INVESTIGATION OF PUTATIVE GENETIC FACTORS ASSOCIATED WITH SOYBEAN [*GLYCINE MAX* (L.) MERR.] SEED QUALITY TRAITS

Jeffrey Allen Skoneczka

Dissertation submitted to the faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

In

CROP AND SOIL ENVIRONMENTAL SCIENCES

M. A. Saghai Maroof, Chairman

Glenn. R. Buss

Katy M. Rainey

Richard E. Veilleux

Carol A. Wilkinson

September 29, 2009
Blacksburg, Virginia

Keywords: Soybean, sucrose, raffinose, stachyose, phytate, seed weight, maturity, linkage map, QTL, whole genome shotgun sequence

Investigation of Putative Genetic Factors Associated with Soybean [*Glycine* max (L.) Merr.] Seed Quality Traits

by

Jeffrey Allen Skoneczka

ABSTRACT

Soybeans are an economically important plant, with an annual crop value that consistently exceeds 20 billion dollars in the United States alone. A recent increase in demand for soybeans, stemming from its diverse applications in products such as animal feed, oil, and biofuel, has created an emphasis for soybean breeders in value added cultivars. These cultivars, have improved, or altered, agronomic or seed composition traits, allowing them to be efficiently utilized in a specific niche of the processing industry. Facilitating the development of such cultivars requires a thorough understanding of the genetic factors that affect the manifestation of value added traits. Value added traits investigated in this study include seed sucrose, raffinose, stachyose, and phytate content, seed weight, and maturity.

The objective of the first part of this project was to characterize the source of low seed stachyose in soybean line PI200508. Two F_2 populations, developed from PI200508 and soybean introductions which exhibited higher seed stachyose content were utilized in a QTL analysis approach that incorporated the use of the Williams82 whole genome shotgun (WGS) sequence (http://www.phytozome.org) in a candidate gene mapping approach. A predicted soybean galactosyltransferase gene was established as a candidate gene due to its observed segregation with the single low stachyose QTL observed on molecular linkage group (MLG) C2 in both populations. Sequencing of this putative

gene revealed a unique 3 bp deletion in PI200508. A marker developed to exploit this deletion accounted for 88% and 94% of the phenotypic variance for seed stachyose content in the two experimental populations, highlighting its potential for use in marker assisted selection of the PI200508 source of low raffinose and stachyose.

The second part of this project involved QTL analysis of seed sucrose, raffinose, stachyose, and phytate content, as well as seed weight in a linkage map for a F_8 RIL population developed from the *Glycine max* line V71-370 and the *Glycine soja* introduction PI40712. Analysis across all 20 soybean MLG identified 25 QTL for these traits on MLG A1, A2, C2, D1b, D2, F, G, H, I, L, M, O. Nine of these QTL were supported across multiple environments, indicating that they, and their associated markers, could be useful to breeders working with these traits.

The third part of this project used the same F_8 RIL linkage map to investigate time to maturity (Reproductive stage R8). V71-370 and PI407162 differ in time to maturity when grown in Virginia, and the RILs developed from this cross displayed a wide range in maturity. Two major QTL were identified on MLG H and L. Examination of the Williams82 WGS sequence in these QTL regions revealed two predicted genes with homology to Arabidopsis thaliana light response and photoperiodism genes which were investigated as candidate soybean maturity genes. Markers developed from these predicted genes showed close association with the observed QTL, and could facilitate the further investigation of this complex trait.

Dedicated to my parents, Marilyn and Dennis,

and my brother, Jeremy, without whose love and support

I could not have hoped to succeed in this undertaking.

ACKNOWLEDGEMENT

I take this opportunity to thank my committee members, Dr. Glenn R. Buss, Dr. Katy M. Rainey, Dr. Richard E. Veilleux, and Dr. Carol A. Wilkinson, for their advice and guidance over the past five years. Also, I express my gratitude to my advisor and committee chair, Dr. M.A. Saghai Maroof, whose counsel has been critical to both the pursuit of this degree, and in my development as a researcher.

