Assessing the efficiency of phenotypic and molecular genotype selection methods for complex traits in Soybean

A Dissertation Presented for the Doctor of Philosophy Degree
The University of Tennessee, Knoxville

Catherine Nyaguthii Nyinyi
December 2011

Acknowledgements

I would like to express my gratitude to all those who have supported me throughout my studies, research and in the preparation of this dissertation.

My major professor Dr. Vincent Pantalone for offering me the opportunity to continue my post graduate education under his tutelage at the University of Tennessee. Thank you for your generosity, patience and for allowing me to broaden my horizons and satisfy my many curiosities by working on many different projects. My committee members Dr Fred Allen, Dr. Dean Kopsell, Dr. Arnold Saxton, and Dr. John Sorochan for their willingness to serve and their seemingly endless patience whilst helping me through my research.

The Bean Team: Benjamin Fallen, Deborah Landau-Ellis, Elizabeth Meyer, Christopher Smallwood, Benjamin and Susannah Wiggins, Nicole Tacey ,and Greg Allen for their friendship, ideas and help in making many tasks easier. Other folks at University of Tennessee, Research and Education Centers for all their help overseeing my research plots. Tetsuaki Ishibashi and James Klein III for overseeing my research plots and collecting data for me at the University of Arkansas and Southern Illinois University, respectively.

Dr. David Hyten for the tremendous effort needed to genotype my population. Dr. Karl Broman at the University of Wisconsin for helping me with the construction of my molecular map. Dr. William Klingeman III for helping overcome many administrative hurdles.

I would also like to thank my family and friends for patience and understanding in my seemingly long journey in academia.

Support from the Tennessee Soybean Promotion Board, the United Soybean Promotion Board, and the Tennessee Agricultural Experiment Station is greatly appreciated.

Abstract

Soybean [*Glycine max* (L.) Merrill] is an important source of protein and oil for both nutritional and industrial applications. Increasing seed yield and protein concentration is the main goal of many soybean breeders to meet market demands. Soybean breeders have occasionally succeeded in producing high yielding cultivars with increased protein content using conventional means despite the negative correlation that exists between these two traits. The efficiency of breeding for seed yield and protein concentration improvement in soybean could be increased using marker assisted selection (MAS) breeding strategies to select genotypes containing favorable alleles for faster cultivar development. The objective of this study was to identify quantitative trait loci (QTL) associated with seed yield, and separately, seed protein concentration and then compare phenotypic selection (PHE) and MAS approaches for seed yield and protein concentration improvement. Two hundred and eighty two F_5 derived recombinant inbred lines (RILs) were developed from a cross of Essex × Williams 82 and genotyped with 1586 single nucleotide polymorphism (SNP) markers. The population was divided by days to maturity (10 days) into three tests (early, mid and late) each with 94 genotypes, with one genotype overlapping in maturity in the mid and late tests. In 2009, the three tests, parents and checks were grown in a randomized complete block design (RCBD) in: Fayetteville, AR; Harrisburg, IL and, Knoxville, TN replicated three times, and evaluated for seed yield and protein concentration. Data were combined within each test across three locations and analyzed using the MIXED procedure of SAS to determine that there were significant genotypic differences among RILs. Composite interval mapping (CIM) detected nine seed yield and ten protein concentration QTL which may be good candidates for MAS as they were environmentally stable. Selections to compare PHE, and MAS for seed yield and protein concentration provided 8 replicated field tests in four relative maturity groups grown in a RCBD

replicated three times in three locations in Tennessee, in 2010. We demonstrated that both MAS and PHE may be used to select quantitative traits; however, more studies are required to optimize MAS for quantitative trait improvement.

Table of Contents

List of Tables

List of figures

Part 1

Introduction and Literature Review

Introduction

Soybean [*Glycine max* (L.) Merrill] is an annual legume originally domesticated in China (ca.1700-1100 B.C.), and it is now grown worldwide. Soybean belongs to the genus *Glycine,* which is divided into subgenera *Glycine* (perennials) and *Soja* (annuals). Subgenus *Soja* contains the species *Glycine max* (L.) Merr. and *Glycine soja* Seib. and Zucc commonly referred to as cultivated and wild soybeans respectively (Hymowitz, 2004).

Currently, the world's largest producers of soybean are the United States of America (U.S.), Brazil, Argentina, China and India (Wilcox, 2004). The U.S. soybean crop in 2010 was valued at $38.9 billion dollars, up $7.1 billion from the previous year. In that same year, 69% of the world's vegetable protein consumption was soybean meal (SBM) while 29% of vegetable oil consumption was soybean oil (American Soybean Association, 2011). The adaptability of soybean to many environments and the intensive soybean breeding efforts have made it an important component of human nourishment, animal feedstock ,and is used in many industrial processes and products.

Soybean seed contains about 40% protein and 20% oil, both of which are important commercially. In addition, soybeans also contain phytochemicals (e.g., phytoestrogens), which have received much attention due to claims of numerous health benefits (Liu, 1997). Soybean components are used in the production of food and for industrial applications. Traditional soybean foods originated in Asia where the crop was first domesticated and cultivated. Asian soy foods can be divided into two groups, non-fermented and fermented. Non- fermented foods include *tofu,* soymilk*, yuba,* sprouts, *kinako* (dry, roasted, and dehulled flour*)* and *edamame* (green immature soybeans). Fermented soy foods include *miso, tempeh, natto* and soy sauce (Hymowitz and Newell, 1981). Historically, soybean has been a staple in the Asian diet, but it

has been adopted into Western cuisine because of the nutritive value and the health benefits (Liu, 1997). Soybean processing results in the production of soybean meal (SBM) and oil. Soybean meal is primarily used as animal feed, but soy protein is also processed into protein isolates, bakery mixes, baby foods, meat analogs and extenders, protein beverages, pastas, etc. (Lusas, 2004). Soybean oil can be processed into margarine, soy vegetable oil, mayonnaise, biodiesel, lubricants, paints, inks and many more edible or industrial oil products.

Seeds may be classified as commodity beans and as food grade beans, which differ in characteristics based on the desired end product. Food grade beans, which also include the green vegetable type *edamame,* are grown for direct human consumption or for processing into edible products. In general, food grade soybeans have a lighter seed coat, clear hilum, high-protein and low oil content (Liu, 1997). There are also some other specifications for taste, sucrose content, palatability etc. For example, higher sucrose content in *edamame* soybean would make it more palatable than a cultivar with more starch.

Commodity beans are processed into SBM and oil is produced as a byproduct of the pressing process. Soybean meal is an important source of animal protein. Poultry and swine industries rely heavily on SBM as a source of protein for confined animal feeding operations (CAFOs). The use of SBM to provide an affordable source of protein in CAFOs has also brought to attention the limitations of SBM because of its phytate content as well as its low amounts of the amino acids methionine (MET) and cysteine (CYS) (Almquist et al., 1942; Krober, 1956; Raboy et al., 1984). Methionine is an essential amino acid that must be provided in the diet while CYS is not. However, because CYS can be synthesized from MET, low CYS content further reduces the amount of MET that can be provided by SBM (Clarke and Wiseman, 2000). Panthee

et al. (2006a; 2006b) identified genomic regions associated with MET and CYS in soybean seeds which could be helpful in improving protein quality.

Monogastric animals such as fowl and swine are not able to digest the compound phytate, which is the secreted in the animal waste (Pallauf and Rimbach, 1997). Large accumulations of phytate can have undesirable environmental repercussions when excessive phosphorus loads pollute water bodies and result in the destruction of wildlife. Hill et al. (2009) found that feeding swine a diet of low phytate corn and low phytate soybeans resulted in a reduction of water soluble phosphorus excreted by swine as compared to animals fed a normal corn or normal soybean diet. In addition, they found that animals fed low or high phytate diet supplemented with phytase also had reduced water soluble phosphorus in excrement.
Plant breeding efforts to improve SBM by reducing the amount of phytate have succeeded in identifying genomic regions that control phytate content (Scaboo et al., 2009; Walker et al., 2006). These tools will be useful in genotypic selection of low phytate lines.

Soybean oil primarily contains the following fatty acids: palmitic acid (16:0), stearic acid (18:0), oleic acid (18:1), linoleic acid (18:2), and linolenic acid (18:3) (Wilson, 2004). Linolenic acid in soybean oil reduces its shelf life, causing the oil to become rancid. The common solution to lengthen shelf life is to partially hydrogenate the oil to reduce the linolenic acid content. Hydrogenation of soybean oil results in the formation of trans-fatty acids which are coronary disease risk factors. Through plant breeding, low linolenic acid soybean cultivars have been developed to counter this problem. Plant breeders have been successful in releasing cultivars with less than 1% linolenic acid to eliminate the need for hydrogenation and produce healthier oils that have a longer shelf life (Clemente and Cahoon, 2009; Cober et al., 2010; Pantalone et al., 2004).

Another important application of soybean oil is biodiesel, which is a good alternative to fossil fuel and is now available in many fueling stations. Pradhan et al.(2009) conducted a life-cycle assessment (LCA) to quantify and compare the environmental and energy flows associated with biodiesel and petroleum-based diesel. The LCA assesses the energy yield of biodiesel or petroleum based diesel per unit of fossil energy consumed. They used a fossil energy ration (FER), defined as the ratio of energy output of the final biofuel product to the fossil energy used to produce it. They performed a life-cycle inventory (LCI) which consisted of: feedstock production and transportation, soybean processing with biodiesel conversion, and product distribution. Pradhan et al. (2009) found that the FER improved from the previously reported 3.2 to 4.56. The improvement was attributed to improved cultural practices such as no-till farming, improvement of factories and plants, and an increase in soybean yield over time. Their study shows that the production of soybeans for biodiesel production is a sustainable means of reducing dependence on fossil fuels for energy. Soybean breeders have been able to produce cultivars which produce soybean oil whose profile is suitable for the production of biodiesel.

Fallen et al. (2011) evaluated the fuel properties of recombinant inbred lines (RILs) from a cross of early generation Roundup Ready™ lines and a high oleic acid line (540 g kg^{-1}). Oleic acid and palmitic acid are monosaturated fatty acids known to increase the oxidative stability and cold flow properties of biodiesel. Oil extracted from the highest oleic acid content line, TN07-93RR, and a commercial cultivar, AG3906, was converted into biodiesel and tested against the American biodiesel standard (ASTM D6751) and the European Biodiesel Standard (EN 14214). TN07-93RR, and was acceptable for both biodiesel standards, while the commercial cultivar AG3906 was not acceptable for the European standard. This study is an example of successful soybean breeding efforts to produce cultivars that meet industry and consumer needs.

To meet the demands of the market, producers, plant breeders, agronomists and other professionals in the industry are charged with finding new ways to increase soybean production. There are many cultural methods that can be employed to increase yield. These methods include chemical weed and disease control, manipulating seeding rates at planting, reducing abiotic stress, and growing adapted high-yielding cultivars. Plant breeders have been successful in increasing soybean yield while simultaneously improving resistance to abiotic and biotic and abiotic stress factors (Morrison et al., 2000; Specht et al., 1999).

Literature Review

The main goal of soybean breeding is to increase yield while improving other characteristics such as protein and oil content. Conventional breeding methods rely on phenotypic data to make selections for the development of superior genotypes. Such methods have been used for a long time and have been successful in the improvement of many traits such as yield, protein, oil, seed size, biotic and abiotic stress tolerance (Pathan and Sleper, 2008). High-yielding cultivars can produce more soybeans per hectare and therefore earn the producer more income with fewer resources. In 2010, U.S. farmers produced 90.6 million metric tons (MMT) of soybeans and received on average $430 per metric ton (American Soybean Association, 2011). Clearly, soybeans are an important crop in the USA economy and this makes yield a very important trait.

Several studies have been conducted to determine the amount of yield increase that has occurred over the years due to plant breeding efforts. Yield trends in the USA and in other countries have been examined together with the effects of breeding efforts on other traits. Specht et al. (1999) conducted a study to investigate yield increases over time in Nebraska (NE), Iowa (IA) and Missouri (MO) of the U.S. They obtained annual U.S. yield estimates from the

U.S. National Agricultural Research Service (NASS) for the years 1924 to 1998. Regression analysis of that data showed linear yield increase of 22.6 ± 0.7 kg ha^{-1} within the period and 40% faster yield improvement (31.2 ± 4.8 kg ha^{-1}) in 1972-1998 than twenty-five years earlier. The comparison between irrigated and rainfed soybeans in NE showed an absolute difference of 800 kg ha^{-1} yr^{-1} with the irrigated yields having about a 40% improvement over the rainfed yields. The authors note a large yield increase when the breeding cultivars shifted from being plant introduction-based to hybridization-based around 1943. Further analysis of "newer" cultivars before and after 1976 showed yield improvement at 30 kg ha^{-1}. The effect of increasing yield on seed protein content was not reported in that study.

In another study, Ustun et al. (2001) evaluated yield increases and other important agronomic traits in the mid south U.S. They used two ancestral lines (S-100, CNS) and six cultivars (Lee, Ogden, Hill, Essex, Hutcheson, and TN 5-85) representing first, third and fifth generation soybean cultivars. They recorded maturity, plant height, lodging, protein and oil content and yield. Their results showed a 24.4% yield increase from ancestral lines to fifth generation cultivars and yield maintained an increasing trend at 14 kg ha^{-1} yr^{-1}. The study also showed that the newer cultivars were more responsive to environmental changes, i.e. improved growing conditions resulted in higher yield. They also reported a decrease in protein from ancestral southern lines to fifth generation cultivars that corresponded with an increase in oil content. The data collected however showed the possibility of producing cultivars that were high-yielding with relatively high-protein and oil content such as Essex (Smith and Camper, 1973).

Recently, Jin et al. (2010) conducted a study to gauge genetic progress in NE China over a period of 56 yrs (1950—2006). They placed four constraints on the methodology of the study

based on previous research namely: 1) experiments must be conducted under field conditions; 2) measurements were collected from comparable plots; 3) cultivars released at different times must be compared simultaneously and, 4) cultivars must be grown within their maturity group. They collected 45 cultivars in maturity groups (MG) 00 and 0 and grew them for 3 consecutive years to determine the contribution of the breeding programs to yield increase, agronomic and physiological changes as well as yield stability and its relation to genetic change over time. The results of the study showed an average yield increase of 32.5% or 0.58 % yr^{-1}which was attributed to breeding efforts. In addition to the yield increase they noted improved resistance to stress, reduced plant height, increased lodging resistance as well as reductions in disease and pest infestations in the newer cultivars. Although they showed a difference in both protein (37— 45.5%), and oil (16.7— 22 %) content, the researchers report that there was no significant association of either trait with yield during the 56 yr period examined.

Morrison et al. (2000) conducted a study to examine changes in agronomic traits and seed quality traits associated with 58 years of soybean breeding for short season cultivars in Canada. The correlation of the traits showed an increase of seeds per plant, oil concentration and disease resistance with an increase in yield and a decrease in protein content. Similar results were reported by Jin et al. (2010). The examination of yield improvement over time showed that yield improvement efforts have been successful in soybeans. However, the reported average yield increase of about 0.5% might be improved upon with the use of new tools such as molecular breeding techniques that are now available to breeders. In addition breeders could attempt to increase or maintain current protein concentration without significantly reducing yield.

Recurrent selection is an example of a conventional breeding method employed to increase the number of favorable alleles in a given population enabling subsequent selection of

lines or individuals containing the best combination of alleles. Using recurrent selection, Wilcox (1998),) successfully selected for higher seed protein content and increased seed protein from 438 to 484 g kg^{-1}. The C0 population was developed from two F_2 populations from the crosses 'Beeson' (ms2ms2) x 'Corsoy' and 'Wells'(ms2ms2) x 'Hark' that were blended with a high-protein selection from the cross 'Cutler 71' x 'Pando'. The blended seed was planted in an intermating block pollinated by bees and only seed from the male sterile plants (ms2ms2) was harvested. After cycle six (C6) of intermating and selection for high-protein, there was no gain in protein content which suggested that the population had no further genetic variability to increase protein.

Piper and Fehr (1987) tested different strategies of recurrent selection for yield improvement in soybean. The results from three generations of intermating showed no significant gain after the first generation. Furthermore they reported that the more conventional breeding method for soybeans, which involves one season of intermating i.e., one hybridization, with more than one cycle of selection, was the most effective in the selection for yield. These results may indicate that it is not necessary to make many crosses to increase recombination with the expectation of greater genetic gain for yield. Moreover, increasing the number of intermating generations lengthens the cultivar development time period.

Molecular breeding tools have improved the accuracy and reduced the time required for trait selection. A quantitative trait such as yield, height or seed size is controlled by many genes. This makes it difficult to map or even select for individual genes affecting inheritance and expression. There can be greater breeding success when chromosomal regions identified with the trait of interest, referred to as quantitative trait loci (QTL) are mapped. Molecular DNA markers are differences in genetic sequences that can be used to differentiate between individuals.

Examples of molecular DNA markers are: Restriction Fragment Length Polymorphisms (RFLPs), Amplified Fragment Length Polymorphisms (AFLPs), Simple Sequence Repeats (SSRs) and Single Nucleotide Polymorphisms (SNPs). DNA markers can be used to identify the location (locus) of a gene or genes on a chromosome or to identify QTL.

Single locus SSRs are polymorphic and abundant in the soybean genome and enabled the construction of the first marker-dense soybean map (Cregan et al., 1999). This map consisted of 606 SSRs, 689 RFLPs, 79 RAPDs, 11 AFLPs, 10 isozymes and 26 classic loci. Although this map was informative, it had large gaps where SSR polymorphisms did not exist. Song et al. (2004).) used expressed sequence tags (ESTs) to develop more SSRs for the regions in the genome where gaps existed. Subsequently, JoinMap analysis of 1019 SSRs, 749 RFLPs, 13 AFLPs, 90 RAPDs, ten isozymes and 30 other markers resulted in 20 linkage groups that spanned 2523.6 cM Kosambi mapping distance.

Choi et al. (2007) created the first genetic transcript map of soybean by mapping one SNP in each of 1141 genes. They developed sequence tagged sites (STS) using EST and 3′-unigene sequence and then used the STSs for SNP discovery. They found 4241 STSs from which they developed 5551 SNPs and placed 1141 SNPs on the genetic map using segregating SNPs mapped in one or more of three mapping populations (the University of Utah 'Minsoy' × 'Noir1' (M×N), Minsoy × 'Archer' and the University of Minnesota 'Evans' × 'Peking'.). The new map placed 291 SNPs in 72 of the 112 gaps between 5 and 10 cM and another 111 SNPs were mapped in 19 of 26 gaps greater than 10cM in the Song et al. (2004) map.

Hyten et al. (2010) genotyped the three mapping populations used by Choi et al. (2007) with 3456 SNPs using the Illumina GoldenGate™ assay. The Illumina GoldenGate™ assay is an accurate high throughput genotyping platform that is capable of genotyping 96—1536 SNPs in

single reaction over 3 days (Hyten et al., 2008). Mapping of SNPs from this study as well as those from previous maps created the Consensus Map 4.0. In addition, the "Universal Soybean Linkage Panel" (USLP 1.0), containing 1536 SNPs developed from diverse germplasm, was developed as a tool to facilitate easier genotyping of many mapping populations. The integration of the USLP 1.0 and the consensus map resulted in a map that is 2296.4 cM containing 3792 SNPs, 1006 SSRs, 664 RFLPs, and 38 other markers.

The availability of dense molecular maps further broadens the possibilities of using molecular markers to directly select for a trait if the markers are tightly linked to the genes governing the trait. Marker assisted selection (MAS) is a breeding strategy that utilizes molecular genetic markers to select for a trait. For this method to be efficient and useful, the marker must correctly identify with the trait. Lande and Thompson (1990) examined the potential for MAS in reducing the time required to successfully select for a single trait. They determined that MAS is most efficient in selection for traits with low heritability and moderate to large additive genetic variance. In addition, larger population samples would probably have better results because they would likely have larger estimated additive genetic variance.

Bernardo (2008) reported 1200 QTL mapping studies for 12 major crops which included barley, bean, corn or maize, cotton, oat, potato, rice, sunflower, sorghum, tomato, wheat and soybean. These studies on average detected 3 to 5 QTL per trait tested. The author attributes this number of QTL reported in the literature to be due to the ease with which a researcher can perform a QTL detection study.

Creation of relatively small mapping populations (N=100—150), availability of molecular markers spaced about 10—15cM, analysis of phenotypic and molecular data with the correct statistical model on relatively user friendly computer systems and software has improved the

efficiency with which QTL can be detected and confirmed. The availability of confirmed QTL (cqQTL) is still relatively small in comparison to reported QTL in Soybase. Soybase data base contains 55 yield QTL, 16 yield/height QTL, 7 yield/seed weight QTL, 2 protein/oil QTL and 86 protein QTL as opposed to 1 yield and 2 protein cqQTL.

The application of QTL as a tool for the selection of complex traits by plant breeders has been minimal, mostly as tests for one major QTL controlling a large proportion of a trait's variation. Studies have shown successful introgression of QTL of interest into different background with mixed results. For example, Sebolt et al. (2000) conducted a study to test the effect of two *Glycine soja* alleles (previously mapped by Diers et al. (1992) on LGs E and I) for their effect on protein and oil as well as other agronomic traits. They used a $BC3F_4$-derived line from a backcross (BC) population created by crossing A81-356022 (recurrent parent) with an F_2 derived line from the cross A81-356022 by PI468916 (*Glycine soja*). They were able to confirm one of the QTL on LG I but not the one on LG E. Further, introgression of the protein QTL into three cultivars showed an effect in two of the three backgrounds tested. In a similar study, Reyna and Sneller (2001) sought to evaluate three protein QTL from the cultivar Archer, previously mapped by Orf et al. (1999) for their usefulness in southern genetic backgrounds for the increase of protein. They created F_6 derived near isogenic lines (NILs) from Archer and two southern commercial cultivars (Asgrow A 5403 and Pioneer 9641), and used them to test for yield differences. The results after testing showed that the yield difference between the NILs (Archer) and the NILs (southern) were not statistically significant.

The authors speculate that the Archer allele was not able to increase yield in those NILs because the QTL was environment specific. Moreover, the two commercial lines used as parents in the

study were southern elite parents and may have outperformed Archer and so the yield increase observed was not shown to be significant.

These results show that it is indeed possible to introgress QTL to improve traits in a breeding program. QTL studies however require extensive testing to ensure stability in different environments and genetic backgrounds to for this to be a truly useful tool. Moreover, there remains a need to confirm QTL to counter the "Beavis effect" which occurs when QTL with small effects are not detected in small populations which can lead to over estimation of the QTL that are detected (Beavis, 1998; 1994)

Objectives

The main objective of this research project is to test the usefulness of detected QTL for selections targeting the improvement of quantitative traits. The specific aims of this research project were:

1. Detect QTL for yield and protein content in an `Essex` × `Williams 82 cross population.
2. Test the effectiveness of marker assisted selection of QTL versus phenotypic selection for seed yield and protein concentration.

