuals Homozygous for *G. soja* allele			Mean of Individuals Homozygous for *G. max* allele			Mean of Individuals Homozygous for *G. soja* allele			Prob (F)		
					1	2	T	1	2	T	1	2	T	1	2	T	1	2	T
BARC-013699-01240	4	C1	97.4	PRO	49	21	70	6	28	34	429	441	432	420	423	422	0.2269	0.0005	0.0029
BARC-040651-07808	5	A1	2.5	%CY	16	34	50	38	19	57	64	68	67	71	72	72	0.0031	0.3298	0.0575
BARC-023411-05376	5	A1	4.1	OIL	6	48	54	49	5	54	163	164	164	175	171	174	0.0166	0.5966	0.0005
BARC-031311-07043	5	A1	4.1	%CY	6	43	49	49	10	59	60	68	67	70	76	71	0.0012	0.0484	0.0331
BARC-031311-07043	5	A1	4.1	OIL	6	43	49	49	10	59	163	166	165	175	170	174	0.0352	0.5476	0.0014
BARC-040459-07745	5	A1	36.1	%CY	52	43	95	8	7	15	67	68	67	72	73	73	0.1905	0.1316	0.0109
BARC-040459-07745	5	A1	36.1	SDWT	52	43	95	8	7	15	7.8	7.8	7.8	9.3	9.2	9.3	0.0067	0.0611	0.0013
BARC-018011-02495	5	A1	51.0	%CY	48	49	97	9	4	13	70	70	70	60	66	62	0.0286	0.7732	0.0242
BARC-026129-05274	5	A1	53.0	%CY	47	50	97	10	3	13	70	70	70	59	64	60	0.0126	0.6126	0.0219
BARC-035145-07126	5	A1	64.9	%CY	46	46	92	8	5	13	70	70	70	59	68	62	0.0237	0.8267	<0.0001
BARC-052097-11347	5	A1	71.8	PRO	34	20	54	25	22	47	429	436	431	425	419	422	0.2025	0.0033	0.0037
BARC-041257-07953	5	A1	78.8	LOD	14	23	37	43	30	73	2.8	2.6	2.7	3.1	2.7	3.0	0.0355	0.3728	0.0086
BARC-058665-17450	5	A1	80.5	%CY	25	22	47	28	15	43	72	69	71	65	68	66	0.0364	0.8134	0.0063
BARC-043209-08557	5	A1	80.5	%CY	27	45	72	28	10	38	71	70	70	65	67	65	0.0768	0.5987	0.0541
BARC-058653-17430	5	A1	80.6	%CY	27	33	60	28	17	45	71	71	71	65	68	66	0.0624	0.5049	0.0586
BARC-020855-03964	5	A1	80.7	%CY	27	44	71	28	10	38	71	70	70	65	67	65	0.0768	0.4863	0.0537
BARC-058783-15432	6	C2	14.5	SDWT	39	42	81	16	9	25	8.0	8.8	8.4	8.0	6.0	7.3	0.9621	<0.0001	0.0009
BARC-018563-02977	6	C2	30.8	%CY	30	27	57	21	18	39	72	72	72	64	64	64	0.0217	0.0047	0.0002
BARC-018563-02977	6	C2	30.8	SDWT	30	27	57	21	18	39	8.3	8.9	8.6	7.6	7.2	7.4	0.246	0.0008	0.0006
BARC-059985-16274	6	C2	34.4	%CY	40	28	68	16	16	32	68	72	70	67	63	65	0.8637	0.0071	0.0045
BARC-059985-16274	6	C2	34.4	SDWT	40	28	68	16	16	32	8.2	9.0	8.5	7.7	7.3	7.5	0.516	<0.0001	0.0004
BARC-044639-08743	6	C2	36.8	%CY	38	20	58	13	26	39	71	73	72	63	66	65	0.0048	0.0217	0.0301
BARC-024429-04882	6	C2	70.9	OIL	58	20	78	3	32	35	173	171	172	163	165	165	0.1778	0.2373	0.0094
BARC-048543-10663	6	C2	80.2	%CY	34	35	69	23	19	42	71	71	71	63	68	65	0.0054	0.0701	0.0326

† Chromosome number.
‡ Linkage Group.
§ Position of SNP on chromosome in centimorgans, based on USLP 1.0 map positions.
¶ Traits: LOD=Lodging score, %CY=Yield as Percent of Check, SDWT=100 seed weight in grams, OIL=Seed oil content, PRO=Seed protein content, MG=Relative Maturity.

Table 2.2 cont.

SNP	Chr †	LG‡	Pos§	Trait¶	Number of Individuals Homozygous for *G. max* allele			Number of Individuals Homozygous for *G. soja* allele			Mean of Individuals Homozygous for *G. max* allele			Mean of Individuals Homozygous for *G. soja* allele			Prob (F)		
					1	2	T	1	2	T	1	2	T	1	2	T	1	2	T
BARC-013837-01254	6	C2	86.3	%CY	51	45	96	4	6	10	70	71	70	57	62	60	0.0037	0.0839	0.0007
BARC-013837-01254	6	C2	86.3	LOD	51	45	96	4	6	10	3.0	2.6	2.8	3.6	2.9	3.2	0.0046	0.1972	0.0086
BARC-021735-04194	6	C2	90.9	%CY	44	41	85	17	12	29	71	71	71	59	69	63	<0.0001	0.035	0.0008
BARC-021735-04194	6	C2	90.9	LOD	44	41	85	17	12	29	2.9	2.6	2.7	3.4	2.9	3.2	<0.0001	0.0039	<0.0001
BARC-021735-04194	6	C2	90.9	OIL	44	41	85	17	12	29	176	169	173	163	160	162	0.0002	0.0028	<0.0001
BARC-031099-06997	6	C2	102.1	OIL	52	41	93	5	12	17	173	169	171	172	154	159	0.8896	0.0067	0.005
BARC-042161-08193	6	C2	102.6	OIL	52	40	92	5	12	17	173	169	171	172	154	159	0.6568	0.0063	0.0048
BARC-049601-09082	6	C2	102.9	OIL	52	40	92	6	12	18	173	169	171	171	154	160	0.7871	0.0063	0.006
BARC-038885-07387	6	C2	114.1	%CY	51	2	53	9	27	36	66	75	67	77	70	72	0.0052	0.5096	0.0548
BARC-038885-07387	6	C2	114.1	LOD	51	2	53	9	27	36	3.1	2.4	3.1	2.9	2.6	2.7	0.1336	0.5039	<0.0001
BARC-038885-07387	6	C2	114.1	SDWT	51	2	53	9	27	36	7.8	5.4	7.7	9.6	8.5	8.8	0.0004	0.0222	0.0021
BARC-039741-07564	7	M	12.7	LOD	43	47	90	10	1	11	2.9	2.7	2.8	3.4	3.3	3.4	0.0063	0.3391	0.0002
BARC-035447-07202	7	M	15.1	LOD	43	49	92	8	2	10	2.9	2.7	2.8	3.4	3.3	3.4	0.0068	0.1048	0.0002
BARC-900461-00929	7	M	39.1	MG	38	25	63	21	27	48	6.4	6.8	6.5	6.8	6.8	6.8	0.0106	0.5423	0.0089
BARC-044841-08824	7	M	45.1	SDWT	17	24	41	34	26	60	8.4	8.8	8.6	7.9	7.1	7.6	0.4877	0.0017	0.0055
BARC-012945-00406	7	M	47.4	SDWT	46	37	83	5	15	20	8.2	8.3	8.3	7.3	6.9	7.0	0.3387	0.0147	0.0066
BARC-060941-16977	7	M	57.3	MG	16	14	30	37	36	73	6.1	6.8	6.4	6.7	6.8	6.8	<0.0001	0.7208	0.0055
BARC-019987-03748	7	M	64.7	SDWT	16	16	32	30	42	72	7.6	6.9	7.3	8.5	8.4	8.4	0.1723	0.0024	0.002
BARC-062193-17703	7	M	65.8	SDWT	26	16	42	24	41	65	7.6	6.9	7.3	8.7	8.5	8.6	0.0173	0.0032	0.0001
BARC-028853-06030	8	A2	14.6	%CY	47	42	89	14	13	27	66	68	67	73	71	72	0.0314	0.1547	<0.0001
BARC-028853-06030	8	A2	14.6	SDWT	47	42	89	14	13	27	7.7	7.7	7.7	9.0	9.0	9.0	0.0035	0.0772	0.0013
BARC-021463-04108	8	A2	14.7	%CY	47	42	89	14	13	27	66	68	67	73	71	72	0.0314	0.1547	0.0221
BARC-021463-04108	8	A2	14.7	SDWT	47	42	89	14	13	27	7.7	7.7	7.7	9.0	9.0	9.0	0.0035	0.0772	0.0013
BARC-045199-08906	8	A2	31.2	MG	14	37	51	36	12	48	6.7	6.8	6.8	6.4	6.7	6.5	0.0261	0.812	0.0027

† Chromosome number.
‡ Linkage Group.
§ Position of SNP on chromosome in centimorgans, based on USLP 1.0 map positions.
¶ Traits: LOD=Lodging score, %CY=Yield as Percent of Check, SDWT=100 seed weight in grams, OIL=Seed oil content, PRO=Seed protein content, MG=Relative Maturity.

Table 2.2 cont.

SNP	Chr †	LG‡	Pos§	Trait¶	Number of Individuals Homozygous for *G. max* allele			Number of Individuals Homozygous for *G. soja* allele			Mean of Individuals Homozygous for *G. max* allele			Mean of Individuals Homozygous for *G. soja* allele			Prob (F)		
					1	2	T	1	2	T	1	2	T	1	2	T	1	2	T
BARC-045047-08867	8	A2	45.6	OIL	37	50	87	10	5	15	173	165	168	180	185	182	0.0391	0.0089	0.001
BARC-040339-07715	8	A2	47.0	OIL	40	46	86	9	5	14	173	165	168	183	185	184	0.001	0.0051	<0.0001
BARC-050171-09440	8	A2	47.3	OIL	26	36	62	16	12	28	172	164	167	179	177	178	0.0574	0.0271	0.0013
BARC-010097-00518	8	A2	47.6	OIL	29	47	76	15	5	20	172	165	167	178	185	179	0.1596	0.0086	0.0012
BARC-014325-01321	8	A2	63.8	%CY	57	45	102	4	10	14	68	68	68	72	74	73	0.409	0.0754	0.0474
BARC-022387-04319	8	A2	81.2	%CY	35	35	70	21	9	30	65	68	67	72	78	73	0.0535	0.024	0.0051
BARC-022387-04319	8	A2	81.2	OIL	35	35	70	21	9	30	168	164	166	177	169	175	0.0065	0.2468	0.0033
BARC-063283-18296	8	A2	94.1	OIL	58	35	93	-	15	15	172	170	171	-	155	155	-	0.0058	<0.0001
BARC-062129-17664	8	A2	103.9	%CY	45	37	82	7	11	18	70	70	70	60	63	62	0.0043	0.0348	0.052
BARC-062129-17664	8	A2	103.9	OIL	45	37	82	7	11	18	175	170	173	162	149	154	0.0222	<0.0001	<0.0001
BARC-050015-09290	8	A2	110.5	%CY	45	38	83	15	14	29	71	70	70	60	66	63	0.0004	0.1567	0.0204
BARC-050015-09290	8	A2	110.5	OIL	45	38	83	15	14	29	175	170	173	164	156	160	0.0021	0.0069	<0.0001
BARC-050015-09290	8	A2	110.5	SDWT	45	38	83	15	14	29	8.2	8.4	8.3	7.6	6.7	7.2	0.1435	0.0049	0.0044
BARC-055141-13084	9	K	6.9	SDWT	5	18	23	55	39	94	7.3	7.1	7.1	8.1	8.4	8.2	0.2342	0.0111	0.0045
BARC-055301-13192	9	K	23.3	LOD	11	31	42	41	22	63	2.9	2.6	2.7	3.1	2.8	3.0	0.2375	0.2878	0.0035
BARC-048623-10678	9	K	39.9	%CY	29	35	64	25	17	42	71	71	71	65	69	67	0.1082	0.1539	0.0285
BARC-047863-10415	9	K	43.8	SDWT	33	38	71	17	15	32	7.8	7.8	7.8	8.5	9.0	8.8	0.301	0.0115	0.0092
BARC-055355-13228	9	K	45.7	SDWT	35	34	69	12	16	28	7.7	7.6	7.7	9.0	9.0	9.0	0.0291	0.0041	0.0006
BARC-030181-06822	9	K	74.6	%CY	35	33	68	16	15	31	71	70	71	63	69	66	0.0166	0.8479	0.0045
BARC-052125-11359	9	K	81.3	%CY	26	14	40	33	38	71	64	67	65	71	71	71	0.046	0.1551	0.0273
BARC-051769-11236	9	K	93.8	SDWT	12	16	28	42	35	77	7.7	6.9	7.2	8.1	8.7	8.4	0.5993	0.0021	0.0038
BARC-062909-18153	10	O	11.0	LOD	41	20	61	10	26	36	3.1	2.8	3.0	2.8	2.5	2.6	0.3032	0.1018	0.0019
BARC-035255-07160	10	O	41.8	%CY	47	15	62	9	29	38	66	65	65	76	73	73	0.0054	0.0203	0.0576
BARC-035255-07160	10	O	41.8	LOD	47	15	62	9	29	38	3.1	2.7	3.0	3.0	2.6	2.7	0.0961	0.6277	0.0011

† Chromosome number.
‡ Linkage Group.
§ Position of SNP on chromosome in centimorgans, based on USLP 1.0 map positions.
¶ Traits: LOD=Lodging score, %CY=Yield as Percent of Check, SDWT=100 seed weight in grams, OIL=Seed oil content, PRO=Seed protein content, MG=Relative Maturity.

Table 2.2 cont.

SNP	Chr †	LG‡	Pos§	Trait¶	Number of Individuals Homozygous for *G. max* allele			Number of Individuals Homozygous for *G. soja* allele			Mean of Individuals Homozygous for *G. max* allele			Mean of Individuals Homozygous for *G. soja* allele			Prob (F)		
					1	2	T	1	2	T	1	2	T	1	2	T	1	2	T
BARC-019105-03305	10	O	50.1	%CY	53	23	76	5	22	27	67	68	67	82	72	74	0.0019	0.3173	0.0308
BARC-063365-18347	10	O	51.0	%CY	51	22	73	10	22	32	66	68	67	75	72	73	0.0118	0.3784	0.0423
BARC-063365-18347	10	O	51.0	SDWT	51	22	73	10	22	32	7.8	7.8	7.8	9.2	8.4	8.6	0.0031	0.1108	0.0047
BARC-059885-16192	10	O	51.4	%CY	51	23	74	9	22	31	66	68	67	77	72	74	0.002	0.3301	0.0578
BARC-059885-16192	10	O	51.4	LOD	51	23	74	9	22	31	3.1	2.7	3.0	2.7	2.6	2.7	0.0213	0.7978	0.0081
BARC-059885-16192	10	O	51.4	SDWT	51	23	74	9	22	31	7.8	7.7	7.8	9.6	8.4	8.7	0.0003	0.1274	0.0026
BARC-060257-16508	10	O	69.8	LOD	60	26	86	-	24	24	3.0	2.7	2.9	-	2.6	2.6	-	0.2288	0.0057
BARC-065663-19627	10	O	99.0	%CY	16	35	51	28	16	44	71	71	71	67	66	67	0.2643	0.1869	0.0377
BARC-008241-00035	10	O	116.3	MG	26	40	66	19	10	29	6.2	6.7	6.5	6.8	7.0	6.8	0.0015	0.1789	0.0091
BARC-054375-12539	10	O	120.7	%CY	23	37	60	32	19	51	73	70	71	65	68	66	0.0054	0.7755	<0.0001
BARC-054375-12539	10	O	120.7	LOD	23	37	60	32	19	51	2.8	2.7	2.7	3.2	2.7	3.0	0.001	0.9383	0.0055
BARC-050205-09457	11	B1	76.9	LOD	53	39	92	4	13	17	3.0	2.8	2.9	3.0	2.5	2.6	0.7189	0.0688	0.0099
BARC-042299-08241	11	B1	92.5	LOD	46	16	62	12	36	48	3.1	2.7	3.0	2.8	2.7	2.7	0.0606	0.4501	0.0021
BARC-062275-17736	11	B1	109.9	%CY	58	37	95	1	15	16	69	72	70	51	65	64	0.136	0.0434	0.0222
BARC-050545-09732	11	B1	112.0	LOD	47	17	64	12	24	36	3.1	2.7	3.0	2.8	2.7	2.7	0.2717	0.7122	0.0083
BARC-027560-06613	12	H	9.8	%CY	31	34	65	23	17	40	70	70	70	63	69	65	0.0081	0.9044	0.0009
BARC-027560-06613	12	H	9.8	LOD	31	34	65	23	17	40	2.9	2.6	2.8	3.2	2.8	3.1	0.03	0.2519	0.0045
BARC-052799-11625	12	H	31.4	%CY	52	42	94	5	9	14	70	70	70	59	68	65	0.0054	0.7452	0.0156
BARC-030145-06814	12	H	33.9	%CY	52	43	95	7	9	16	70	70	70	57	68	63	0.0026	0.7873	0.0485
BARC-030145-06814	12	H	33.9	SDWT	52	43	95	7	9	16	8.1	8.4	8.3	7.2	6.6	6.9	0.1981	0.0172	0.0034
BARC-056205-14140	12	H	57.8	OIL	60	27	87	-	13	13	172	168	171	-	157	157	-	0.1035	0.004
BARC-029055-06058	12	H	60.3	OIL	60	31	91	-	12	12	172	167	171	-	156	156	-	0.054	0.0022
BARC-029055-06058	12	H	60.3	PRO	60	31	91	-	12	12	427	428	427	-	439	439	-	0.0414	0.0057
BARC-055907-13843	12	H	61.7	%CY	43	22	65	10	30	40	65	67	66	81	72	74	<0.0001	0.2613	0.0008

† Chromosome number.
‡ Linkage Group.
§ Position of SNP on chromosome in centimorgans, based on USLP 1.0 map positions.
¶ Traits: LOD=Lodging score, %CY=Yield as Percent of Check, SDWT=100 seed weight in grams, OIL=Seed oil content, PRO=Seed protein content, MG=Relative Maturity.

Table 2.2 cont.

SNP	Chr †	LG‡	Pos§	Trait¶	Number of Individuals Homozygous for *G. max* allele			Number of Individuals Homozygous for *G. soja* allele			Mean of Individuals Homozygous for *G. max* allele			Mean of Individuals Homozygous for *G. soja* allele			Prob (F)		
					1	2	T	1	2	T	1	2	T	1	2	T	1	2	T
BARC-055907-13843	12	H	61.7	LOD	43	22	65	10	30	40	3.1	2.7	3.0	2.6	2.7	2.7	0.0043	0.4672	0.0026
BARC-055907-13843	12	H	61.7	SDWT	43	22	65	10	30	40	7.8	7.6	7.8	9.5	8.5	8.8	0.0018	0.0405	0.0008
BARC-061985-17608	12	H	62.1	%CY	30	22	52	19	30	49	64	67	65	75	72	73	0.0012	0.3097	0.0088
BARC-900926-00961	13	F	5.1	SDWT	29	34	63	27	18	45	8.1	7.4	7.7	8.1	9.4	8.6	0.9021	<0.0001	0.0054
BARC-055229-13122	13	F	71.9	PRO	47	27	74	11	26	37	427	424	426	429	436	434	0.6423	0.0209	0.009
BARC-028583-05961	13	F	75.9	SDWT	31	31	62	22	23	45	7.9	7.3	7.6	8.3	8.7	8.5	0.5962	0.0057	0.008
BARC-065455-19481	14	B2	39.4	%CY	47	36	83	7	17	24	69	73	71	59	64	63	0.0449	0.0094	0.0316
BARC-044549-08718	14	B2	40.0	%CY	47	36	83	7	14	21	69	73	71	59	67	64	0.0505	0.0177	0.002
BARC-048959-10760	14	B2	63.5	%CY	47	40	87	10	14	24	70	70	70	63	65	64	0.1391	0.1349	0.0008
BARC-048959-10760	14	B2	63.5	OIL	47	40	87	10	14	24	175	168	172	159	158	159	0.0007	0.1078	0.0002
BARC-052789-11619	14	B2	65.2	OIL	44	29	73	14	22	36	174	167	172	165	162	163	0.0199	0.1455	0.0017
BARC-064301-18614	14	B2	86.9	%CY	36	32	68	11	14	25	66	68	67	73	71	72	0.0868	0.3777	0.014
BARC-018959-03045	15	E	5.7	%CY	24	43	67	22	10	32	70	71	71	65	64	65	0.2615	0.1314	0.0066
BARC-055329-13210	15	E	15.6	%CY	33	38	71	17	14	31	69	72	71	67	65	66	0.4045	0.0145	0.0039
BARC-027786-06670	15	E	33.2	LOD	18	36	54	38	12	50	2.9	2.6	2.7	3.1	2.6	3.0	0.1618	0.1173	0.0065
BARC-025663-04988	15	E	47.9	%CY	18	6	24	37	43	80	72	73	72	66	68	67	0.1014	0.1903	0.0054
BARC-025663-04988	15	E	47.9	OIL	18	6	24	37	43	80	178	176	178	170	164	167	0.0451	0.122	0.0024
BARC-025663-04988	15	E	47.9	PRO	18	6	24	37	43	80	421	420	420	430	432	431	0.0201	0.1888	0.0044
BARC-025663-04988	15	E	47.9	SDWT	18	6	24	37	43	80	9.0	9.1	9.0	7.7	7.9	7.8	0.003	0.3115	0.0033
BARC-050109-09389	15	E	54.9	%CY	38	22	60	13	10	23	70	72	71	61	69	64	0.012	0.3497	0.0214
BARC-050109-09389	15	E	54.9	OIL	38	22	60	13	10	23	175	172	174	163	159	162	0.0084	0.0152	0.0003
BARC-053201-11762	15	E	76.6	PRO	48	39	87	2	10	12	427	428	427	440	441	441	0.1198	0.0504	0.0063
BARC-056287-14214	15	E	76.9	PRO	50	40	90	2	10	12	426	428	427	440	441	441	0.2801	0.0735	0.0065
BARC-059221-15678	15	E	77.4	PRO	50	40	90	2	10	12	426	428	427	440	441	441	0.2042	0.0735	0.0067

† Chromosome number.
‡ Linkage Group.
§ Position of SNP on chromosome in centimorgans, based on USLP 1.0 map positions.
¶ Traits: LOD=Lodging score, %CY=Yield as Percent of Check, SDWT=100 seed weight in grams, OIL=Seed oil content, PRO=Seed protein content, MG=Relative Maturity.