I also thank Dr. Ruslan Biyashev, for providing both important research support, and good conversation, my graduate student and post-doc colleagues, for their assistance and friendship, as well as my grandparents, Ralph and Phyllis Schelling, for instilling in me a love of agriculture, and the recognition of its importance. There are countless others who have assisted and supported me along the way, and I am extremely grateful to you all.

TABLE OF CONTENTS

LIST OF FIGURES

Chapter 2

Chapter 3

Chapter 4

LIST OF TABLES

Chapter 2

Chapter 3

Chapter 4

Chapter 1

Introduction

Because of its many uses, soybean [*Glycine max* (L.) Merr.] has become an extremely important crop both at home and abroad. The United States, the leader in soybean production, planted over 75 million acres, with a crop value that exceeded 19 billion dollars in 2006. Despite these lofty numbers, it is important to note that the US soy market now faces abundant competition in the international marketplace, with production in South America and Asia increasing yearly. Additionally, the planted acreage of soybeans seems to be leveling off, after increasing steadily each year prior to 2000. Because of this, it is critical that American growers are provided with the most productive, highest quality soybean cultivars that breeders can provide.

Like many plants in the family Fabaceae, soybean is an important crop worldwide, serving as a major protein source for both humans and livestock. Because of this, it is an excellent dietary complement to cereals, which are high in carbohydrates, but relatively low in protein. There are additional benefits to soy consumption. Compounds such as isoflavones abound in soyfoods and have reported anti-carcinogenic properties (Jacobsen, 1998). Furthermore, there are many industrial uses for soy and soy by-products, such as manufacture of plastics (Paetau, 1994), biodiesel (Mahajan et al., 2006), and as components in adhesives, inks, dyes, and pharmaceuticals (USB, 2007).

As mentioned, the high levels of protein found in soy complement a cereal based diet. It is this balance that is used in many animal feeds. Cereals, although an excellent energy source, lack the necessary amount of protein, as well as the proper amino acid composition in the protein they provide. The essential amino acid lysine, for example, is much higher in soy protein than protein found in cereals (Rockland, 1981). Animal feed

can provide adequate protein, with a more acceptable amino acid profile if soy meal is included as an ingredient. With currently available soybean cultivars, the percent of soy meal used in animal feed is limited by naturally occurring, anti-nutritional compounds, including phytate, and raffinose family oligosaccharides (RFOs). The presence of these compounds therefore imposes, albeit indirectly, a threshold on the amount of protein that an animal feed can contain. A soy line with low levels of stachyose and phytate, that maintains high sucrose levels and agronomic performance would be highly desirable to the animal feed industry, as it would ultimately increase the amount of soy meal that can be used.

Phytate

Also known as phytic acid and *myo*-inositol-1,2,3,4,5,6-hexa*kis*phosphate, phytate is a seed storage form of phosphorous, typically accumulating in seeds as phytate and phytin salts of potassium and magnesium. It has been shown to account for as much as 85% of seed total phosphorous (Raboy, 1997). Non-phytate P is usually found as inorganic P, and that found in nucleic acids, proteins, or cellular membranes (Lopez et al., 2002). Phytate is a chelator of positively charged cations, and when ingested, can bind minerals such as calcium, iron, and zinc, and positively charged proteins (Raboy, 2002).

The phytin salts that are formed between phytate and these mineral cations are assumed to play an important role in mineral storage. Present in seeds, phytins are degraded by enzymes called phytases, which releases phosphate, mineral cations, and myo-inositol for use by the growing seedling (Lopez et al., 2002). When ingestion

occurs in monogastric animals, the chelated compounds are generally excreted, providing no nutritional value. Because of this, feed must be supplemented with additional phosphorous, or phytase. An additional concern that then arises beyond the nutrition of the animal is the environmental impact of the excreted phytate phosphorous, as it can contribute to water pollution and eutrophication.

The biosynthesis of phytate begins with the synthesis of D-*myo*-inositol-3 phosphate ($Ins(3)P_1$), from D-glucose 6-P. The enzyme D-*myo*-inositol-3-phosphate synthase (MIPS) facilitates this process (Chappell et al., 2006), which is irreversible, and the end product is involved in several plant metabolic pathways (Loewus and Murthy, 2000). In addition to being a phytate precursor, *myo*-inositol also plays a role in synthesis of pectins, hemicelluloses, and other cell wall components (Loewus and Loewus, 1983), raffinose family oligosaccharides (Horbowicz and Obendorf, 1994), and in signaling pathways (Stevenson-Paulik et al., 2005).