References

Almquist, H.J., E. Mecchi, C.R. Kratzer, C.R. Grau. 1942. Soybean protein as a source of amino acids for the chick. J. Nutr. 24:385-392.

American Soybean Association. 2011. A reference guide to important soybean facts and figures.

Beavis, W.D. 1998. QTL analyses: Power, precision, and accuracy. *In* A. H. Paterson (Ed.). Molecular dissection of complex traits. CRC Press LLC, Boca Raton, FL.

Beavis, W.D. 1994. The power and deceit of QTL experiments: Lessons from comparative QTL studies., 49th Annual corn and sorghum industry research conference. American Seed Trade Association, Washington, DC. pp. 250-266.

Bernardo, R. 2008. Molecular markers and selection for complex traits in plants: learning from the last 20 years. Crop Sci. 48:1649-1664.

Choi, I.-Y., D.L. Hyten, L.K. Matukumalli, Q. Song, J.M. Chaky, C.V. Quigley, K. Chase, K.G. Lark, R.S. Reiter, M.-S. Yoon, E.-Y. Hwang, S.-I. Yi, N.D. Young, R.C. Shoemaker, C.P. van Tassell, J.E. Specht, P.B. Cregan. 2007. A soybean transcript map: gene distribution, haplotype and single-nucleotide polymorphism analysis. Genetics 176:685-696.

Clarke, E.J., J. Wiseman. 2000. Developments in plant breeding for improved nutritional quality of soya beans I. Protein and amino acid content. J. Agric. Sci. 134:111-124.

Clemente, T.E., E.B. Cahoon. 2009. Soybean oil: genetic approaches for modification of functionality and total content. Plant Physiol. 151:1030-1040.

Cober, E.R., S.R. Cianzio, V.R. Pantalone, I. Rajcan. 2010. Soybean, *In* J. Vollmann and I. Rajcan (Eds.), pp. 57-90. Oil crops. Springer Dordrecht, Heidlberg, London, New York.

Cregan, P.B., J. Mudge, E.W. Fickus, D. Danesh, R. Denny, N.D. Young. 1999. Two simple sequence repeat markers to select for soybean cyst nematode resistance coditioned by the rhg1 locus. Theor. Appl. Genet. 99:811-818.

Fallen, B., V. Pantalone, C. Sams, D. Kopsell, S. Vaughn, B. Moser. 2011. Effect of soybean oil fatty acid composition and selenium application on biodiesel properties. J. Am. Oil Chem. Soc.:1-10.

Hill, B.E., A.L. Sutton, B.T. Richert. 2009. Effects of low-phytic acid corn, low-phytic acid soybean meal, and phytase on nutrient digestibility and excretion in growing pigs. J. Anim. Sci. 87:1518-1527.

Hymowitz, T., C.A. Newell. 1981. Taxonomy of the genus glycine, domestication and uses of soybeans. Econ. Bot. 35:272-288.

Hyten, D., Q. Song, I.-Y. Choi, M.-S. Yoon, J. Specht, L. Matukumalli, R. Nelson, R. Shoemaker, N. Young, 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., I.-Y. Choi, Q. Song, J.E. Specht, T.E. Carter, R.C. Shoemaker, E.-Y. Hwang, L.K. Matukumalli, 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.

Jin, J., X. Liu, G. Wang, L. Mi, Z. Shen, X. Chen, S.J. Herbert. 2010. Agronomic and physiological contributions to the yield improvement of soybean cultivars released from 1950 to 2006 in northeast China. Field Crops Res. 115:116-123.

Krober, O.L. 1956. Nutritive quality of crops, methionine content of soybeans as influenced by location and season. J. Agric. Food Chem. 4:254-257.

Lande, R., R. Thompson. 1990. Efficiency of marker-assisted selection in the improvement of quantitative traits. Genetics 124:743-756.

Liu, K. 1997. Soybeans: chemistry, technology and utilization. International Thompson publishing, New York.

Lusas, E.W. 2004. Soybeans, processing and utilization, *In* H. R. Boerma and J. E. Specht (Eds.), pp. 949-1036. Soybeans, improvement, production and uses.ASA, CSSA, SSSA, Madison, Wiscosin, USA.

Morrison, M.J., H.D. Voldeng, E.R. Cober. 2000. Agronomic changes from 58 years of genetic improvement of short-season soybean cultivars in canada. Agron. J. 92:780-784.

Pallauf, J., G. Rimbach. 1997. Nutritional significance of phytic acid and phytase. Arch Tierernahr. 50:301 - 319.

Pantalone, V.R., D.R. Walker, R.E. Dewey, I. Rajcan. 2004. DNA Marker-assisted selection for improvement of soybean oil concentration and quality, *In* R. F. Wilson, H.T. Stalker and E.C. Brummer (Eds.) pp. 283-311. Legume Crop Genomics.

Panthee, D., V. Pantalone, C. Sams, A. Saxton, D. West, J. Orf, A. Killam. 2006a. Quantitative trait loci controlling sulfur containing amino acids, methionine and cysteine, in soybean seeds. Theor. App. Genet. 112:546-553.

Panthee, D., V. Pantalone, A. Saxton, D. West, C. Sams. 2006b. Genomic regions associated with amino acid composition in soybean. Mol. Breed. 17:79-89.

Pathan, M.S., D.A. Sleper. 2008. Advances in soybean breeding, *In* G. Stacey (Ed.), pp. 113-133. Genetics and genomics of soybean, Springer New York.

Piper, T.E., W.R. Fehr. 1987. Yield improvement in a soybean population by utilizing alternative strategies of recurrent selection 1. Crop Sci. 27:172-178.

Pradhan, A., D.S. Shrestha, A. McAloon, W. Yee, M. Haas. 2009. Energy life-cycle assessment of soybean biodiesel, United States Department of Agriculture.

Raboy, V., D.B. Dickinson, F.E. Below. 1984. Variation in seed total phosphorus, phytic acid, zinc, calcium, magnesium, and protein among lines of Glycine max and G. soja 1. Crop Sci. 24:431-434.

Scaboo, A.M., V.R. Pantalone, D.R. Walker, H.R. Boerma, D.R. West, F.R. Walker, C.E. Sams. 2009. Confirmation of molecular markers and agronomic traits associated with seed phytate content in two soybean ril populations. Crop Sci. 49:426-432.

Smith, T.J., H.M. Camper. 1973. Registration of Essex soybean. Crop Sci. 13:495-495.

Song, Q., L. Marek, R. Shoemaker, K. Lark, V. Concibido, X. Delannay, J. Specht, P. Cregan. 2004. A new integrated genetic linkage map of the soybean. Theor. App. Genet. 109:122-128.

Specht, J.E., D.J. Hume, S.V. Kumudini. 1999. Soybean yield potential—a genetic and physiological perspective. Crop Sci. 39:1560-1570.

Ustun, A., F.L. Allen, B.C. English. 2001. Genetic progress in soybean of the u.s. midsouth. Crop Sci. 41:993-998.

Walker, D.R., A.M. Scaboo, V.R. Pantalone, J.R. Wilcox, H.R. Boerma. 2006. Genetic mapping of loci associated with seed phytic acid content in cx1834-1-2 soybean. Crop Sci. 46:390-397.

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

Wilcox, J.R. 2004. World Distribution and Trade of Soybean, *In* H. R. Boerma and J. E. Specht (Eds.), pp. 1-13. Soybeans: improvement, production and uses. ASA, CSSA, SSSA, Madison, Wisconsin, USA.

Wilson, R.F. 2004. Seed composition, *In* H. R. Boerma and J. E. Specht (Eds.), pp. 621-669. Soybeans: improvement, production and uses. ASA, CSSA, SSSA, Madison, Wiscosin, USA.

Part 2

Soybean Quantitative Trait Loci Associated with Seed Yield

Abstract

The efficiency of breeding for yield improvement in soybean [*Glycine max* (L.) Merrill] could be increased by using marker assisted selection (MAS) breeding strategies to introgress favorable alleles for yield improvement into adapted germplasm. The objective of this study was to identify quantitative trait loci (QTL) associated with seed yield for use in MAS breeding in the southern U.S. A population of 282 F_5 derived recombinant inbred lines (RILs) was developed from a cross of Essex × Williams 82 and genotyped with 1586 single nucleotide polymorphism (SNP) markers from the Universal Soybean Linkage panel 1.0 (U.S.L.P 1.0). The population was divided by days to maturity (10 days) into three tests (early, mid and late) each with 94 genotypes, with one genotype overlapping in maturity in the mid and late tests. In 2009, the three tests, parents and checks were evaluated in a randomized complete block design (RCBD), replicated three times in three environments for plant height, lodging, maturity and seed yield. Data for each test were combined across the three locations and analyzed using the MIXED procedure of SAS software to determine whether there were significant genotypic differences among RILs. Four hundred and eighty polymorphic SNPs and growth habit were mapped onto 21 linkage groups (LGs). Composite interval mapping detected seed yield QTL in the early test on chromosome (chr) 6 (LG C2) and chr 19 (LG L); mid test chrs 12, 15 and, 19 (LGs H, E and, L) and the late test chrs 5, 6, 10, 13, and 19) (LGs A1, C2, F, O and L). The nine seed yield QTL detected in this population explaining 8 — 40% of the phenotypic variation in seed yield, and may be good candidates for MAS as they were stable across environments.

Introduction

Soybean [*Glycine max* (L.) Merrill] is an important oil seed crop that is grown worldwide for protein and oil (Wilcox, 2004). Soybean breeding efforts have succeeded in increasing yield at an average rate of about 0.5 % per year over a 50 yr period (Jin et al., 2010; Morrison et al., 2000). Increased yield would complement the improvement of cultural practices to meet the increased demand for soybean food and industrial products.

Seed yield is a complex trait that is governed by many genes. The expression of these genes is influenced by genetic variation (G), environmental variation (E) and genetic by environmental variation (G × E). Despite the complexity of the trait, gains could be made if genes with genetic effects can be identified and combined through breeding. Improvement of cultivars over broad, general areas is a goal of most U.S. soybean breeding programs. Molecular markers and computer technology have equipped researchers with the tools needed to identify quantitative trait loci (QTL), or regions on the genome that are significantly associated with the trait of interest which may be used as selection tools for soybean improvement (Bernardo, 2008). The different marker types that are available for use in soybean genotyping are: classical markers, isozyme marker loci, amplified fragment length polymorphisms (AFLPs), random amplified polymorphic DNA (RAPDs), single nucleotide polymorphisms (SNPs), simple sequence repeats (SSRs) and restriction fragment length polymorphisms (RFLPs) (Akkaya et al., 1992; Choi et al., 2007; Cregan et al., 1999). The availability of molecular markers in soybean has enabled the construction of a dense molecular map (Consensus Map 4.0) that covers twenty linkage groups (LG) and 2296.4 cM map distance. The markers mapped to create the Consensus map 4.0 were: 3792 SNPs, 1006 SSRs, 664 RFLPs and 38 other markers (classical, RAPDs, AFLPs and isozyme markers) (Hyten et al., 2010). The Consensus map 4.0 which was

developed using diverse germplasm, including 96 landraces and elite × elite mapping populations, genotyped and mapped by Hyten et al. (2010), created a revised for use as a genetic tool by U.S soybean breeders.

Studies to identify QTL associated with yield have been conducted with varying results because of the complex inheritance and expression of the trait. Panthee et al. (2007) conducted a study to detect agronomic trait QTL in a population derived from a high-yielding parent (TN93-99) and a high-protein parent (N87-984-16). The population used in that study consisted of 101 F_6 derived recombinant inbred lines (RIL) evaluated in six environments in the southern U.S region. The diversity of environments enabled the detection of stable and environment specific QTL. Three QTL associated with seed yield [chr 5/LG A1 (Satt042, R^2= 13.8%); chr 2/LG D1b (Satt412, R^2 = 13.3%) and chr 19/LG L (Satt076, R^2 =13.8%)] were detected, of which only Satt076 was significant in multiple environments. These QTL collectively explained about 41% of the yield variation observed. Thus, they could be useful in selecting high-yielding alleles in the southern U.S region.

In a similar study Kassem et al. (2006) detected QTL underlying 29 agronomic traits in an 'Essex' × 'Forrest' (E × F) RIL population designed for molecular mapping purposes to create a molecular map which in included southern germplasm. They reported seed yield QTL in the presence or absence of sudden death syndrome (SDS). Four of the yield QTL identified in this study were stable across environments while three were environment specific. In addition, they found 19 possible marker inversions on 12 of 20 linkage groups (LGs) when comparing the E × F map to the composite interval map previously published by Cregan et al. (1999). The differences between the two maps suggested that future use of more diverse germplasm including southern cultivars in the construction of maps, would expand the genetic information

available to more U.S. soybean breeders. The Consensus map 4.0 developed using diverse germplasm, including 96 landraces and elite × elite mapping populations, genotyped created a revised molecular for use by U.S soybean breeders which meets this need (Hyten et al., 2010).

Detection of QTL and subsequent testing to determine stability across environments is an important step in marker assisted selection studies. Marker assisted selection for quantitative traits has successfully been demonstrated, but it has not been practiced extensively. A QTL increasing yield was found in a wild-type soybean [*Glycine soja* Sieb. and Zucc.] and was shown to have a 9% yield advantage (Concibido et al., 2003). However, this QTL was only stable in two of the six commercial cultivars it was backcrossed into, reiterating the importance of testing QTL in many backgrounds before they can be used extensively for MAS. Perhaps as new knowledge on yield QTL is gained, breeders can begin to understand which loci are favorable for general or specific combinations for the improvement of adapted elite genotypes.

The purpose of this study was to detect yield QTL in an Essex × 'Williams 82' F_5 derived population. Subsequently, the results will be applied in a MAS selection strategy for yield improvement in the southern U.S.

Materials and Methods

Plant materials and field experiments

The initial crosses for the Essex (Smith and Camper, 1973) × Williams 82 (Bernard and Cremeens, 1988) population were made at the University of Tennessee Knoxville Plant Science Farm (KPSF) of the East Tennessee Research and Education Center (ETREC) in Knoxville, TN in the summer of 2005. The F_1 seeds were harvested in the fall of 2005 and F_1 single plants grown in Puerto Rico at the Tropical Agricultural Research Station (TARS) in Isabela, PR in the winter of 2005-06. The population was advanced through single seed descent (Brim, 1966) from

the F_2 to the F_5 generation as follows: at the ETREC location, summers of 2006 and 2007 (F_2 to F_3), and at the TARS location, winter 2007 (F_4) and spring 2008 (F_5). In the summer of 2008 the two parents and 284 individual $F_{5:6}$ RILs were planted in 3.1m single plant rows in ETREC for agronomic data collection, leaf collection for DNA extraction, and seed increase.

Based on maturity data collected in 2008, the population was divided by days to maturity (10 days) into three tests (early, mid and late) each with 94 genotypes, with one genotype overlapping in maturity the mid and late tests. In 2009, the three tests, two parents and checks were planted in two 6.1m row plots in a randomized complete block design (RCBD), replicated three times, in three locations: Knoxville, TN (ETREC), Harrisburg, IL (HA) and Fayetteville, AR (FA). Checks in the 2009 field studies were assigned by maturity group as follows: Early ('IA4004', LD00-2817P, LD00-3309 and, 'Macon'), mid (TN05-4008, TN06-189, and TN06-196) and late (JTN-5203, 'OSAGE', '5002T' and '5601T').

Phenotypic traits

Flower color was recorded when more than 95% of the plants in a plot were in full bloom (R1 to R2). Maturity was recorded at R8 growth stage when 95% of the pods in a plot showed their mature color (Fehr and Caviness, 1977). Pubescence color and growth habit were also recorded at maturity. Maturity was calculated as the number of days from date of planting to R8. Plant height was measured as the average plant height from the base of the stem at the soil surface to the top of the plant at maturity within a plot. Lodging was scored at maturity with a score range from 1 (all plants erect) to 5 (all plants prostrate). Plots were end-trimmed to 4.9 m prior to harvest. Yield per plot (kg ha^{-1}) and percent moisture were recorded at harvest. For data analysis, yield per plot was adjusted to 13% moisture.

DNA Extraction and GoldenGate™ Assay

DNA was extracted from the leaves of RILs and parents utilizing the Qiagen Plant DNeasy Extraction Kit (Qiagen, Valencia, CA). Parents and RILs were screened with the Universal Soybean Linkage Panel (U.S.L.P. 1.0) developed by Hyten et al. (2010). which contains 1536 SNPs using the IlluminaBead™ and the GoldenGate assay™. The GoldenGate™ assay was performed as per manufacturer's protocol and as described by Fan et al. (2003) and Hyten et al. (2008).

Data Analysis

The 2009 data for each test (early, mid, and late) were combined across three locations (ETREC, HA, and FA). Phenotypic data were analyzed using the MIXED procedure of SAS (SAS Institute Inc., 2004) software to determine that there were significant genotypic differences among RILs. Locations and replications were considered blocking factors in the model.

Heritability for each trait was calculated on an entry mean basis for three environments and three replications according to Holland et al. (2010) as follows:

$$h^2 = \frac{\sigma_g^2}{\sigma_g^2 + \left(\frac{\sigma_{ge}^2}{e}\right) + \left(\frac{\sigma^2}{re}\right)}$$

where h^2 represents the heritability, σ^2_g is genetic variance, σ^2_{ge} is genotype by environment variance, σ^2 is error variance, r is number of replications, and e is the number of environments. REML estimation with PROC MIXED in SAS 9.1.3 (SAS Institute, 2004) produced variance components for heritability calculations. Genotypic and phenotypic correlations were estimated and declared significant when the 95% confidence interval did not contain zero according to Holland (2006).

Molecular Mapping and QTL Analysis

After excluding genotypes with missing data 274 RILs were mapped with 480 SNP markers and growth habit. The R/QTL package (Broman et al., 2003) in the R language and environment for statistical computing (R Core Development Team, 2009) was used to estimate recombination fractions, SNP marker distances, and linkage.

Composite interval mapping (CIM) performed to detect and map seed yield QTL, using Windows QTL Cartographer v. 2.5 (Wang et al., 2011) for the early, mid and late tests across three locations using seed protein concentration least squares mean (LSMEAN) estimates. Heterozygote marker loci were excluded in the analyses as we had an F_5 derived population and were only interested in additive genetic effects for the study. One thousand permutations were performed to establish empirical likelihood of odds (LOD) thresholds at the 5% probability level (Churchill and Doerge, 1994). The standard *Zmapqtl* model 6 with a 10cM window and a 2cM walking speed was used for the CIM procedure. Confidence intervals (CI) for the QTL were established by finding 1-LOD support interval.

Results and Discussion

Phenotypic traits

There were significant genotypic differences ($P < 0.0001$) in the total population of RILs for seed yield, maturity, and lodging. However, no significant difference ($P = 0.9996$) for plant height was observed.

Analysis for genotypic differences in the individual tests had similar results except that there were significant genotypic differences ($P < 0.0001$) for plant height in the mid and late tests but not in the early test. Non-significant plant height differences have previously been reported by Kabelka et al. (2004) who found significance in their mid maturity test but not in the early or

late tests. Environment and G × E were significant for all traits ($P < 0.0001$). The high E and G × E variation contributed to the non-significant difference in plant height observed in this study.

Agronomic descriptive statistics and heritability estimates for the early, mid and, late tests and the parents are presented in table 2.1. Seed yield means in the RIL population did not differ greatly from the parental means and no transgressive segregation was observed (Table 2.1). There was an average of a 10 day difference in maturity between the early and the mid test and a two day difference between the mid and late tests in 2009 (Table 2.1). The close gap in maturity between the Mid and Late was caused by earlier maturity of the late tests at the ETREC and AR locations (data not shown).

Heritability estimates for seed yield, maturity, plant height and lodging calculated on an entry mean basis were similar to those reported in other studies (Table 2.1) (Brummer et al., 1997; Wang et al., 2004). Seed yield heritability estimates varied from moderately low (0.51) in the late test to moderately high (0.71) in the early test. Maturity heritability estimates were high in this population at 0.96, 0.82 and, 0.95 for the early, mid and late tests, respectively. Plant height heritability estimates were high for the mid and late tests (0.96, 0.96) and low for the early test (0.0). Lodging heritability estimates were 0.58, 0.83 and, 0.83 for the early, mid and, late tests, respectively. There was a strong genetic correlation, 0.89, between seed yield and maturity in the early test while the mid and late tests had weak positive genetic correlations at 0.51, and 0.53, respectively (Table 2.2).

Table 2.1 Descriptive statistics for agronomic traits, heritability estimates, and their standard errors (in parenthesis) in an F_5 derived soybean RIL population from an Essex × Williams 82 cross, averaged across three environments (Fayetteville, AR; Harrisburg, IL and, Knoxville, TN) within the early, mid and late tests in 2009.

Trait	Min	Mean	Max	Std. Dev	h^2
		Early Test			
Yield (kg ha^{-1})	1803.1	3350.5	4404.9	551.3	0.70 (0.04)
Maturity (days)	110.9	119.7	139.8	6.4	0.96 (0.01)
Height (cm)	62.7	81.7	103.3	8.4	0.00 (0.00)
Lodging[a]	1.303030	2.2	3.2	0.5	0.58 (0.01)
		Mid Test			
Yield (kg ha^{-1})	2640.0	3771.0	4853.1	378.0	0.56 (0.08)
Maturity (days)	121.3	130.9	147.7	3.7	0.82 (0.03)
Height (cm)	59.4	88.2	122.2	17.2	0.96 (0.08)
Lodging[a]	1.6	2.7	4.4	0.7	0.83 (0.03)
		Late Test			
Yield (kg ha^{-1})	2532.7	3575.0	4294.2	351.7	0.51 (0.09)
Maturity (days)	121.8	132.6	138.2	2.8	0.76 (0.04)
Height (cm)	62.8	97.3	121.1	16.5	0.95 (0.01)
Lodging[a]	1.2	2.6	4.6	0.7	0.83 (0.03)
		Parents			
		Essex			
Yield (kg ha^{-1})		4036.8 ± 200.2			
Maturity (days)		132.7 ± 1.9			
Height (cm)		72.9 ± 3.0			
Lodging[a]		1.9 ± 0.2			
		Williams 82			
Yield (kg ha^{-1})		2677.4 ± 146.8			
Maturity (days)		117.2 ± 1.8			
Height (cm)		84.9 ± 2.6			
Lodging[a]		2.02 ± 0.2			

[a]Lodging scored on a 1□5 scale, 1 = all plants erect, and 5 = all plants prostrate.