Table 2.2 cont.

SNP	Chr †	LG‡	Pos§	Trait¶	Number of Individuals Homozygous for *G. max* allele			Number of Individuals Homozygous for *G. soja* allele			Mean of Individuals Homozygous for *G. max* allele			Mean of Individuals Homozygous for *G. soja* allele			Prob (F)		
					1	2	T	1	2	T	1	2	T	1	2	T	1	2	T
BARC-040185-07678	15	E	77.7	PRO	50	40	90	2	10	12	426	428	427	440	441	441	0.251	0.0735	0.0068
BARC-028221-05799	15	E	79.3	OIL	43	28	71	8	27	35	172	173	173	170	160	162	0.7702	0.001	0.0008
BARC-023181-03804	15	E	81.1	OIL	52	24	76	2	27	29	172	173	172	165	160	160	0.4992	0.0024	0.0002
BARC-055571-13451	15	E	81.6	SDWT	30	10	40	25	43	68	7.8	6.2	7.4	8.4	8.5	8.4	0.4308	<0.0001	0.0041
BARC-052379-11435	15	E	86.5	PRO	32	27	59	24	30	54	426	424	425	428	436	433	0.2421	0.007	0.0024
BARC-020425-04614	15	E	86.6	PRO	33	27	60	24	30	54	426	424	425	428	436	433	0.6412	0.009	0.0084
BARC-022009-04249	15	E	91.3	LOD	54	34	88	4	11	15	3.1	2.7	2.9	2.7	2.5	2.6	0.293	0.4338	0.0089
BARC-023739-03474	16	J	11.1	SDWT	44	31	75	13	15	28	7.9	7.4	7.7	8.6	9.0	8.8	0.2678	0.0188	0.0055
BARC-045157-08897	16	J	25.3	%CY	26	19	45	29	30	59	63	67	65	72	71	71	0.0047	0.4698	0.0045
BARC-045157-08897	16	J	25.3	SDWT	26	19	45	29	30	59	7.6	7.3	7.5	8.4	8.4	8.4	0.1604	0.0638	0.0098
BARC-054099-12340	16	J	25.8	SDWT	13	18	31	34	30	64	7.4	7.0	7.2	8.3	8.7	8.5	0.2071	0.0057	0.0011
BARC-060265-16514	16	J	26.0	%CY	43	37	80	13	18	31	69	71	70	67	66	66	0.0702	0.1577	0.0009
BARC-064455-18689	16	J	71.6	OIL	23	34	57	38	12	50	169	165	167	174	176	175	0.1462	0.0467	0.0045
BARC-041945-08143	17	D2	24.4	SDWT	51	42	93	4	6	10	8.1	8.1	8.1	7.6	5.4	6.3	0.8015	0.0001	0.0009
BARC-058841-15463	17	D2	33.0	SDWT	50	35	85	5	12	17	8.1	8.3	8.2	7.2	6.6	6.8	0.4264	0.0012	0.0004
BARC-056481-14397	17	D2	38.0	PRO	32	36	68	19	11	30	432	434	433	419	426	421	0.0006	0.0854	0.0004
BARC-062955-18179	17	D2	39.5	PRO	32	37	69	20	13	33	432	434	433	419	427	422	0.001	0.1702	0.0018
BARC-012687-00367	17	D2	42.4	PRO	35	30	65	17	17	34	431	434	432	418	423	420	0.0013	0.0793	0.0005
BARC-025885-05138	17	D2	55.6	SDWT	24	46	70	25	4	29	8.6	8.4	8.5	7.5	6.9	7.4	0.0436	0.0159	0.0016
BARC-017525-03061	17	D2	56.3	SDWT	26	48	74	19	3	22	8.5	8.3	8.4	7.2	7.3	7.2	0.0178	0.1177	0.0033
BARC-013969-01290	17	D2	62.8	%CY	9	27	36	46	22	68	78	69	72	66	69	67	0.0015	0.9859	0.0497
BARC-013969-01290	17	D2	62.8	SDWT	9	27	36	46	22	68	9.3	8.5	8.7	7.7	7.6	7.7	0.01	0.2082	0.0078
BARC-021991-04246	17	D2	72.2	SDWT	40	37	77	13	7	20	8.2	8.7	8.4	7.4	7.1	7.3	0.1788	0.0014	0.0004
BARC-042295-08238	17	D2	72.6	%CY	40	35	75	12	11	23	70	71	70	64	67	65	0.0927	0.4446	0.0307

† Chromosome number.
‡ Linkage Group.
§ Position of SNP on chromosome in centimorgans, based on USLP 1.0 map positions.
¶ Traits: LOD=Lodging score, %CY=Yield as Percent of Check, SDWT=100 seed weight in grams, OIL=Seed oil content, PRO=Seed protein content, MG=Relative Maturity.

Table 2.2 cont.

SNP	Chr †	LG‡	Pos§	Trait¶	Number of Individuals Homozygous for *G. max* allele			Number of Individuals Homozygous for *G. soja* allele			Mean of Individuals Homozygous for *G. max* allele			Mean of Individuals Homozygous for *G. soja* allele			Prob (F)		
					1	2	T	1	2	T	1	2	T	1	2	T	1	2	T
BARC-042295-08238	17	D2	72.6	SDWT	40	35	75	12	11	23	8.2	8.7	8.5	7.4	7.2	7.3	0.2217	<0.0001	<0.0001
BARC-056215-14153	17	D2	73.7	SDWT	41	37	78	12	8	20	8.2	8.5	8.4	7.4	7.0	7.2	0.2332	0.0041	0.0015
BARC-061049-17016	17	D2	77.4	OIL	8	17	25	36	31	67	177	158	164	173	172	172	0.0416	0.0051	0.0063
BARC-013709-01242	17	D2	99.7	SDWT	40	49	89	9	5	14	8.4	8.2	8.3	7.2	6.8	7.1	0.04	0.1622	0.0097
BARC-048855-10738	17	D2	109.3	LOD	24	37	61	31	17	48	2.9	2.6	2.7	3.1	2.8	3.0	0.0276	0.1688	0.0014
BARC-055551-13421	18	G	8.2	%CY	30	37	67	23	15	38	71	70	70	64	65	64	0.0731	0.0385	0.0001
BARC-055551-13421	18	G	8.2	OIL	30	37	67	23	15	38	176	170	173	166	157	163	0.0045	0.0133	0.0011
BARC-048275-10534	18	G	9.9	%CY	23	35	58	32	21	53	72	71	72	65	66	66	0.0391	0.1629	0.003
BARC-048275-10534	18	G	9.9	LOD	23	35	58	32	21	53	2.8	2.6	2.7	3.2	2.9	3.0	0.0371	0.0199	0.0004
BARC-G01477-00243	18	G	10.0	%CY	17	36	53	40	15	55	71	70	70	68	64	67	0.3114	0.0231	0.0536
BARC-015067-02556	18	G	12.0	OIL	30	21	51	24	22	46	175	174	174	171	162	167	0.4229	0.0037	0.0015
BARC-901121-00988	18	G	40.4	LOD	7	30	37	52	16	68	2.7	2.6	2.6	3.1	2.8	3.0	0.0598	0.2048	<0.0001
BARC-063581-18909	18	G	41.2	LOD	6	27	33	53	19	72	3.1	2.6	2.7	3.1	2.7	3.0	0.0345	0.5528	0.0011
BARC-051587-11167	18	G	44.1	%CY	17	20	37	42	28	70	65	66	66	70	72	71	0.2084	0.054	0.006
BARC-063705-18440	18	G	54.4	%CY	18	26	44	38	14	52	64	68	66	70	75	71	0.0575	0.0417	0.0394
BARC-047504-12947	18	G	55.6	LOD	6	22	28	53	30	83	2.8	2.7	2.7	3.1	2.7	2.9	0.1678	0.1483	0.0032
BARC-062847-18118	18	G	56.2	%CY	24	35	59	25	12	37	71	71	71	65	67	66	0.0907	0.1573	0.0167
BARC-062847-18118	18	G	56.2	LOD	24	35	59	25	12	37	2.9	2.6	2.7	3.2	2.9	3.1	0.171	0.0349	0.0014
BARC-057809-14935	18	G	56.7	%CY	24	37	61	27	12	39	72	70	71	63	67	64	0.0036	0.3156	0.0004
BARC-065333-19350	18	G	56.7	%CY	22	38	60	28	13	41	72	70	70	64	67	65	0.0128	0.3998	0.0499
BARC-057809-14935	18	G	56.7	LOD	24	37	61	27	12	39	2.9	2.6	2.7	3.2	2.9	3.1	0.0553	0.0969	0.0002
BARC-065333-19350	18	G	56.7	LOD	22	38	60	28	13	41	2.9	2.6	2.7	3.2	2.9	3.1	0.0596	0.0908	0.0006
BARC-057809-14935	18	G	56.7	SDWT	24	37	61	27	12	39	8.5	8.5	8.5	7.7	6.9	7.5	0.1785	0.0241	0.0099
BARC-015633-02774	18	G	64.3	LOD	23	34	57	24	13	37	2.8	2.6	2.7	3.1	2.8	3.0	0.0541	0.22	0.004

† Chromosome number.
‡ Linkage Group.
§ Position of SNP on chromosome in centimorgans, based on USLP 1.0 map positions.
¶ Traits: LOD=Lodging score, %CY=Yield as Percent of Check, SDWT=100 seed weight in grams, OIL=Seed oil content, PRO=Seed protein content, MG=Relative Maturity.

Table 2.2 cont.

SNP	Chr †	LG‡	Pos§	Trait¶	Number of Individuals Homozygous for *G. max* allele			Number of Individuals Homozygous for *G. soja* allele			Mean of Individuals Homozygous for *G. max* allele			Mean of Individuals Homozygous for *G. soja* allele			Prob (F)		
					1	2	T	1	2	T	1	2	T	1	2	T	1	2	T
BARC-044363-08678	18	G	100.4	MG	44	25	69	5	22	27	6.5	6.7	6.5	6.8	6.9	6.9	0.2481	0.153	0.0076
BARC-021603-04153	18	G	104.2	MG	44	40	84	4	11	15	6.5	6.7	6.6	7.1	7.0	7.0	0.0791	0.1818	0.0073
BARC-044415-08701	19	L	55.9	%CY	28	25	53	20	27	47	69	72	71	66	66	66	0.3488	0.0207	0.003
BARC-030101-06809	19	L	84.1	%CY	44	54	98	12	3	15	70	70	70	62	62	62	0.0255	0.4018	0.0268
BARC-030101-06809	19	L	84.1	LOD	44	54	98	12	3	15	2.9	2.6	2.8	3.3	3.2	3.3	0.0486	0.0631	0.0003
BARC-029419-06181	19	L	84.5	%CY	44	55	99	10	3	13	70	70	70	62	62	62	0.0301	0.3887	0.0142
BARC-029419-06181	19	L	84.5	LOD	44	55	99	10	3	13	3.0	2.6	2.8	3.2	3.2	3.2	0.2548	0.0664	0.0038
BARC-055107-13809	19	L	89.8	%CY	33	43	76	23	7	30	72	70	71	62	61	62	0.0004	0.0241	0.0442
BARC-055107-13809	19	L	89.8	LOD	33	43	76	23	7	30	2.8	2.6	2.7	3.4	3.0	3.3	<0.0001	0.0453	<0.0001
BARC-029051-06057	19	L	90.3	%CY	36	43	79	22	6	28	72	70	71	62	64	62	<0.0001	0.326	0.0134
BARC-029051-06057	19	L	90.3	LOD	36	43	79	22	6	28	2.8	2.6	2.7	3.4	3.0	3.3	<0.0001	0.0545	<0.0001
BARC-013129-01447	19	L	90.4	LOD	42	51	93	12	2	14	2.9	2.6	2.8	3.3	2.9	3.3	0.0029	0.3503	<0.0001
BARC-056173-14137	19	L	95.3	LOD	38	50	88	14	4	18	2.9	2.6	2.7	3.2	3.2	3.2	0.1051	0.0045	0.0002
BARC-065047-19054	20	I	1.0	%CY	43	40	83	16	13	29	71	70	70	62	69	65	0.0022	0.5808	0.0109
BARC-027552-06609	20	I	18.9	%CY	27	23	50	23	26	49	64	70	67	72	70	71	0.0297	0.1155	0.0265
BARC-046420-12581	20	I	25.8	PRO	43	46	89	11	8	19	426	426	426	435	448	440	0.071	0.0005	0.0003
BARC-029827-06444	20	I	29.6	PRO	51	46	97	4	6	10	426	426	426	441	449	445	0.051	0.002	<0.0001
BARC-060115-16397	20	I	29.6	PRO	51	46	97	4	6	10	426	426	426	441	449	445	0.051	0.0004	<0.0001
BARC-023131-03782	20	I	30.0	OIL	45	39	84	9	12	21	173	171	172	169	153	159	0.6029	0.0005	0.001
BARC-023131-03782	20	I	30.0	PRO	45	39	84	9	12	21	426	423	424	434	447	442	0.093	<0.0001	<0.0001
BARC-029803-06418	20	I	50.8	PRO	49	43	92	7	8	15	426	426	426	430	446	439	0.5192	0.0027	0.0033

† Chromosome number.
‡ Linkage Group.
§ Position of SNP on chromosome in centimorgans, based on USLP 1.0 map positions.
¶ Traits: LOD=Lodging score, %CY=Yield as Percent of Check, SDWT=100 seed weight in grams, OIL=Seed oil content, PRO=Seed protein content, MG=Relative Maturity.

Table 2.3 Results of model selection and multiple linear regression of loci associated with increased percent of check yield derived from *G. soja*. R-square value of the model was 0.37.

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	59.2	1.8	32.98	<.0001
BARC-031311-07043	6.5	1.9	3.46	0.0009
BARC-063365-18347	2.5	2.2	1.16	0.2514
BARC-017645-02642	8.1	2.0	4.02	0.0001
BARC-035255-07160	8.0	2.0	3.97	0.0002

Table 2.4 Results of model selection and multiple linear regression of loci associated with increased 100-seed weight derived from *G. soja*. R –square value of the model was 0.32.

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	6.0	0.4	13.54	<.0001
BARC-029185-06108	1.0	0.4	2.87	0.0056
BARC-055141-13084	1.6	0.4	3.93	0.0002
BARC-028853-06030	1.2	0.4	3.11	0.0028
BARC-062193-17703	0.2	0.2	0.73	0.4653

Table 2.5 Results of model selection and multiple linear regression of loci associated with increased seed oil content derived from *G. soja*. R-square value of the model was 0.34.

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	157	2.8	56.34	<.0001
BARC-061049-17016	7	3.0	2.43	0.0174
BARC-031311-07043	6	2.6	2.11	0.0379
BARC-022387-04319	6	3.0	2.05	0.0434
BARC-050171-09440	10	2.8	3.71	0.0004

Table 2.6 Results of model selection and multiple linear regression of loci associated with increased seed protein content derived from *G. soja*. R-sqaure of the model was 0.43.

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	415	2.5	163.17	<.0001
BARC-055229-13122	10	2.7	3.63	0.0005
BARC-025663-04988	9	3.0	3.22	0.0018
BARC-053201-11762	13	4.0	3.14	0.0023
BARC-060115-16397	17	4.0	4.26	<.0001

Table 2.7 Results of model selection and multiple linear regression of loci associated with lodging tolerance derived from *G. soja*. R-square of the model was 0.21

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	3.1	0.1	45.97	<.0001
BARC-055907-13843	-0.2	0.1	-2.34	0.0211
BARC-059885-16192	-0.3	0.1	-2.56	0.0120
BARC-050205-09457	-0.2	0.1	-1.51	0.1333
BARC-050545-09732	-0.2	0.1	-1.80	0.0748

Table 2.8 Summary of SNP diversity collection created from JCD-1

Percent Check Yield	Number of Unique SNPs	Number of Breeding Lines
>90%	261	10
>80%	362	14
>70%	520	46
>60%	539	87
>50%	548	102

Table 2.9 Summary of SNP diversity collection created from JCD-2

Percent Check Yield	Number of Unique SNPs	Number of Breeding Lines
>90%	114	1
>80%	472	13
>70%	541	45
>60%	557	81
>50%	557	88

Table 2.10 Summary of SNP diversity collection created from both populations

Percent Check Yield	Number of Unique SNPs	Number of Breeding Lines
>90%	287	11
>80%	513	27
>70%	556	91
>60%	558	168

Table 2.11 Summary of SNP diversity collection created by breeding lines that yield within 70% of the *G. max* checks. Collection contains at least 1 homoyzgote for the *G. soja* allele at 556 SNP loci of a total of 558.

Breeding Line	Number of Homozygous *G. soja* SNPs	Percent *G. soja*	Percent Check Yield
NMS5-115-4-247	173	36	74
NMS4-20-273	176	35	74
NMS5-101-5-206	177	37	78
NMS4-97-502	141	33	72
NMS5-85-2-147	153	31	86
NMS4-37-305	180	37	70
NMS5-108-2-226	168	37	72
NMS5-98-3-192	169	36	83
NMS5-37-10-70	149	32	80
NMS4-22-280	176	35	76
NMS4-67-462-1	171	34	79
NMS5-102-1-207	159	34	75
NMS5-253-1-537	177	40	86
NMS4-44-329	160	37	83
NMS4-38-316	155	34	79
NMS4-123-651	163	35	77

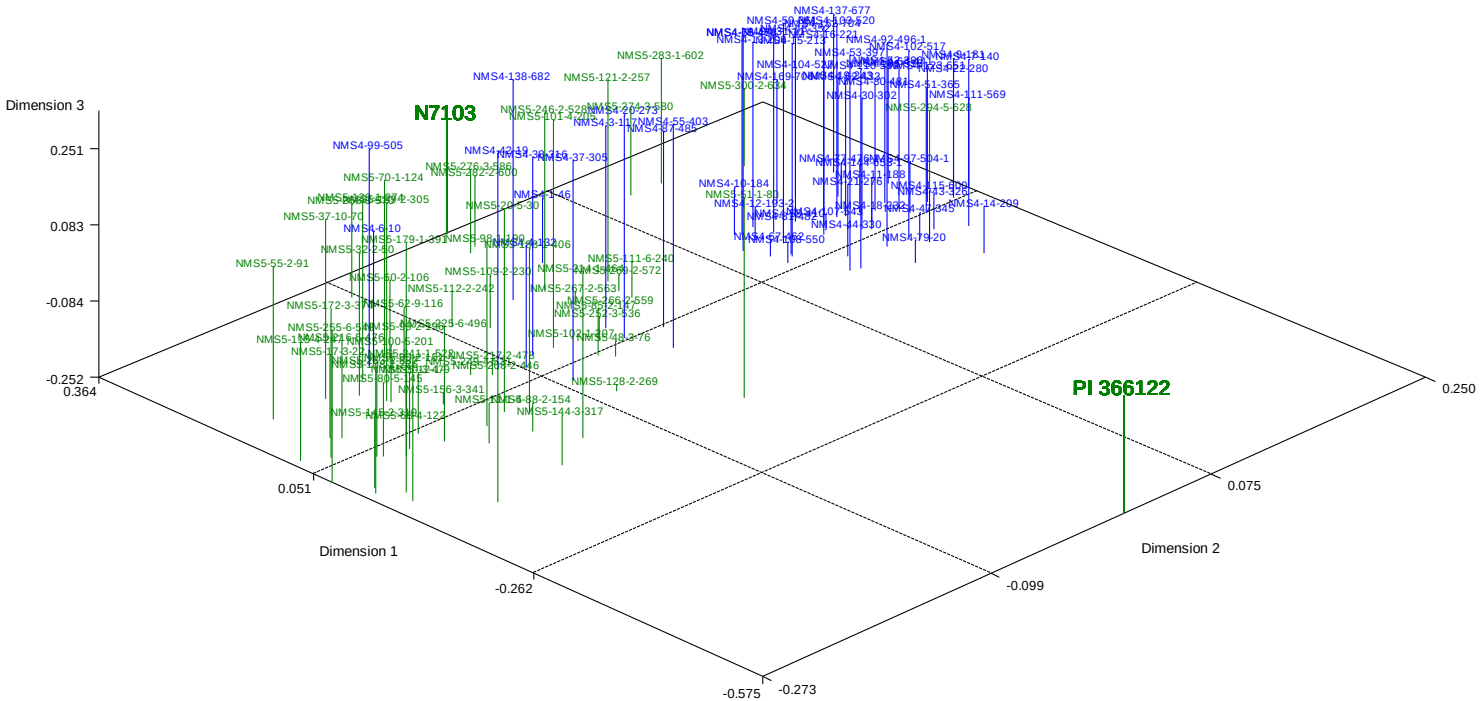
FIGURES

Figure 2.1 MDS plot of 120 breeding lines from population JCD-1(Blue) and JCD-2 (Green) and parental lines N7103 and *G. soja* PI 366122.