MIPS, and more specifically, the gene (*MIPS*) that encodes the protein has become of great interest to molecular breeders attempting to identify the genetic basis for phytate synthesis. A recent study identified a soybean line harboring a *MIPS* mutation in which seed phytate levels were reduced by 50% (Hitz et al., 2002). Of further interest, the mutant genotype appeared to increase sucrose levels, while decreasing raffinose and stachyose, which are all desirable soybean phenotypes. There does, however, appear to be a significant difference between altered *MIPS* expression and silenced expression of the gene. In a study where interfering RNA technology was used to silence *GmMIPS1*, it was seen that although primary transformants exhibited normal phenology, nearly every seed (98.55%) that had inherited the transgene cassette aborted before reaching maturity

(Nunes et al., 2006b). Because the numerous signaling and metabolic pathways involving *myo*-inositol and its derivatives are still poorly understood, it is difficult to determine the effect of altered *MIPS* expression. It would seem, however, that complete silencing or a null mutation has largely deleterious effects. Nunes et al. (2006) were only able to produce T1 plants that inherited their transgene in one case where *GmMIPS1* appeared to be only partially silenced. The resultant progeny (T2) from this transformation event showed decreased phytate levels.

Initial success with low-phytate soybeans has come from generating lines that overexpress the enzyme phytase (Chiera et al., 2005), or through altering the activity of *myo*-inositol 1-phosphate synthase (MIPS) genes using RNAi technology (Nunes et al., 2006a) or mutagenesis (Hitz et al., 2002). However, no low phytate quantitative trait loci (QTL) in soybean have been reported until recently. Exceptions lie in two recent studies that applied QTL analysis to locate sources of the low phytate phenotype in soybean line CX1834 on soybean molecular linkage groups (MLG) L and N (Gao et al., 2008; Walker et al., 2006). The cause of these QTL were later identified as putative mutations, not in the *MIPS* class of genes, but in soybean multidrug resistance-associated protein genes (*MRP*) (Saghai Maroof et al., 2009). These genes were previously shown to be deficient in low phytate *Zea mays* mutants (Shi et al., 2007). Along with *MIPS* genes, the association of *MRP* genes with seed phytate content offers important information regarding the manifestation of this trait.

Sugars

Oligosaccharides (sugars) found in soybean seed include the primary photosynthate sucrose, as well as the oligosaccharides raffinose and stachyose. The former being the most desirable seed oligosaccharide as the latter two, although ubiquitous in plant seeds, provide little nutritional value for monogastric animals, and are considered anti-nutritional. Raffinose and stachyose oligosaccharides must first be digested by α-galactosidase before they can be metabolized. This enzyme is not synthesized in appreciable levels by most monogastric animals. It is critical to consider the oligosaccharide makeup when developing a soybean that, when ground into meal and incorporated into animal feed, provides an optimal level of carbohydrate energy.

The bulk of soybean seed oligosaccharide content is comprised of sucrose, while stachyose and raffinose make up the remaining fraction (Hymowitz and Collins, 1974). Additionally, sucrose is the dominant oligosaccharide found in developing fruit during seed fill (Thorne, 1980). Sucrose is a disaccharide made up of two monosaccharides, D-glucose and D-fructose, joined by a glycosidic bond. Although work has been done to identify genomic regions which affect sucrose content in soybeans (Kim et al., 2006; Maughan et al., 2000) the incorporation of this knowledge into a marker assisted selection (MAS) breeding strategy has not been realized.