Table 2.2 Genetic (upper right diagonal) and phenotypic (lower left diagonal) correlations for agronomic traits of an F_5 derived recombinant inbred line population from an Essex x Williams 82 cross averaged across three locations (Fayetteville, AR; Harrisburg, IL and, Knoxville, TN) within the early, mid and late tests in 2009.

Test	Trait	Yield	Maturity	Height	Lodging
Early	Yield		0.89* (0.01)	0.93* (0.72)	0.15* (0.00)
	Maturity	*0.06ns (0.00)*		-0.91* (0.79)	0.01* (0.16)
	Height	*0.08ns (0.04)*	*0.21* (0.04)*		0.00 ns (0.00)
	Lodging	*0.06ns(0.00)*	*-0.26* (0.04)*	*0.32* (0.03)*	
Mid		Yield	Maturity	Height	Lodging
	Yield		0.51* (0.12)	0.21* (0.14)	0.08 ns (0.00)
	Maturity	*0.41* (0.05)*		0.04ns (0.00)	0.34* (0.11)
	Height	*0.08* (0.06)*	*0.03n (0.00)*		0.76* (0.06)
	Lodging	*0.13* (0.00)*	*0.24* (0.07)*	*0.58* (0.05)*	
Late		Yield	Maturity	Height	Lodging
	Yield		0.53* (0.13)	0.37* (0.05)	-0.01* (0.16)
	Maturity	*0.31* (0.05)*		0.27* (0.11)	0.12* (0.13)
	Height	*0.22* (0.05)*	*0.17* (0.07)*		0.73* (0.06)
	Lodging	*-0.01* (0.06)*	*0.14* (0.06)*	*0.60* (0.04)*	

, Significant at P=0.05 when r ± 1.96 S.E did not overlap 0.

The phenotypic correlations between seed yield and maturity were weak and positive (mid and late tests) or non-significant (early test) (Table 2.2). The genetic and phenotypic correlations observed in the early test indicate that there was a strong G effect on yield while those correlations observed in the mid and late test show an effect of both G and G × E.

Genetic and phenotypic correlations between yield, and plant height, and lodging were weak or non-significant for all tests (Table 2.2). The genetic correlation between seed yield and plant height for the late test seems strongly positive at 0.93; however, the large standard error associated with it (0.72) and the zero heritability estimate leads to the conclusion that we could not estimate the genetic effect on seed and plant height in the early test. This could be for various reasons such as sampling errors (Falconer, 1989), small population sizes and low heritability estimates (Cheverud, 1988).

QTL Analysis

After excluding genotypes with missing data, 274 RILs were mapped with 480 polymorphic SNP markers and growth habit placed on 21 linkage groups (Figure 1.1). The Essex × Williams 82 map covered a distance of 3035.4 cM and had larger gaps than the Consensus Map 4.0. However, there was similarity in SNP order and linkage group assignments except that in this population chromosome (chr) 13 (LG F) was split in two (designated 13a and 13b), and one marker, BARC-015435-01966, was moved from chr 6 (LG C2), to chr 13. This population was genotyped only using only the U.S.L.P 1.0 SNP panel, therefore in the future many more SNP and SSR markers available on the consensus map 4.0 could be used to test for potential polymorphisms between Essex and Williams 82 to fill in the gaps observed (Figure 1.1). Filling in the large gaps would give better map coverage and improve the accuracy of QTL detection (Figure 1.1).

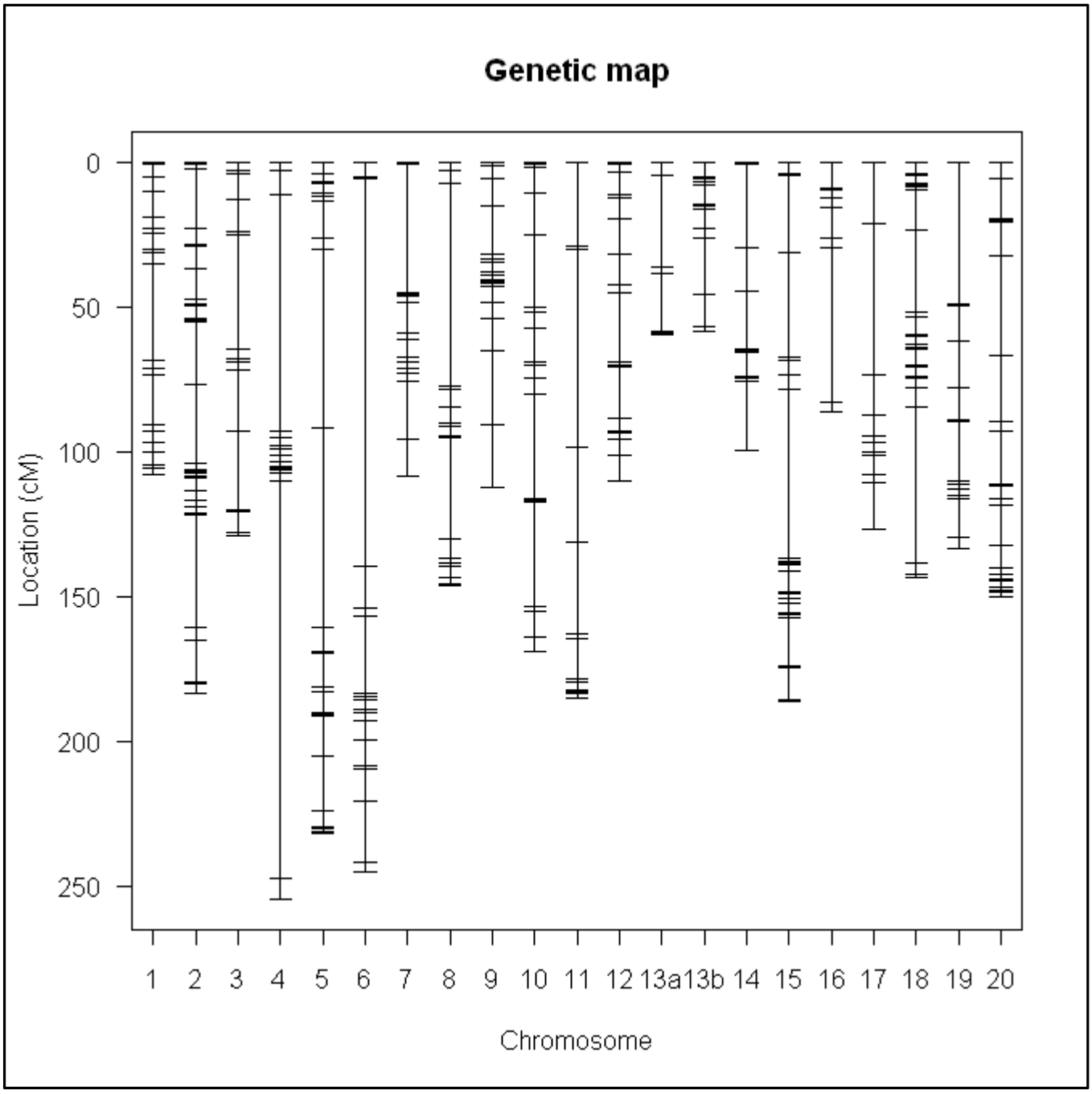

Figure 1.1 Molecular map of the Essex × Williams 82 population of 274 recombinant inbred lines mapped with 480 SNPs and growth habit (Dt 1 locus).

We detected two, three and four significant ($P < 0.05$) seed yield QTL at the established threshold values in the early, mid and late tests, respectively (Table 2.3). Illustrations of seed yield QTL reported on table 2.3 are shown on figure1.2. Negative effects indicate that the Essex allele did not increase yield at that locus, but, selection for the alternative Williams 82 allele may increase protein concentration. Yield QTL in the early test were mapped on chr 6 (LG C2) and chr 19 (LG L); in the mid test on chrs12, 15 and, 19 (LGs H, E and, L) and in the late test on chrs 5, 6, 10, 13, and 19 (LGs A1, C2, F, O and L).

The seed yield QTL intervals, Y1 and Y7, mapped on chr 6 (LG C2) in the Early and Late tests ($R^2 = 40$ % and 8 %, respectively) are close to the E1 locus on the Consensus map 4.0. Although the Y1 and Y7 loci are close to the E1 locus on the Consensus map 4.0, only Y1 was significantly ($P < 0.05$) associated with maturity in this population (data not shown). Seed yield close to the E1 locus have previously been reported (Kabelka et al., 2004; Kassem et al., 2006; Mansur et al., 1996; Orf et al., 1999; Wang et al., 2004). The QTL reported by Mansur et al. (1996) and Wang et al. (2004) were significantly associated with maturity while those reported by Orf et al. (1999) and Kabelka et al. (2004) were not. The study by Kassem et al. (2006) did not analyze data for maturity; however there was an association between yield and SDS. Symptoms of SDS were observed in our population but the disease did not manifest itself possibly due to the dry weather experienced in 2009. The seed yield QTL intervals, Y2 and Y11, mapped on chr 19 (LG L) in the Early and Late tests ($R^2 = 10$ % and 15 %, respectively) occur in the same interval which suggests that this is one QTL rather than two. The close proximity of this interval to the Dt1 locus further suggests that growth habit affects yield in this population.

Table 2.3 Quantitative trait Loci associated with seed yield measured in three environments (Fayetteville, AR; Harrisburg, IL and Knoxville, TN) in an F_5 derived Essex × Williams 82 population of 274 recombinant inbred lines in 2009.

Test	Chromosome	LG	QTL	Molecular Marker	QTL Position(cM)	95% Confidence Interval of QTL Interval	R^2	Effect[a] (kg ha^{-1})
Early	6	C2	Y1	BARC-063661-18416	206	203□213	0.44	430
	19	L	Y2	BARC-035235-07156	136	128□143	0.10	-215
Mid	12	H	Y3	BARC-065503-19514	9.7	8.7□24.4	0.16	152
	15	E	Y4	BARC-059221-15678	169.4	165.8□172.4	0.10	125
	19	L	Y5	BARC-047428-12928	42.2	27.5□54.5	0.31	-214
Late	5	A1	Y6	BARC-058647-17392	230.8	226.4□>230.8	0.08	107
	6	C2	Y7	BARC-062515-17881	222	215.6□235.2	0.06	102
	10	O	Y8	BARC-065789-19751	26.9	16.8□42.5	0.08	101
	13b	F	Y9	BARC-028583-05961	56.7	43.4□>58.4	0.04	83
	15	E	Y10	BARC-027480-06591	89.6	86.1□95.6	0.18	146
	19	L	Y11	BARC-060587-16731	132.2	122.9□148.6	0.15	-146

[a]Effect with respect to the Essex allele. Negative sign indicates that Essex did not contribute the favorable allele.

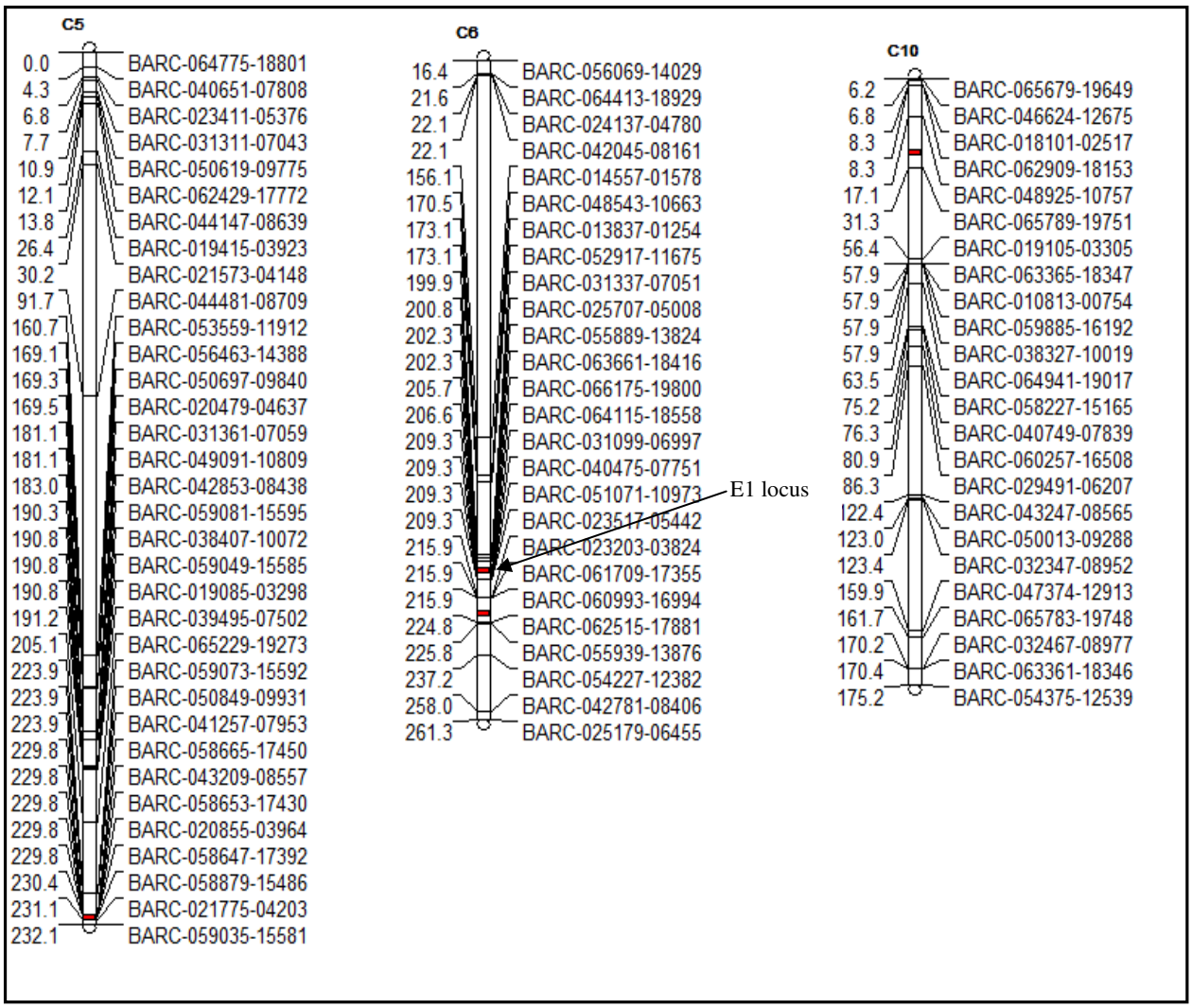

Figure 2.1 Illustration of seed yield QTL(red bars) measured in Fayetteville, AR; Harrisburg, IL and Knoxville, TN environments in an F_5 derived Essex × Williams 82 population of 274 recombinant inbred lines.

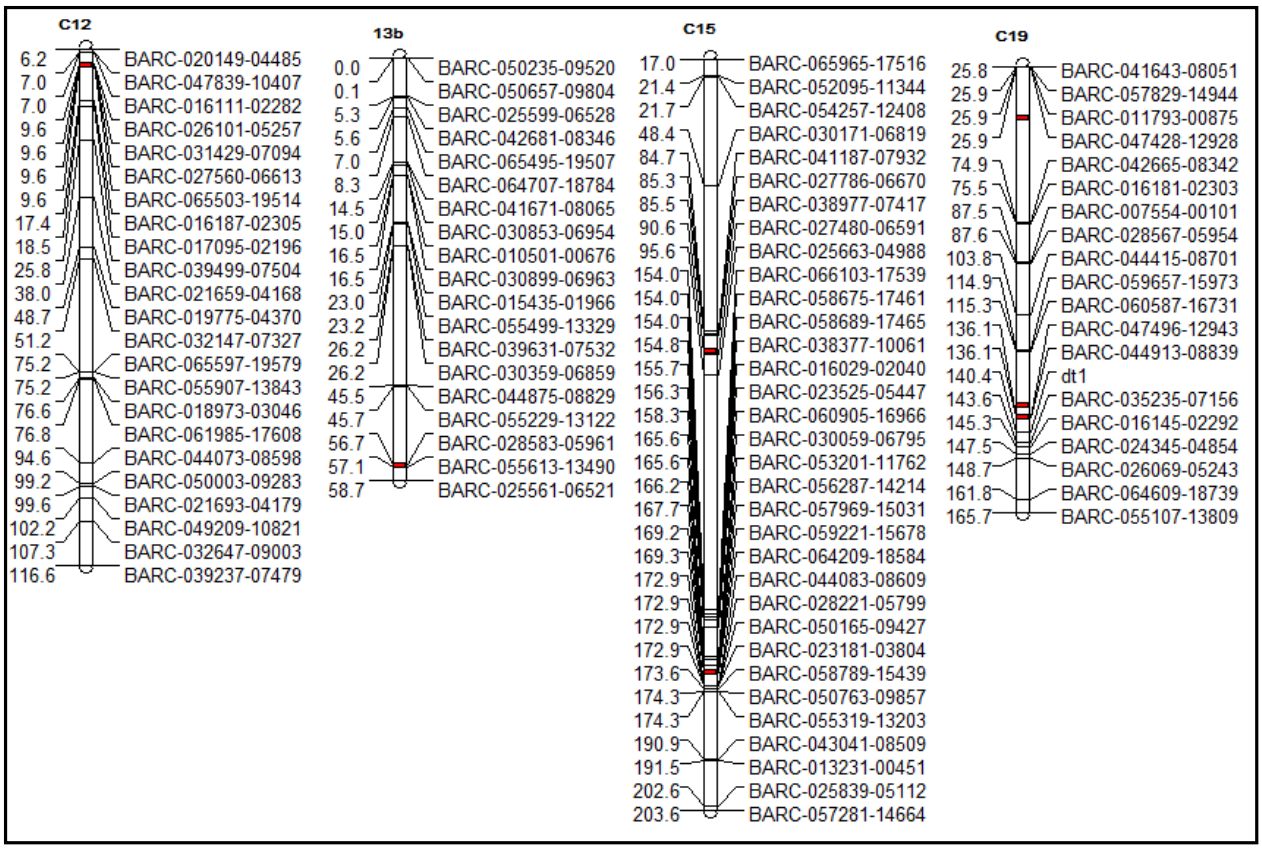

Figure 2.1Continued.

Specht et al. (2001) detected a yield QTL associated with growth habit on chr 19 (LG L) close to Y2 and Y11.

All other seed yield QTL detected in this study were not significantly associated with the major genes E1or Dt1 whose pleiotropic effect would largely be the reason for yield QTL detection. Moreover, yield QTL in this study have previously been reported in Soybase (Grant et al., 2010) as yield QTL and our findings support those results. Yield QTL reported in this population had reasonably high R-square values and large effects and may be good candidates for utilization in a marker assisted selection scheme.

Conclusion

Minor seed yield QTL, with R-squares less than 10%, and major seed yield QTL, R-squares greater than 10% were detected in this study. These results were stable across all environments. Four QTL intervals on chr 6 (LG C2) and chr 19 (LG L) were detected in the early and late tests. These intervals likely represent one QTL on chr 6 and another on chr 19. There was an association between the E1 locus (earliness) and the Dt1 locus (growth habit). This is significant because both loci influence maturity and yield in soybean. Therefore, selection for QTL by maturity or growth habit could be utilized in this population for yield improvement. Nine total seed yield QTL that have been reported by other studies were detected in this study and are good candidates for marker assisted selection strategies in soybean breeding programs in the southern United States.

References

Akkaya, M.S., A.A. Bhagwat, P.B. Cregan. 1992. Length polymorphisms of simple sequence repeat DNA in soybean. Genetics 132:1131-1139.

Bernardo, R. 2008. Molecular markers and selection for complex traits in plants: learning from the last 20 years. Crop Sci. 48:1649-1664.

Broman, K.W., H. Wu, Ś. Sen, G.A. Churchill. 2003. R/qtl: QTL mapping in experimental crosses. Bioinformatics 19:889-890.

Brummer, E.C., G.L. Graef, J. Orf, J.R. Wilcox, R.C. Shoemaker. 1997. Mapping qtl for seed protein and oil content in eight soybean populations. Crop Sci. 37:370-378.

Cheverud, J.M. 1988. A comparison of genetic and phenotypic correlations. Evolution 42:958-968.

Choi, I.-Y., D.L. Hyten, L.K. Matukumalli, Q. Song, J.M. Chaky, C.V. Quigley, K. Chase, K.G. Lark, R.S. Reiter, M.-S. Yoon, E.-Y. Hwang, S.-I. Yi, N.D. Young, R.C. Shoemaker, C.P. van Tassell, J.E. Specht, P.B. Cregan. 2007. A soybean transcript map: gene distribution, haplotype and single-nucleotide polymorphism analysis. Genetics 176:685-696.

Churchill, G.A., R.W. Doerge. 1994. Empirical threshold values for quantitative trait mapping. Genetics 138:963-971.

Concibido, V. La, McLaird, Pineda, Meyer, Hummel, Yang, Wu, Delannay. 2003. Introgression of a quantitative trait locus for yield from glycine soja into commercial soybean cultivars. Theor. Appl. Genet. 106:575-582.

Cregan, P.B., T. Jarvik, A.L. Bush, R.C. Shoemaker, K.G. Lark, A.L. Kahler, N. Kaya, T.T. VanToai, D.G. Lohnes, J. Chung, J.E. Specht. 1999. An integrated genetic linkage map of the soybean genome. Crop Sci. 39:1464-1490.

Falconer, D.S. 1989. Introduction to quantitative genetics . Third Edition ed. Oliver & Boyd, Edinburgh.

Fan, J.B., A. Oliphant, R. Shen, B.G. Kermani, F. Garcia, K.L. Gunderson, M. Hansen, F. Steemers, S.L. Butler, P. Deloukas, L. Galver, S. Hunt, C. McBride, M. Bibikova, T. Rubano, J. Chen, E. Wickham, D. Doucet, W. Chang, D. Campbell, B. Zhang, S. Krugylak, D. Bentley, J. Haas, P. Rigault, L. Zhou, J. Stuelpnagel, M.S. Chee. 2003. Highly parallel SNP genotyping. Cold Spring Harbor Symp. Quant. Biol. 68:69-78.

Fehr, W.R., C.E. Caviness. 1977. Stages of Soybean Development, Iowa State Univ., Ames. Grant, D., R.T. Nelson, S.B. Cannon, R.C. Shoemaker. 2010. SoyBase, the USDA-ARS soybean genetics and genomics database. Nucleic Acids Res. 38:843-846.

Hyten, D., Q. Song, I.-Y. Choi, M.-S. Yoon, J. Specht, L. Matukumalli, R. Nelson, R. Shoemaker, N. Young, 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., I.-Y. Choi, Q. Song, J.E. Specht, T.E. Carter, R.C. Shoemaker, E.-Y. Hwang, L.K. Matukumalli, 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.

Holland, J.B. 2006. Estimating genotypic correlations and their standard errors using multivariate restricted maximum likelihood estimation with sas Proc MIXED. Crop Sci. 46:642-654.