MDS Plot of Populations JCD-1 and JCD-2

558 SNPs

Tilt 30
Rotate 45



MDS Plot of Populations JCD-1 and JCD-2

558 SNPs

Tilt 45
Rotate 70

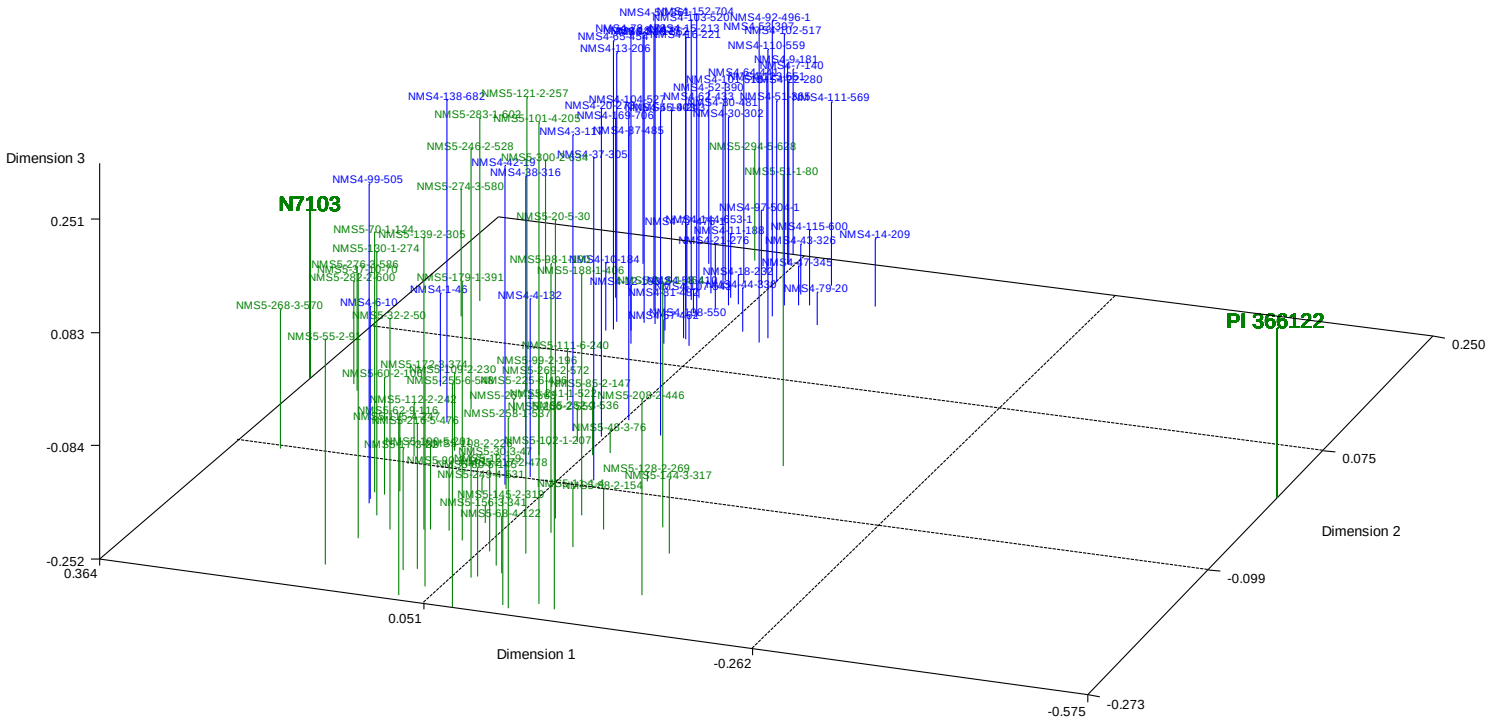
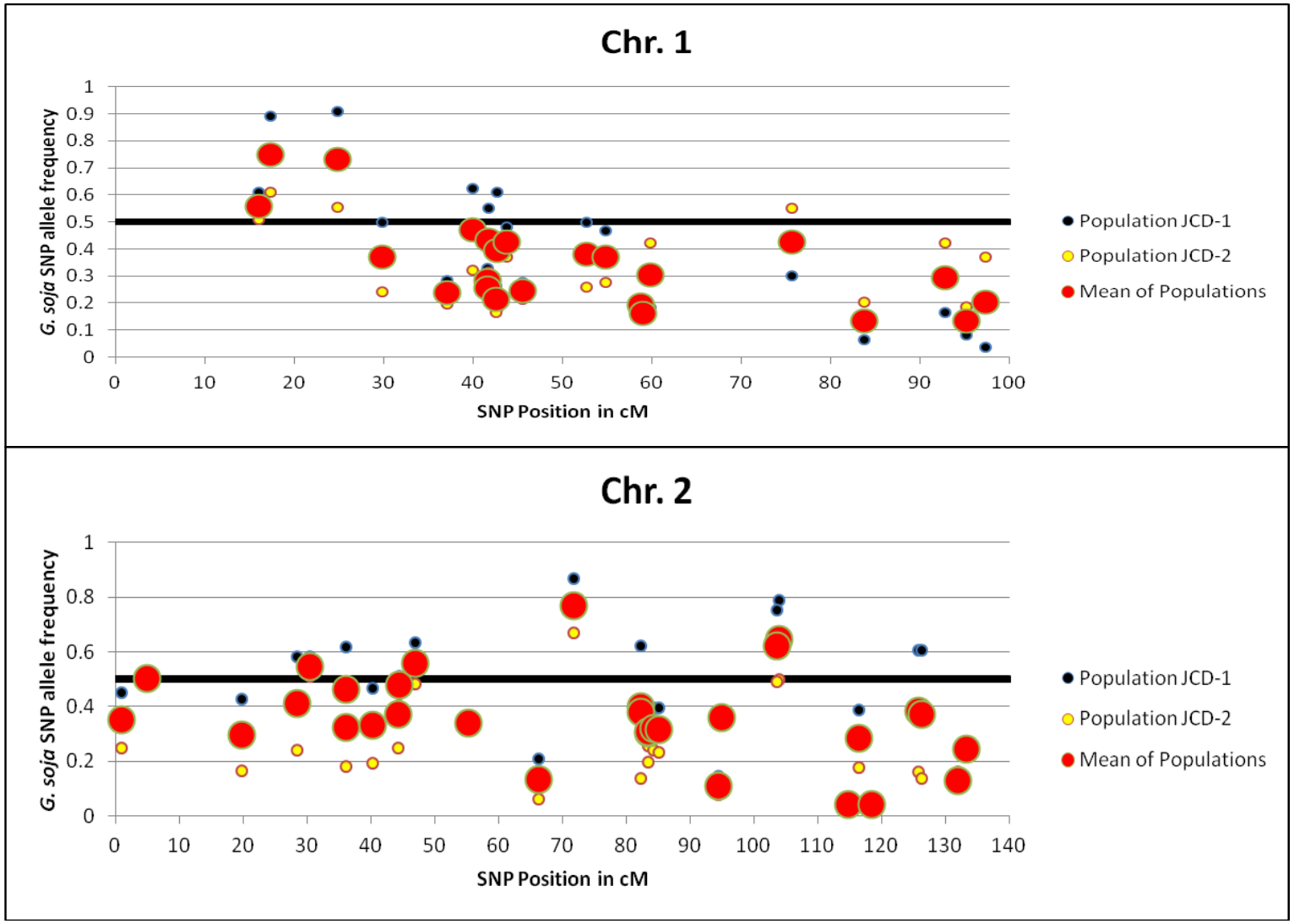
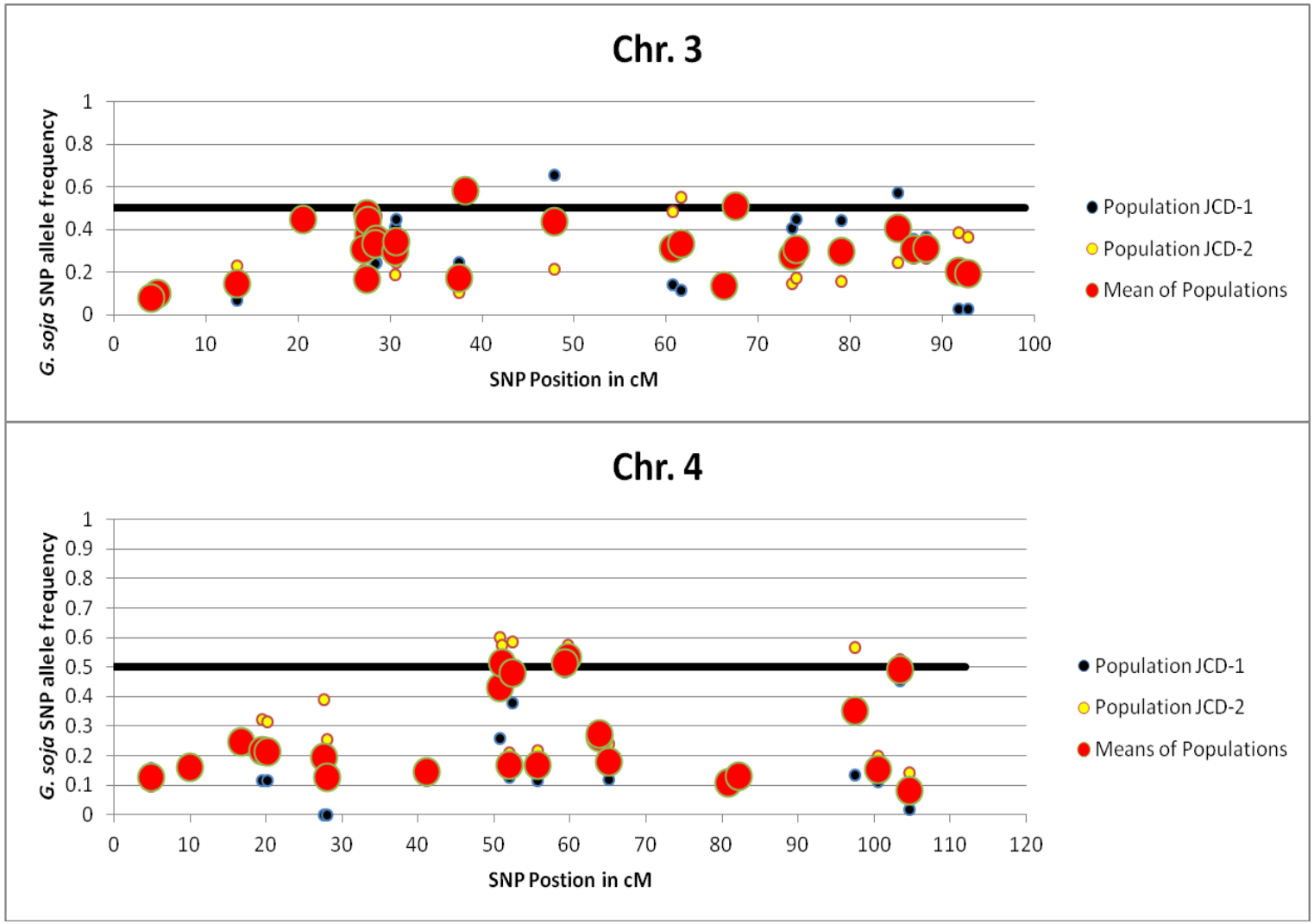
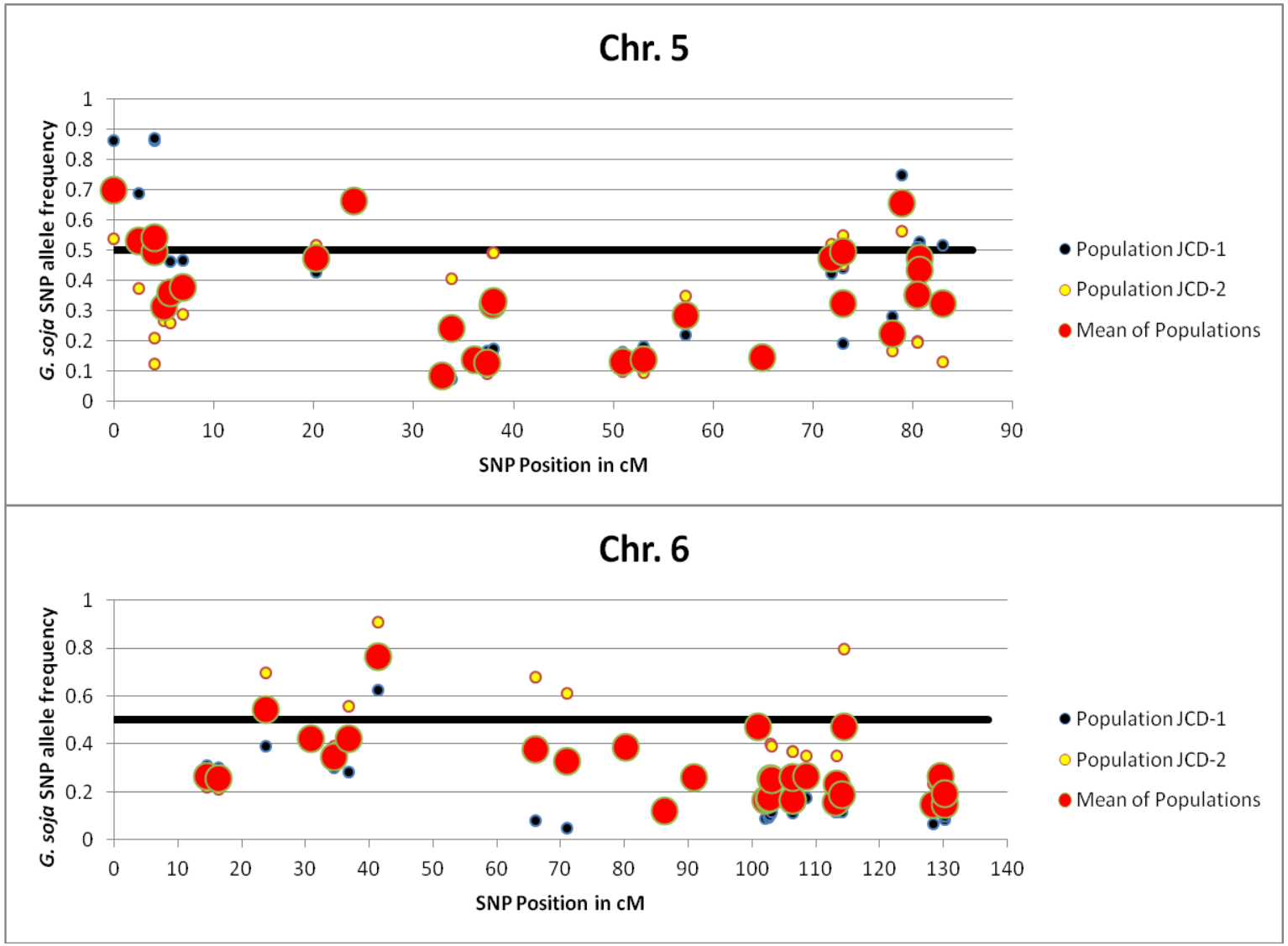
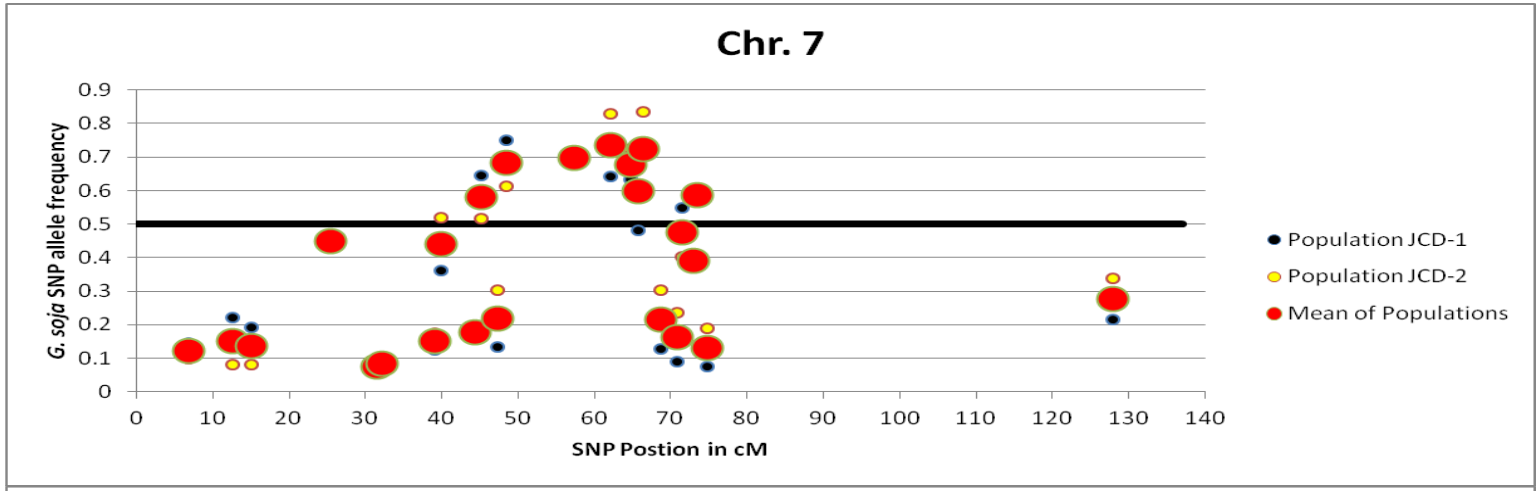


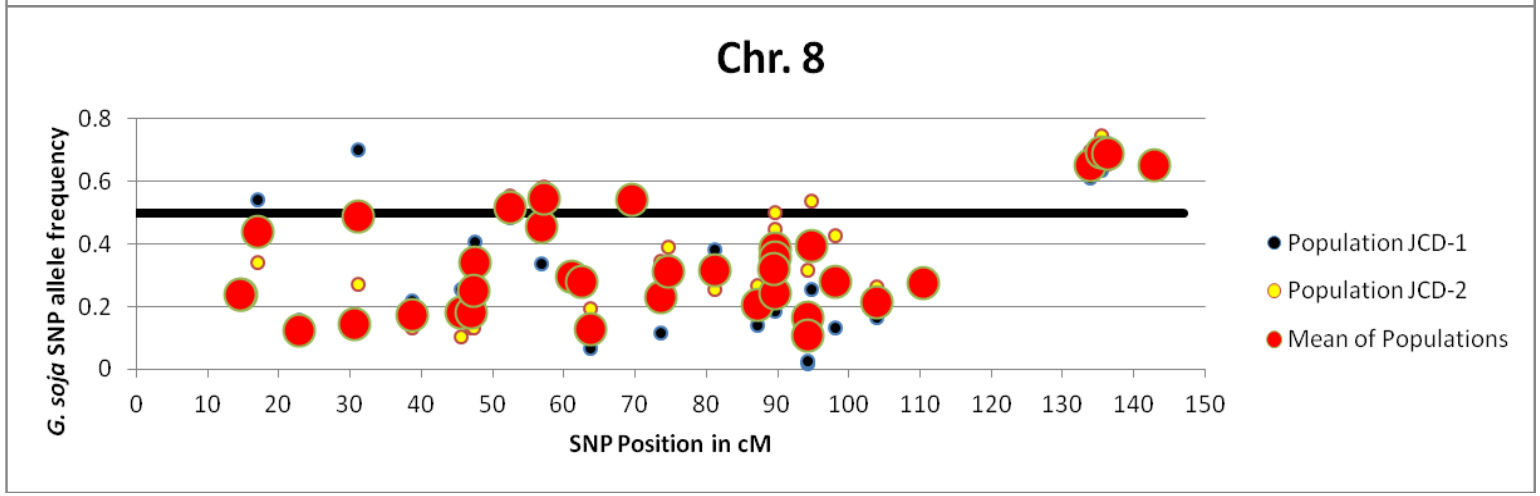
Figure 2.2 Graphical representation of allele frequencies of SNPs on Chromosomes 1-20. Frequencies below black line represent skewness towards N7103 allele, while frequencies above represent skewness toward *G. soja* PI 366122.