Raffinose family oligosaccharides (RFOs), raffinose and stachyose, have been reported to play a role in cold acclimation, desiccation tolerance, and seed viability (Horbowicz and Obendorf, 1994). They are related in that they both have sucrose as their precursor and they are tri- and tetra- saccharides, respectively, differing primarily in the number of galactose residues sequentially linked to D-glucose and D-fructose (the latter

two residues comprise a sucrose molecule). Synthesis of both molecules is dependent on the transfer of a galactosyl group to sucrose from galactinol through the action of transferase enzymes (Pharr and Sox, 1984). Galactinol concentrations are positively correlated with the activity of galactinol synthase activity, which in turn reaches its peak directly before RFO seed concentrations begin to accumulate (Saravitz et al., 1987). The biosynthetic pathway of these RFOs are as follows (Dey, 1985):

UDP-galactose + *myo*-inositol → galactinol + UDP

Sucrose + galactinol → raffinose + *myo*-inositol

Raffinose + galactinol → stachyose + *myo*-inositol

To date, little has been made public regarding breeding for oligosaccharide levels. Maughan et al. (2000) reported RFLP, RAPD, and SSR markers associated with seed sucrose content on seven MLG in a F_2 population derived from an interspecific cross. Sucrose accumulation, like phytate accumulation, manifests through a complex pathway in which many genes or QTL are involved. Investigation into the genetic components of this pathway would provide breeders with additional tools to select from and develop soy lines with lower stachyose and raffinose. The result would benefit primarily the animal feed industry, which relies heavily on soy meal as the primary protein component of feed. Soy meal with lower anti-nutritional compounds will provide more carbohydrate-derived energy, while still providing the desired protein amount.

Agronomic traits

Agronomic traits of soybean include: yield, seed size, and maturity. Because these are the basic contributors to plant performance and quality, they have been traits on which traditional breeding methods of selection have always been applied. However, these traits are strongly influenced by the environment and genetic background (Lee et al., 1996a; Orf et al., 1999). QTL analysis will allow us to apply molecular marker technology to discover the underlying genetic pathways contributing to these traits, facilitating MAS.

Increasing plant yield has always been a goal of breeding programs. In order to increase production, either the yield per plant needs to be increased, or the quality of the product needs to be improved (as in developing value added soybean lines). Mansur et al. (1996) and Zhang et al. (2004) mapped 15 and 10 agronomic traits, respectively, and both found that QTL for these traits seemed to exist in clusters, rather than spread throughout the genome. This is not surprising, as selecting for yield will generally mean that you are positively selecting larger plants, plants with larger seed, more nodes per stem, etc. A general trait such as yield will have numerous physiological components contributing to it, and each of those components can be under the influence of their own genetic factors, making yield a very complex trait to analyze. As a result, its components are often analyzed individually for use in breeding schemes.

Seed weight is a commonly examined trait affecting yield. A recent study showed QTL for seed weight on MLG D2, H, M, and O (Liu et al., 2007). Previous studies have shown similar QTL on MLG D2 and O (Chen et al., 2007; Zhang et al., 2004), which in each case, explained over 10 percent of the phenotypic variation. A QTL on MLG L was

reported by Hyten et al. (2004a), which explained as much as 43% of the phenotypic variation. Additionally, a study using single marker analysis revealed significant associations between seed size and markers that reside within MLG M and L (Csanadi et al., 2001). These reports are encouraging when considering the possibility of using MAS to select for seed size. A major step in increasing soy production has been the development of larger seeded cultivars throughout the twentieth century, however, the development of smaller seeded varieties is also desirable for certain markets. Incorporating knowledge of seed size QTL into a breeding scheme could provide a valuable boost to breeding for this trait.

Maturity

Maturity has been a trait that has affected the domestication of soybean greatly. Because it is a multigenic trait, there exists a great diversity for time to maturity, allowing soybean to be grown in a wide range of locations. Soybean varieties are divided into maturity groups, which are largely determined by daylength variation caused by latitude. A given cultivar's response to daylength is thought to be largely under the control of maturity genes of the *E* class. There are currently seven known *E* class maturity genes, *E1* and *E2* (Bernard, 1971), *E3* (Buzzell, 1971), *E4* (Buzzell and Voldeng, 1980), *E5* (McBlain and Bernard, 1987), *E6* (Bonato and Vello, 1999), and *E7* (Cober and Voldeng, 2001). An additional locus (*J*) is responsible for long juvenility (Ray et al., 1995). Allelic differences at these loci are largely responsible for affecting soybean maturity, yet there has been little correlation between maturity QTL positions and *E* genes. Exceptions lie on MLG C2 where a maturity QTL linked to the *t* locus (tawny pubescence) and

microsatellite marker Satt205 explained as much as 47% of the phenotypic variation, and resides at the putative location of *E1 (Tasma et al., 2001)*. Additional minor QTL for maturity have been reported on numerous MLG; however, close association with the above maturity genes has not been established. Increased knowledge regarding the genetic regions surrounding these loci could facilitate the cloning of these genes, allowing the time to maturity to be altered. Existing cultivars could be "customized" to grow in a wider range of environments, and developing cultivars could be tailored to grow in any desired region.