Holland, J.B., W.E. Nyquist, C.T. Cervantes-Martínez. 2010. Estimating and interpreting heritability for plant breeding: An update, pp. 9-112. Plant Breed. Rev., John Wiley & Sons, Inc.

Jin, J., X. Liu, G. Wang, L. Mi, Z. Shen, X. Chen, S.J. Herbert. 2010. Agronomic and physiological contributions to the yield improvement of soybean cultivars released from 1950 to 2006 in Northeast China. Field Crops Res. 115:116-123.

Kabelka, E.A., B.W. Diers, W.R. Fehr, A.R. Leroy, I.C. Baianu, T. You, D.J. Neece, R.L. Nelson. 2004. Putative alleles for increased yield from soybean plant introductions. Crop Sci. 44:8.

Kassem, M., J. Shultz, K. Meksem, Y. Cho, A. Wood, M. Iqbal, D. Lightfoot. 2006. An updated 'Essex' by 'Forrest' linkage map and first composite interval map of QTL underlying six soybean traits. Field Crops Res. 113:1015-1026.

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

Morrison, M.J., H.D. Voldeng, E.R. Cober. 2000. Agronomic changes from 58 years of genetic improvement of short-season soybean cultivars in canada. Agron. J. 92:780-784.

Nyquist, W.E. 1991. Estimation of heritability and prediction of selection response in plant populations. Crit. Rev. Plant Sci. 10:235 - 322.

Orf, J.H., K. Chase, F.R. Adler, L.M. Mansur, K.G. Lark. 1999. Genetics of soybean agronomic traits: II. interactions between yield quantitative trait loci in soybean. Crop Sci. 39:1652-1657.

Panthee, D.R., V.R. Pantalone, A.M. Saxton, D.R. West, C.E. Sams. 2007. Quantitative trait loci for agronomic traits in soybean. Plant Breeding 126:51-57.

R Core Development Team. 2009. R: A language and environment for statistical computing. r foundation for statistical computing. Vienna, Austria.

Sas Institute. 2004. SAS/STAT software. Release 9.1.3. , SAS Inst., Cary, N.C.

Specht, J.E., K. Chase, M. Macrander, G.L. Graef, J. Chung, J.P. Markwell, M. Germann, J.H. Orf, K.G. Lark. 2001. Soybean response to water. Crop Sci. 41:493-509.

Wang, S., J. Basten, Z.B. Zeng. 2011. Windows QTL cartographer 2.5.

Wang, D., G.L. Graef, A.M. Procopiuk, B.W. Diers. 2004. Identification of putative QTL that underlie yield in interspecific soybean backcross populations. Theor. App. Genet. 108:458-467

Part 3

Soybean Quantitative Trait Loci Associated with Seed Protein Concentration

Abstract

Soybean seed contains 40 % protein and 20 % oil and is an important source of protein for both human nutrition and as livestock feed. Improvement of soybean cultivars to increase protein concentration without significant oil reduction is a goal of many breeding programs. The negative correlation between these two traits makes it difficult to achieve this goal. A better understanding of the genomic regions, termed quantitative trait loci (QTL), associated with protein could provide useful molecular marker tools for marker assisted selection (MAS) for protein. The objective of this study was to identify protein concentration QTL for use in marker assisted selection in the southern U.S. Two hundred and eighty two F_5 derived recombinant inbred lines (RILs) were developed from a cross of Essex × Williams 82 and genotyped with 1586 single nucleotide polymorphism (SNP) markers from the Universal Soybean Linkage panel 1.0 (U.S.L.P 1.0). The population was divided by days to maturity (10 days) into three tests (early, mid and late) each with 94 genotypes, with one genotype overlapping in maturity in the mid and late tests. In 2009, the three tests, including parents and checks, were evaluated for in a randomized complete block design (RCBD), replicated three times, in three environments and evaluated for seed protein and oil concentration, and seed weight. Data were combined within each test across the three environments and analyzed using the MIXED procedure of SAS to determine that there were significant genotypic differences among RILs. Four hundred and eighty polymorphic SNPs were mapped onto 21 linkage groups (LGs). Composite interval mapping detected ten protein concentration QTL, one of which was new in this study, which explained 4% to 36% of the phenotypic variation for seed protein concentration. These QTL may be good candidates for MAS as they were stable across environments.

Introduction

Soybean [*Glycine max* (L.) Merrill] is an important source of protein for human consumption and livestock feed (Liu, 1997).). Soybean seeds contain 40% protein and 20% oil on a dry mass basis (Wilson, 2004).). The current industry standard for high-protein soybean meal (SBM) is 48% crude protein. To maintain the competitiveness of SBM as a vegetable protein source, plant breeders must produce soybean cultivars that contain about 44 □ 45% crude protein and no less than 18% oil to meet the industry standard (Wilson, 2004).

Several studies have documented the negative correlations between soybean protein and oil concentration with yield (Burton, 1987; Hartwig and Hinson, 1972; Mansur et al., 1993; Pantalone et al., 1996; Wilcox, 1998; Wilson, 2004). Although the negative correlation between seed yield and protein content makes it difficult to select and develop high-protein and high-yielding cultivars, plant breeders have occasionally been successful in increasing protein without significantly reducing yield (Jin et al., 2010). Wilcox and Cavins (1995) backcrossed a high-protein line, Pando (donor parent), to a high-yielding cultivar, 'Cutler 71' (recurrent parent). They determined that high seed protein genes could be backcrossed into a high-yielding cultivar with full recovery of the seed yield of the recurrent parent. Burton and Carver (1993) and Burton et al. (1999) used the recurrent selection approach to create 'Prolina,' a high-protein cultivar that averaged 461 g kg^{-1} protein and 198 g kg^{-1} oil seed concentration. Prolina is a bulk of two F_8 derived lines selected from the first cycle of recurrent selection following matings of 10 high-protein lines. Panthee and Pantalone (2006) released two high-protein, high-yield germplasm lines, from a mapping population developed from the cross Prolina × TN93-99. The two high-protein germplasm lines, TN03-350 and TN04-5321, had higher protein concentrations than the low parent, TN93-99 (416 g kg^{-1}), but were lower than Prolina.

Cober and Voldeng (2000) also found that crosses utilizing a high yielding parent and a high-protein parent produced some progeny with higher protein than the high-yielding parent, but none of the improved protein content lines had higher protein concentration than the high-protein parent.

The availability of molecular markers and dense molecular maps has enabled researchers to identify quantitative trait loci (QTL) associated with protein in soybean. The QTL, available in the literature and in Soybase (Grant et al., 2010),), may be useful tools in marker assisted selection (MAS) strategies for increasing seed protein concentration. Qiu et al. (1999) conducted a study to identify DNA markers associated with soybean cyst nematodes (SCN) as well as seed protein and oil concentration, using three $F_{2:3}$ derived families from a cross of 'Essex' (high-protein) × Peking (SCN resistance). Two major QTL ($R^2 = 0.32$, and $R^2 = 0.17$) significantly associated ($P < 0.01$) with protein concentration were detected on LG H, now noted as chromosome (chr) 12 by the RFLP marker B072, and chr 13 (LG F) by the RFLP marker B148. The QTL on chr 12 (LG H) was also significantly associated ($P < 0.01$) with oil concentration ($R^2 = 0.21$). These QTL combined explained 33% of the total phenotypic variation observed for protein. Essex, a southern adapted cultivar, was the source of all favorable alleles for protein concentration. This implies that it could be used in a breeding program as a protein source.

Chapman et al. (2003) identified a protein concentration QTL on chr 11 (LG B1) associated with Satt251 ($R^2= 0.03$, $P < 0.05$). This QTL was identified in both the F_2 and $F_{2:4}$ generations of the population derived from an Essex × 'Williams' cross evaluated in environments one and two.

The authors acknowledge that selections based on this QTL would have limited success because of the small R^2 value. However, more research conducted in multiple locations utilizing more molecular markers could reveal QTL with higher R^2 values.

In a separate study, Hyten et al. (2004) detected protein concentration QTL on chr 6 (LG C2), chr 13 (LG F), chr 9 (LG K) and chr 7 (LG M) in an F_6 derived population of Essex × Williams that was evaluated in three environments. Three of the protein QTL reported in that study on chr 6 (LG C2), chr 13 (LG F) and, chr 9 (LG K), were stable in all three environments, while the QTL on chr 7 (LG M) was only significant at one location. Protein concentration QTL detected Hyten et al. (2004) explained 4 – 28 % of the phenotypic variation observed and could be useful in MAS selection strategy for protein. That study found that three of the protein QTL detected had favorable alleles from Essex while only one, which was attributed to maturity on chr 6 (LG C2), was contributed by Williams. Essex and Williams belong to the southern and northern germplasm pools, respectively, and are important ancestors of many cultivars (Sneller, 1994). It is then not surprising that protein concentration QTL close to those detected by Chapman et al. (2003), Hyten et al. (2004) and Qiu et al. (1999) have previously been reported by other researchers and are available in Soybase. Therefore, QTL detected in our population, which is derived from an Essex × 'Williams 82' cross, may be useful to many plant breeders because of the broad north × south genetic base.

The purpose of this study was to detect protein concentration QTL in an Essex × Williams 82 F_5 derived recombinant inbred line (RIL) population for later use in a MAS selection strategy for increased protein concentration in the southern United States.

Materials and Methods

Plant materials and field experiments

The initial crosses for the Essex (Smith and Camper, 1973) × Williams 82 (Bernard and Cremeens, 1988) population were made at the University of Tennessee Knoxville Plant Science Farm (KPSF) of the East Tennessee Research and Education Center (ETREC) in Knoxville, TN in the summer of 2005. The F_1 seeds were harvested in the fall of 2005 and F_1 single plants were grown in Puerto Rico at the Tropical Agricultural Research Station (TARS) in Isabela, PR in the winter of 2005-06. The population was advanced through single seed descent (Brim, 1966) from the F_2 to the F_5 generation as follows: At the ETREC location, summers of 2005 and 2007 (F_2 to F_3), and at the TARS location, winter 2007 (F_4) and spring 2008 (F_5). In the summer of 2008 the two parents and 284 individual $F_{5:6}$ RILs were planted in 3.1m single plant rows in ETREC for agronomic data collection, leaf collection for DNA extraction, and seed increase.

Based on maturity data collected in 2008, the population was divided by days to maturity (10 days) into three tests (early, mid and late) each with 94 genotypes, with one genotype overlapping in the mid and late tests. In 2009, the three tests, the two parents and, checks were planted in two 6.1m row plots in a randomized complete block design (RCBD), replicated three times, in three locations: Knoxville, TN (ETREC), Harrisburg, IL (HA) and Fayetteville, AR (FA). Checks in the 2009 field studies were assigned by maturity group as follows: Early ('IA4004', LD00-2817P, LD00-3309 and, 'Macon'), mid (TN05-4008, TN06-189, and TN06-196) and late (JTN-5203, 'OSAGE', '5002T' and '5601T').

Phenotypic traits

Approximately 25 g of soybean seeds were ground in a water-cooled Knifetec 1095 Sample Mill (FOSS Tecator, S-26321, Hogana, Sweden) for 20 sec.

This setting produced soybean flour with a uniform particle size. Approximately 15 g of ground sample were analyzed for seed protein and oil concentration using a near infrared spectroscopy (NIRS) instrument (NIRS 6500, FOSS North America). The NIRS instrument was warmed up for 2 h after turning on the lamp, and auto-diagnostic tests were run. Diagnostics tests performed every day until sample scanning was completed ensured the instrument passed three different tests for instrument response, wavelength accuracy, and NIRS repeatability. A room dehumidifier was used throughout the analysis, setting the humidity to 40%. Room temperature was approximately 20 ºC. Approximately 15 g of ground soybean sample per plot was scanned using ISIScan software version 2.85.3. Samples were analyzed according to their replication in the field tests and the instrument was not turned off between reps. Sample scanning produced the predicted levels of protein and oil on a 13% moisture basis. Seed size was determined by weighing 100 seeds per plot.

DNA Extraction and GoldenGate™ Assay

DNA was extracted from the young leaves of RILs and parents utilizing the Qiagen Plant DNeasy Extraction Kit (Qiagen, Valencia, CA). Parents and RILs were screened with the Universal Soybean Linkage Panel (U.S.L.P. 1.0) developed by Hyten et al. (2010) which contains 1536 SNPs using the IlluminaBead™ and the GoldenGate assay™. The GoldenGate™ assay was performed as per manufacturer's protocol and as described in Fan et al., 2003) and Hyten et al. (2008).

Data Analysis

The 2009 data for each test (early, mid, and late) were combined across three locations (ETREC, HA, and FA). Phenotypic data were analyzed using the MIXED procedure of SAS (SAS Institute Inc., 2004) to determine significant genotypic differences among RILs. Locations and replications were considered blocking factors in the model.

Heritability estimates and their standard errors for each trait on an entry mean basis for three environments and three replications were calculated according to Holland et al. (2010) as follows:

$$h^2 = \frac{\sigma_g^2}{\sigma_g^2 + \left(\frac{\sigma_{ge}^2}{e}\right) + \left(\frac{\sigma^2}{re}\right)}$$

where $\boldsymbol{h^2}$ represents the heritability, $\sigma^2_{\boldsymbol{g}}$ is genetic variance, $\sigma^2_{\boldsymbol{ge}}$ is genotype by environment variance, σ^2 is error variance, $\boldsymbol{r}$ is number of replications, and $\boldsymbol{e}$ is the number of environments. REML estimation with PROC MIXED in SAS 9.1.3 (SAS Institute, 2004) produced variance components for heritability calculations. Genotypic and phenotypic correlations were estimated and declared significant when the 95% confidence interval was different from zero according to Holland (2006).

Molecular Mapping and QTL Analysis

After excluding genotypes with missing data, 274 RILs were mapped with 480 SNP markers. The R/QTL package (Broman et al., 2003) in the R language and environment for statistical computing (R Core Development Team, 2009) was used to estimate recombination fractions, SNP marker distances, and linkage.

Composite interval mapping (CIM) for seed protein concentration QTL was performed to detect and map QTL, using Windows QTL Cartographer v. 2.5 (Wang et al., 2011) for the early, mid and late tests across three locations using seed protein concentration least squares mean (LSMEAN) estimates.

Heterozygote marker loci were excluded in the analyses as we had an F_5 derived population and were only interested in additive genetic effects for the study. One thousand permutations were performed to establish empirical likelihood of odds (LOD) thresholds at the 5% probability level

(Churchill and Doerge, 1994). The standard *Zmapqtl* model 6 with a 10cM window and a 2cM walking speed was used for the CIM procedure. Confidence intervals (CI) for the QTL were established by finding 1-LOD support interval.

Results and Discussion

Phenotypic traits

There was significant ($P < 0.05$) genotype (G), environmental (E) and genotype by environment variation (G × E) observed among all tests for seed protein and oil concentration and seed weight in this population. There was a weak but significant ($P < 0.05$) positive genetic correlation between seed protein and oil concentration in the early test; however, there were weak negative genetic correlations between seed protein and oil concentration in the mid and late tests (Table 3.1). Weak positive phenotypic correlations between protein concentration and seed weight (early and late tests), and oil concentration and seed weight (mid and late tests) were also observed (Table 3.1). The negative correlation between protein and oil in soybean that has been reported by many studies (Brummer et al., 1997; Burton, 1987; Hartwig and Hinson, 1972; Panthee et al., 2005; Wilcox and Shibles, 2001) was not observed in this population.

Small genetic and phenotypic covariance for seed protein and oil concentration as well as large environmental variation and a large genetic × environmental effect may be the reason for the genetic and phenotypic correlation coefficients between the two traits observed in this population. According to Falconer (1989, p. 317) “estimates of genetic correlations are subject to large sampling errors and as such may not be accurate.”

The average seed protein and oil concentrations, and seed weight were similar to the parental means (Table 3.2). However, transgressive segregation (± 2 std. dev. of high or low parent) for seed protein concentration and seed weight was observed in the RILs.

Table 3.1 Genetic (upper right diagonal) and phenotypic (lower left diagonal, italicized) correlations and their standard errors (in parenthesis) for seed protein and oil concentration, and seed weight of an F_5 derived recombinant inbred line population from an Essex × Williams 82 cross, averaged across three environments (Fayetteville, AR; Harrisburg, IL and, Knoxville, TN), within the early, mid and late tests in 2009.

Test	Trait	Protein	Oil	Seed Weight
Early	Protein		0.04* (0.11)	0.40* (0.17)
	Oil	*0.30* (0.06)*		0.22* (0.18)
	Seed Weight	*0.08* (0.06)*	*0.06* (0.07)*	
Mid		Protein	Oil	Seed Weight
	Protein		-0.30* (0.13)	0.45* (0.02)
	Oil	*0.09* (0.06)*		0.31* (0.03)
	Seed Weight	*0.36* (0.01)*	*0.14* (0.01)*	
Late		Protein	Oil	Seed Weight
	Protein		-0.13* (0.17)	0.29* (0.01)
	Oil	*0.13* (0.06)*		0.37* (0.05)
	Seed Weight	*0.21^{ns} (0.00)*	*0.12* (0.01)*	

, Significant at P=0.05 when r ± 1.96 S.E did not overlap 0.

Extreme values for protein concentration may be attributed to recombination as the parents only differed by 1 g kg^{-1} (Table 3.2). Moderately high heritability estimates on an entry mean basis for protein and oil concentration were similar for all three tests (Table 3.2). Heritability estimates for seed weight were high except in the Early test where low G and high G × E produced a low estimate of 0.45 (Table 3.1). Similar heritability estimates have been reported in other studies (Brummer et al., 1997; Chung et al., 2003; Csanádi et al., 2001b; Mansur et al., 1996; Panthee et al., 2005). Moderately high heritability estimate s for seed protein and oil concentration suggest that the variation for seed protein and oil concentration as well as seed weight (except in the late test) is largely genetic and supports our notion that the low genetic and phenotypic correlations observed were due to environmental variation.

QTL analysis

After excluding genotypes with missing data, 274 RILs were mapped with 480 polymorphic SNP markers placed on 21 linkage groups (Figure 1.1, p. 27).

The Essex × Williams 82 map covered a distance of 3035.4 cM and had larger gaps than the Consensus Map 4.0. However, there was similarity in SNP order and linkage group assignments except that in this population chromosome (chr) 13 (LG F) was split in two (designated 13a and 13b), and one marker, BARC-015435-01966, was moved from chr 6 (LG C2), to chr 13. This population was genotyped using only the U.S.L.P 1.0 SNP panel, therefore in the future many more SNP and SSR markers available on the consensus map 4.0 could be used to test for potential polymorphisms between Essex and Williams 82 to fill in the gaps observed which would give better map coverage and improve the accuracy of QTL detection (Figure 1.1, p. 27).

Table 3.2 Descriptive statistics for seed protein and oil concentration, and seed weight on a 13% moisture basis, heritability estimates, and their standard errors (in parenthesis) in an F_5 derived soybean RIL population from an Essex × Williams 82 cross grown in three environments in 2009 for the early, mid and late tests.

Trait	Test	Min	Mean	Max	Std. Dev	h^2
	Early Test					
Protein (g kg^{-1})		344.4	371.6	397.0	12.0	0.79 (0.04)
Oil (g kg^{-1})		175.9	188.7	201.3	5.2	0.80 (0.04)
Seed weight (mg $seed^{-1}$)		126.7	146.0	177.7	9.3	0.45 (0.10)
	Mid Test					
Protein (g kg^{-1})		347.5	364.6	387.1	8.8	0.72 (0.05)
Oil (g kg^{-1})		174.7	185.6	195.3	5.3	0.77 (0.04)
Seed weight (mg $seed^{-1}$)		108.8	142.6	274.9	19.5	0.94 (0.01)
	Late Test					
Protein (g kg^{-1})		349.7	366.7	401.5	7.9	0.61 (0.07)
Oil (g kg^{-1})		173.6	185.2	194.7	4.7	0.64 (0.06)
Seed weight (mg $seed^{-1}$)		116.9	142.6	173.0	13.0	0.88 (0.02)
	Essex					
Protein (g kg^{-1})			371.9 ± 1.6			
Oil (g kg^{-1})			180.0 ± 0.9			
Seed weight (mg $seed^{-1}$)			134.3 ± 2.8			
	Williams 82					
Protein (g kg^{-1})			370.9 ± 5.5			
Oil (g kg^{-1})			201.8 ± 3.5			
Seed weight (mg $seed^{-1}$)			149.5 ± 2.2			

Seed protein concentration QTL in the early test were mapped on chrs 5, 6 and, 7 (LGs A1, C2 and, M); in the mid test on chrs 4, 6 and, 11 (LGs C1, C2 and, B1) and in the late test chrs 1, 6, 8, 12, 16, and 20) (LGs D1a, C2, A2, H, J and, I) (Table 3.3). Illustrations of seed protein concentration QTL reported on table 3.3 are shown on figure 2.1. The numeric difference in map position between our map and the Consensus map 4.0 observed in this population was due to the gaps in our map (Figure 1.1, p. 31). The term effect, in table 3.3, describes the additive effect of allele substitution of the Essex allele on seed protein concentration. Negative effects indicate that the Essex allele did not increase protein at that locus, but selection for the alternative Williams 82 allele would increase protein concentration.

Seed protein concentration QTL P1 and P2 on chr 5 (LG A1) detected in the early test are 15 cM apart, and most likely represent one, not two QTL. A protein QTL in the P1 – P2 interval on chr 5 (LG 5) has previously been reported by Orf et al. (1999) in two populations of 'Minsoy' ×'Noir' and Minsoy ×'Archer'. The protein concentration QTL detected on chr 6 (LG C2) P3, P7, P8 and P10 have overlapping intervals, lie close to the E1 maturity locus and most likely represent the same QTL detected in all three tests. The seed protein concentration QTL, P6, is about 20 cM upstream from the P3, P7, P8 and P10 interval and is most the same QTL. This QTL on chr 6 is likely an effect of maturity on protein due to its proximity to the E1 locus. Moreover, this interval was significantly associated with maturity in our study (data not shown). These QTL on chr 6 have previously been reported by other researchers who did not make the association maturity (Csanádi et al., 2001a; Hyten et al., 2004; Kabelka et al., 2004). Seed protein concentration QTL P5 on chr 4 (LG C1) was also previously identified by Orf et al. (1999).

Table 3.3 Quantitative trait Loci associated with seed yield measured in Fayetteville, AR; Harrisburg, IL and Knoxville, TN in an F_5 derived Essex × Williams 82 population of 274 recombinant inbred lines.