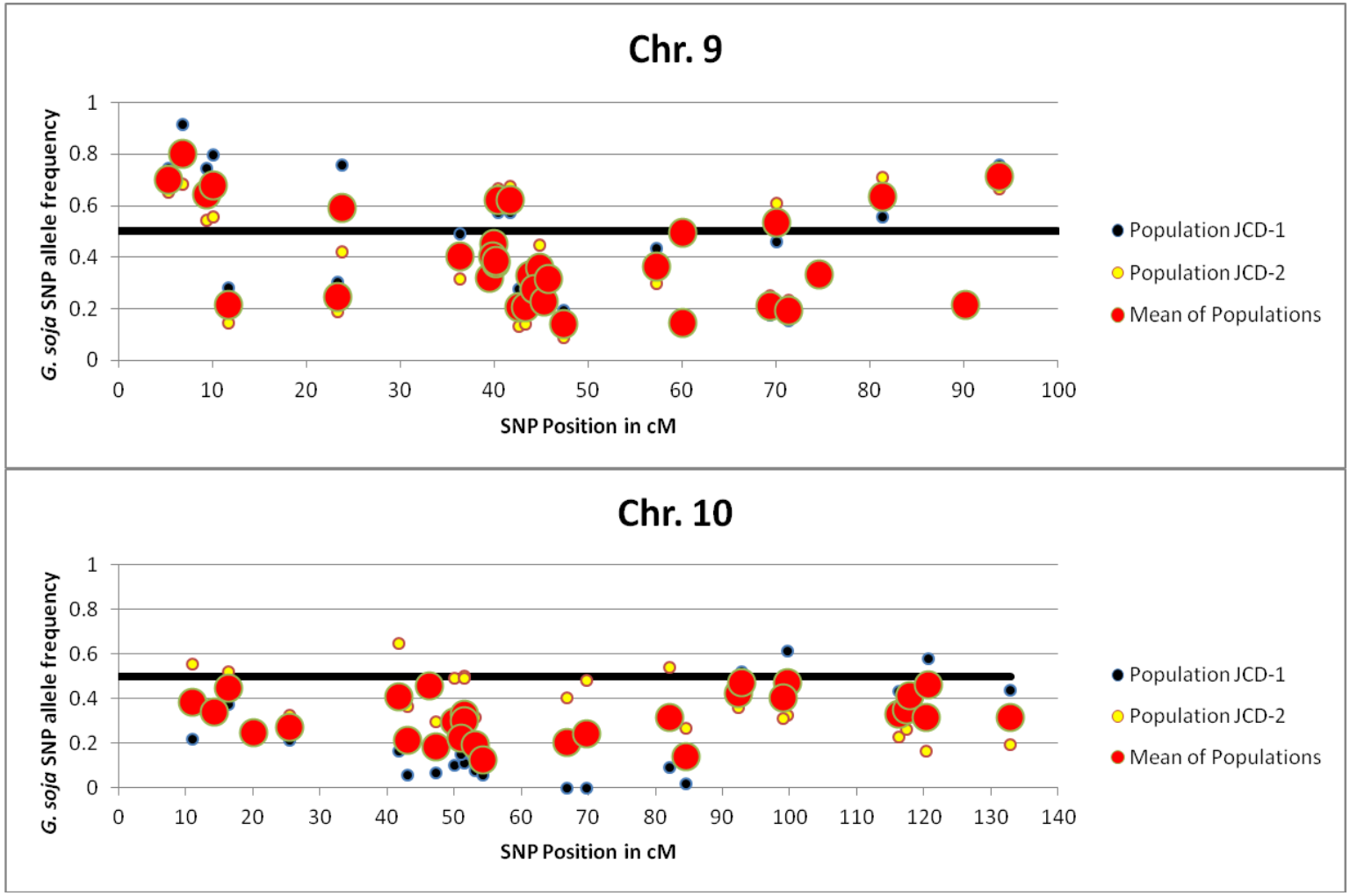

Chr. 1
G. soja SNP allele frequency
SNP Position in cM
Population JCD-1
Population JCD-2
Mean of Populations
Chr. 2
G. soja SNP allele frequency
SNP Position in cM
Population JCD-1
Population JCD-2
Mean of Populations

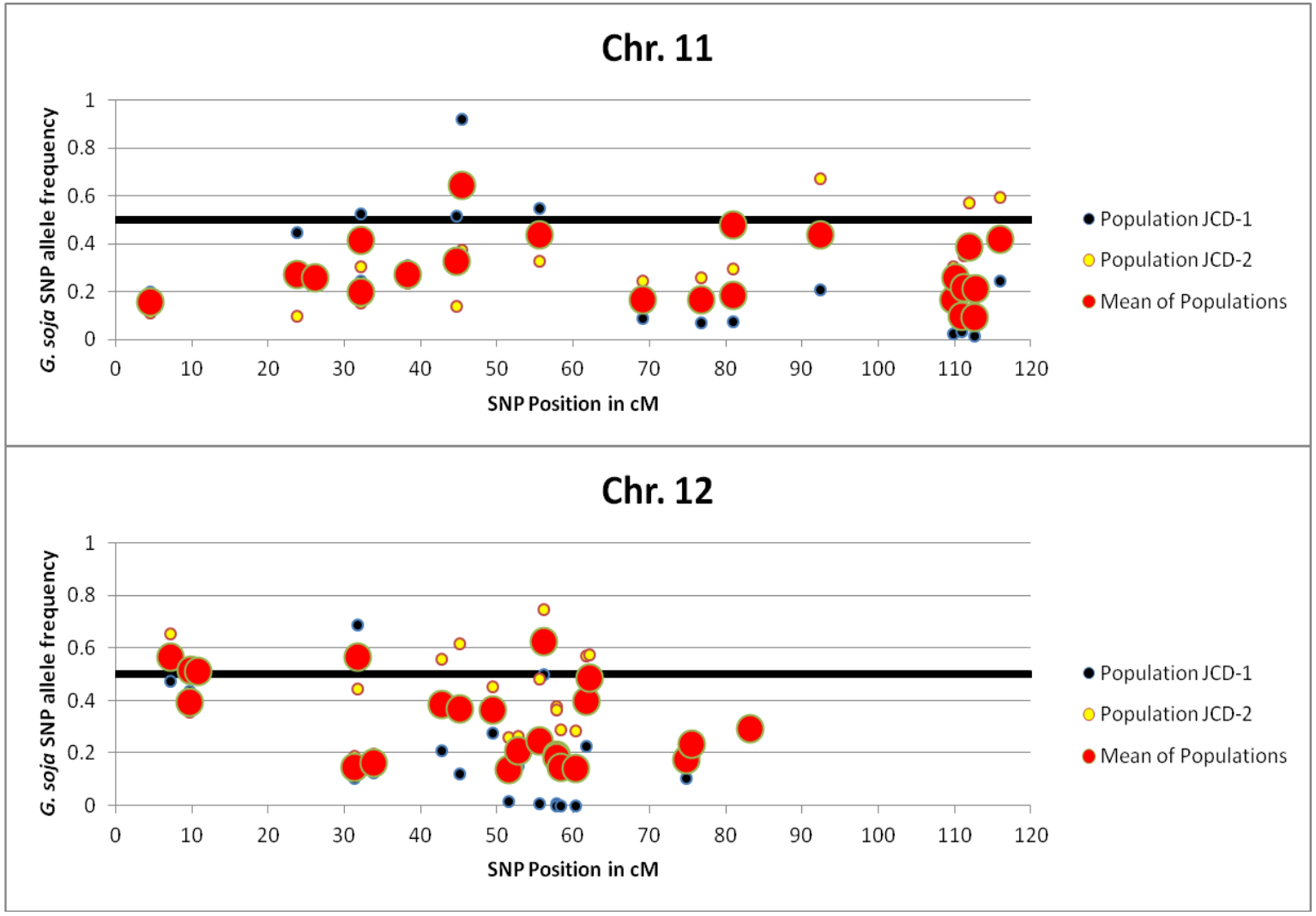

Chr. 3
G. soja SNP allele frequency
1
0.8
0.6
0.4
0.2
0
0
10
20
30
40
50
60
70
80
90
100
SNP Position in cM
Population JCD-1
Population JCD-2
Mean of Populations
Chr. 4
G. soja SNP allele frequency
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
10
20
30
40
50
60
70
80
90
100
110
120
SNP Postion in cM
Population JCD-1
Population JCD-2
Means of Populations

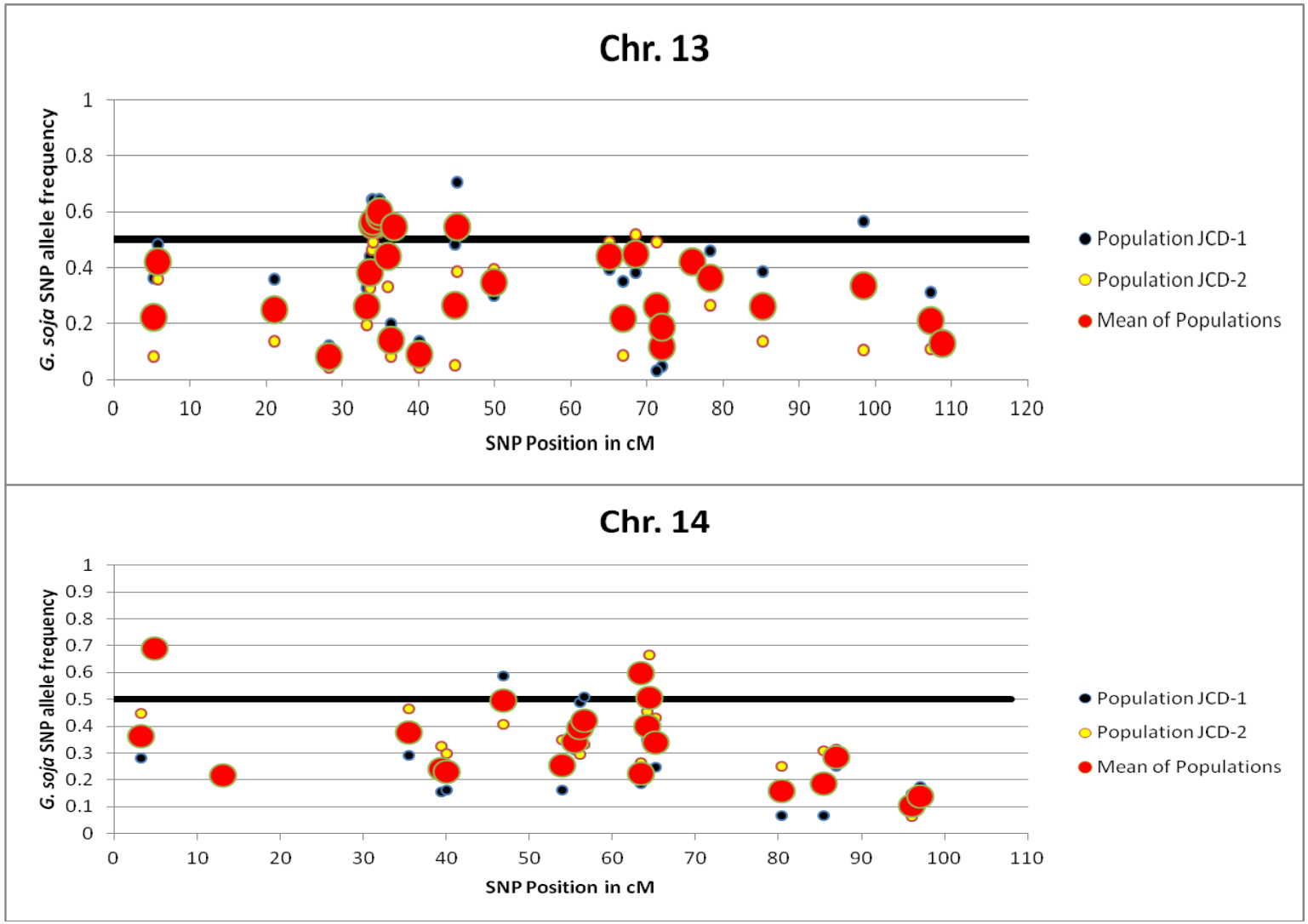

Chr. 5
G. soja SNP allele frequency
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
10
20
30
40
50
60
70
80
90
SNP Position in cM
Population JCD-1
Population JCD-2
Mean of Populations
Chr. 6
G. soja SNP allele frequency
1
0.8
0.6
0.4
0.2
0
0
10
20
30
40
50
60
70
80
90
100
110
120
130
140
SNP Position in cM
Population JCD-1
Population JCD-2
Mean of Populations

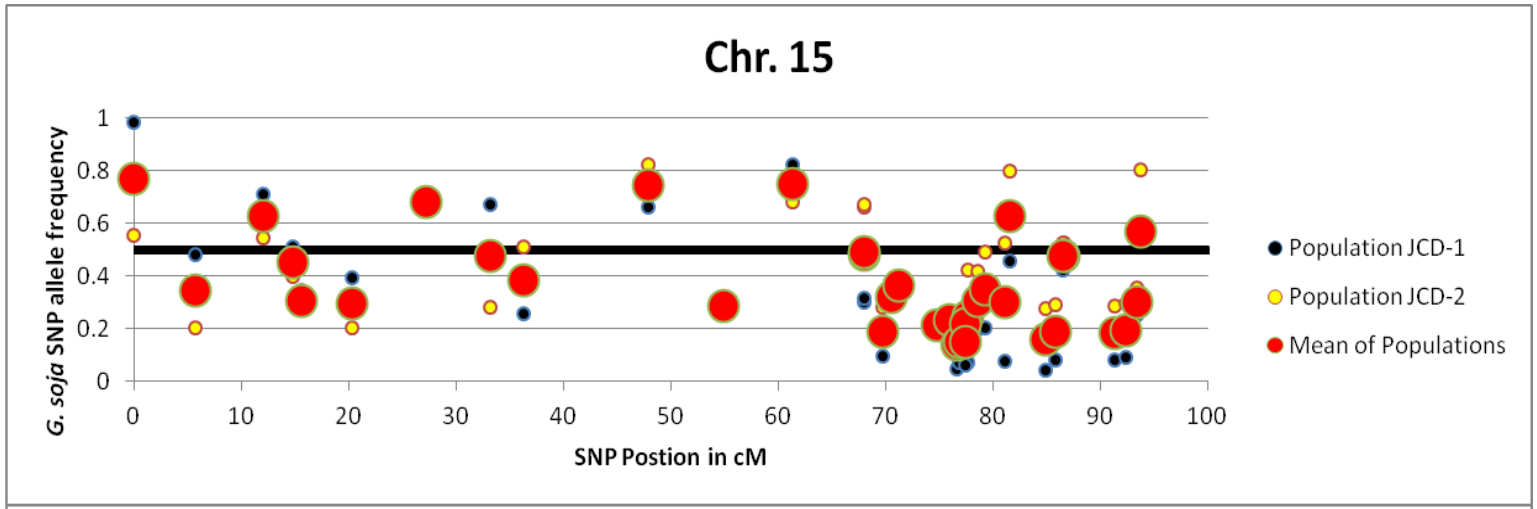

Chr. 7
G. soja SNP allele frequency
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0 10 20 30 40 50 60 70 80 90 100 110 120 130 140
SNP Postion in cM
Population JCD-1
Population JCD-2
Mean of Populations

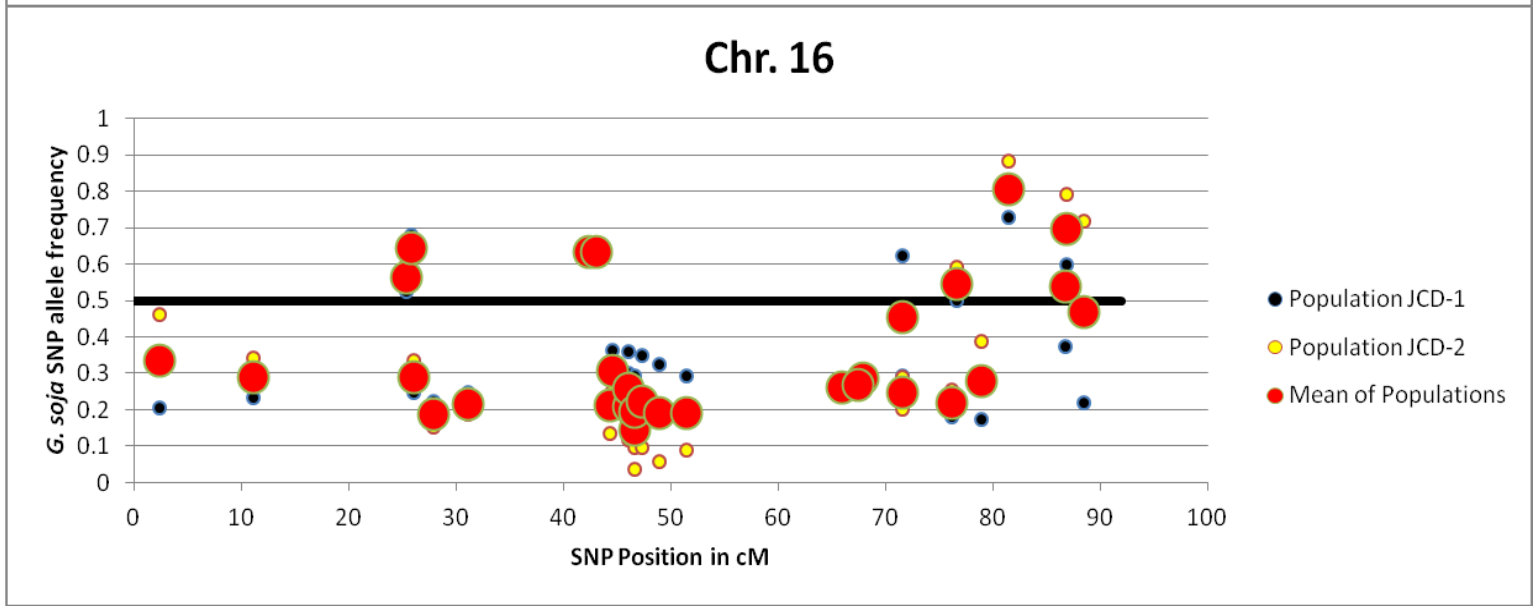

Chr. 8
G. soja SNP allele frequency
0.8
0.6
0.4
0.2
0
0 10 20 30 40 50 60 70 80 90 100 110 120 130 140 150
SNP Position in cM
Population JCD-1
Population JCD-2
Mean of Populations