Oil and Protein Content

Soybean seed can be composed of over 40% protein and 20% oil. Traditional breeding methods have been highly successful in increasing protein and oil content due to the high heritability of the traits. However, like most traits, these can be greatly affected by environment, making QTL analysis and establishment of linkage to molecular markers useful in breeding schemes. A problem facing any program breeding for these traits lies in that oil is strongly, negatively correlated with protein content, making selection for high levels of both difficult. It has also resulted in reports of pleiotropic QTL affecting both traits (Diers et al., 1992; Mansur et al., 1993). Early studies identified significantly associated molecular markers on numerous linkage groups (Brummer et al., 1997; Csanadi et al., 2001; Lee et al.); however, variation across generations, environments, and populations remains an issue. An exception is a *G. soja* high protein/low oil QTL on MLG I first identified by Diers et al. (1992). The region was successfully introgressed into a *G. max* line, where it increased protein content and decreased oil content in the

resultant progeny (Sebolt et al., 2000). This QTL has been fine mapped recently (Chung et al., 2003). The ability of the environment to mask QTL, as well as the inconsistency between different populations make it necessary to further establish a molecular basis to track and modify this trait. Some of the inconsistency observed in previous studies is merely due to differences in linkage maps, with marker density creating gaps that may interfere with detection of smaller QTL. This is an issue needing to be addressed when considering QTL as candidates for gene discovery studies, and makes a single high density, comprehensive linkage map in a diverse population desirable for analysis of this, and any trait.

Fatty Acid Composition

In addition to seed oil content, the composition of the oil is also important as it determines the quality of the oil. Soy oil is typically 10% palmitic acid, 4% stearic acid, 22% oleic acid, 54% linoleic acid, and 10% linolenic acid (Wilson, 2004). Palmitic and stearic acids are undesirable as they are saturated fatty acids, while linoleic and linolenic acids lack oxidative stability, and soybean oil high in these fatty acids must be hydrogenated. Additionally, hydrogenation results in the partial synthesis of trans fat. Saturated and hydrogenated oils, and recently, trans fat, have been shown to have negative health effects when consumed.

Genetic variation in seed content of the four undesirable fatty acids is limited, and sources conferring low-levels of undesirable fatty acids were initially found resulting from gene mutations. Reduced palmitic acid levels have been conferred through the action of three independent genes of the *fap* class (Erickson et al., 1988; Fehr et al., 1991;

Wilcox et al., 1994) and shown not to reduce yield (Rebetzke et al., 1998b). Similarly, low linolenic acid is conferred through mutation in the *fan* class of genes, which may exist at as many as three independent genetic loci (Fehr et al., 1992; Rahman and Takagi, 1997; Ross et al., 2000; Wilcox and Cavins, 1987). Oleic acid, the most desirable fatty acid due to its stability and lack of negative health implications, has been increased through recurrent selection of a natural mutation (Burton et al., 1983). Aside from these mutations, however, effective genetic sources to modify these traits may be found in fatty acid modifier genes, which can behave like minor QTL, contributing to the quantitative expression of the trait (Rebetzke et al., 1998a). Exploiting these modifying genes has led to a few reported QTL; however, in some cases they cosegregate with maturity QTL which raises questions of whether the observations have been confounded by differences in time to maturity (Hyten et al., 2004b). Based on the difficulties in identifying genetic diversity affecting this trait, the most likely source of polymorphism for these modifying elements may be in a population derived from an interspecific cross in which genetic diversity affecting the synthesis and accumulation of fatty acids would be maximized.

QTL analysis

The current emphasis for plant breeders is in "value added" cultivars, or lines that possess superior agronomic ability, combined with seed characteristics that have been optimized for their desired use. Because the majority of these characteristics is controlled by many genetic factors, they exhibit continuous variation, and are difficult to analyze and modify. A powerful tool available to breeders for characterizing multigenic traits is QTL analysis.