Test	Chromosome	LG	QTL[b]	Molecular Marker	QTL Position (cM)	95% Confidence Interval of QTL position	R^2	Effect[a] (g kg^{-1})
Early	5	A1	P1	BARC-031311-07043	7.8	4.9 - 10.9	0.05	-2.8
	5	A1	P2	BARC-019415-03923	23.9	15.9 - 29	0.06	-3.0
	6	C2	P3	BARC-062515-17881	220.0	212 - 225	0.36	-7.3
	7	M	P4	BARC-047995-10452	93.6	85.4 - 118.6	0.07	3.0
Mid	4	C1	P5	BARC-058277-15192	156.2	114.6 - 205.6	0.28	-4.7
	6	C2	P6	BARC-025707-05008	201.0	195.7 - 202.5	0.12	-4.2
	6	C2	P7	BARC-055939-13876	224.8	219.5 - 233.3	0.13	-3.8
	11	B1	P8	BARC-021459-04106	187.6	187.4 - 188.1	0.07	2.3
Late	1	D1a	P9	BARC-011057-00831	45.1	38.6 - 54.8	0.04	-1.7
	6	C2	P10	BARC-055939-13876	228.0	217.7 - 235	0.12	-3.3
	8	A2	P11	BARC-060405-16674	153.6	148.7 - 155.3	0.13	-3.0
	12	H	P12	BARC-019775-04370	44.0	38.1 – 49.9	0.05	1.7
	16	J	P13	BARC-022453-04332	26.8	12.7 – 301.1	0.13	-3.0
	20	I	P14	BARC-021887-04232	19.1	8.9 - 29	0.05	-2.1

[a]Effect with respect to the Essex allele. The negative sign indicates that Essex did not contribute the favorable allele.

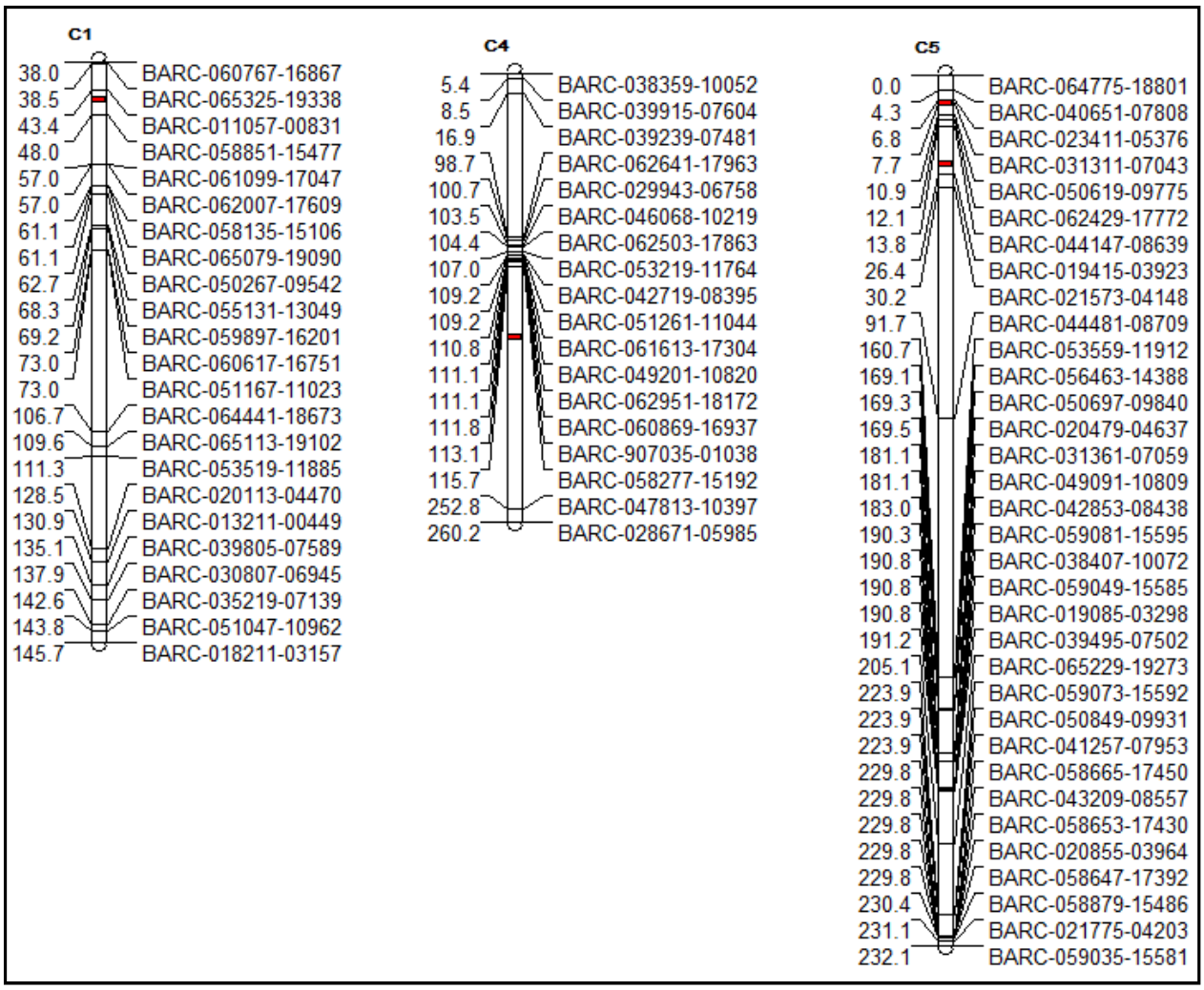

Figure 3.1 Illustration of seed protein concentration QTL (red bars) measured in Fayetteville, AR; Harrisburg, IL and Knoxville, TN environments in an F_5 derived Essex × Williams 82 population of 274 recombinant inbred lines.

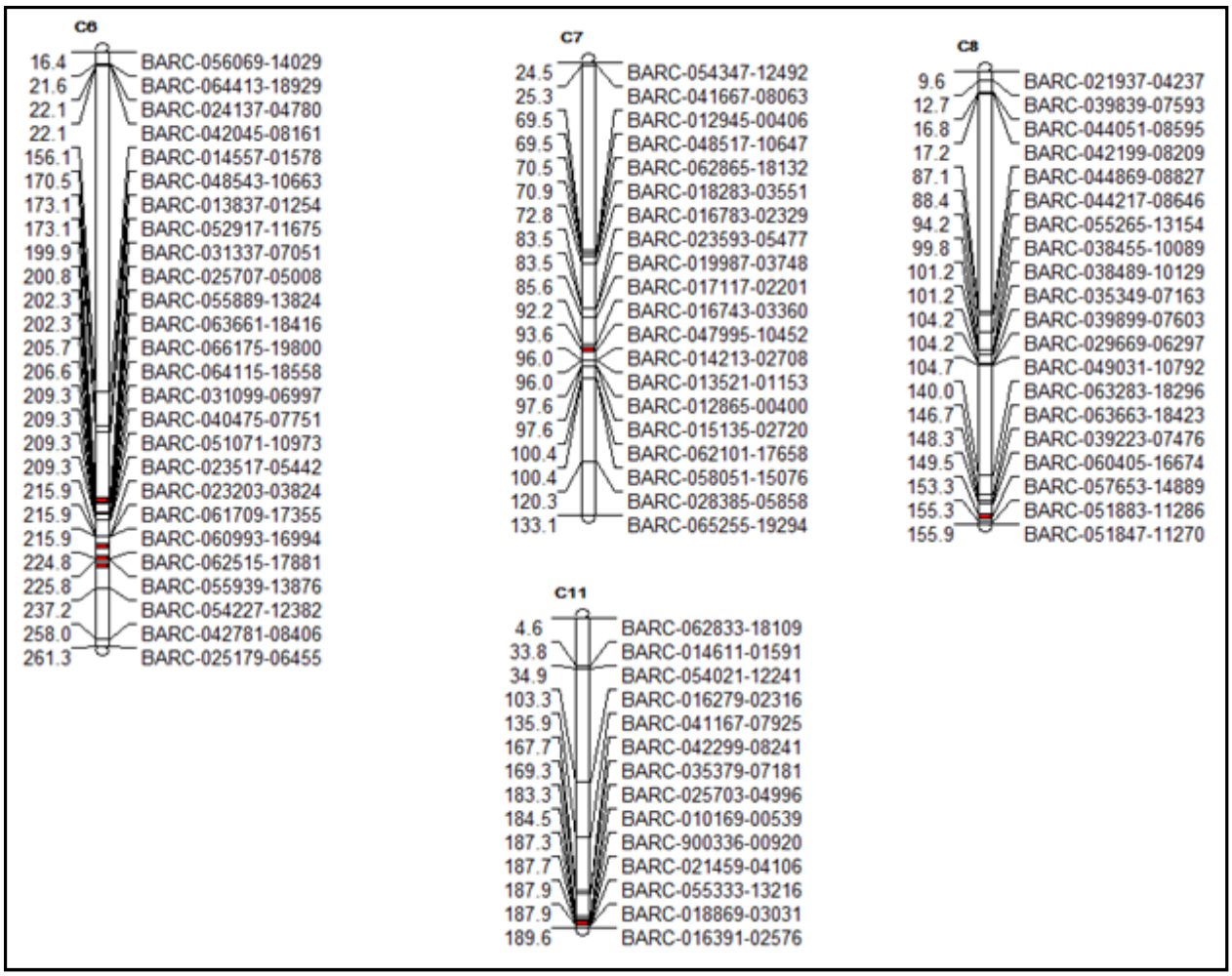

Figure 3.1 Continued.

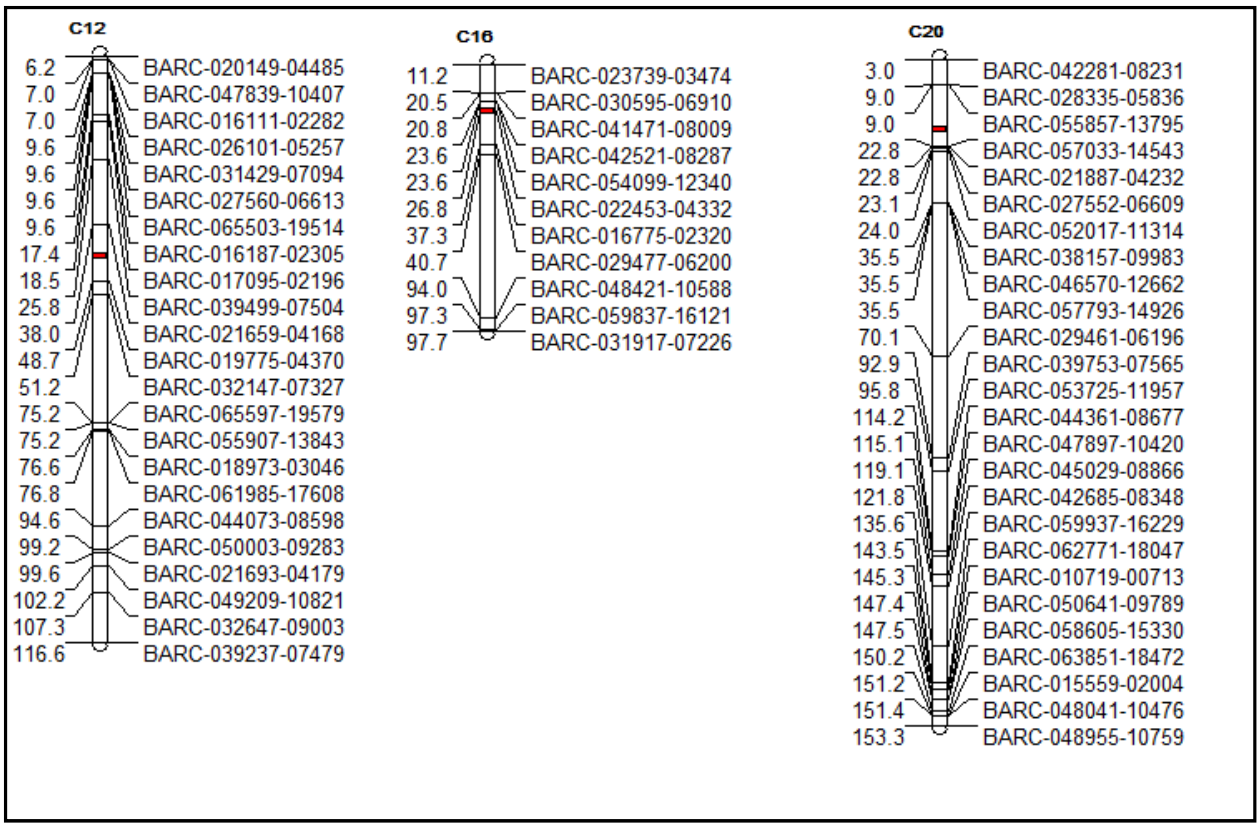

Figure 3.1 Continued.

Hyten et al. (2004) also detected a seed protein concentration QTL on chr 7 close to P4 which explained 13.3% of the variation in one environment. The P4 QTL only explained 7% of the phenotypic correlation but was detected in multiple environments. The Essex × Williams cross used by Hyten at al. (2004) is similar to our Essex × Williams 82 in that Williams 82 is a near isogenic line (NIL) of Williams which contains resistance to *Phytophthora* spp., which could mean that this QTL on chr 7 in both studies is environment specific. The seed protein concentration QTL on chr 11 (LG B1) is most likely new as there are no other protein QTL reported in this region. Csanadi et al. (2001) identified a protein QTL close to P9 on chr 1 (LG D1a) that had similar effect and explained about the same amount of variation, 4.1% in their study vs. 4.0% in ours. P11 on chr 8 (LG A2), and P12 on chr 12 (LG H) are about 10 cM upstream and 10 cM downstream of two protein QTL detected by Brummer et al. (1997) on those LGs and are most likely not new QTL. Diers et al. (1992) detected a seed protein QTL on chr 16 (LG J) which lies in close proximity to P13 and Lee et al. (1996) detected a QTL almost at the same position as P14 on chr 20 (LG I),), therefore, our results support their findings

All but one of the 10 seed protein concentration QTL reported in this study were detected in three environments and have previously been reported in Soybase and may be good candidates for MAS.

Conclusion

Several QTL associated with soybean seed protein concentration were detected in this study, account for 4 - 36% of the phenotypic variation. The QTL detected on chr 6 which were associated with maturity at the E1 locus, but may still be good candidates for MAS as this population is segregating for maturity. In contrast to these QTL on chr 6, a significant association between maturity and protein concentration was not found in any of the other QTL.

In addition, a new seed protein QTL on chr 11 (LG B1) with small effects was detected. Upon confirmation this seed protein could be used in addition to other QTL to increase seed protein concentration. The stable protein concentration QTL detected in this study should be further tested to ascertain whether selection for protein by MAS would be effective.

References

Bernard, R.L., C.R. Cremeens. 1988. Registration of 'Williams 82' Soybean. Crop Sci. 28:1027-1028.

Broman, K.W., H. Wu, Ś. Sen, G.A. Churchill. 2003. R/qtl: QTL mapping in experimental crosses. Bioinformatics 19:889-890.

Brummer, E.C., G.L. Graef, J. Orf, J.R. Wilcox, R.C. Shoemaker. 1997. Mapping QTL for seed protein and oil content in eight soybean populations. Crop Sci. 37:370-378.

Burton, J.W. 1987. Quantitative genetics: Results relevant to soybean breeding., *In* J. R. Wilcox (Ed.), pp. 211-242. Soybeans: Improvement, production, and uses., ASA, CSSA, and SSSA, Madison, WI.

Burton, J.W., T.E. Carter, R.F. Wilson. 1999. Registration of 'Prolina' Soybean. Crop Sci. 39:294-227.

Burton, J.W., B.F. Carver. 1993. Selection among s1 families vs. selfed half-sib or full-sib families in autogamous crops. Crop Sci. 33:21-28.

Chapman, A, Pantalone, R. V, Ustun, Allen, Fl, E. Landau, D, Trigiano, N. R, Gresshoff, M. P. 2003. Quantitative trait loci for agronomic and seed quality traits in an F_2 and $F_{4:6}$ soybean population. Euphytica 129:7.

Chung, J., H.L. Babka, G.L. Graef, P.E. Staswick, D.J. Lee, P.B. Cregan, R.C. Shoemaker, J.E. Specht. 2003. The Seed Protein, Oil, and Yield QTL on Soybean Linkage Group I. Crop Sci. 43:1053-1067.

Churchill, G.A., R.W. Doerge. 1994. Empirical threshold values for quantitative trait mapping. Genetics 138:963-971.

Cober, E.R., H. D Voldeng. 2000. Developing high-protein, high-yield soybean populations and lines. Crop Sci. 40:39-42.

Csanádi, G., J. Vollmann, G. Stift, T. Lelley. 2001a. Seed quality QTL identified in a molecular map of early maturing soybean. Theor. Appl. Genet. 103:912-919.

Csanádi, G., J. Vollmann, G. Stift, T. Lelley. 2001b. Seed quality QTL identified in a molecular map of early maturing soybean. Theor. App. Genet. 103:912-919.

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

Falconer, D.S. 1989. Introduction to quantitative genetics . Third Edition ed. Oliver & Boyd, Edinburgh.

Fan, J.B., A. Oliphant, R. Shen, B.G. Kermani, F. Garcia, K.L. Gunderson, M. Hansen, F. Steemers, S.L. Butler, P. Deloukas, L. Galver, S. Hunt, C. McBride, M. Bibikova, T. Rubano, J. Chen, E. Wickham, D. Doucet, W. Chang, D. Campbell, B. Zhang, S. Krugylak, D. Bentley, J. Haas, P. Rigault, L. Zhou, J. Stuelpnagel, M.S. Chee. 2003. Highly parallel SNP genotyping. Cold Spring Harbor Symp. Quant. Biol. 68:69-78.

Grant, D., R.T. Nelson, S.B. Cannon, R.C. Shoemaker. 2010. SoyBase, the USDA-ARS soybean genetics and genomics database. Nucleic Acids Res. 38:D843-D846.

Hartwig, E.E., K. Hinson. 1972. Association between chemical composition of seed and seed yield of soybeans 1. Crop Sci. 12:829-830.

Holland, J.B. 2006. Estimating genotypic correlations and their standard errors using multivariate restricted maximum likelihood estimation with SAS Proc MIXED. Crop Sci. 46:642-654.

Holland, J.B., W.E. Nyquist, C.T. Cervantes-Martínez. 2010. Estimating and interpreting heritability for plant breeding: An Update, pp. 9-112. Plant Breed. Rev., John Wiley & Sons, Inc.

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

Jin, J., X. Liu, G. Wang, L. Mi, Z. Shen, X. Chen, S.J. Herbert. 2010. Agronomic and physiological contributions to the yield improvement of soybean cultivars released from 1950 to 2006 in Northeast China. Field Crops Res. 115:116-123.

Kabelka, E.A., B.W. Diers, W.R. Fehr, A.R. Leroy, I.C. Baianu, T. You, D.J. Neece, R.L. Nelson. 2004. Putative alleles for increased yield from soybean plant introductions. Crop Sci. 44:8.

Lee, S., M. Bailey, M. Mian, T. Carter, E. Shipe, D. Ashley, W. Parrott, R. Hussey, H. Boerma. 1996. RFLP loci associated with soybean seed protein and oil content across populations and locations. Theor. App. Genet. 93:649-657.

Liu, K. 1997. Soybeans: chemistry, technology and utilization. International Thompson Publishing, New York.

Mansur, L.M., K.G. Lark, H. Kross, A. Oliveira. 1993. Interval mapping of quantitative trait loci for reproductive, morphological, and seed traits of soybean (Glycine max L.). Theor. App. Genet. 86:907-913.

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

Orf, J.H., K. Chase, T. Jarvik, L.M. Mansur, P.B. Cregan, F.R. Adler, K.G. Lark. 1999. Genetics of soybean agronomic traits: i. comparison of three related recombinant inbred populations. Crop Sci. 39:1642-1651.

Pantalone, V.R., J.W. Burton, T.E. Carter. 1996. Soybean fibrous root heritability and genotypic correlations with agronomic and seed quality traits. Crop Sci. 36:1120-1125.

Panthee, D.R., V.R. Pantalone. 2006. Registration of soybean germplasm lines TN03–350 and TN04–5321 with improved protein concentration and quality. Crop Sci. 46:2328-2329.

Panthee, D.R., V.R. Pantalone, D.R. West, A.M. Saxton, C.E. Sams. 2005. Quantitative trait loci for seed protein and oil concentration, and seed size in soybean. Crop Sci. 45:2015-2022.

Qiu, B.X., P.R. Arelli, D.A. Sleper. 1999. RFLP markers associated with soybean cyst nematode resistance and seed composition in a 'Peking'×'Essex' population. Theor. App. Genet. 98:356-364.

R Core Development Team. 2009. R: A Language and Environment for Statistical Computing., in: R. Foundation for Statistical Computing. (Ed.), Vienna, Austria.

Sas Institute. 2004. SAS/STAT software. Release 9.1.3. , SAS Inst., Cary, N.C.

Smith, T.J., H.M. Camper. 1973. Registration of Essex Soybean. Crop Sci. 13:495-495.

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

Wang, S., J. Basten, Z.B. Zeng. 2011. Windows QTL cartographer 2.5.

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

Wilcox, J.R., J.F. Cavins. 1995. Backcrossing high seed protein to a soybean cultivar. Crop Sci. 35:1036-1041.

Wilcox, J.R., R.M. Shibles. 2001. Interrelationships among seed quality attributes in soybean. Crop Sci. 41:11-14.

Wilson, R.F. 2004. Seed Composition, *In* H. R. Boerma and J. E. Specht (Eds.), pp. 621-669. Soybeans: Improvement, Production and Uses. ASA, SSSA, CSSA. Madison, WI.

Part 4

Comparison of Phenotypic and Marker Assisted Selection for Quantitative Traits in Soybean

Abstract

Breeding for quantitative traits such as seed yield and protein concentration in soybean [*Glycine max* (L.) Merrill] could be accelerated using marker assisted selection (MAS). The objective of this study was to compare the efficiency of MAS with conventional phenotypic selection (PHE) methods to improve seed yield and protein concentration using seed yield and protein concentration QTL, and phenotypic trait data previously collected from an 'Essex × 'Williams 82' F_5 derived population evaluated in the southern U.S.. Phenotypic selection for each trait separately involved selection for high and low RILs representing the top and bottom 2.5 % of the population. For the MAS selection RILs containing all 9 yield QTL and separately10 protein concentration QTL previously reported were selected. This provided 16 replicated field tests in four relative maturity groups RM 3 late (3L), RM 4 early (4E), RM 4 late (4L), RM 5 early (5E) grown in a RCBD replicated three times in three locations in Tennessee. Tests compared MAS and PHE for high and low protein concentration, and high and low seed yield. Phenotypic data for each test were combined across the three locations, and analyzed methods details not needed to evaluate differences of fixed effects. There were significant ($P < 0.05$) differences between MAS and PHE for yield selection only in the RM 3L test. Differences between yield selections (high vs. low) were significant ($P < 0.05$) for the RM 3L, 4L, and 5E but not in the 4E test. No significant ($P <0.05$) yield differences were observed in the treatment × selection interaction in all tests. Difference in selection ($P < 0.05$), (high vs. low), was observed in all four tests while treatment × selection was only significant ($P < 0.05$) in the RM 3L test for protein. These results show that both MAS and PHE may be used to select quantitative traits; however, more studies are required to optimize MAS for quantitative trait improvement.