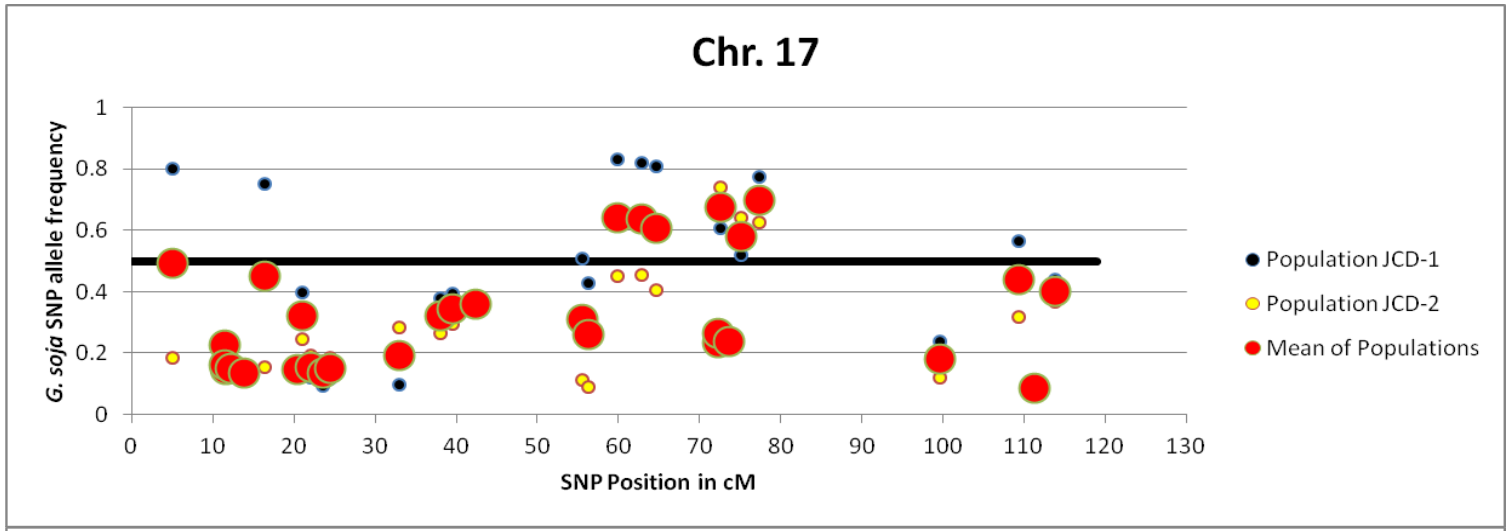

Chr. 9
G. soja SNP allele frequency
1
0.8
0.6
0.4
0.2
0
0
10
20
30
40
50
60
70
80
90
100
SNP Position in cM
Population JCD-1
Population JCD-2
Mean of Populations
Chr. 10
G. soja SNP allele frequency
1
0.8
0.6
0.4
0.2
0
0
10
20
30
40
50
60
70
80
90
100
110
120
130
140
SNP Position in cM
Population JCD-1
Population JCD-2
Mean of Populations

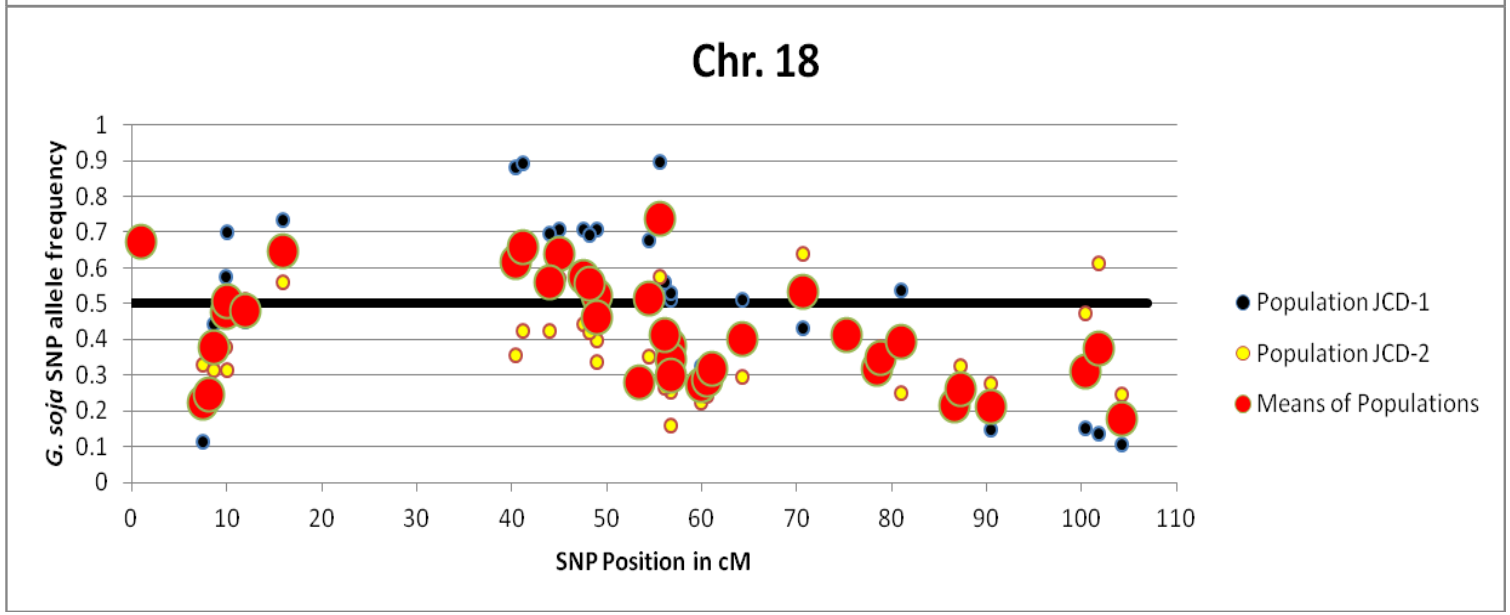

Chr. 11
G. soja SNP allele frequency
1
0.8
0.6
0.4
0.2
0
0
10
20
30
40
50
60
70
80
90
100
110
120
SNP Position in cM
Population JCD-1
Population JCD-2
Mean of Populations
Chr. 12
G. soja SNP allele frequency
1
0.8
0.6
0.4
0.2
0
0
10
20
30
40
50
60
70
80
90
100
110
120
SNP Position in cM
Population JCD-1
Population JCD-2
Mean of Populations

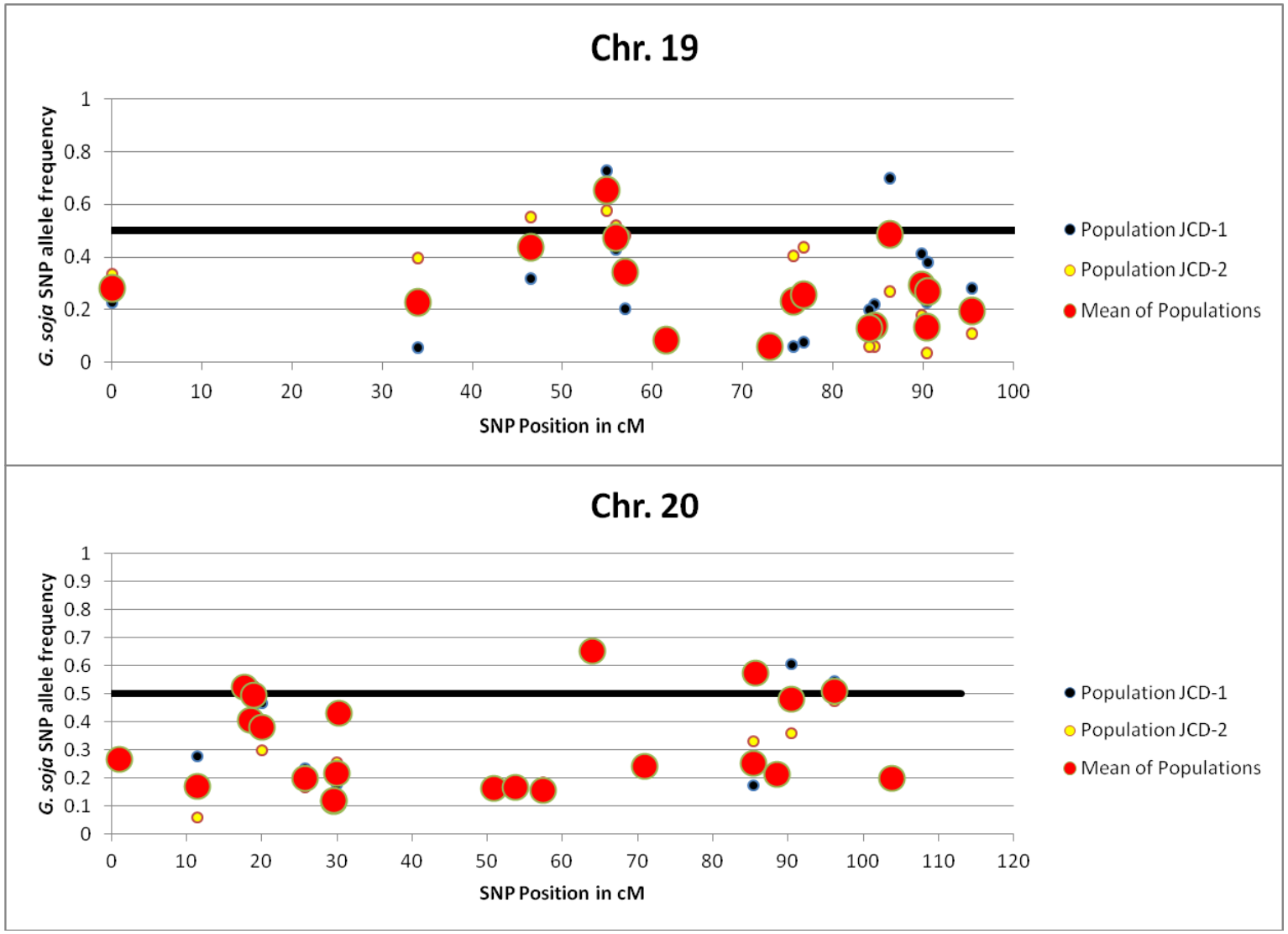

Chr. 13
G. soja SNP allele frequency
1
0.8
0.6
0.4
0.2
0
0
10
20
30
40
50
60
70
80
90
100
110
120
SNP Position in cM
Population JCD-1
Population JCD-2
Mean of Populations
Chr. 14
G. soja SNP allele frequency
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
10
20
30
40
50
60
70
80
90
100
110
SNP Position in cM
Population JCD-1
Population JCD-2
Mean of Populations


Chr. 15
G. soja SNP allele frequency
1
0.8
0.6
0.4
0.2
0
0
10
20
30
40
50
60
70
80
90
100
SNP Postion in cM
Population JCD-1
Population JCD-2
Mean of Populations


Chr. 16
G. soja SNP allele frequency
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
10
20
30
40
50
60
70
80
90
100
SNP Position in cM
Population JCD-1
Population JCD-2
Mean of Populations


Chr. 17
G. soja SNP allele frequency
1
0.8
0.6
0.4
0.2
0
0
10
20
30
40
50
60
70
80
90
100
110
120
130
SNP Position in cM
Population JCD-1
Population JCD-2
Mean of Populations


Chr. 18
G. soja SNP allele frequency
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
10
20
30
40
50
60
70
80
90
100
110
SNP Position in cM
Population JCD-1
Population JCD-2
Means of Populations


Chr. 19
G. soja SNP allele frequency
1
0.8
0.6
0.4
0.2
0
0
10
20
30
40
50
60
70
80
90
100
SNP Position in cM
Population JCD-1
Population JCD-2
Mean of Populations
Chr. 20
G. soja SNP allele frequency
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
10
20
30
40
50
60
70
80
90
100
110
120
SNP Position in cM
Population JCD-1
Population JCD-2
Mean of Populations

CHAPTER 3

STRATEGIES FOR EFFICIENT SELECTION OF UPRIGHT PLANTS IN A SINGLE CROSS OF *GLYCINE MAX* AND *GLYCINE SOJA*

ABSTRACT

Wild soybean (*Glycine soja* (Subb. and Zucc.) has long been regarded as a novel source of variation for soybean [*Glycine max* (L.) Merr.] breeding, but few cultivars are derived from it, because of wild soybean's undesirable viny growth habit. Recently, Delheimer (2012) demonstrated the accessibility of wild soybean genome to applied breeding by developing agronomic breeding lines from a single cross of soybean x the wild progenitor, using a very large F_3 population (greater than 1 million plants) combined with visual selection for the *G.max* type. For future breeding with wild soybean, it is desirable to refine this procedure by optimizing population size. However, optimization cannot be achieved easily because, although clearly multigenic in nature, the inheritance of upright plant architecture is not well understood. To resolve this issue, approximately 2,000 random $F_{2:3}$ families from the cross of Black Raleigh soybean by *G. soja* plant introduction 366122 were evaluated over two years for upright growth habit, employing two replications near Kinston, NC in 2009 and 2010. Selected $F_{3:4}$ families from F_3 plants exhibiting upright growth habit were also evaluated in 2010 and 2011 No F_2 plant exhibited an erect main stem growth habit, and no $F_{2:3}$ family exhibited a uniform or predominately erect plant architecture. However, across years, 118 $F_{2:3}$ families produced at least one F_3 plant with an erect main stem. In 2010, most of the 123 $F_{3:4}$ families had a preponderance of erect plants and 70% of the $F_{3:4}$ families showed segregation patterns that could be explained by the segregation of one or two genes where dominance favored plant erectness. Overall observed segregation

patterns were fit to various genetic models using the binomial distribution and chi square goodness of fit tests. The best fit models were found to be a seven gene model in which plant erectness is epistatically recessive, a ten gene model in which dominance for plant erectness is observed at one locus, and a twelve gene model in which dominance for plant erectness is observed at two loci. Based on these results, selection of upright plants in the F_3 generation after a generation of bulk breeding may be an efficient method to implement when using breeding material derived from an interspecific cross.

INTRODUCTION

Wild soybean has long been regarded as a novel source of variation for traits of interest to a soybean breeder, including improved fatty acid profile, soybean cyst nematode resistance, salt tolerance and improved seed yield (Rebetzke et al., 1997; Pantalone et al., 1997; Wang et al., 2001; Luo et al., 2005, Li et al., 2008,). Recent genomic studies reveal that *G. soja* is a substantial reservoir of allelic diversity as well (Hyten et al. 2006, Stupar 2010). These attributes make wild soybean a potentially desirable parental stock for applied breeding.

A long-standing obstacle to the successful use of *G. soja* in applied breeding is the non-agronomic appearance and poor performance of its progeny in the F_1, F_2, F_3, and F_4 generations. (Williams 1948, Weber 1950, Ertl and Fehr 1986). Typically, breeders have resorted to the backcrossing breeding method to ‘breed away’ from the poor agronomic plant type of wild soybean type while retaining specific desirable alleles from wild soybean (Concibido et al., 2003) However, even this approach has produced few germplasm releases or cultivars to the present.

Recently, Delheimer (2012) demonstrated the accessibility of wild soybean genome to applied breeding by developing agronomic breeding lines from a single cross of domesticated x wild soybean, using a very large F_3 population (greater than 1 million plants derived from 5000 random F2 plants) combined with intense visual selection for the *G.max* type. This approach, coupled with pedigree selection in the F_4 generation, produced more than 100 F_4-derived breeding lines with acceptable agronomic performance and diverse allelic combinations. Breeding lines from this study are now being used as parental stock in applied breeding programs, and this approach is being applied to additional wild x domesticated populations (T.E. Carter, 2012, personal communication)

The optimum resource allocation for selection in a bi-parental single-cross is a topic of interest for soybean breeders who are considering the use of wild soybean in applied breeding. Two key determinants of efficient resource allocation are the number of genes controlling plant uprightness trait and gene action. A theoretical basis for estimating the minimum number of genes controlling a quantitative trait was first proposed by Wright (1968), based on equations that involve the calculation of variance components within a population. Variations to Wright's equation have increased the accuracy of these estimates (Lande 1981, Cockherham 1986). Although this approach is useful for estimating gene numbers for quantitative metric traits, the practicality of applying them to plant architecture is open to question, in populations derived from domesticated x wild soybean. The '*G. max*-like' segregants from such crosses tend to have an obvious appearance; however, visual rating of intermediate plant types is much more subjective and easily influenced by population density in the field. Given the extremely viny nature of most *G. max* X *G. soja*

progeny, such measurements would be subject to error. A metric versus visual rating approach to architecture, although clearly advantageous, would be tedious, and subject to the same errors related to viny growth habit

The stochastic model proposed by Isleib (1999) offers a more realistic approach to estimating gene numbers affecting upright max-like growth habit in a cross, using the binomial distribution and the equation: $\frac{n}{N} = (f)^L$, where *n* is the number of selected individuals, *N* is the total population size, f is the frequency of the desired genotype at a particular locus based on the generation of inbreeding of the population, and *L* is the number of segregating loci between the parents for the trait of interest. The only assumptions are: i. parents are completely homozygous, so that each may contribute only one allele to each segregating locus, ii. genes assort independently, and iii. there is no selection in the segregating population. With this approach, it is possible to evaluate alternative models for the number of genes controlling upright plant growth in *G. max* X *G. soja* crosses. Manipulation of this equation also offers insight into optimal resource allocation for soybean breeders interested in using the breeding method described by Delheimer (2012) in wild x domesticated breeding populations.

Definitive answers are not available currently regarding the number of loci controlling plant erectness in wild x domesticated populations. Clearly, however, plant erectness is controlled by many genes and progeny tend to favor the wild soybean much more than the domesticate. For example, a RIL population derived from nearly 1000 individual F_2 plants produced no F_4-derived lines which had erect main stems (P. Cregan, 2012, personal communication), indicating that at least 6 genes control upright growth habit.

Other groups have worked toward developing breeding material from selection within interspecific crosses, but selection is often hindered by the poor agronomic appearance of the F_2 material (R. Nelson, 2012, personal communication). Delheimer (2012) identified nearly 600 individual F_3 plants from among approximately 1,000,000 which had an erect main stem. In the F_3 generation, the probability of a completely homozygous class at an allele is 3/8. Therefore, if we solve the equation above for *L*, one would conclude that the trait of interest, an erect main stem, is controlled by eight genes, assuming that an erect "*Max*-like" type was observed only when all controlling genes are in one plant in a homozygous state. The homozygous state was likely the case in that study, because progeny rows from the F_3 plants did not produce a wild soybean phenotype, although some rows looked more agronomic than others.

The objective of the present research is to i) estimate the number of loci controlling plant erectness in a cross of *G. soja* PI 366122 to a black seeded mutant strain of 'NC-Raleigh' (Burton et al., 2006) (designated as Black NC-Raleigh) by evaluating segregation patterns in $F_{2:3}$ and $F_{3:4}$ families, and ii) develop guidelines for efficient resource allocation in developing upright breeding lines from *G. max* x *G. soja* hybridization.

MATERIALS AND METHODS

POPULATION DEVELOPMENT

In the summer of 2007, *G. soja* PI 366122 was crossed to the soybean genotype Black NC-Raleigh. This black-seeded mutation in NC-Raleigh was used as a parent in order to employ seed coat color as a protection against bias introduced by obvious outcrossing in later generations. Black seed coat color is a recessive trait in soybean, and both parents are black-

seeded. Thus, any yellow seed found in progeny would likely be the result of out-crossing from insect cross pollination or otherwise mechanical contamination. The F_1 plants were grown at the USDA-ARS Tropical Agriculture Research Station (TARS) winter nursery in Isabella, Puerto Rico during the winter nursery from 2007-2008. The F_1 plants were harvested as a bulk. Approximately 5,000 seed from this bulk were planted at the Central Crops Research Station in Clayton, NC in 2008. Seed were planted in a 1-m row width and at approximately 4 seed per m of row. Later in the season, plant stands were thinned back to two plants per m in order to facilitate harvest of individual plants from the population in the fall. In the fall of 2008, approximately 1,700 single F_2 plants were harvested to form $F_{2:3}$ families. An additional 1,500 F_2 seed were sent to the winter nursery at TARS in December of 2008, and 400 plants were harvested individually to produce additional $F_{2:3}$ families. Individual plants that did not have enough seed for further field evaluation were dropped from the study, leaving 1,935 $F_{2:3}$ lines available for field studies. Seed of these lines were scarified using sandpaper in order to ensure germination. Seed quantity was sufficient for each F_2 plant so that each $F_{2:3}$ family could be evaluated both in 2009 and 2010. In December of 2009, 1000 additional F_2 seed was sent to the TARS station in Puerto Rico, and single plants were harvested to develop an additional 400 $F_{2:3}$ families. These families were evaluated in 4 additional replicated tests at the Lower Coastal Plain Research Station on the Hugo farm near Kinston, NC in 2010, because no uniformly upright $F_{2:3}$ families were observed in 2009.

PLOT TECHNIQUE AND VISUAL EVALUATION OF $F_{2:3}$ FAMILIES

The families were separated into sets and evaluated in 14 individual tests using a

randomized complete block experimental design and two replications. Each test included the control cultivars 'Stressland' (maturity group (MG) IV, Cooper et al., 1999), '5601T' (MG V, Pantalone et al., 2003), 'NC-Roy' (MG VI, Burton et al., 2005), and both 'NC-Raleigh' (MG VII) and the parental Black NC-Raleigh. Seed size appeared to vary among families, and, thus, ten-seed weight measurements were taken on all $F_{2:3}$ families prior to planting. $F_{2:3}$ families were then randomly assigned to field tests in such a way that a range of seed size was represented in each test. The $F_{2:3}$ families were planted in single-row observational plots at the Lower Coastal Plain Research Station on the Hugo farm near Kinston, NC in 2009 and 2010. The plots were planted at 1.8 m length in 97 cm row spacing. Families were seeded at 33 seed per m of row.