QTL analysis uses statistical algorithms to determine cosegregation between marker data and phenotypic trait data. In the case of single marker analysis, ANOVA or regression analysis can be performed to determine whether there is a significant association between a marker and trait. Additionally, more thorough analysis, such as interval mapping can be performed using a linkage map to determine the most likely position of a QTL. In general, we use single marker analysis to determine regions of importance during marker selection for construction of a linkage map, and interval mapping to determine the QTL position. Although Mapmaker 3.0 (Lander, 1987) is occasionally used, the bulk of our analysis is done with JoinMap 4 (Van-Oiogen, 2006) for linkage analysis, and MapQTL (Van-Oiogen, 2004) for QTL analysis.

Although it can be useful in incorporating marker assisted selection into breeding programs, QTL identification is merely an early step in characterizing and cloning a QTL, a process ultimately leading to gene discovery. Knowledge of a QTL position through the above statistical methods yields no knowledge of sequence information, whether a QTL is a gene, or if so, what kind of gene. Characterizing a QTL on this level requires cloning and sequencing, which can be difficult. To date, there are few QTL that have been successfully cloned. Mapping expressed sequence tags (ESTs) and gene sequences obtained from databases has provided new opportunities to identify QTL; however, the aforementioned steps to clone and sequence are still necessary. Our expanding pool of sequence information will allow us to better characterize QTL and use them to facilitate crop improvement. It is our research goal to gather phenotypic data for as many useful traits as possible, identify QTL through conventional methods, while incorporating the use of new sequence information, and clone those QTL that could be

used in a molecular breeding program. Many QTL have been reported in soybean, affecting yield, disease resistance, and seed quality traits, to name a few.

Summary

The growing economic importance of soy, as well as its diverse uses, makes it important that value-added soybeans be developed to optimize their usefulness and maximize their profitability for grower and industry alike. The development of such lines that are specialized for their niche in industry depends on isolating and identifying the genetic components that are responsible for the trait of interest. A significant step in gene discovery is QTL analysis, which identifies regions on a genetic linkage map that affect the manifestation of a trait. A number of important traits with reported QTL are outlined here; however, these are a small fraction of the total number of soybean QTL that have been reported, and that remain undiscovered. Future knowledge, combined with sequence information, the volume of which increases daily, makes the large scale identification of genes in a highly profitable crops a reality.

In the face of increasing competition in the global market, it is important that emphasis be placed on developing soybean lines with biochemical traits that are optimized for industrial use. While United States' soy production has shown little growth over the last three years, continued growth is projected from South American sources, most notably Brazil and Argentina (USDA, 2007b). Although the United States out-produces both of these countries in total soybean production, this is primarily as an oilseed crop. There is currently more soybean meal, an important commodity to the animal feed industry, produced in South America on a yearly basis (USDA, 2007a).

The objectives of the studies detailed in this dissertation include the evaluation of the following seed traits in various experimental populations: seed sucrose, raffinose, stachyose, and phytate content, seed size, and time to maturity. Seed for evaluation was harvested in bulk when the plants reached maturity and were stored at room temperature. Additional agronomic traits such as maturity and seed weight were also evaluated as they are important components of yield and geographic distribution of soybean. The investigation of the genetic basis behind traits such as low phytate, low stachyose, and high sucrose is also important, as their incorporation into a high-yielding, consistently performing commercial cultivar would be worthwhile.

REFERENCES

Bernard, R.L. 1971. Two major genes for time of flowering and maturity in soybeans. Crop Sci 11:242-244.

Bonato, E.R., and N.A. Vello. 1999. *E6*, a dominant gene conditioning early flowering and maturity in soybeans. Genet Mol Biol 22:229-232.

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

Burton, J.W., R.F. Wilson, and C.A. Brim. 1983. Recurrent selection in soybean IV. Selection for increased oleic acid percentage in seed oil. Crop Sci 23:744-747.

Buzzell, R.I. 1971. Inheretance of a soybean flowering response to fluorescent-daylength conditions Can J Genet Cytol 13:703-707.