Introduction

Soybean [*Glycine max* (L.) Merrill] is an important crop grown worldwide. Soybean is a protein and oil source that is used for nutritional and industrial applications (Liu, 1997; Wilcox, 2004). Soybean breeding efforts are concentrated on increasing yield as well as improving seed quality traits such as protein concentration, amino acid profile, oil concentration and fatty acid profile manipulation to suit market demand and consumer preferences (Pantalone et al., 2004; Wilson, 2004).

Understanding the genetics underlying important quantitative traits and the application of that knowledge into soybean breeding programs could reduce the time and resources required for rapid development of high-yielding, high-protein cultivars. Identification of genomic regions associated with quantitative traits and, or quantitative trait loci (QTL), is the first step in a marker assisted selection (MAS) strategy for yield and protein improvement. Many QTL associated with protein and oil in soybean, could be useful tools for MAS. These QTLs are published and available in Soybase (Grant et al., 2010). Lande and Thompson (1990) examined the potential for MAS in reducing the time required to successfully select for a single trait. They determined that MAS is most efficient in selection for traits with low heritability and moderate to large additive genetic variance. In addition larger population samples would likely yield better results because they would likely have larger additive genetic variance.

Hospital et al. (2000) performed a computer simulation study to determine the effectiveness of what they call marker-based recurrent selection (MBRS). This concept takes for granted that QTL detection has occurred and been confirmed for accuracy. The next step would is similar to recurrent selection but uses molecular markers to select for individuals with the most favorable alleles. The hypothetical situation proposed by the authors assumes that the QTL are

flanked by markers upstream and downstream so that selection is not limited by marker availability. The authors discuss truncation selection and complementary selection as the two primary methods that maybe employed. Truncation selection (MTS) selects for individuals with the best marker score at various QTL while complementary selection (QCS) takes into account the existence of unfavorable alleles at the QTL. The authors found that while both methods of selection were successful, QCS was superior after ten cycles. In addition, the authors recommended QCS in experimental smaller populations that have fewer markers at some loci as well as variable marker-QTL and QTL-QTL distances.

Backcrossing of major QTL and genes controlling traits such as disease resistance and herbicide tolerance has proved to be a successful technique for introgressing beneficial alleles into adapted germplasm. For example, soybean cyst nematode (SCN) resistance genes (rhg1 and Rg4) from plant introduction sources have been identified and molecular markers associated with these genes may be used to confirm greenhouse assays for resistance to various SCN races (Concibido et al., 2004; Cregan et al., 1999; Ruben et al., 2006). In contrast, backcrossing of beneficial yield and protein alleles has shown mixed results because of the complex nature of quantitative traits and known genetic (G), environmental (E) and G × E interaction.

Sebolt et al., (2000) backcrossed one of two *Glycine soja* QTL for increased protein concentration on chromosomes (chr) 15 and 20 (LGs E and I, respectively) into adapted soybean cultivars. They were able to confirm the QTL on chr 20 (LG I) and showed an increase in two of the three population backgrounds tested. In a similar study, a QTL increasing yield was found in a wild-type soybean [*Glycine soja* Sieb. and Zucc.] and was shown to have a 9% yield advantage (Concibido et al., 2003). In the study by Concibido et al. (2003), the QTL of interest was only stable in two of the six commercial cultivars into which it was backcrossed. Other researchers

have shown that QTL may be specific to a particular genetic background and subject to QTL × E interactions, reiterating the importance of testing QTL in many backgrounds and environments before they can be used extensively for MAS (Fasoula et al., 2004; Orf et al., 1999; Reyna and Sneller, 2001).

Sebastian et al. (2010) outlined the difficulty of correctly measuring yield, even within one environment, due to confounding non-genetic factors such as soil type, seed quality, plot size and disease pressure. Although it is possible to correct for most of these effects through replication and proper yield trial design, but it is likely that some confounding factors remain. Moreover, increased replications require more resources such as time, seed stocks, land, machinery and personnel.

Sebastian et al. (2010) used the term "target populations of environments" (TPE) in reference to the fact that plant breeders seek to identify the highest yielding cultivars for their specific region. Many different populations are tested in TPE to develop different cultivars. To this end, Sebastian et al. (2010) coined the term "context specific MAS" (CSM) to address the issue of QTL being subject to G and G × E. A study was conducted by Sebastian et al. (2010) to detect favorable yield QTL haplotypes for a specific TPE, and test whether those QTL could be useful for CSM of superior yielding progeny for that TPE. Nine elite commercial F_3 or F_4 derived 'motherlines' that could be expected to retain within line heterogeneity were selected and genotyped to confirm heterogeneity. Commercial lines were selected for this study because they are already suitable for a given TPE. Motherlines were grown in field plots, and approximately 300 single plants were genotyped, self-pollinated pollinate and harvested to form populations of higher yielding RIL. Sublines were selected based on confirmed yield QTL (cqQTL) and yield tested against their commercial motherlines. In multi-year, multi-environment

trials, three of the five selected haplotypes were significantly higher yielding than their motherlines. Thus, they were able to demonstrate that CSM could be used to increase yield for specific TPE.

The purpose of this study was to compare the efficiency of MAS with conventional phenotypic selection (PHE) using yield and protein concentration QTL previously detected in an 'Essex × 'Williams 82' F_5 derived population for the southern United States TPE.

Materials and Methods

Plant materials and field experiments

The initial crosses for the Essex (Smith and Camper, 1973) × Williams 82 (Bernard and Cremeens, 1988) population were made at the University of Tennessee Knoxville Plant Science Farm (KPSF) of the East Tennessee Research and Education Center (ETREC) in Knoxville, TN in the summer of 2005. The F_1 seeds were harvested in the fall of 2005 and F_1 single plants grown in Puerto Rico at the Tropical Agricultural Research Station (TARS) in Isabela, PR in the winter of 2005-06. The population was advanced through single seed descent (Brim, 1966) from the F2 to the F5 generation as follows: At the ETREC location, summers of 2005 and 2007 (F_2 to F_3), and at the TARS location, winter 2007 (F_4) and spring 2008 (F_5).

In the summer of 2008 the two parents and 284 individual $F_{5:6}$ RILs were planted in 3.1m single plant rows in ETREC for agronomic data collection, leaf collection for DNA extraction, and seed increase. Based on maturity data collected in 2008, the population was divided by days to maturity (10 days) into three tests (early, mid and late) each with 94 genotypes, with one genotype overlapping in maturity in the mid and late tests. In 2009, the three tests, the two parents and checks were planted in two 6.1m row plots in a randomized complete block design

(RCBD), replicated three times, in three locations: Knoxville, TN (ETREC), Harrisburg, IL (HA) and Fayetteville, AR (FA). Checks in the 2009 field studies were assigned by maturity group as follows: early ('IA4004', LD00-2817P, LD00-3309 and, 'Macon'), mid (TN05-4008, TN06-189, and TN06-196) and late (JTN-5203, 'OSAGE', '5002T' and '5601T').

Selections for the 2010 experiments were made from yield and protein concentration QTL data as well as phenotypic data from 2009 experiments. The extensive 2009 data allowed delineation of relative maturity groups (RM) assigned as follows: RM III late (3L), RM IV early (4E), RM IV late (4L), and RM V early (5E). For the phenotypic selection, the top (high selection) and bottom (low selection) 2.5% RILs for protein concentration and yield were selected separately. Marker assisted selections for RILs carrying favorable (high selection) and unfavorable (low selection) alleles were made from QTL detected by composite interval mapping (CIM) of the 2009 seed yield and protein concentration data within each test averaged over three environments (Tables 2.3 and 3.3), and from these subtests of RILs by MAS, the top and bottom 2.5% were randomly selected. This was done to ensure that same number of RILs were used for each test to make comparisons between the MAS selections and the phenotypic selections within their maturity group tests.

Table 4.1 Essex × Williams82 RIL selections for the 2010 field tests based on MAS for nine seed yield QTL for favorable (high-yield) or unfavorable (low yield) alleles. Seed yield, protein and oil concentration means on a 13% moisture basis of three replications over three locations grown in 2009 are reported.

Maturity Group	Selection	RIL	Yield (kg ha^{-1})	Protein (g kg^{-1})	Oil (g kg^{-1})
3L	Low-yield	ExW3-559	1957.2	389.0	180.0
3L	Low-yield	ExW3-683	3392.1	366.1	190.4
3L	High-yield	ExW3-157	3510.6	369.9	193.2
3L	High-yield	ExW3-518	3722.3	376.0	192.9
4E	Low-yield	ExW3-264	2161.3	388.9	180.0
4E	Low-yield	ExW-3217	2446.4	364.6	187.6
4E	High-yield	ExW3-082	3551.5	383.4	186.0
4E	High-yield	ExW3-203	3924.6	363.1	187.8
4L	Low-yield	ExW3-390	2738.3	378.0	187.1
4L	Low-yield	ExW3-700	3041.3	375.3	190.3
4L	Low-yield	ExW3-286	3064.5	365.7	177.7
4L	Low-yield	ExW3-363	3266.7	371.7	183.0
4L	Low-yield	ExW3-265	3454.6	378.8	174.7
4L	Low-yield	ExW3-169	3676.4	347.9	193.6
4L	High-yield	ExW3-331	3219.0	373.0	185.3
4L	High-yield	ExW3-363	3266.6	371.70	183.0
4L	High-yield	ExW3-420	3488.4	371.6	184.0
4L	High-yield	ExW3-144	3676.6	362.6	189.1
4L	High-yield	ExW3-242	3702.0	365.3	188.1
4L	High-yield	ExW3-413	3715.9	370.7	182.8
5E	Low-yield	ExW3-079	3367.6	369.6	185.4
5E	Low-yield	ExW3-188	3513.8	359.4	185.2
5E	Low-yield	ExW3-389	3562.2	375.0	183.8
5E	Low-yield	ExW3-015	3740.3	354.2	192.6
5E	High-yield	ExW3-446	3703.9	363.1	184.1
5E	High-yield	ExW3-206	3709.9	368.5	189.9
5E	High-yield	ExW3-190	4098.0	367.4	180.3
5E	High-yield	ExW3-177	4152.0	351.1	193.6

Table 4.2 Essex × Williams 82 RIL selections for the 2010 tests based on phenotypic high or low seed yield. Seed yield, protein and oil concentration means on a 13% moisture basis of three replications over three locations grown in 2009 are reported.

Maturity Group Test	Selection	RIL	Yield (kg ha^{-1})	Protein (g kg^{-1})	Oil (g kg^{-1})
3L	Low-yield	ExW3-673	1803.1	379.1	184.5
3L	Low-yield	ExW3-559	1957.2	388.9	180.0
3L	High-yield	ExW3-664	3567.6	376.2	192.3
3L	High-yield	ExW3-518	3764.4	376.0	192.9
4E	Low-yield	ExW3-264	2161.3	388.9	180.0
4E	Low-yield	ExW3-652	2255.2	369.29	186.9
4E	High-yield	ExW3-021	4073.2	361.1	188.8
4E	High-yield	ExW3-091	4404.9	354.4	196.9
4L	Low-yield	ExW3-001	2640.0	356.0	184.8
4L	Low-yield	ExW3-390	2744.7	378.0	187.1
4L	Low-yield	ExW3-065	2788.9	378.0	185.4
4L	Low-yield	ExW3-651	2921.5	385.3	184.3
4L	Low-yield	ExW3-154	2961.7	363.1	188.4
4L	Low-yield	ExW3-552	3054.5	362.0	185.6
4L	High-yield	ExW3-338	4244.4	366.5	192.7
4L	High-yield	ExW3-716	4294.2	366.7	190.3
4L	High-yield	ExW3-259	4314.9	3259.6	191.5
4L	High-yield	ExW3-380	4352.2	365.9	174.9
4L	High-yield	ExW3-315	4383.9	360.5	180.7
4L	High-yield	ExW3-230	4400.7	370.1	189.8
5E	Low-yield	ExW3-473	2534.5	375.0	183.1
5E	Low-yield	ExW3-467	3206.2	383.3	182.0
5E	Low-yield	ExW3-437	3222.7	367.6	186.7
5E	Low-yield	ExW3-371	3232.5	356.6	183.8
5E	High-yield	ExW3-231	4376.3	350.2	183.9
5E	High-yield	ExW3-309	4407.6	366.8	187.1
5E	High-yield	ExW3-375	4491.6	357.8	180.1
5E	High-yield	ExW3-388	4707.9	365.1	186.6

Table 4.3 Essex × Williams 82 RIL selections for the 2010 tests based on MAS for ten seed protein concentration QTL for favorable (high-protein) or unfavorable (low protein) alleles. Seed yield, protein and oil concentration means on a 13% moisture basis of three replications over three locations grown in 2009 are reported.

Maturity Group	Selection	RIL	Yield (kg ha^{-1})	Protein (g kg^{-1})	Oil (g kg^{-1})
3L	Low-protein	ExW3-598	3522.5	373.4	187.3
3L	Low-protein	ExW3-667	3198.6	382.0	189.9
3L	High-protein	ExW3-613	2414.9	368.6	196.4
3L	High-protein	ExW3-353	2959.7	393.3	185.5
4E	Low-protein	ExW3-019	3411.4	362.3	194.4
4E	Low-protein	ExW3-378	3425.4	387.1	194.9
4E	High-protein	ExW3-283	3456.5	388.0	185.4
4E	High-protein	ExW3-642	3696.5	390.6	189.4
4L	Low-protein	ExW3-001	4180.9	356.9	178.7
4L	Low-protein	ExW3-332	3987.2	357.7	191.8
4L	Low-protein	ExW3-340	3864.4	358.1	186.0
4L	Low-protein	ExW3-242	3702.0	365.3	188.2
4L	High-protein	ExW3-310	3569.2	350.4	187.9
4L	High-protein	ExW3-286	3064.5	365.7	177.7
4L	High-protein	ExW3-386	3517.7	374.0	178.4
4L	High-protein	ExW3-256	3987.3	380.7	175.9
5E	Low-protein	ExW3-329	3882.7	352.6	177.7
5E	Low-protein	ExW3-361	3764.9	361.7	190.0
5E	Low-protein	ExW3-447	3479.1	366.3	190.5
5E	High-protein	ExW3-138	3566.4	356.1	184.2
5E	High-protein	ExW3-309	4407.6	366.8	187.1
5E	High-protein	ExW3-467	3206.2	383.3	182.0

Table 4.4 Essex × Williams 82 RIL selections for the 2010 tests based on phenotypic high or low seed protein concentration. Seed yield, protein and oil concentration means on a 13% moisture basis of three replications over three locations grown in 2009 are reported.

Maturity Group	Selection	RIL	Yield (kg ha^{-1})	Protein (g kg^{-1})	Oil (g kg^{-1})
3L	Low-protein	ExW3-117	2886.0	363.7	196.7
3L	Low-protein	ExW3-709	3194.2	365.7	188.6
3L	High-protein	ExW3-660	3221.0	395.2	189.8
3L	High-protein	ExW3-119	2563.2	397.0	191.2
4E	Low-protein	ExW3-091	4404.9	354.4	196.9
4E	Low-protein	ExW3-290	3659.9	355.0	191.5
4E	High-protein	ExW3-164	3801.4	395.6	186.4
4E	High-protein	ExW3-618	2812.8	401.5	194.7
4L	Low-protein	ExW3-407	3962.3	344.4	194.7
4L	Low-protein	ExW3-169	3676.4	347.9	193.7
4L	Low-protein	ExW3-239	3852.8	349.1	195.3
4L	Low-protein	ExW3-107	3070.8	350.4	192.5
4L	Low-protein	ExW3-310	3569.2	350.4	187.9
4L	Low-protein	ExW3-685	3979.7	350.5	188.9
4L	High-protein	ExW3-390	2738.3	378.0	187.1
4L	High-protein	ExW3-265	3454.6	378.8	174.7
4L	High-protein	ExW3-256	3987.3	380.7	175.9
4L	High-protein	ExW3-222	3341.7	380.7	184.7
4L	High-protein	ExW3-013	3832.5	381.1	177.6
4L	High-protein	ExW3-651	2921.5	385.3	184.3
5E	Low-protein	ExW3-118	4119.4	347.5	184.2
5E	Low-protein	ExW3-246	3923.1	349.2	191.0
5E	Low-protein	ExW3-078	3561.8	349.9	183.7
5E	Low-protein	ExW3-231	4376.3	350.2	183.9
5E	High-protein	ExW3-009	3891.8	375.4	184.2
5E	High-protein	ExW3-313	3892.8	375.7	173.6
5E	High-protein	ExW3-637	3407.6	378.4	180.4
5E	High-protein	ExW3-467	3206.2	383.3	182.0

This approach provided 8 replicated field tests comparing high and low protein, and high and low yield, for both MAS and phenotypic selection. The RIL selections for the MAS and phenotypic selections for yield are presented in tables 4.1 and 4.2, and those for protein are in Tables 4.3 and 4.4, respectively. All tests were grown in two rows, with three replications in a RCBD at ETREC, Highland Rim Research and Education Center (HRREC) and the Research and Education Center at Milan (RECMLN) in Tennessee.

Phenotypic traits

Flower color was recorded when more than 95% of the plants in a plot were in full bloom (R1 to R2). Maturity was recorded when 95% of the pods in a plot showed their mature color at R8 growth stage (Fehr and Caviness, 1977). Pubescence color was also recorded at maturity. Maturity was calculated as the number of days from date of planting to R8. Plant height was measured as the average plant height from the base of the stem at the soil surface to the top of the plant at maturity within a plot. Lodging was scored at maturity with a score range from 1 (all plants erect) to 5 (all plants prostrate). Plots were end-trimmed to 4.9 m prior to harvest. Yield per plot (kg ha^{-1}) and percent moisture were recorded at harvest. For data analysis, yield per plot was adjusted to 13% moisture.

Sample Preparation for Protein and Oil NIR Analyses

Approximately 25 g of soybean seeds were ground in a water-cooled Knifetec 1095 Sample Mill (FOSS Tecator, S-26321, Hogana, Sweden) for 20 sec. This testting produced soybean flour with a uniform particle size. Approximately 15 g of ground sample was analyzed for seed protein and oil concentration using a near infrared spectroscopy (NIRS) instrument (NIRS 6500, FOSS North America). The NIRS instrument was warmed up for 2 h after turning on the lamp, and auto-diagnostic tests were run.

Diagnostics tests were performed every day until sample scanning was completed, ensuring the instrument passed three different tests for instrument response, wavelength accuracy, and NIRS repeatability. A room dehumidifier was used throughout the analysis, testting the humidity to 40%. Room temperature was approximately 20ºC. 15 gduplicates above using ISIScan software version 2.85.3. Samples were analyzed according to their replication in the field tests and the instrument was not turned off between reps. Sample scanning produced the predicted levels of protein and oil on a 13% moisture basis.

DNA Extraction and GoldenGate™ Assay

DNA was extracted from the young leaves of RILs and parents utilizing the Qiagen Plant DNeasy Extraction Kit (Qiagen, Valencia, CA). Parents and RILs were screened with the Universal Soybean Linkage Panel (U.S.L.P. 1.0) developed by (Hyten et al., 2010) which contains 1536 SNPs using the IlluminaBead™ and the GoldenGate assay™. The GoldenGate™ assay was performed as per manufacturer's protocol and as described in (Fan et al., 2003; Hyten et al., 2008).

Data Analysis

The 2010 phenotypic data for each RM test (3L, 4E, 4L, and 5E) were combined across the three locations (ETREC, HRREC, and RECMLN). Phenotypic data were analyzed using the MIXED procedure of the SAS statistical software package (release 9.2) (SAS Institute, 2008) to evaluate the significance ($p < 0.05$) of fixed effects.

Treatment (MAS or PHE), selection (High and Low), treatment × selection and genotypes nested within treatment × selection were considered to be fixed effects in the models. Locations and replications were considered blocking factors in the model. The LSMEANS statement of the MIXED procedure of the SAS statistical software package (release 9.2) (SAS

Institute, 2004) was used to calculate the means of the fixed effects. Tukey mean separation was used to separate mean differences at $P < 0.05$ for all fixed effects. To assess the efficiency of MAS vs. PHE selection, the correlation between genomic scores and observed phenotypic values for both seed yield and protein concentration was estimated. The genomic score was calculated by adding all the QTL effects to the population mean for each RIL. In addition we later determined how many seed yield and protein concentration QTL were carried by the RILs selected on the basis of phenotypic data. These two methods would indicate the efficiency of selection between the MAS and PHE selection methods.

Results and discussion

There were significant ($P < 0.05$) differences between the two treatments, (MAS and PHE), for yield selection only in the RM 3L test. Differences between yield selection (high vs. low) were significant for the RM 3L, 4L and 5E but, not in the 4E test. That is, high selections within the PHE or MAS yield tests always yielded higher than the low selections except in the RM 4E test. No significant ($P <0.05$) yield differences were observed in the treatment × selection interaction in all tests. Yield means for the RM 3L, 4E, 4L and 5E are presented in tables 4.5, 4.6, 4.7 and 4.8. Mean yield for the MAS treatment was only significantly higher than PHE in the RM 3L test (Table 4.5). In the RM 3L yield test the MAS high selection was higher yielding than the PHE high test but this increase was not statistically significant (Table 4.5). However in the RM 4L and 5E tests the only differences observed were in the treatment × selection interaction because RILs selected for high-yield in either treatment always yielded more than low yield selections (Table 4.7 and 4.8).

Table 4.5 Essex × Williams 82 relative maturity 3L recombinant inbred line MAS and phenotypic selection means, standard errors and mean separation letter groups for seed yield (g kg $^{-1}$) of three replications averaged in three environments in Tennessee in 2010.