Flower color and maturity date ratings were taken on all plots. Maturity was rated as first day after October 1 on which 95% of the pods reached harvest maturity. At maturity, stand counts were taken on 50 randomly selected individual plots in order to estimate an average population density for the entire study. As plots reached maturity they were visually phenotyped for the presence of "Max-like" plants present within a single row. A "Max-like" plant type was defined using three criteria: i. the main stem did not contact the ground, ii. no lateral branches were longer in length than the main stem, and iii. at least 5 pods were found on the main stem. Any $F_{2:3}$ family that had at least one single plant that fit this criteria was also subjected to a stand count at maturity Individual plants that fit the aforementioned "Max-like" criteria were threshed for evaluation in a $F_{3:4}$ nursery the following season**.**

PLOT TECHNIQUE AND VISUAL EVALUATION OF $F_{3:4}$ FAMILIES

Seed from the selected F_3 plants from the field study in 2009 were evaluated in

single-row plots at the Lower Coastal Plain Station in 2010 using a randomized complete block experimental design and two replications. Seed from individual F_3 plants selected in 2010 were grown in similar fashion near Clayton, NC at the Central Crop Research Station in 2011. The ingle-rows plots were 1.2 m in length with 97-cm row spacing. Families were planted at 23 seed per m of row. Lines were evaluated for flower color, maturity date, as well as phenotyped for the plant type criteria outlined above.

DATA ANALYSIS

To estimate of the number of genes controlling an upright "Max-like" phenotype, the binomial distribution was used to develop expected numbers of "Max-like" progeny for the population sizes employed. We first fit the F_2 data (i.e. number of F_2 rows that bred true for upright growth habit) to various genes models (i.e. five gene, six gene model, etc.) and rejected gene models which did not fit. Using stand counts to estimate total plants grown in the $F_{2:3}$ family tests, we next fit the F_3 segregation data plants to various gene models and rejected those that did not fit the observed data, ignoring the $F_{2:3}$ family structure and assuming that the number of F_2 plants represented was sufficiently large to overcome any effects of finite population size. Subsequently, we fit the $F_{3:4}$ family data for upright growth habit to dominance and epistastic models where presumably many genes for uprightness were fixed, but some continued to segregate. Chi-Square was used to determine which gene action and number did not fit. Lastly, we pooled results from these sources to determine which models best matched overall segregation patterns. The most likely gene number and action models were used to estimate the distribution of various gene combinations in a F_2 population from the cross.

RESULTS

OBSERVATIONAL RESULTS OF PLANT TYPE IN $F_{2:3}$ and $F_{3:4}$ FAMILIES

No F_2 plants exhibited upright max-like growth habit during family development. In 2009 and 2010, no single $F_{2:3}$ family of the 1,936 planted produced a progeny in which all plants exhibited upright growth habit similar to that of the *G. max* parent. Although no family bred true for upright growth habit or had predominantly upright growth habit in 2009, a total of 123 individual F_3 plants were identified as having upright growth habit (75 from replication one, and 48 from replication 2) and traced to 59 $F_{2:3}$ families. Of these 59 families, 12 produced at least one upright plant in both replications. An estimated total of 135,450 plants were grown in 2009, based on stand counts. The 1,936 $F_{2:3}$ families were evaluated again in 2010 with similar results. A total of 173 single F_3 plants upright plants were identified (76 from replication one and 97 from replication two) from among an estimated 116,110 plants. Desirable plants traced to 76 individual $F_{2:3}$ families. Of these 76 families, 14 produced desirable plants in both replications. Across years, 17 $F_{2:3}$ families exhibited at least one desirable plant in both years. Three $F_{2:3}$ families had at least one single desirable plant in both replications in both years. A total of 296 F_3 plants derived from 118 F_2 plants were identified over years.

As previously stated, 400 additional $F_{2:3}$ families were studied in 2010. Of these 400 families, ten produced at least one desirable F_3 plant. A total of 16 F_3 plants were identified from an estimated population of 24,000.

Seed from the 123 selected F_3 plants from 2009 were grown as replicated $F_{3:4}$ progeny rows in 2010. The number of upright and visually poorer plants was determined for

each row at maturity and converted to a percentage of upright plants for each row, revealing a good agreement between replications (r=0.66 on a plot basis). Thus, the raw data were combined over replications for an overall analysis of segregation patterns. Of the 123 $F_{3:4}$ families 55 were identified in which at least 70% of the plants exhibited the desirable plant architecture (Fig. 3.1). An additional 33 $F_{3:4}$ families had at least 50% of the plants with the desirable phenotype. No $F_{3:4}$ family exhibited 100% upright growth habit in both replications, although 3 individual plots were identified in which all plants exhibited the upright plant architecture.

Individual F_3 plants selected from the study in 2010 were grown in single $F_{3:4}$ in 2011. Of the 2,537 F_4 plants grown, only 34 were identified as "Max-like" and traced to 14 $F_{3:4}$ families.

GENETIC MODEL FITTING OF OBSERVATIONAL RESULTS

No single F_2 plant or entire $F_{2:3}$ family exhibited the "Max-like" plant criteria outlined above, therefore models were fit based on segregation within $F_{2:3}$ families and the entire F_3 population. When the data was combined over years, the segregation patterns observed in the F_3 generation fit a seven gene model in which all loci are homozygous, a ten gene model in which 9 loci are homozygous for the recessive allele and dominance is observed at the last locus, and a 12 gene model in which 10 loci are homozygous for the recessive allele and dominance is observed at the last loci (Tables 3.3 and 3.4). The within $F_{2:3}$ family segregation patterns were observed and fitted to various ratios (3:1, 15:1, etc.). The data fitted was on families in which stand counts were taken on both plots within a year. All of these patterns fit a 15:1 to segregation ratio ($p<0.01$, Table 3.5).

The $F_{3:4}$ progeny rows grown in 2010 showed segregation for the erect main stem phenotype. Of the 123 $F_{3:4}$ families tested, 88 exhibited a segregation pattern in which 50% or greater of the plants had an erect main stem (Figure 3.1). Fifty-nine of these families fit a 3:1 ratio, 78 families fit a 9:7 ratio, and 58 families fit a 27:37 ratio (Table 3.6).

DISCUSSION

ESTIMATING THE NUMBER OF GENES CONTROLLING AN ERECT MAIN STEM

Delheimer (2012) showed that agronomically acceptable soybean germplasm could be developed from single cross *G. max* X *G. soja* germplasm using large F_3 populations and visual selection for upright plant growth. Based on their results, it appeared that the *G. max* upright plant type could be best explained by segregation at eight loci. However, F_2 families were not identified as such in that study. To more carefully consider the question of the number of loci controlling upright *G.max*-like growth habit, we observed the plant phenotype of 1936 $F_{2:3}$ families. For F_2 plants, the probability of recovering a homozygous locus is ¼. Assuming a completely additive gene action model for erect plant growth, the probability of recovering a completely homozygous individual at multiple loci, exponentially decreases as the number of genes segregating in the population increases (Table 3.1). This can limit the effectiveness of selection in the F_2 generation. Based on 2009 and 2010 data, we did not observe a single $F_{2:3}$ family in which all plants exhibited upright plant type, even though all *G. max* control cultivars exhibited 100% upright growth habits in all replications. No F_2 plants were observed as having upright growth habit during family development, as well, but no control genotypes were available for direct comparison, so the lack of upright growth habit for F_2 plants is best inferred from the $F_{2:3}$ family observations. Based on genetic

models of five or fewer completely additive genes, we should have observed at least one $F_{2:3}$ family in which all plants exhibited erect plant architecture (Table 3.1). If the plant architecture were controlled by six additive genes, 4096 F_2 plants would need to be grown to observe the phenotype. Ignoring F_2 family structure and looking at the ratio of upright F_3 plants to others over replications and years, 296 plants exhibited upright plant growth from an estimated population size of 251550. These data fit a 7 gene model in which all loci are fixed, a ten gene model in which nine loci are fixed and one loci exhibits dominance, and a twelve gene model in which ten loci are fixed and two loci exhibit dominance. When the progeny of these selected F_3 plants were grown in 2010, the majority of these plants segregated in a 3:1 or 9:7 ratio or in some cases a 27:37 ratio that favored the erect main stem phenotype. The selected $F_{3:4}$ progeny grown in 2011 did not show any discernable pattern in segregation for plant type. In 2011, these progeny were grown in a well irrigated environment which led to greater vininess and in turn led to much less upright growth habit. However, taken altogether, it appears that, after selection of F_3 plants only a few genes remained segregating within the population.

IMPLICATIONS TO SOYBEAN BREEDING

To our knowledge, no breeder has ever observed a F_2 plant from a domesticted x wild hybridization that appeared agronomically identical to the *G. max* parent. The findings of Delheimer (2012) showed that by delaying selection to the F_3 generation and applying visual selection to a large F3 population, large amounts of allelic diversity can be introgressed from *G. soja* into upright "*G. max*-like" soybean breeding lines. In that study the pedigree of selected F_3 plants was unknown and, thus it was not determined if the rare F_3 plants came

from several or very few F_2 plants. In the present study, we identified F_3 plants from 118 individual F_2 –derived rows out of 1,936 evaluated. Previous studies (Williams 1948; Weber 1950; Ertl and Fehr 1986) have shown that the distribution of plant type in interspecific crosses tend to favor the wild soybean. The plants in this study were no exception, as the distribution of plant types appeared to be skewed in favor of the *G. soja* parent. Of the 1,936 F_2 plants harvested none appeared to have an erect main stem, and only about 5% produced at least one progeny with this attribute. Our results indicate that the plant type that we selected upon best fit a seven gene additive model, a 10 gene model in which one locus exhibits dominance, and a 12 gene model in which two loci exhibit dominance. If a breeder wanted to produce 200 single plants with an upright main stem, they would need to grow at least 3.3 million F_2 plants under the 7 gene scenario, 6.9 million in the 10 gene scenario, and 5.6 million in the 12 gene scenario. However, if a breeder waits to make selections until the F_3 generation, the numbers shrink to about 191,000, 218,200, and 141,000 F_3 plants respectively. A population of this size could be easily managed by any breeder looking to increase diversity within their respective programs. Combined with an offseason nursery, an F_3 bulk could be easily and quickly obtained. One could argue that backcrossing may provide a more efficient way to handle exotic material. However, there are a few things to consider with that regard. Our data points to the erect plant architecture being controlled by as little as seven genes or as many as 12 genes. If there are seven genes different between the two parents, 0.7% of BC_1F_1 progeny, 13% BC_2F_1 progeny, and 39% of the BC_3F_1 progeny would be fixed for the seven genes. These percentages decrease as additional genes are considered. These percentages could be increased if visual selection is used, but visual ratings of plant

type are obtained only after plants have reached full maturity. Empirical studies would have to be conducted in order to compare the efficiency of either method directly. If these possible genomic regions were linked to genes with a negative effect on agronomic performance on yield, larger population sizes would be necessary. However, due to the success of the breeding lines developed in the study by Delheimer (2012), it appears that manageable population sizes are easily developed. This research shows some of the selection tools that a breeder may employ in working with material from an interspecies cross. As more *G. soja* PIs are studied and used in this breeding methodology, the results should prove clearer and perhaps provide more information on the domestication of soybean.

REFERENCES

Burton, J.W., T.E. Carter, and D.T. Bowman 2005. Registration of 'NC-Roy' soybean. 2005. Crop Sci. 45: 2654.

Burton, J.W., T.E. Carter, M.O. Fountain, and D.T. Bowman. 2006. Registration of 'NC-Raleigh' soybean. Crop Sci. 46: 2710-2711.

Cockerham, C.C. 1986. Modifications in estimating the number of genes for a quantitative character. Genetics. 114: 659-664.

Cooper, R.L., R.J. Martin, S.K. St. Martin, A. Calip-DuBois, R.J. Fioritto, and A.F. Schmitthenner. 1999. Registration of 'Stressland' soybean. Crop Sci. 39: 590-591.

Delheimer, J.C. 2012. Agronomic performance of breeding lines developed from a single cross of *Glycine max* X *Glycine soja*. Dissertation, Chapter 1.

Ertl, D.S., and W.R. Fehr. 1985. Agronomic performance of soybean genotypes from *Glycine max* X *Glycine soja* crosses. Crop Sci 25: 589-592.

Hyten, D.L., Q. Song, Y. Zhu, I. Choi, R.L. Nelson, J. M. Costa, J. E. Specht, R. C. Shoemaker, and P.B. Cregan. 2006. Impacts of genetic bottlenecks on soybean genome diversity. PNAS. Vol. 103. 45: 16666-16671.

Isleib, T. G. 1999. Recovery of superior homozygous progeny from biparental crosses and backcrosses. Crop Sci. 39: 558-563.

Lande, R. 1981. The minimum number of genes contributing to quantitative variation between and within populations. Genetics. 99: 541-553.

Li, D., T.W. Pfeiffer, and P.L. Cornelius. 2008. Soybean QTL for yield and yield components associated with *Glycine soja* alleles. Crop Sci. 48: 571-581.

Luo, Q., B. Yu, Y. Liu. 2005. Differential sensitivity to chloride and sodium ions in seedlings of *Glycine max* and *G. soja* under NaCl stress. Journal of Plant Phys. 162: 1003-1012.

Pantalone, V.R., F. L. Allen, and D. Landau-Ellis. 2003. Registration of '5601T' soybean. Crop Sci. 43: 1123-1124.

Pantalone, V.R., G.J. Rebetzke, R.F. Wilson, and J.W. Burton. 1997. Genetic regulation of linolenic acid concentration in wild soybean *Glycine soja* accessions. J. Am. Oil. Chem. Soc. Vol. 74 No. 2: 159-163.

Rebetzke, G.J., V. R. Pantalone, J. W. Burton, T. E. Carter, Jr., and R. F. Wilson. 1997. Genotypic variation for fatty acid content in selected *Glycine max* x *Glycine Soja* populations. Crop Sci. 37: 1636-1640.

Stupar, R.M. 2010. Into the wild: the soybean genome meets its undomesticated relative. PNAS. Vol. 107 No. 51: 21947-21948.

USDA-ARS National Genetic Resources Program (USDA-ARS NGRP). 2012c. *Glycine soja* Sieb. And Zucc. PI 366122. Germplasm Resources Information Network (GRIN). [Online database] National Germplasm Resources Laboratory, Beltsville, MD. Available at http://www.arsgrin.gov/cgi-bin/npgs/acc/display.pl?1274061 (verified 09 Feb. 2012).

Wang, D., P.R. Arelli, R.C. Shoemaker, B.W. Diers. 2001. Loci underlying resistance to Race 3 of soybean cyst nematode in *Glycine* soja plant introduction 468916. Theor. Appl. Genet. 103: 561-566.

Weber, C.R. 1950. Inheritance and interrelation of some agronomic and chemical characters in an interspecific cross in soybeans, *Glycine max* x *Glycine ussuriensis*. Res. Bull. 374. Agricultural Experiment Station, Iowa State College. Ames, IA.

Williams, L.F. 1948. Inheritance in a species cross in the soybean (Abstr.). Genetics 33: 131-132.

Wright, S. 1968. Evolution and the genetics of populations, vol.1: genetic and biometric foundations. Chicago, IL: The University of Chicago Press.

TABLES

Table 3.1. Theoretical number of F2 plants needed to recover, on average, one "Max-Like" plant from a wild x domesticated soybean hybridization, for various genetic models. Calculations are based on the binomial distribution and the assumption that upright growth habit is a minority phenotype conditioned by 1 to 10 segregating genes. For each gene number scenario, the effect of dominance on recovery of upright plants is examined. For example, given that one gene controls upright growth habit, one in four F_2 plants would be expected to be upright on average. If the locus exhibits dominance for upright growth habit, then 3 of 4 F_2 plants (1 homozygous plant and 2 heterozygotes on average) would be expected to be upright.

Theoretical number of segregating genes controlling upright growth habit	Theoretical number of dominant loci controlling trait (i.e. number of loci which can be heterozygous and still recover upright growth habit)												
	0	1	2	3	4	5	6	7	8	9	10	11	12
1	4†	1											
2	16	3	2										
3	64	7	2	2									
4	256	21	5	2	3								
5	1024	68	11	4	3	4							
6	4096	228	30	8	3	3	6						
7	16384	780	87	17	6	3	3	7					
8	65536	2731	260	43	12	5	3	4	10				
9	262144	9709	809	116	26	9	4	3	4	13			
10	1048576	34953	2589	324	62	17	7	4	4	5	18		
11	4194304	127100	8473	941	157	37	12	6	4	4	6	24	
12	16777216	466034	28244	2824	418	87	25	10	5	4	4	8	32

† Calculated as $1/(\frac{n!}{p!q!})p^k q^{n-k}$, where n is the total number of genes, p is the frequency of homozygous recessive loci (0.25), q is the frequency of loci with at least a single dominant allele (0.75), k is the number of homozygous recessive loci, and n-k is the number of loci with at least 1 dominant allele.

Table 3.2 Theoretical number of F3 plants needed to recover, on average, one "Max-Like" plant from a wild x domesticated soybean hybridization, for various genetic models. Calculations are based on the binomial distribution and the assumption that upright growth habit is a minority phenotype conditioned by 1 to 10 segregating genes. For each gene number scenario, the effect of dominance on recovery of upright plants is examined.

Theoretical number of segregating genes controlling upright growth habit	Theoretical number of dominant loci controlling trait (i.e. number of loci which can be heterozygous and still recover upright growth habit												
	0	1	2	3	4	5	6	7	8	9	10	11	12
1	3†	2											
2	7	2	3										
3	19	4	2	4									
4	51	8	3	3	7								
5	135	16	5	3	3	10							
6	360	36	9	4	3	5	17						
7	959	82	16	6	4	4	6	27					
8	2557	192	33	10	5	4	4	9	43				
9	6819	455	68	18	7	4	4	5	13	69			
10	18184	1091	145	33	11	6	4	4	7	18	110		
11	48490	2645	317	63	19	8	5	4	5	9	27	176	
12	129308	6465	705	127	34	13	7	5	4	6	12	39	281

† Calculated as $1/(\frac{n!}{p!q!})p^k q^{n-k}$, where n is the total number of genes, p is the frequency of homozygous recessive loci (0.375), q is the frequency of loci with at least a single dominant allele (0.625), k is the number of homozygous recessive loci, and n-k is the number of loci with at least 1 dominant allele.

Table 3.3 Expected number of plants to be identified with an erect main stem in an F_3 population of 251,550. An estimated 251,550 F_3 plants were grown from 2009-2010, from which 296 upright plants were recovered.

Total Number of genes	Number of Heterozygous Loci												
	0	1	2	3	4	5	6	7	8	9	10	11	12
1	94331†	157219											
2	35374	117914	98262										
3	13265	66327	110544	61414									
4	4974	33163	82908	92120	38383								
5	1865	15545	51818	86363	71969	23990							
6	700	6995	29147	64772	80965	53977	14994						
7	262	3060	15302	42507	70845	70845	39358	9371					
8	98	1312	7651	25504	53133	70845	59037	28113	5857				
9	37	553	3689	14346	35865	59775	66417	47441	19767	3661			
10	14	231	1729	7685	22416	44831	62266	59301	37063	13727	2288		
11	5	95	793	3963	13209	30822	51369	61154	50962	28312	9437	1430	
12	2	39	357	1981	7430	19814	38527	55038	57332	42468	21234	6435	894

† Calculated by multiplying probabilities derived from binomial distribution times the total population size.

Table 3.4 Results of Chi Square Tests for Model Fitting 2009-2010 $F_{2:3}$ observational data. 296 F_3 plants were selected from an estimated population of 251,550.

Total Number of Genes	Number of Heterozygous Loci	Desirable Plants Observed	Undesirable Plants Observed	Desirable Plants Expected†	Undesirable Plants Expected	p-value
5	0	296	251254	1865	249685	<0.00001
6	0	296	251254	700	250850	<0.00001
7	0	296	251254	262	251288	0.03559
8	0	296	251254	98	251452	<0.00001
9	0	296	251254	37	251513	<0.00001
10	0	296	251254	14	251536	<0.00001
10	1	296	251254	231	251319	0.00002
12	2	296	251254	357	251193	0.00123

† Based on binomial probabilities and the estimated population size of 251,550.

Table 3.5 Segregation patterns observed within $F_{2:3}$ families in 2009 and 2010.

F2 Plant	F3 Plants Selected	F3 Plants Not Selected	p-value associated with 3:1 model	p-value associated with 15:1 model	p-value associated with 63:1 model	Year
49	2	68	<0.0001	0.2409	0.3825	2009
88	1	43	0.0005	0.2758	0.7040	2009
100	3	57	0.0003	0.6892	0.0318	2009
131	5	29	0.1657	0.0417	<0.0001	2010
134	1	51	0.0001	0.1974	0.8339	2009
152	1	54	0.0001	0.1745	0.8785	2009
224	5	45	0.0143	0.2733	<0.0001	2009
256	6	38	0.0817	0.0430	<0.0001	2010
259	1	85	<0.0001	0.0513	0.7650	2009
267	5	53	0.0040	0.4557	<0.0001	2009
329	1	31	0.0043	0.4652	0.4760	2009
361	1	78	<0.0001	0.0672	0.8316	2009
367	1	47	0.0002	0.2330	0.7711	2009
405	3	38	0.0089	0.7777	0.0030	2010
408	3	59	0.0002	0.6462	0.0375	2009
413	5	58	0.0018	0.5803	<0.0001	2009
415	1	73	<0.0001	0.0817	0.8836	2009
417	1	60	<0.0001	0.1368	0.9614	2009
497	1	25	0.0127	0.6126	0.3478	2009
508	3	44	0.0032	0.9700	0.0077	2009
515	3	37	0.0106	0.7440	0.0025	2009
523	1	50	0.0001	0.2057	0.8186	2009
543	3	51	0.0010	0.8330	0.0180	2009

Table 3.5 cont.