Trait	Treatment	Mean	Std. Error	Letter groups
Seed Yield	MASYLD	2195.2	641.4	A
(kg ha-1)	PHEYLD	1625.1	641.6	B
	High	2466.1	641.4	A
	Low	1354.2	641.6	B
	MASYLD High	2623.5	652.8	A
	PHEYLD High	2308.8	652.8	AB
	MASYLD Low	1767.0	652.8	B
	PHEYLD Low	941.4	653.4	C
Maturity	PHEYLD	104.2	0.6	A
(days after planting)	MASYLD	102.1	0.6	B
	Low	103.3	0.6	A
	High	103.0	0.6	A
	PHEYLD Low	104.3	0.8	A
	PHEYLD High	104.1	0.8	A
	MASYLD Low	102.3	0.8	A
	MASYLD High	101.9	0.8	A
Height	MASYLD	67.2	11.7	A
(cm)	PHEYLD	55.7	11.7	B
	High	76.3	11.7	A
	Low	46.6	11.7	B
	MAS High	79.2	12.0	A
	PHEYLD High	73.4	12.0	A
	MASYLD Low	55.2	12.0	B
	PHEYLD Low	38.0	12.0	C
Lodging[†]	MASYLD	1.7	0.4	A
	PHEYLD	1.7	0.4	A
	High	1.9	0.4	A
	Low	1.5	0.4	B
	PHEYLD High	1.9	0.4	A
	MASYLD High	1.8	0.4	AB
	MASYLD Low	1.6	0.4	AB
	PHEYLD Low	1.4	0.4	B

[†] Lodging scored on a 1□5 scale, 1 = all plants erect, and 5 = all plants prostrate.

Table 4.6 Essex × Williams 82 relative maturity 4E recombinant inbred line MAS and phenotypic selection means, standard errors and mean separation letter groups for seed yield (g kg $^{-1}$) of three replications averaged in three environments in Tennessee in 2010.

Trait	Treatment	Mean	Std. Error	Letter Group
Seed Yield	PHEYLD	1987.6	597.1	A
(kg ha-1)	MASYLD	1823.3	647.9	A
	High	1966.3	616.2	A
	Low	1844.6	616.9	A
	PHEYLD High	2222.7	658.9	A
	MASYLD Low	1936.6	729.6	A
	PHEYLD Low	1752.5	662.9	A
	MASYLD High	1709.9	729.6	A
Maturity	PHEYLD	105.4	1.2	A
(days after planting)	MASYLD	104.9	1.6	A
	High	106.9	1.4	A
	Low	103.4	1.4	A
	MASYLD High	107.8	2.3	A
	PHEYLD High	105.9	1.7	A
	PHEYLD Low	104.8	1.7	A
	MASYLD Low	101.9	2.3	A
Height	PHEYLD	64.9	13.5	A
(cm)	MASYLD	62.3	14.9	A
	High	83.7	14.1	A
	Low	43.5	14.1	B
	MASYLD High	85.6	16.9	A
	PHEYLD High	81.8	15.1	A
	PHEYLD Low	47.9	15.1	A
	MASYLD Low	39.1	16.9	A
Lodging[†]	PHEYLD	1.9	0.4	A
	MASYLD	1.6	0.4	A
	High	2.0	0.4	A
	Low	1.4	0.4	B
	PHEYLD High	2.3	0.4	A
	MASYLD High	1.8	0.4	AB
	MASYLD Low	1.4	0.4	B
	PHEYLD Low	1.4	0.4	B

[†]Lodging scored on a 1□5 scale, 1 = all plants erect, and 5 = all plants prostrate.

Table 4.7 Essex × Williams 82 relative maturity 4L recombinant inbred line MAS and phenotypic selection means, standard errors and mean separation letter groups for seed yield (g kg $^{-1}$) of three replications averaged in three environments in Tennessee in 2010.

Trait	Treatment	Mean	Std. Error	Letter Group
Seed Yield	PHEYLD	2570.1	479.0	A
(kg ha-1)	MASYLD	2495.1	479.0	A
	High	2720.6	479.0	A
	Low	2344.6	479.0	B
	PHEYLD High	2814.4	483.9	A
	MASYLD High	2626.8	483.9	AB
	MASYLD Low	2363.4	483.9	B
	PHEYLD Low	2325.7	483.9	B
Maturity	MASYLD	111.4	1.3	A
	PHEYLD	110.9	1.3	A
	High	111.3	1.3	A
	Low	110.9	1.3	A
	MASYLD High	112.0	1.3	A
	PHEYLD Low	111.1	1.3	A
	MASYLD Low	110.8	1.3	A
	PHEYLD Low	110.6	1.3	A
Height	MASYLD	87.9	15.2	A
(cm)	PHEYLD	77.4	15.2	B
	High	86.8	15.2	A
	Low	78.6	15.2	B
	MASYLD High	94.4	15.2	A
	MASYLD Low	81.5	15.2	B
	PHEYLD High	79.3	15.2	B
	PHEYLD Low	75.7	15.2	B
Lodging[†]	PHEYLD	2.2	0.5	A
	MASYLD	2.0	0.5	A
	Low	2.2	0.5	A
	High	2.2	0.5	A
	PHEYLD Low	2.4	0.5	A
	MASYLD High	2.3	0.5	AB
	PHEYLD High	2.1	0.5	AB
	MASYLD Low	1.9	0.5	B

[†]Lodging scored on a 1□5 scale, 1 = all plants erect, and 5 = all plants prostrate.

Table 4.8 Essex × Williams 82 relative maturity 5E recombinant inbred line MAS and phenotypic selection means, standard errors and mean separation letter groups for seed yield (g kg $^{-1}$) of three replications averaged in three environments in Tennessee in 2010.

Trait	Treatment	Mean	Std. Error	Letter groups
Seed Yield (kg ha-1)	MASYLD	2447.8	402.0	A
	PHEYLD	2404.7	402.3	A
	High	2533.8	402.0	A
	Low	2318.7	402.3	B
	PHEYLD High	2613.0	407.3	A
	MASYLD High	2454.5	407.3	AB
	MASYLD Low	2441.0	407.3	AB
	PHEYLD Low	2196.4	408.2	B
Maturity (days)	MASYLD	114.1	2.5	A
	PHEYLD	113.1	2.5	A
	Low	113.6	2.5	A
	High	113.6	2.5	A
	MASYLD High	115.5	2.5	A
	PHEYLD Low	114.5	2.5	AB
	MASYLD High	112.7	2.5	BC
	PHEYLD Low	111.7	2.5	C
Height (cm)	MASYLD	91.1	16.5	A
	PHEYLD	89.5	16.5	A
	High	90.8	16.5	A
	Low	89.7	16.5	A
	MASYLD High	95.9	16.6	A
	PHEYLD Low	93.1	16.6	AB
	MASYLD Low	86.2	16.6	B
	PHEYLD Low	85.8	16.6	B
Lodging [†]	MASYLD	2.6	0.8	A
	PHEYLD	2.5	0.8	A
	High	2.7	0.8	A
	Low	2.4	0.8	A
	MASYLD High	2.9	0.8	A
	PHEYLD Low	2.5	0.8	A
	PHEYLD High	2.4	0.8	A
	MASYLD Low	2.4	0.8	A

[†]Lodging scored on a 1□5 scale, 1 = all plants erect, and 5 = all plants prostrate.

There was a difference in maturity between the MASYLD and the PHEYLD treatments in the RM 3L tests where the PHEYLD treatment was about 2 days later in maturity than the MASYLD. Selection by MASYLD was successful in producing RILs with higher yield and earlier maturity than the PHE treatment only in the RM 3L.These results suggest that MAS was only successful in the earlier maturing RILs in this population which may be explained by the major maturity effect on yield detected and selected for on chromosome (Chr 6) (LG C2) (Table 2.3).

There were significant ($P < 0.05$) differences between the MASPRO and PHEPRO treatments observed only in the RM 3L test. Difference in selection ($P < 0.05$), (high vs. low), was observed in all four tests while treatment × selection was only significant ($P < 0.05$) in the RM 3L test. The PHEPRO treatment had significantly higher protein than the MASPRO treatment in the RM 3L test (Table 4.9). Protein, and oil concentration and seed weight means for the RM 3L, 4E, 4L and 5E, are shown in tables 4.9, 4.10, 4.11 and 4.12, respectively. Selection for protein concentration in the RM 3L test also resulted in higher seed weight but had no effect on the oil concentration (Table 4.9). Selection for high-protein resulted in significantly ($P < 0.05$) lower oil in the RM 4E, and 4L tests but not in the RM 3L or 5E (Tables 4.10, 411, 4.9 and 4.12, respectively). These observations are acceptable because very weak but significant ($P < 0.05$) genetic and phenotypic correlations between protein, and oil concentration and weak positive genetic and phenotypic correlations between seed weight and protein concentration were observed in the Essex × Williams 82 RIL population from which the RM 3L was selected (Table 3.1, p.47). Results from this study indicate that there are no statistically significant differences between MAS and PHE selection approaches for seed yield increase in this population.

Table 4.9 Essex × Williams 82 relative maturity 3L RIL selections based on MAS or phenotypic selection for seed protein concentration. Seed protein and oil concentration on a 13% moisture basis, and seed weight means, standard errors and mean separation letter groups of three replications in three locations grown in 2010 are reported.

Trait	Treatment	Mean	Std. Error	Letter Group
Protein g kg^{-1}	PHEPRO	377.6	4.2	A
	MASPRO	371.4	4.2	B
	High	382.7	4.2	A
	Low	366.3	4.2	B
	PHEPRO High	389.7	4.7	A
	MASPRO High	375.7	4.6	B
	MASPRO Low	367.2	4.6	B
	PHEPRO Low	365.5	4.6	B
Oil g kg^{-1}	PHEPRO	186.8	1.1	A
	MASPRO	185.6	1.1	A
	Low	186.5	1.1	A
	High	185.8	1.1	A
	PHEPRO Low	187.7	1.4	A
	PHEPRO High	185.8	1.5	A
	MASPRO High	185.8	1.4	A
	MASPRO Low	185.3	1.4	A
Seed weight mg seed^{-1}	PHEPRO	146.5	5.8	A
	MASPRO	117.1	5.8	B
	High	135.0	5.8	A
	Low	128.6	5.8	B
	PHEPRO High	152.1	5.9	A
	PHEPRO Low	140.9	5.9	B
	MASPRO High	117.9	5.9	C
	MASPRO Low	116.2	5.9	C

Table 4.10 Essex × Williams 82 relative maturity 4E RIL selections based on MAS or phenotypic selection for seed protein concentration. Seed protein and oil concentration on a 13% moisture basis, and seed weight means, standard errors and mean separation letter groups of three replications over three locations grown in 2010 are reported.

Trait	Treatment	Mean	Std. Error	Letter Groups
Protein g kg^{-1}	MASPRO	364.3	8	A
	PHEPRO	358.3	8	A
	High	367.5	8	A
	Low	355.1	8	B
	MASPRO High	373.8	9	A
	PHEPRO High	361.2	9	A
	PHEPRO Low	355.3	9	A
	MASPRO Low	354.9	9	A
Oil g kg^{-1}	PHEPRO	198.4	2.9	A
	MASPRO	195.5	2.9	A
	Low	200.6	2.9	A
	High	193.3	2.9	B
	MASPRO Low	200.9	3.5	A
	PHEPRO Low	200.2	3.5	A
	PHEPRO High	196.5	3.6	A
	MASPRO High	190.1	3.6	A
Seed weight mg seed^{-1}	MASPRO	129.1	3.3	A
	PHEPRO	123.7	3.3	A
	High	133.6	3.3	A
	Low	119.4	3.3	B
	MASPRO High	135.7	3.9	A
	PHEPRO High	131.6	3.9	AB
	MASPRO Low	122.5	3.9	BC
	PHEPRO Low	115.8	3.9	C

Table 4.11 Essex × Williams 82 relative maturity 4L RIL selections based on MAs or phenotypic selection for seed protein concentration. Seed protein and oil concentration on a 13% moisture basis, and seed weight means, standard errors and mean separation letter groups of three replications over three locations grown in 2010 are reported.

Trait	Treatment	Mean	Std. Error	Letter groups
Protein g kg^{-1}	PHEPRO	361.3	2.5	A
	MASPRO	360.7	2.5	A
	HIGH	367.3	2.5	A
	LOW	354.7	2.5	B
	PHEPRO High	369.8	2.6	A
	MASPRO High	364.9	2.8	A
	MASPRO Low	356.6	2.8	B
	PHEPRO Low	352.7	2.6	B
Oil g kg^{-1}	MASPRO	193.5	1.3	A
	PHEPRO	193.4	1.2	A
	LOW	197.1	1.3	A
	HIGH	189.8	1.3	B
	PHEPRO Low	197.3	1.3	A
	MASPRO Low	197.0	1.4	A
	MASPRO High	190.0	1.4	B
	PHEPRO High	189.6	1.3	B
Seed weight mg $seed^{-1}$	PHEPRO	132.2	4.9	A
	MASPRO	122.9	4.9	B
	HIGH	128.8	4.9	A
	LOW	126.4	4.9	A
	PHEPRO High	134.5	4.9	A
	PHEPRO High	129.8	4.9	A
	MASPRO High	123.0	5.0	B
	MASPRO Low	122.9	5.0	B

Table 4.12 Essex × Williams 82 relative maturity 3L RIL selections based on MAs or phenotypic selection for seed protein concentration. Seed protein and oil concentration on a 13% moisture basis, and seed weight means, standard errors and mean separation letter groups of three replications over three locations grown in 2010 are reported.

Trait	Treatment	Mean	Std. Error	Letter Groups
Protein g kg^{-1}	PHEPRO	361.2	6.7	A
	MASPRO	359.2	6.8	A
	HIGH	366.8	6.7	A
	LOW	353.6	6.7	B
	PHEPRO High	371.6	6.7	A
	MASPRO High	361.9	7.1	AB
	MASPRO Low	356.5	7.3	B
	PHEPRO Low	350.8	7.1	B
Oil g kg^{-1}	MASPRO	196.4	1.8	A
	PHEPRO	194.9	1.8	A
	LOW	196.6	1.8	A
	HIGH	194.7	1.8	A
	MASPRO Low	197.0	2.1	A
	PHEPRO Low	196.2	2.0	A
	MASPRO High	195.9	2.1	A
	PHEPRO High	193.5	2.0	A
Seed weight mg $seed^{-1}$	PHEPRO	132.9	7.7	A
	MASPRO	130.2	7.7	A
	HIGH	132.9	7.7	A
	LOW	130.2	7.8	A
	PHEPRO High	138.4	7.8	A
	MASPRO Low	133.0	7.8	AB
	MASPRO High	127.4	7.8	B
	PHEPRO Low	127.3	7.8	B

There was a weak positive correlation (r= 0.41) between the seed yield genomic scores and the observed phenotypic values and, a moderate positive correlation (r= 0.63) between seed protein concentration genomic scores and phenotypic score as illustrated in figures 4.1 and 4.2, respectively. In addition QTL survey of all RIL selected on a phenotypic basis for both seed yield and protein concentration revealed that they all contained at least one QTL controlling seed yield or protein concentration but none contained all the QTL detected in this study for either trait (Tables 4.13 and, 4.14).

Results from this study indicate that there are no statistically significant differences between MAS and PHE selection approaches for seed yield increase in this population. In particular, the interactions of interest, MAS high and PHE high, did not show significant yield difference in any of our tests. The same observations were made in the seed protein concentration tests except in the RM 3L tests, where PHE high selection was significantly higher than MAS high (Table 4.9).

Marker assisted selection studies for quantitative traits in soybean are lacking in the literature. A few researchers have published work indicating that selection for QTL underlying a certain trait, for example seed yield, may not lead to yield increases for several reasons such as the QTL already being fixed in the population of interest, or genetic background specific and QTL being a false positive or environmentally specific (Brummer et al., 1997; Guzman et al., 2007; Kabelka et al., 2004; Orf et al., 1999; Reyna and Sneller, 2001; Wang et al., 2004). A QTL increasing yield found in a wild type soybean [*Glycine soja* Sieb. and Zucc.) was shown to have a 9% yield advantage, but it was stable in only two of the six commercial cultivars it was backcrossed into when tested in multiple environments (Concibido et al., 2003).

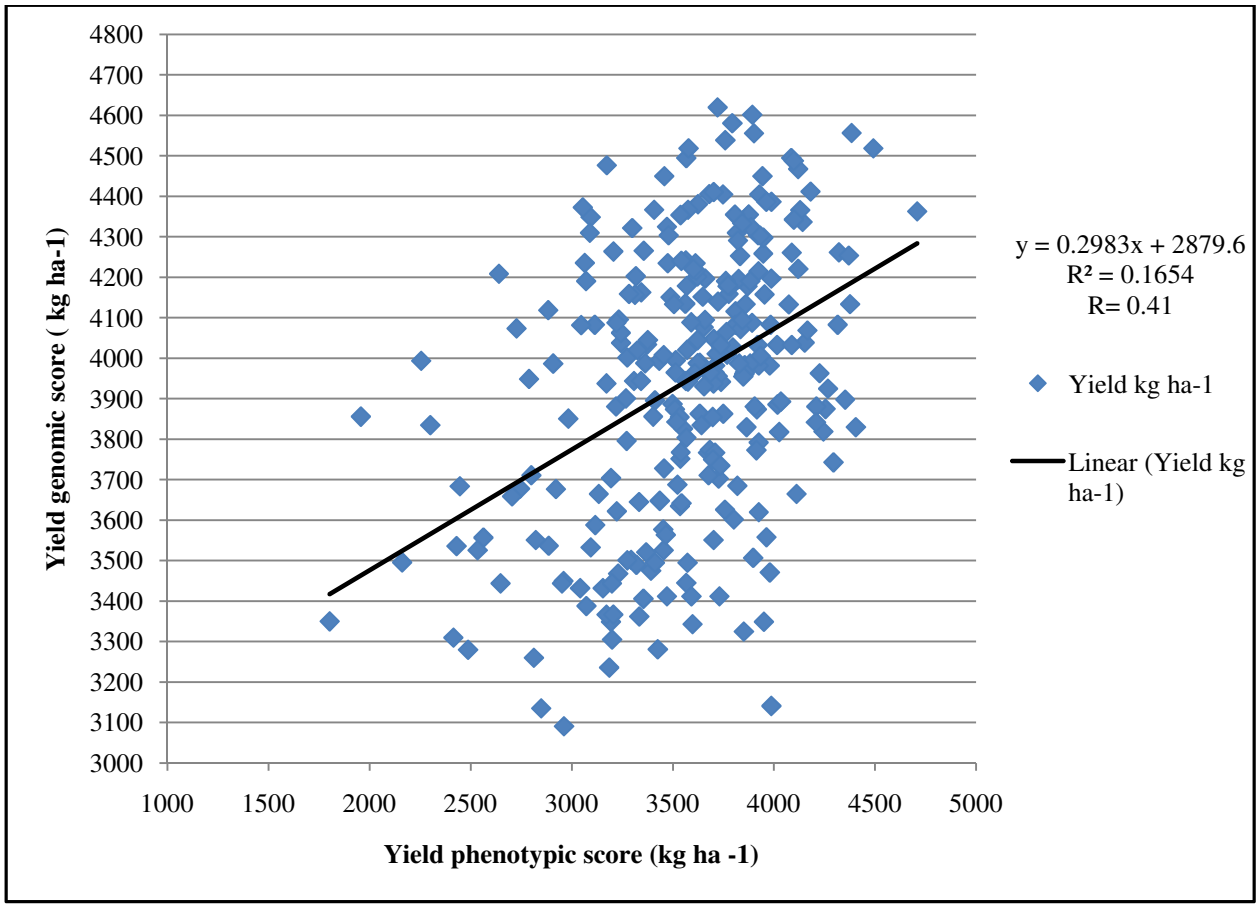

Figure 4.1 Seed yield correlation between genomic and phenotypic scores of 274 RILs from an Essex × Williams 82 cross, averaged across three environments (Fayetteville, AR; Harrisburg, IL and Knoxville, TN) in 2009.

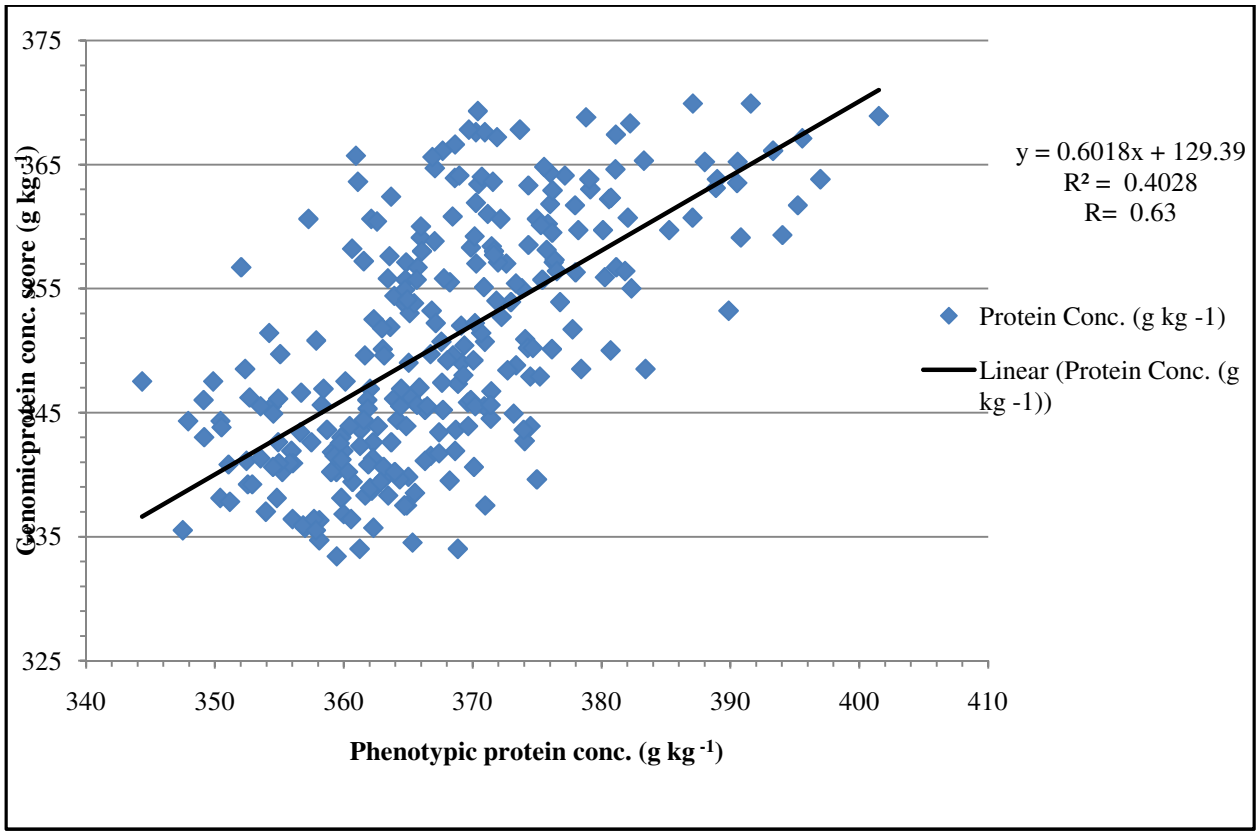

Figure 4.2 Seed protein concentration correlation between genomic scores and phenotypic data of 274 RILs from an Essex × Williams 82 cross, averaged across three environments (Fayetteville, AR; Harrisburg, IL and Knoxville, TN) in 2009.