F2 Plant	F3 Plants Selected in 2009	F3 Plants Not Selected in 2009	p-value associated with 3:1 model	p-value associated with 15:1 model	p-value associated with 63:1 model	Year
573	1	69	<0.0001	0.0956	0.9280	2009
579	1	74	<0.0001	0.0786	0.8729	2009
683	4	42	0.0107	0.4932	0.0001	2009
701	4	76	<0.0001	0.6442	0.0132	2010
745	1	21	0.0267	0.7412	0.2593	2009
760	2	43	0.0015	0.6168	0.1190	2009
782	2	34	0.0071	0.8633	0.0534	2009
797	6	46	0.0250	0.1152	<0.0001	2009
808	1	40	0.0008	0.3134	0.6509	2009
820	1	51	0.0001	0.1974	0.8339	2009
831	1	64	<0.0001	0.1166	0.9875	2009
831	7	64	0.0032	0.2090	<0.0001	2010
832	1	32	0.0036	0.4448	0.4966	2009
948	2	32	0.0100	0.9294	0.0423	2010
970	6	70	0.0006	0.5536	<0.0001	2010
975	7	47	0.0411	0.0416	<0.0001	2009
1023	2	60	0.0001	0.3252	0.2910	2009
1044	1	58	<0.0001	0.1483	0.9346	2009
1045	4	49	0.0033	0.6964	0.0004	2009
1082	1	66	<0.0001	0.1077	0.9632	2009
1105	4	44	0.0077	0.5510	0.0002	2010
1107	10	57	0.0569	0.0034	<0.0001	2009
1120	1	53	0.0001	0.1818	0.8639	2009

Table 3.5 cont.

F2 Plant	F3 Plants Selected in 2009	F3 Plants Not Selected in 2009	p-value associated with 3:1 model	p-value associated with 15:1 model	p-value associated with 63:1 model	Year
1136	6	47	0.0215	0.1272	<0.0001	2009
1146	1	48	0.0002	0.2235	0.7872	2009
1204	1	41	0.0007	0.3003	0.6689	2009
1221	1	72	<0.0001	0.0850	0.8944	2009
1228	6	47	0.0215	0.1272	<0.0001	2010
1238	1	67	<0.0001	0.1035	0.9513	2009
1284	1	59	<0.0001	0.1425	0.9481	2009
1400	1	67	<0.0001	0.1035	0.9513	2009
1438	3	60	0.0002	0.6256	0.0406	2009
1450	2	22	0.0593	0.6733	0.0075	2009
1471	1	75	<0.0001	0.0756	0.8623	2009
1485	1	51	0.0001	0.1974	0.8339	2009
1505	1	72	<0.0001	0.0850	0.8944	2009
1576	8	59	0.0136	0.0543	<0.0001	2010
1741	1	70	<0.0001	0.0919	0.9166	2009
1858	1	52	0.0001	0.1894	0.8490	2009
1868	2	41	0.0021	0.6649	0.1024	2009
1868	4	38	0.0205	0.3808	<0.0001	2010
1874	1	71	<0.0001	0.0884	0.9054	2009
1876	1	65	<0.0001	0.1120	0.9753	2009
1930	3	45	0.0027	1.0000	0.0088	2009
1930	12	44	0.5371	<0.0001	<0.0001	2010

Table 3.6 Segregation patterns of $F_{3:4}$ families grown in 2010.

$F_{3:4}$ Family	Max Like Plants	Non Max Like Plants	p-value for 3:1 Segregation Pattern	p-value for 15:1 Segregation Pattern	p-value for 9:7 Segregation Pattern	p-value for 63:1 Segregation Pattern	p-value for 27:37 Segregation Pattern
49-1	33	8	0.4171	0.0005	0.0018	<0.0001	<0.0001
49-2	14	3	0.4838	0.0522	0.0300	<0.0001	0.0008
88-1	1	27	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001
100-1	2	22	<0.0001	<0.0001	<0.0001	<0.0001	0.0008
100-2	18	22	<0.0001	<0.0001	0.1515	<0.0001	0.7187
134-1	16	22	<0.0001	<0.0001	0.0788	<0.0001	0.9918
152-1	22	10	0.4142	<0.0001	0.1540	<0.0001	0.0023
224-1	28	8	0.7003	0.0001	0.0092	<0.0001	<0.0001
224-2	13	15	0.0005	<0.0001	0.2948	<0.0001	0.6495
224-3	13	16	0.0002	<0.0001	0.2150	<0.0001	0.7734
224-4	2	9	<0.0001	<0.0001	0.0109	<0.0001	0.1069
224-5	13	21	<0.0001	<0.0001	0.0342	<0.0001	0.6408
267-1	28	13	0.3213	<0.0001	0.1201	<0.0001	0.0007
267-2	31	10	0.9282	<0.0001	0.0125	<0.0001	<0.0001
267-3	22	14	0.0543	<0.0001	0.5566	<0.0001	0.0215
267-4	23	14	0.0713	<0.0001	0.4685	<0.0001	0.0139
267-5	15	20	<0.0001	<0.0001	0.1102	<0.0001	0.9361
329-1	16	24	<0.0001	<0.0001	0.0383	<0.0001	0.7794
361-1	19	12	0.0779	<0.0001	0.5716	<0.0001	0.0313
367-1	14	18	<0.0001	<0.0001	0.1540	<0.0001	0.8580
408-1	34	5	0.0790	0.0900	0.0001	<0.0001	<0.0001
408-2	11	4	0.8815	0.0011	0.1823	<0.0001	0.0146

Table 3.6 cont.

$F_{3:4}$ Family	Max Like Plants	Non Max Like Plants	p-value for 3:1 Segregation Pattern	p-value for 15:1 Segregation Pattern	p-value for 9:7 Segregation Pattern	p-value for 63:1 Segregation Pattern	p-value for 27:37 Segregation Pattern
408-3	16	0	0.0209	0.3017	0.0004	0.6143	<0.0001
413-1	34	10	0.7277	<0.0001	0.0049	<0.0001	<0.0001
413-2	29	10	0.9263	<0.0001	0.0226	<0.0001	<0.0001
413-3	39	3	0.0075	0.8111	<0.0001	0.0035	<0.0001
413-4	39	6	0.0707	0.0496	<0.0001	<0.0001	<0.0001
413-5	34	17	0.1693	<0.0001	0.1337	<0.0001	0.0004
415-1	19	17	0.0021	<0.0001	0.6745	<0.0001	0.1982
417-1	18	19	0.0002	<0.0001	0.3513	<0.0001	0.4261
497-1	23	8	0.9174	<0.0001	0.0440	<0.0001	0.0003
508-1	18	15	0.0067	<0.0001	0.8435	<0.0001	0.1506
508-2	5	10	0.0002	<0.0001	0.0736	<0.0001	0.4875
508-3	19	18	0.0009	<0.0001	0.5481	<0.0001	0.2590
515-1	25	10	0.6256	<0.0001	0.0703	<0.0001	0.0005
515-2	16	20	<0.0001	<0.0001	0.1533	<0.0001	0.7839
523-1	13	29	<0.0001	<0.0001	0.0010	<0.0001	0.1404
543-1	20	21	0.0001	<0.0001	0.3350	<0.0001	0.3927
543-2	34	10	0.7277	<0.0001	0.0049	<0.0001	<0.0001
543-3	34	7	0.2411	0.0042	0.0006	<0.0001	<0.0001
573-1	19	22	<0.0001	<0.0001	0.2009	<0.0001	0.5902
579-1	32	12	0.7277	<0.0001	0.0276	<0.0001	<0.0001
683-1	27	15	0.1088	<0.0001	0.2938	<0.0001	0.0037
683-2	1	13	<0.0001	<0.0001	0.0002	<0.0001	0.0079
683-3	15	29	<0.0001	<0.0001	0.0030	<0.0001	0.2768

Table 3.6 cont.

$F_{3:4}$ Family	Max Like Plants	Non Max Like Plants	p-value for 3:1 Segregation Pattern	p-value for 15:1 Segregation Pattern	p-value for 9:7 Segregation Pattern	p-value for 63:1 Segregation Pattern	p-value for 27:37 Segregation Pattern
683-4	13	13	0.0032	<0.0001	0.5206	<0.0001	0.4199
745-1	10	30	<0.0001	<0.0001	0.0001	<0.0001	0.0277
760-1	9	26	<0.0001	<0.0001	0.0003	<0.0001	0.0485
760-2	16	17	0.0004	<0.0001	0.3685	<0.0001	0.4639
782-1	0	33	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001
782-2	9	25	<0.0001	<0.0001	0.0005	<0.0001	0.0635
797-1	29	7	0.4414	0.0011	0.0033	<0.0001	<0.0001
797-2	14	4	0.7855	0.0051	0.0656	<0.0001	0.0022
797-3	24	19	0.0037	<0.0001	0.9540	<0.0001	0.0704
797-4	17	21	<0.0001	<0.0001	0.1525	<0.0001	0.7503
797-5	18	22	<0.0001	<0.0001	0.1515	<0.0001	0.7187
797-6	25	9	0.8430	<0.0001	0.0423	<0.0001	0.0002
808-1	23	23	0.0001	<0.0001	0.3928	<0.0001	0.2833
820-1	32	9	0.6521	<0.0001	0.0049	<0.0001	<0.0001
831-1	20	12	0.1025	<0.0001	0.4760	<0.0001	0.0200
832-1	15	4	0.6911	0.0077	0.0461	<0.0001	0.0012
856-1	4	28	<0.0001	<0.0001	<0.0001	<0.0001	0.0007
975-1	30	11	0.7868	<0.0001	0.0290	<0.0001	0.0001
975-2	28	7	0.4945	0.0008	0.0046	<0.0001	<0.0001
975-3	10	29	<0.0001	<0.0001	0.0001	<0.0001	0.0364
975-4	19	18	0.0009	<0.0001	0.5481	<0.0001	0.2590
975-5	27	9	1.0000	<0.0001	0.0233	<0.0001	0.0001
975-6	0	13	<0.0001	<0.0001	<0.0001	<0.0001	0.0021
975-7	32	7	0.3092	0.0025	0.0012	<0.0001	<0.0001

Table 3.6 cont.

$F_{3:4}$ Family	Max Like Plants	Non Max Like Plants	p-value for 3:1 Segregation Pattern	p-value for 15:1 Segregation Pattern	p-value for 9:7 Segregation Pattern	p-value for 63:1 Segregation Pattern	p-value for 27:37 Segregation Pattern
1023-1	10	8	0.0568	<0.0001	0.9526	<0.0001	0.2508
1023-2	11	10	0.0167	<0.0001	0.7208	<0.0001	0.3442
1044-1	20	18	0.0015	<0.0001	0.6530	<0.0001	0.1924
1045-1	17	22	<0.0001	<0.0001	0.1110	<0.0001	0.8593
1045-2	1	15	<0.0001	<0.0001	0.0001	<0.0001	0.0036
1045-3	0	32	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001
1045-4	26	19	0.0076	<0.0001	0.8363	<0.0001	0.0342
1082-1	13	12	0.0079	<0.0001	0.6684	<0.0001	0.3205
1107-1	28	13	0.3213	<0.0001	0.1201	<0.0001	0.0007
1107-2	21	8	0.7477	<0.0001	0.0793	<0.0001	0.0010
1107-3	25	8	0.9199	<0.0001	0.0239	<0.0001	0.0001
1107-4	31	7	0.3490	0.0019	0.0016	<0.0001	<0.0001
1107-5	25	11	0.4414	<0.0001	0.1105	<0.0001	0.0009
1107-6	19	20	0.0002	<0.0001	0.3430	<0.0001	0.4089
1107-7	35	5	0.0679	0.1025	0.0001	<0.0001	<0.0001
1107-8	28	6	0.3221	0.0060	0.0022	<0.0001	<0.0001
1107-9	25	11	0.4414	<0.0001	0.1105	<0.0001	0.0009
1107-10	27	9	1.0000	<0.0001	0.0233	<0.0001	0.0001
1120-1	20	18	0.0015	<0.0001	0.6530	<0.0001	0.1924
1136-1	1	32	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001
1136-2	4	9	0.0002	<0.0001	0.0640	<0.0001	0.4045
1136-3	0	29	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001
1136-4	8	21	<0.0001	<0.0001	0.0019	<0.0001	0.1113
1136-5	0	33	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001

Table 3.6 cont.

$F_{3:4}$ Family	Max Like Plants	Non Max Like Plants	p-value for 3:1 Segregation Pattern	p-value for 15:1 Segregation Pattern	p-value for 9:7 Segregation Pattern	p-value for 63:1 Segregation Pattern	p-value for 27:37 Segregation Pattern
1136-6	0	38	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001
1146-1	1	14	<0.0001	<0.0001	0.0001	<0.0001	0.0053
1204-1	32	10	0.8586	<0.0001	0.0092	<0.0001	<0.0001
1221-1	6	35	<0.0001	<0.0001	<0.0001	<0.0001	0.0004
1238-1	30	13	0.4281	<0.0001	0.0740	<0.0001	0.0003
1284-1	30	16	0.1255	<0.0001	0.2202	<0.0001	0.0016
1400-1	0	12	<0.0001	<0.0001	0.0001	<0.0001	0.0031
1438-1	10	20	<0.0001	<0.0001	0.0114	<0.0001	0.3261
1438-2	9	24	<0.0001	<0.0001	0.0008	<0.0001	0.0828
1438-3	9	27	<0.0001	<0.0001	0.0002	<0.0001	0.0368
1450-1	10	7	0.1235	<0.0001	0.8306	<0.0001	0.1649
1450-2	2	31	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001
1450-3	21	11	0.2207	<0.0001	0.2850	<0.0001	0.0073
1450-4	30	9	0.7815	<0.0001	0.0093	<0.0001	<0.0001
1450-5	31	13	0.4862	<0.0001	0.0575	<0.0001	0.0001
1471-1	24	19	0.0037	<0.0001	0.9540	<0.0001	0.0704
1485-1	32	13	0.5469	<0.0001	0.0445	<0.0001	0.0001
1505-1	23	11	0.3221	<0.0001	0.1804	<0.0001	0.0026
1741-1	2	30	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001
1858-1	23	12	0.2046	<0.0001	0.2590	<0.0001	0.0048
1868-1	24	22	0.0003	<0.0001	0.5773	<0.0001	0.1702
1868-2	14	17	0.0001	<0.0001	0.2133	<0.0001	0.7374
1874-1	13	15	0.0005	<0.0001	0.2948	<0.0001	0.6495
1876-1	26	11	0.5064	<0.0001	0.0856	<0.0001	0.0005

Table 3.6 cont.

$F_{3:4}$ Family	Max Like Plants	Non Max Like Plants	p-value for 3:1 Segregation Pattern	p-value for 15:1 Segregation Pattern	p-value for 9:7 Segregation Pattern	p-value for 63:1 Segregation Pattern	p-value for 27:37 Segregation Pattern
1930-1	29	5	0.1657	0.0417	0.0006	<0.0001	<0.0001
1930-2	1	2	0.0956	<0.0001	0.4236	<0.0001	0.7562
1930-3	4	0	0.2482	0.6056	0.0778	0.8011	0.0192

FIGURES

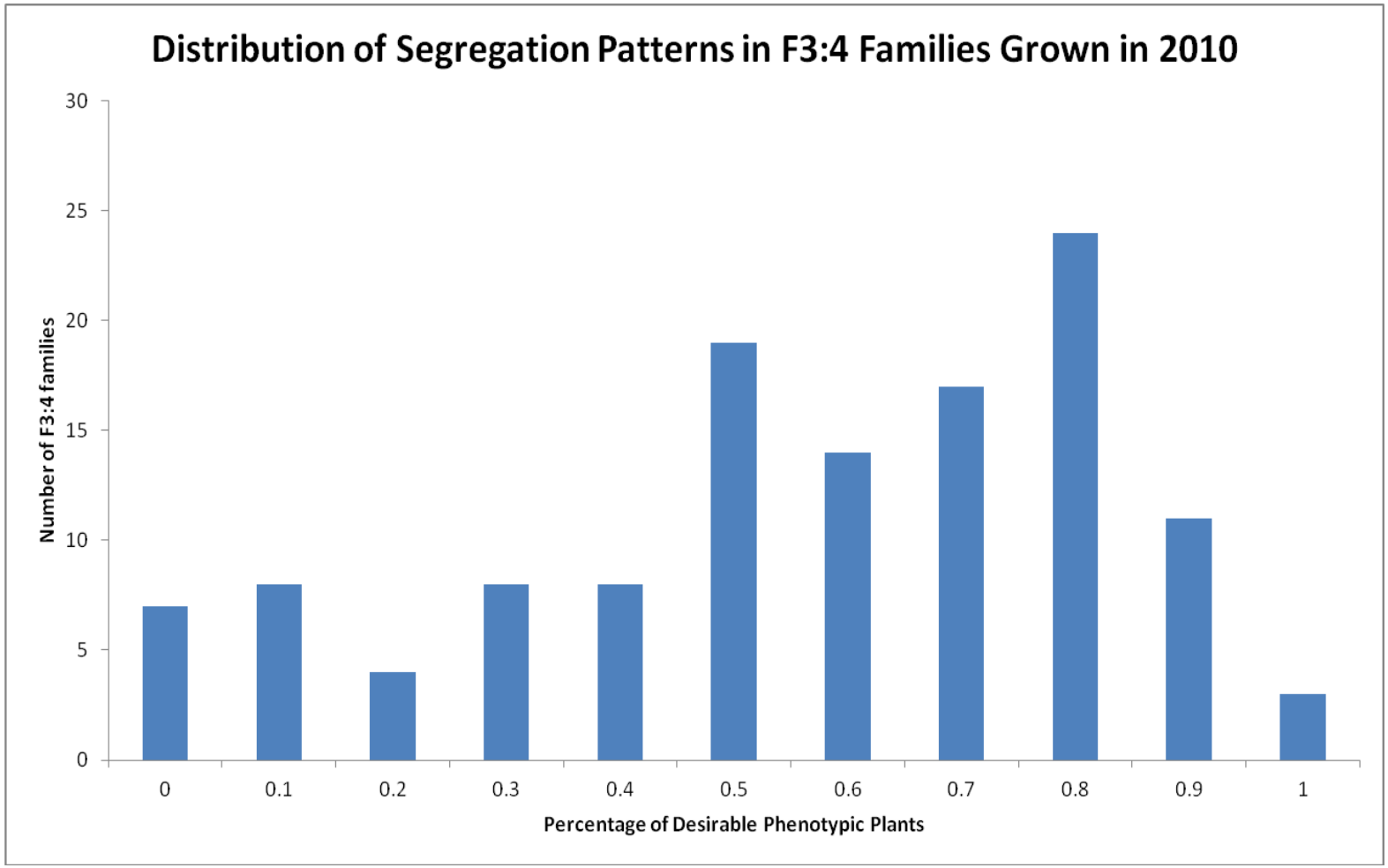


Figure. 3.1 Distribution of the segregation patterns in $F_{3:4}$ families grown in 2010.

CHAPTER 4

CHARACTERIZING THE EFFECTS OF PHENOTYPIC TRAITS OF GLYCINE SOJA PI 366122 ON AGRONOMIC PERFORMANCE THROUGH THE USE OF NEAR ISOGENIC LINES

ABSTRACT

Applied breeding with wild soybean is a new area of interest. It is not known whether incorporation of qualitative traits from wild soybean, *Glycine soja* (Subb. and Zucc.) to adapted breeding lines will be accompanied by linkage drag for yield or other negative associations. To quantify the impact of phenotypic traits from *G. soja* on agronomic performance, we developed near isogenic lines (NILs) for flower color, hilum color, seed coat color, and leaf shape, from the hybridization of 'N7103' soybean [*Glycine max* (L.) Merr.] by wild soybean accession PI 366122. A population of 54 F_7-derived NILs was developed and tested along with control cultivars three separate environments during 2010 and 2011 in replicated trials. A significant ($p<0.05$) association was observed between reduced 100-seed weight and purple flower color in one pair of NILs. In two separate NIL pairs, a significant ($p<0.05$) association was detected between purple flower color and increased lodging. We also identified three pairs of NILs where increased 100-seed weight was associated with round leaf shape, a trait inherited from the *G. soja* parent. We were unable to detect a significant difference in seed yield for any of the near isogenic lines. Results suggest that qualitative traits derived from wild soybean may not be genetically linked to poor agronomic performance or otherwise that these linkages may be easily overcome. Further field testing of the NIL populations is needed to validate these findings.

INTRODUCTION

A recent study (Delheimer 2012) showed that is possible to create agronomically acceptable soybean germplasm from single cross populations of *G. max* x *G. soja*. Of the approximately 200 breeding lines produced, the majority had the *G. max* flower, pubescence and seed coat color (white, gray, and yellow, respectively), even though color was not selected upon directly during the development of the lines. This trend suggests that qualitative traits emanating from the wild soybean may be associated with poor agronomic performance. To aid a breeder in the breeding with wild soybean, it would be helpful to identify possible linkages/pleiotropy between easily identifiable phenotypic traits and agronomic performance.

One method to measure the association of qualitative traits with seed yield is development of near isogenic lines (NILs). In soybean, NILs have been used to evaluate the effects of particular QTL on agronomic performance (Brucker et al. 2005, Delheimer et al 2010). These studies were focused on comparing different sources of soybean cyst nematode resistance and their associated effects on agronomic performance. The focus of the present study is to understand the relationship of phenotypic traits inherited from the *G. soja* parent, plant introduction (PI) 366122 (USDA-ARS, NGRP, 2012), and their associated effects on agronomic performance.

MATERIALS AND METHODS

DEVELOPMENT OF NEAR ISOGENIC LINES

The development of F_4-derived breeding lines from a domesticated x wild soybean hybridization has been described previously (Delheimer, 2012). Briefly, the *G. max* cultivar

‘N7103’(Carter et al., 2003) was crossed to *G. soja* PI 366122 in the summer of 2002. F_1 seed were sent to the winter nursery in Puerto Rico in December of 2002. In May of 2003, the bulk of F_2 seed was planted at Clayton, NC and harvested as a single large block. In 2004 at Kinston, NC, eight acres of F_3 plants were grown, and approximately 200 upright F_3 plants were selected. In 2005, $F_{3:4}$ lines were grown at Clayton, NC and approximately 700 single plants were selected to create $F_{4:5}$ lines. In 2006, these $F_{4:5}$ lines were grown for seed increase at Clayton, NC.

The NILs in the present study were derived from these plant rows. Individual F_5 plants were taken from lines which were segregating for phenotypic traits such flower color, pubescence color, seed color, etc. $F_{5:6}$ seed was planted in Clayton, NC in 2007. Individual F_6 plants were harvested from rows that continued to segregate for traits of interest. The $F_{6:7}$ seed were then planted at Clayton, NC in the spring of 2008. Individual F_7 plants were harvested from segregating rows. In 2009, the $F_{7:8}$ progeny rows were grown at Clayton, NC. Rows that were identified as breeding true for a particular trait (i.e. purple vs. white flower color) were bulk harvested to develop contrasting near isogenic lines that could traced back to a single F_6 plant.

YIELD TESTING

A total of 54 NIL were developed and tested in 2010 and 2011. Twenty-six of these lines were tested in a yield trial designated as TCLP 436 (Table 4.1). These lines consisted of contrasting pairs for the following phenotypic traits: flower color, seed coat color (yellow vs. brown), and hilum color. They were tested along with maturity checks: ‘Manokin’, (Kenworthy et al., 1996), ‘Dillon’ (Shipe et al., 1997), ‘5601T’ (Pantalone et al., 2003), ‘NC-

Roy' (Burton et al., 2005), 'N7002' (Carter et al., 2007), and N7103. TCLP 436 was grown in four replications in a randomized complete block design in 2010 at Clayton, NC and in 2011 at Clayton, NC and Kinston, NC. An additional 28 NILs were developed which contrasted for leaf shape (round vs. narrow leaflets), and they were grown in a yield test designated TCLP 437 (Table 4.2) in a randomized complete block design consisting of four replications at Clayton, NC in 2010 and at both Clayton, NC and Kinston, NC in 2011.

FIELD PLOT TECHNIQUE, TRAITS EVALUATED, AND STATISTICAL ANALYSIS

The USDA breeding program in Raleigh, NC managed all yield plots grown in Clayton and Kinston, NC. All plots were planted in three rows, 97 cm wide, which were end trimmed to a length of approximately 4.5 m near maturity. The center row was harvested from these plots, constituting a final plot area of 4.5 m^2. Plant populations were approximately 306,000 plants ha^{-1}.

Plots were rated for flower color, seed coat color, hilum color, leaf shape, plant height, maturity, lodging, 100-seed weight, and seed yield. Height was recorded for each plot as the mean of 3 plants, and maturity was recorded as the first day on which at least 95% of the pods turned a mature color. Lodging was measured on a 1 to 5 scale, where 1 indicates all plants were erect and 5 indicates all plants were prostrate. For TCLP 437, plants were visually rated as either having narrow or round leaves during the R5 stage (mid-podfill stage). Leaf shape was also quantified as the mean leaf length and leaf width on 5 leaves per plot at approximately half way up the main stem.

Analysis of variance was performed using the GLM procedure of SAS (SAS Institute, 2007). Fischer's protected LSD was used for comparison of genotypes over locations using

the genotype x location mean square as the error estimate. Linear contrasts were also used to compare NILs for particular phenotypic traits.

RESULTS

Across the three environments tested, there was no significant ($p<0.05$) difference within NIL groups for seed yield in TCLP 436 (Table 4.1). Purple flower color, a trait found in the *G. soja* parent, was associated with greater lodging scores and smaller 100-seed weight based on LSD comparison of three of the NIL groups. Combining across NIL groups with common traits revealed no significant ($p<0.05$) associations between agronomic performance and the traits of interest (Table 4.3)

Across the three environments tested, we did not observe any significant ($p<0.05$) differences within NIL groups for seed yield in TCLP 437 (Table 4.2). Round leaf shape, a trait derived from the *G. soja* parent, was associated with larger 100- seed weight based on LSD comparison of three of the NIL groups. Combining the data of the different narrow and round lead NIL groups, a significant ($p<0.05$) association was detected between round leaf shape and increased 100-seed weight. There was also a significant ($p<0.05$) difference between the narrow and round leaf NILs for leaf width and length measurements (Table 4.4).

DISCUSSION

In the development of breeding material from wild by domesticated crosses of soybean, it would be beneficial for breeders to select against easily identifiable traits derived from the wild parent that are associated with negative agronomic performance. We found that the purple flower color derived from *G. soja* was associated with decreased 100-seed in two pair of NILs but not in a third. Thus, it may be possible to break this linkage

through recombination. We noted prior to NIL development that the majority of breeding lines tested by Delheimer (2012), had the white flower color trait, but there appeared to be no association of flower color and seed yield in the NILs tested. Thus, the skewed frequency toward white flower color in the progeny could be related primarily to both conscious and unconscious selection for increased seed size.

The NILs tested in population TCLP 437 had contrasting pairs for leaf shape. The narrow leaf trait (i.e. shorter leaf width and longer leaf length as compared to round leaves), was significantly ($p<0.05$) associated with smaller seed size. The narrow leaf trait is associated with the *G. max* parent N7103. Previous studies have mapped QTL affecting seed weight close to the major leaf shape gene *ln* (Reinprecht et al., 2006; Sebolt et al., 2000). The QTL reported by Sebolt et al. was conducted using a *G. soja* parent, from which smaller seed size was derived. Delheimer (2012) used single marker analysis of variance on a population derived from the cross of the same parents studied here. They reported several significant associations for seed size derived from either parent, but none were located in this region. Overall, the findings of this study provide evidence that the phenotypic traits derived from the *G. soja* process are not associated with negative agronomic performance. As breeders begin to develop exotic material using the wild soybean, it appears it is not necessary to select against phenotypic traits such as purple flower color or brown seed coat color. Further analysis of these populations could further quantify these relationships, and provide even more useful tools for soybean breeders.

REFERENCES

Brucker, E., T. Niblack, F.J. Kopsich-Obuch, and B.W. Diers. 2005. The effect of *rhg1* on reproduction of *Heterodera glycines* in the field and greenhouse and associated effects on agronomic traits. Crop Sci. 45: 1721-1727.

Burton, J.W., T.E. Carter, and D.T. Bowman 2005. Registration of 'NC-Roy' soybean. 2005. Crop Sci. 45: 2654.

Carter, T.E., J.W. Burton, M.R. Villagarcia, Z. Cui, X. Zhou, M.O. Fountain, D.T. Bowman, and A.S. Niewoehner. 2003. Registration of 'N7103' soybean. Crop Sci. 43: 1128.

Carter Jr, T.E., J.W. Burton, M.O. Fountain, P.E. Rzewnicki, M.R. Villagarcia, and D.T. Bowman. 2007. Registration of soybean cultivar 'N7002.' Journal of Plant Reg. 1:93-94.

Delheimer, J.C. T. Niblack, M. Schmidt, G. Shannon, and B.W. Diers. 2010. Comparison of the effects in field tests of soybean cyst nematode resistance genes from different resistance sources. Crop Sci. 50: 2231-2239.

Delheimer, J.C. 2012. Agronomic performance of breeding lines developed from a single cross of *Glycine max* X *Glycine soja*. Dissertation, Chapter 1-2.

Kenworthy, W.J. 1996. Registration of 'Manokin' soybean. Crop Sci. 36: 1079.

Pantalone, V.R., F. L. Allen, and D. Landau-Ellis. 2003. Registration of '5601T' soybean. Crop Sci. 43: 1123-1124.

Reinprecht, Y., V.W. Poysa, K. Yu, I. Rajcan, G.R. Ablett, and K.P. Pauls. 2006. Seed and agronomic QTL in low linolenic acid, lipoxygenase-free soybean (*Glycine max* (L.) Merrill) germplasm. Genome 49: 1510-1527.

Sebolt, A.M., R.C. Shoemaker, and B.W. Diers. 2000. Analysis of a quantitative trait locus allele from wild soybean that increases seed protein concentration in soybean. Crop Sci. 40: 665-669.

Shipe, E.R., J.D. Mueller, S.A. Lewis, P.F. Williams, Jr., and J. P. Tomkins. 1997. Registration of 'Dillon' soybean. Crop Sci. 37: 1983.

USDA-ARS National Genetic Resources Program (USDA-ARS NGRP). 2012. *Glycine soja* Sieb. And Zucc. PI 366122. Germplasm Resources Information Network (GRIN). [Online database] National Germplasm Resources Laboratory, Beltsville, MD. Available at http://www.arsgrin.gov/cgi-bin/npgs/acc/display.pl?1274061 (verified 09 Feb. 2012).

TABLES

Table 4.1 Agronomic Data of NIL populations grown in TCLP 436 averaged across 3 locations from 2010-2011. Lines from same F_7 row trace back to same F_6 plant.

Genotype	F_3 Plant	F_7 Row	Trait of Interest†	Yield	Maturity Date	Lodging‡	100 Seed Weight
				kg ha^{-1}	Oct.1=1		g
MXS-NIL-1	19	11	WF	1881	14	2.5	8.3
MXS-NIL-2	19	11	PF	2009	14	2.6	7.4
MXS-NIL-3	103-1	185	PF	1737	19	3.0	5.9
MXS-NIL-4	103-1	185	WF	1471	21	3.3	6.3
MXS-NIL-5	50-1	242	WF	1539	19	2.7	6.4
MXS-NIL-6	50-1	242	PF	1555	17	3.1	5.8
MXS-NIL-15	50-1	205	YS,WF	1738	20	2.7	6.1
MXS-NIL-16	50-1	205	BS,PF	1513	19	3.1	5.5
MXS-NIL-17	67	547	YS	1791	22	2.6	8.6
MXS-NIL-18	67	547	BS	1693	21	2.3	8.8
MXS-NIL-19	67	565	YS	1886	21	2.5	8.0
MXS-NIL-20	67	565	BS	1961	21	2.5	8.3
MXS-NIL-23	20	437	BLHI	1316	17	2.5	6.6
MXS-NIL-24	20	437	BLHI	1800	20	2.5	7.6
MXS-NIL-25	20	437	BRHI	1662	18	2.4	7.1
MXS-NIL-26	20	437	BRHI	1746	19	2.4	7.4
$LSD_{0.05}$				437	4	0.4	0.7

† Trait of interest corresponds to reason for NIL. WF=White flower color, PF=Purple flower color, YS=Yellow seed coat color, BS=Brown seed coat color, BLHI=Black hilum color, BRHI=Brown hilum color.

‡ 1 indicates all plants erect, 5 indicates all prostrate.

Table 4.2 Agronomic Data of NIL populations grown in TCLP 437 averaged across 3 locations from 2010-2011. Lines from same F_7 row trace back to same F_6 plant

Genotype	F_3 Plant	F_7 Row	Trait of Interest†	Yield	Maturity Date	Lodging‡	100 Seed Weight	Leaf Width	Leaf Length
				kg ha^{-1}	Oct.1=1		g	cm	cm
MXS-NIL-27	77	344	R	1798	25	2.9	8.0	6	11
MXS-NIL-28	77	344	N	1600	27	3.3	7.9	3	10
MXS-NIL-29	77	348	R	1705	26	3.0	8.3	5	9
MXS-NIL-30	77	348	N	1714	26	2.9	7.4	4	11
MXS-NIL-33	77	370	N	1574	23	3.3	7.7	3	10
MXS-NIL-34	77	370	R	1696	25	3.1	8.4	5	10
MXS-NIL-35	77	372	R	1774	28	3.0	8.5	5	10
MXS-NIL-36	77	372	N	1842	26	2.9	8.2	4	11
MXS-NIL-37	81	394	R	1453	25	3.9	6.6	5	9
MXS-NIL-38	81	394	N	1425	24	3.8	5.8	3	9
MXS-NIL-39	81	406	N	1740	25	3.1	7.2	3	11
MXS-NIL-40	81	406	R	1760	26	2.9	7.7	5	9
MXS-NIL-41	81	406	R	1684	25	2.8	7.8	6	10
MXS-NIL-42	81	406	N	1725	27	3.2	6.7	3	11
MXS-NIL-43	81	406	R	1754	26	3.4	7.2	5	9
MXS-NIL-44	81	407	N	1705	26	3.0	7.1	3	11
MXS-NIL-45	81	407	N	1779	25	2.7	7.1	3	11
MXS-NIL-46	81	407	R	1669	26	3.0	7.9	5	9
MXS-NIL-49	20	479	R	1796	20	2.6	8.0	5	10
MXS-NIL-50	20	479	N	1796	19	2.4	7.2	3	11
MXS-NIL-53	67	544	R	1918	19	2.8	8.5	6	10
MXS-NIL-54	67	544	N	1771	19	2.7	8.0	4	11
$LSD_{0.05}$				403	3	0.5	0.7	0.5	0.5

† Trait of interest corresponds to reason for NIL. N=Narrow Leaf R=Round Leaf.

‡ 1 indicates all plants erect, 5 indicates all prostrate.

Table 4.3 Mean comparisons across NIL groups of TCLP 436 for the various traits of interest.

Trait	Number of Genotypes	Yield	Maturity	Lodging†	100 Seed Weight
		kg ha^{-1}	Oct 1=1		g
Purple Flower	7	2100	20	3.2	7.2
White Flower	7	1895	20	3.0	7.3
P>F		0.1666	0.8687	0.1527	0.8032
Brown Seed	3	1801	20	2.6	7.5
Yellow Seed	3	1814	20	2.6	7.6
P>F		0.9504	0.8211	0.7944	0.9844
Black Hilum	2	1606	18	2.5	7.1
Brown Hilum	2	1686	18	2.4	7.2
P>F		0.7808	0.9231	0.7718	0.8920

† 1 indicates all plants erect, 5 indicates all prostrate.

Table 4.4 Mean comparisons of the leaf shape groups of TCLP 437.

Leaf Shape	Number of Genotypes	Yield	Maturity	Lodging†	100 Seed Weight	Leaf Width	Leaf Length
		kg ha^{-1}	Oct 1=1		g	cm	cm
Narrow	13	1713	24	2.9	7.3	4	11
Round	15	1786	24	3.1	7.9	5	10
P>F		0.1983	0.4189	0.1605	0.0235	<0.0001	0.0010

† 1 indicates all plants erect, 5 indicates all prostrate.

APPENDIX

Appendix A. Means of agronomic traits for breeding lines tested in yield test TCLP 384, averaged over 5 environments from 2008 to 2010.

Genotype†	Yield	Protein	Oil	Maturity	Est. MG‡	Lodging§	100-Seed Weight	Yield as % Check¶	% *G. soja* SNP alleles#
	kg ha^{-1}	g kg^{-1}	g kg^{-1}	Oct. 1 = 1			g		
Stressland	1666	422	195	4	4.3	1.7	14.8	-	-
Holladay††	2802	392	207	13	5	1.4	15.2	-	-
5601T‡‡	2641	412	197	15	6	1.6	14.6	-	-
NC-Roy§§	2853	414	193	26	7	2.1	13.0	-	-
NMS4-7-138	1230	449	168	25	6.7	2.8	8.2	42	34
NMS4-15-214	1303	457	153	23	6.5	3.2	5.8	46	37
NMS4-15-219	1088	425	182	18	6.0	4.4	8.2	40	31
NMS4-15-216	1277	454	152	23	6.6	2.8	5.8	44	37
NMS4-17-228	1404	464	150	20	6.1	3.5	7.1	51	39
NMS4-17-226	1727	463	150	21	6.3	2.6	8.0	61	36
NMS4-17-230	1861	450	164	23	6.6	3.1	9.0	65	36
NMS4-20-274-1	1693	423	147	23	6.6	2.8	6.0	59	37
NMS4-20-272	1398	476	145	24	6.6	3.2	8.1	48	39
NMS4-26-293	1693	452	151	15	5.6	3.2	8.3	65	37
NMS4-27-296	1854	460	158	24	6.7	3.1	7.9	64	35
NMS4-28-299	1324	454	165	22	6.4	2.8	8.6	47	34
$LSD_{0.05}$	356	12	9	3		0.5	0.8	-	-

† NMS4 corresponds to lines in population JCD-1. The first number after the hyphen corresponds to F_3 plant.
‡ Estimated from regression of maturity date of conventional check against published maturity group.
§ 1 indicates all plants erect; 5 indicates all plants prostrate.
¶ Based on comparison of yield mean vs. regression of check means.
 # Based on 558 polymorphic SNP markers
†† Represents the mean of three duplicate entries of Holladay, 5601T, and NC-Roy.

Appendix A cont.

Genotype†	Yield	Protein	Oil	Maturity	Est. MG‡	Lodging§	100-Seed Weight	Yield as % Check¶	% *G. soja* SNP alleles#
	kg ha^{-1}	g kg^{-1}	g kg^{-1}	Oct. 1 = 1			g		
NMS4-30-301	1875	440	172	22	6.4	2.4	7.8	66	36
NMS4-32-303-1	1384	444	166	21	6.3	3.3	8.1	49	38
NMS4-32-303-2	1377	442	155	23	6.5	3.1	8.1	48	37
NMS4-32-303	1518	448	149	25	6.7	3.4	7.7	52	37
NMS4-44-333	1351	422	180	27	7.0	2.8	10.0	45	33
NMS4-44-334	1250	407	184	25	6.7	3.1	8.1	43	37
NMS4-45-338	1619	423	173	20	6.2	3.7	7.0	59	35
NMS4-46-340	1579	444	160	20	6.2	3.3	7.4	57	32
NMS4-55-402	1700	440	162	21	6.3	3.1	6.9	61	36
NMS4-64-449	1653	466	148	22	6.5	3	7.1	58	31
NMS4-66-460	1707	427	177	28	7.1	3.4	6.8	57	33
NMS4-74-473	1767	440	165	26	6.9	3.2	6.3	59	32
NMS4-83-21	1411	433	164	25	6.7	3.4	6.9	48	33
NMS4-111-568	1492	469	153	22	6.4	2.8	6.5	53	39
NMS4-111-572	1330	466	156	24	6.6	2.9	6.8	46	41
NMS4-128-660	1082	431	178	24	6.7	4.3	7.5	37	35
NMS4-135-676	1505	452	154	19	6.1	3.2	7.1	55	37
NMS4-147-694	1498	456	159	22	6.5	3.2	8.7	53	43
NMS4-30-301	1875	440	172	22	6.4	2.4	7.8	66	36
$LSD_{0.05}$	356	12	9	3		0.5	0.8		

† NMS4 corresponds to lines in population JCD-1. The first number after the hyphen corresponds to F_3 plant.
‡ Estimated from regression of maturity date of conventional check against published maturity group.
§ 1 indicates all plants erect; 5 indicates all plants prostrate.
¶ Based on comparison of yield mean vs. regression of check means.
 # Based on 558 polymorphic SNP markers.
†† Represents the mean of three duplicate entries of Holladay, 5601T, and NC-Roy.