Table 4.13 Recombinant inbred line selections for yield based on phenotypic data measured in 2009 for the 2010 MAS vs. PHE comparison study and, the yield QTL detected in this study contained in each line.

Maturity Group	Selection	RIL	Yield	Protein	Oil	Yield QTL
3L	High yield	ExW3-518	3764.4	376.0	192.9	Y1, Y2, Y5, Y6, Y11
3L	High yield	ExW3-664	3567.6	376.2	192.3	Y4, Y7, Y9, Y10
3L	Low yield	ExW3-559	1957.2	388.9	180.0	Y1, Y3, Y4, Y7, Y10
3L	Low yield	ExW3-673	1803.1	379.0	184.5	Y2, Y4, Y8, Y9,10
4E	High yield	ExW3-091	4404.88	354.4	196.9	Y1, Y2, Y4, Y5, Y7, Y8, Y9, Y11
4E	High yield	ExW3-021	4073.2	361.1	188.8	Y1, Y2, Y3, Y4, Y5, Y8, Y9
4E	Low yield	ExW3-652	2255.2	369.3	186.9	Y1
4E	Low yield	ExW3-264	2161.3	388.9	180.0	Y5, Y10
4L	High yield	ExW3-230	4400.7	370.1	189.8	Y1, Y2, Y3, Y4, Y5, Y6, Y7, Y9, Y10, Y11
4L	High yield	ExW3-315	4383.9	360.5	180.7	Y1, Y4, Y6, Y7, Y9, Y11
4L	High yield	ExW3-380	4352.2	365.9	174.9	Y1, Y3, Y4, Y7, Y10
4L	High yield	ExW3-259	4314.9	359.6	191.5	Y1, Y2, Y4, Y5, Y6, Y7, Y8, Y9
4L	High yield	ExW3-716	4294.2	366.7	190.3	Y2, Y3, Y4, Y5, Y7, Y9, 10
4L	High yield	ExW3-338	4244.4	366.5	192.7	Y1, Y4, Y6, Y7, Y9, Y10
4L	Low yield	ExW3-552	3054.5	362.0	185.6	Y2, Y3, Y4, Y8, Y9, Y10
4L	Low yield	ExW3-154	2961.7	363.1	188.4	Y2, Y5, Y7, Y11
4L	Low yield	ExW3-651	2921.5	385.3	184.3	Y3, Y4, Y5, Y9, Y10
4L	Low yield	ExW3-065	2788.9	378.0	185.4	Y8, Y11
4L	Low yield	ExW3-390	2744.8	378.0	187.1	Y1, Y2, Y3, Y7, Y8, Y9, Y10
4L	Low yield	ExW3-001	2640.0	356.0	184.8	Y1, Y3, Y6, Y7, Y11
5E	High yield	ExW3-388	4708.0	365.1	186.6	Y3, Y4, Y7, Y8, Y11
5E	High yield	ExW3-375	4491.6	357.8	180.1	Y7, Y10, Y11
5E	High yield	ExW3-309	4407.6	366.8	187.1	Y1, Y7, Y10, Y11
5E	High yield	ExW3-231	4376.3	350.2	183.9	Missing SNP data
5E	Low yield	ExW3-371	3232.5	356.6	183.8	Y5, Y7, Y9, Y11
5E	Low yield	ExW3-437	3222.7	367.6	186.7	Y1, Y2, Y3, Y5, Y6, Y11
5E	Low yield	ExW3-467	3206.2	383.3	182.0	Y1, Y2, Y5, Y7, Y8, Y11
5E	Low yield	ExW3-473	2534.5	375.0	183.1	Y1, Y2, Y3, Y4, Y6, Y8

Table 4.14 Recombinant inbred line selections for yield based on phenotypic data measured in 2009 for the 2010 MAS vs. PHE comparison study and, the yield QTL detected in this study contained in each line.

Maturity Group	Selection	RIL	Yield	Protein	Oil	Protein QTL
3L	High Protein	ExW3-660	3221.0	395.2	189.8	P4, P5, P6, P8, P9
3L	High Protein	ExW3-119	2563.2	397.0	191.2	P5, P14
3L	Low Protein	ExW3-709	3194.2	365.7	188.6	P1, P3, P4, P7, P8, P9, P10, P12
3L	Low Protein	ExW3-117	2886.0	363.7	196.7	P1, P4, P5, P8, P13
4E	High Protein	ExW3-164	3801.4	395.6	186.4	P1, P4, P8, P9, P11, P12
4E	High Protein	ExW3-618	2812.8	401.5	194.7	P4, P9
4E	Low Protein	ExW3-091	4404.9	354.4	196.9	P3, P5, P7, P9, P10, P11, P12
4E	Low Protein	ExW3-290	3659.9	355.0	191.5	P3, P6, P7, P10, P11, P13, P14
4L	High Protein	ExW3-256	3987.3	380.7	175.9	P3, P4, P6, P7, P8, P10, P11, P12, P13
4L	High Protein	ExW3-013	3832.5	381.1	177.6	P6, P8, P11, P12
4L	High Protein	ExW3-265	3454.6	378.8	174.7	P1, P2, P4, P8, P12
4L	High Protein	ExW3-222	3341.7	380.7	184.7	P4, P5, P6, P8, P9
4L	High Protein	ExW3-651	2921.5	385.3	184.3	P5, P6, P8, P2, P13
4L	High Protein	ExW3-390	2738.3	378.0	187.1	P2, P5, P6, P8, P9
4L	Low Protein	ExW3-685	3979.7	350.5	188.9	P1, P3, P6, P7, P8, P9, P10, P11
4L	Low Protein	ExW3-407	3962.3	344.4	194.7	P2, P3, P5, P9, P11, P12, P14
4L	Low Protein	ExW3-239	3852.8	349.1	195.3	P3, P6, P7, P10, P11
4L	Low Protein	ExW3-169	3676.4	347.9	193.7	P3, P6, P7, P9 P10, P13
4L	Low Protein	ExW3-310	3569.2	350.4	187.9	P3, P4, P5, P6, P7, P10, P13
4L	Low Protein	ExW3-107	3070.8	350.4	192.5	P1, P2, P3, P4, P6, P7, P10, P11, P13, P14
5E	High Protein	ExW3-313	3892.8	375.7	173.6	Missing SNP data
5E	High Protein	ExW3-009	3891.8	375.4	184.2	P5, P6, P11, P13
5E	High Protein	ExW3-637	3407.6	378.4	180.4	P1, P2, P3, P4, P6, P7, P8, P10
5E	High Protein	ExW3-467	3206.2	383.3	182.0	P1, P4, P6, P12
5E	Low Protein	ExW3-231	4376.3	350.2	183.9	Missing SNP data
5E	Low Protein	ExW3-118	4119.4	347.5	184.2	P1, P2, P3, P5, P6, P7, P9, P10, P12, P13
5E	Low Protein	ExW3-246	3923.1	349.2	191.0	P3, P5, P6, P7, P10, P11, P12
5E	Low Protein	ExW3-078	3561.8	349.9	183.7	P3, P4, P6, P7, P8, P9, P10, P13, P14

A similar observation was made by Reyna and Sneller (2001) who found that favorable QTL in one environment from a particular parent might not be transferable to other environments where the source of the QTL is inferior. These observations make the concept of CSM for particular TPE worth examining. A better understanding of TPE and available germplasm is important when developing improved cultivars that are commercially viable. Sebastian et. al. (2010) showed that it was possible to use MAS to increase yield for the central U.S. TPE. Sebastian et al. (2010) tested nine RIL sublines derived from different commercial lines which resulted in yield improvement in five of the nine tested. This suggests that testing more populations per TPE whenever possible may increase the number of improved lines for the trait of interest. In addition, a more relaxed probability level which would lead to the detection of more QTL may be more useful in a MAS scheme than a conservative one, even though it would likely cause more Type I errors. The alternative probability level of $P < 0.05$, which we used, is associated with less power and more Type II errors which are means that fewer QTL will be detected (Bernardo, 2008). It is possible that our conservative probability level caused us to miss some QTL in this population associated with yield and seed protein concentration that would have allowed us to test more RILs that had desirable haplotypes. In the future, a study that includes more populations adapted to our TPE, relaxed probability levels, and checks to ensure that MAS RIL selections have phenotypes that actually high or low for the trait might give a better understanding of how to employ MAS for quantitative traits effectively.

Conclusion

There was a significant difference between MAS and PHE selection for yield and protein concentration in the RM 3L tests, which were influenced by maturity in our study. However, because only a small number of genotypes and random RILs with the correct haplotype for MAS

were selected, further testing of more lines and populations is necessary to get more definitive results.

Selection for high or low quantities of either seed yield or protein concentration did not always include lines whose phenotype was high or low. However, since the objective of this study was to compare phenotypic vs. MAS for the purpose of determining the possibility of accurately selecting phenotypes by selecting markers flanking the QTL of interest, ,we did not manually make substitutions to correct the MAS selection. This study and the work of others is an indication that MAS for quantitative traits requires better understanding not only because of the complexity of quantitative traits but also because of the challenges involved in the accurate detection and confirmation of QTL used for selection.

References

Bernard, R.L., C.R. Cremeens. 1988. Registration of 'Williams 82' Soybean. Crop Sci. 28:1027-1028.

Bernardo, R. 2008. Molecular markers and selection for complex traits in plants: learning from the last 20 years. Crop Sci. 48:1649-1664.

Brummer, E.C., G.L. Graef, J. Orf, J.R. Wilcox, R.C. Shoemaker. 1997. Mapping QTL for seed protein and oil content in eight soybean populations. Crop Sci. 37:370-378.

Concibido, V. La, Mclaird, Pineda, Meyer, Hummel, Yang, Wu, Delannay. 2003. Introgression of a quantitative trait locus for yield from Glycine soja into commercial soybean cultivars. Theor. Appl. Genet. 106:575-582.

Concibido, V.C., B.W. Diers, P.R. Arelli. 2004. A decade of QTL mapping for cyst nematode resistance in soybean. Crop Sci. 44:1121-1131.

Cregan, P.B., J. Mudge, E.W. Fickus, D. Danesh, R. Denny, N.D. Young. 1999. Two simple sequence repeat markers to select for soybean cyst nematode resistance coditioned by the rhg1 locus. Theor. App. Genet. 99:811-818.

Fan, J.B., A. Oliphant, R. Shen, B.G. Kermani, F. Garcia, K.L. Gunderson, M. Hansen, F. Steemers, S.L. Butler, P. Deloukas, L. Galver, S. Hunt, C. Mcbride, M. Bibikova, T. Rubano, J. Chen, E. Wickham, D. Doucet, W. Chang, D. Campbell, B. Zhang, S. Krugylak, D. Bentley, J. Haas, P. Rigault, L. Zhou, J. Stuelpnagel, M.S. Chee. 2003. Highly parallel SNP genotyping. Cold Spring Harbor Symp. Quant. Biol. 68:69-78.

Fasoula, V.A., D.K. Harris, R.H. Boerma. 2004. Validation and designation of quantitative trait loci for seed protein, seed oil, and seed weight from two soybean populations. Crop Sci. 44:8.

Fehr, W.R., C.E. Caviness. 1977. Stages of soybean development, Iowa State Univ., Ames.

Grant, D., R.T. Nelson, S.B. Cannon, R.C. Shoemaker. 2010. SoyBase, the USDA-ARS soybean genetics and genomics database. Nucleic Acids Res. 38:D843-D846.

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

Hyten, D., Q. Song, I.-Y. Choi, M.-S. Yoon, J. Specht, L. Matukumalli, R. Nelson, R. Shoemaker, N. Young, 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., I.-Y. Choi, Q. Song, J.E. Specht, T.E. Carter, R.C. Shoemaker, E.-Y. Hwang, L.K. Matukumalli, 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.

Kabelka, E.A., B.W. Diers, W.R. Fehr, A.R. Leroy, I.C. Baianu, T. You, D.J. Neece, R.L. Nelson. 2004. Putative alleles for increased yield from soybean plant introductions. Crop Sci. 44:8.

Liu, K. 1997. Soybeans: chemistry, technology and utilization. International Thompson Publishing, New York.

Orf, J.H., K. Chase, T. Jarvik, L.M. Mansur, P.B. Cregan, F.R. Adler, K.G. Lark. 1999. Genetics of Soybean Agronomic Traits: I. Comparison of Three Related Recombinant Inbred Populations. Crop Sci. 39:1642-1651.

Pantalone, V.R., D.R. Walker, R.E. Dewey, I. Rajcan. 2004. DNA Marker-Assisted Selection for Improvement of Soybean Oil Concentration and Quality, *In* R. F. Wilson, et al. (Eds.), pp. 283-311.

Reyna, N., C.H. Sneller. 2001. Evaluation of marker-assisted introgression of yield QTL alleles into adapted soybean. Crop Sci. 41:1317-1321.

Ruben, E., A. Jamai, J. Afzal, V. Njiti, K. Triwitayakorn, M. Iqbal, S. Yaegashi, R. Bashir, S. Kazi, P. Arelli, C. Town, H. Ishihara, K. Meksem, D. Lightfoot. 2006. Genomic analysis of the *rhg1* locus: candidate genes that underlie soybean resistance to the cyst nematode. Mol. Gen. Genom. 276:503-516.

Sas Institute. 2004. SAS/STAT software. Release 9.1.3. , SAS Inst., Cary, N.C.

Smith, T.J., H.M. Camper. 1973. Registration of Essex Soybean. Crop Sci. 13:495-495.

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, J.A. Hoeck. 2010. Context-specific marker-assisted selection for improved grain yield in elite soybean populations. Crop Sci. 50:1196-1206.

Sebolt, A.M., R.C. Shoemaker, B.W. Diers. 2000. Analysis of a quantitative trait locus allele from wild soybean that increases seed protein concentration in soybean. Crop Sci. 40:1438-1444.

Wang, D., G.L. Graef, A.M. Procopiuk, B.W. Diers. 2004. Identification of putative QTL that underlie yield in interspecific soybean backcross populations. Theor. App. Genet. 108:458-467.

Wilcox, J.R. 2004. World Distribution and Trade of Soybean, *In* H. R. Boerma and J. E. Specht (Eds.), pp. 1-13. Soybeans: Improvement, Production and Uses. ASA, CSSA, SSSA, Madison, WI.

Wilson, R.F. 2004. Seed Composition, *In* H. R. Boerma and J. E. Specht (Eds.), pp. 621-669. Soybeans: Improvement, Production and Uses. ASA, CSSA, SSSA, Madison, WI.

Part 5

Conclusion

Conclusion

Soybean quantitative traits such as seed yield, protein and oil concentration have many industrial and nutritional applications. The U.S. is currently the largest producer of soybeans followed by Brazil, Argentina, China and India (Wilcox, 2004). The U.S. soybean crop in 2010 was valued at $38.9 billion dollars up $7.1 billion from the previous year. In that same year, 69% of the world's vegetable protein consumption was soybean meal (SBM) while 29% of vegetable oil consumption was soybean oil (American Soybean Association, 2011).

Cultivar improvement to increase yield while maintaining protein content or with modest seed protein content increase to about 44 □ 45 % with no less than 18 % oil concentration is the goal of many soybean breeders (Wilson, 2004). The negative correlation that exists between soybean seed yield and protein concentration presents a challenge to attaining this goal (Burton, 1987; Wilson, 2004). Soybean breeders have occasionally been successful in overcoming this challenge and have produced high yielding high protein cultivars (Jin et al., 2010; Ustun et al., 2001).

Conventional methods such as recurrent selection have been used successfully to improve quantitative traits (Burton et al., 1999; Wilcox, 1998; Wilcox and Cavins, 1995). Such methods require extensive phenotyping which requires a lot of resources such as land, personnel in addition to the time needed to test a cultivar before it is released. Moreover, traits such as yield are difficult to measure accurately even with replication due not non genetic confounding effects in the environment such as soil type, disease pressure, plot size and seed quality (Sebastian et al., 2010) .

The availability of molecular markers and dense molecular maps has enabled researchers to identify many genomic regions associated with quantitative traits (QTL) such as seed yield

and protein concentration (Bernardo, 2008). Currently, there are 55 yield and 86 protein QTL reported in Soybase which could be useful tools for marker assisted selection (MAS) breeding (Grant et al., 2010). However, studies investigating the potential of MAS for the improvement of quantitative traits are lacking in the literature. Recently, (Sebastian et al., 2010) were able to show that MAS strategies could be used successfully to increase yield in a targeted population of environments.

The objective of this study was to identify quantitative trait loci (QTL) associated with seed yield, and separately protein concentration, and then compare phenotypic selection (PHE) and MAS approaches for seed yield and protein concentration increase.

Two hundred and eighty two F_5 derived recombinant inbred lines (RILs) were developed from a cross of Essex × Williams 82 and genotyped with 1586 single nucleotide polymorphism (SNP) markers. The population was divided by days to maturity (10 days) into three tests (Early, Mid and Late) each with 94 genotypes, with one genotype overlapping in maturity days in the mid and late tests. In 2009, the three tests, parents and, checks were evaluated in a randomized complete block design (RCBD), in three environments replicated three times, and evaluated for seed yield and protein concentration. Data was combined within each test across the three locations and analyzed using the MIXED procedure of SAS software to determine if there were significant genotypic differences among RILs.

Composite interval mapping (CIM) detected nine minor and major seed yield QTL, R-squares less than or greater than 10 %, respectively, which were stable across environments, and have been reported by other studies. We also detected ten seed protein concentration QTL, accounting for 4 — 36% of the phenotypic variation observed. One of the ten seed protein QTL detected in this study has not been reported in Soybase (Grant et al., 2010) and maybe a new

QTL for seed protein concentration. The seed yield and protein concentration QTL in this study are good candidates for MAS as they were stable across environments.

Selections to compare phenotypic selection (PHE), and MAS for seed yield and protein concentration provided 8 replicated field tests for each trait, in four relative maturity (RM) groups (3L, 4E, 4L and 5E) grown in a RCBD, replicated three times in three locations in Tennessee, in 2010.For the phenotypic selection, the top (high selection) and bottom (low selection) 2.5% RILs for protein concentration and yield were selected separately. Marker assisted selections for RILs carrying favorable (high selection) and unfavorable alleles (low selection) were made from the nine seed yield QTL, and the ten protein concentration QTL detected by composite interval mapping (CIM) of the 2009 data. Random selections from the MAS pool of genotypes were made to ensure that we had balanced tests in four relative maturity groups (RM) 3L, 4E, 4L and 5E.

Analysis of the 2010 data showed that there was a significant difference between MAS high, and PHE high selection for yield and protein concentration only in the RM 3L tests which was influenced by maturity in our study. This result indicates that it may be possible to use MAS for the selection of quantitative traits. A small number of genotypes representing only the top and bottom 2.5 % of the total RILs in a test were selected, in addition to selecting random RILs for MAS which had the correct haplotype. Further testing of more lines and populations is necessary to get more definitive results. Selection for high or low quantities of either seed yield or protein concentration did not always select lines whose phenotype was high or low. However, since the objective of this study was to compare phenotypic vs. MAS to see if one could accurately select phenotypes by selecting markers flanking the QTL of interest so we did not manually make substitutions correct the MAS selection.

Our findings in study suggest that MAS for quantitative traits can be successful but requires better understanding not only because of the complexity of quantitative traits but also because of the challenges involved in the accurate detection and confirmation of QTL used for selection.

References

American Soybean Association. 2011. A Reference Guide to Important Soybean Facts and Figures.

Bernardo, R. 2008. Molecular Markers and Selection for Complex Traits in Plants: Learning from the Last 20 Years. Crop Sci. 48:1649-1664.

Burton, J.W. 1987. Quantitative genetics: Results relevant to soybean breeding., *In* J. R. Wilcox (Ed.), pp. 211-242. Soybeans: Improvement, production, and uses., ASA, CSSA, and SSSA, Madison, WI.

Burton, J.W., T.E. Carter, R.F. Wilson. 1999. Registration of 'Prolina' Soybean. Crop Sci. 39:294-227.

Grant, D., R.T. Nelson, S.B. Cannon, R.C. Shoemaker. 2010. SoyBase, the USDA-ARS soybean genetics and genomics database. Nucleic Acids Res. 38:D843-D846.

Hospital, F., I. Goldringer & S. Openshaw. 2000. Efficient marker-based recurrent selection for numerous quantitative trait loci. Genet Res. 75 (3): 357–368.

Jin, J., X. Liu, G. Wang, L. Mi, Z. Shen, X. Chen, S.J. Herbert. 2010. Agronomic and physiological contributions to the yield improvement of soybean cultivars released from 1950 to 2006 in Northeast China. Field Crops Res. 115:116-123.

Lande, R. and R. Thompson. 1990. Efficiency of marker-assisted selection in the improvement of quantitative traits. Genetics, 124: 743–756.

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, J.A. Hoeck. 2010. Context-Specific Marker-Assisted Selection for Improved Grain Yield in Elite Soybean Populations. Crop Sci. 50:1196-1206.

Ustun, A., F.L. Allen, B.C. English. 2001. Genetic Progress in Soybean of the U.S. Midsouth. Crop Sci. 41:993-998.

Wilcox, J.R. 1998. Increasing Seed Protein in Soybean with Eight Cycles of Recurrent Selection. Crop Sci. 38:1536-1540.

Wilcox, J.R. 2004. World Distribution and Trade of Soybean, *In* H. R. Boerma and J. E. Specht (Eds.), pp. 1-13. Soybeans: Improvement, Production and Uses. ASA, CSSA, SSSA, Madison, WI.

Wilcox, J.R., J.F. Cavins. 1995. Backcrossing High Seed Protein to a Soybean Cultivar. Crop Sci. 35:1036-1041. ASA, CSSA, SSSA. Madison, WI.

Wilson, R.F. 2004. Seed Composition, *In* H. R. Boerma and J. E. Specht (Eds.), pp. 621-669. Soybeans: Improvement, Production and Uses, Madison. ASA, CSSA, SSSA, Madison, WI.

Vita

Catherine Nyaguthii Nyinyi was born in Nyeri, Kenya on the 9th of March 1982. She attended Murray State University in Murray, Kentucky and received a Bachelor of Science degree in Agriculture in 2004. In 2005, she worked as a corn breeding intern for Monsanto Company in Arlington, WI.

In 2006 she enrolled at the University of Tennessee and began soybean breeding research under Dr. Vincent Pantalone and received a Master of Science degree in plant sciences in 2008. Catherine continued her PhD graduate work at the University of Tennessee under the Dr. Vincent Pantalone in the summer of 2008. She hopes to begin a new career as a soybean breeder for Monsanto Company after the successful completion of her studies.