Buzzell, R.I., and H.D. Voldeng. 1980. Inheritance of insensitivity to long daylength. Soybean Genetics Newsletter:26-29.

Chappell, A.S., A.M. Scaboo, X. Wu, H. Nguyen, V.R. Pantalone, and K.D. Bilyeu. 2006. Characterization of the MIPS gene family in *Glycine max*. Plant Breed 125:493-500.

Chen, Q., Z. Zhang, C. Liu, D. Xin, H. Qiu, D. Shan, C. Shan, and G. Hu. 2007. QTL analysis of major agronomic traits in soybean. Agricultural Sciences in China 6:399-405.

Chiera, J.M., J.J. Finer, and E.A. Grabau. 2005. Ectopic expression of a soybean phytase in developing seeds of *Glycine max* to improve phosphorus availability. Plant Mol Biol 56:895-904.

Chung, J., H.L. Babka, G.L. Graef, P.E. Staswick, D.J. Lee, P.B. Cregan, R.C. Shoemaker, and J.E. Specht. 2003. The seed protein, oil, and yield QTL on soybean linkage group I. Crop Sci 43:1053-1067.

Cober, E.R., and H.D. Voldeng. 2001. A new soybean maturity and photoperiod-sensitivity locus linked to *E1* and *T*. Crop Sci 41:698-701.

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

Dey, P.M. 1985. D-Galactose Containing Oligosaccharides, p. 53-130, *In* P. M. Dey and R. A. Dixon, eds. Biochemistry of Storage Carbohydrates in Green Plants. Academic, New York.

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

Erickson, E.A., J.R. Wilcox, and J.F. Cavins. 1988. Inheritance of altered palmitic acid percentage in two soybean mutants. J Hered 79:465-468.

Fehr, W.R., G.A. Welke, E.G. Hammond, D.N. Duvick, and S.R. Cianzio. 1991. Inheritance of reduced palmitic acid content in seed oil of soybean. Crop Sci 31:88-89.

Fehr, W.R., G.A. Welke, E.G. Hammond, D.N. Duvick, and S.R. Cianzio. 1992. Inheritance of reduced linolenic acid content in soybean genotypes A16 and A17. Crop Sci 32:903-906.

Gao, Y., R.M. Biyashev, M.A.S. Maroof, N.M. Glover, D.M. Tucker, and G.R. Buss. 2008. Validation of low-phytate QTLs and evaluation of seedling emergence of low-phytate soybeans. Crop Sci 48:1355-1364.

Hitz, W.D., T.J. Carlson, P.S. Kerr, and S.A. Sebastian. 2002. Biochemical and molecular characterization of a mutation that confers a decreased raffinosaccharide and phytic acid phenotype on soybean seeds. Plant Physiol 128:650-660.

Horbowicz, M., and R.L. Obendorf. 1994. Seed desiccation tolerance and storability: Dependence on flatulence-producingoligosaccharides and cyclitols -- review and survey. Seed Science Research 4:385-405.

Hymowitz, T., and F.I. Collins. 1974. Variability of sugar content in seed of *Glycine max* (L.) Merrill and *G. soja* Sieb and Zucc. Agron J 66:239-240.

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

Hyten, D.L., V.R. Pantalone, A.M. Saxton, M.E. Schmidt, and C.E. Sams. 2004b. Molecular mapping and identification of soybean fatty acid modifier quantitative trait loci. Journal of the American Oil Chemists' Society 81:1115-1118.

Jacobsen, B.K., Synnøve F. Knutsen, Gary E. Fraser. 1998. Does High Soy Milk Intake Reduce Prostate Cancer Incidence? The Adventist Health Study (United States). Cancer Causes & Control 9:541-543.

Kim, H.K., S.T. Kang, and K.W. Oh. 2006. Mapping of putative quantitative trait loci controlling the total oligosaccharide and sucrose content of *Glycine max* seeds. J Plant Res 119:533-538.

Lander, E.S., P. Green, J. Abrahamson, A. Barlow, M. J. Daly, S. E. Lincoln, and L. Newberg 1987. Mapmaker: an interactive computer package for constructing primary genetic linkage maps of experimental and natural populations. Genomics 1:174-181.

Lee, S.H., B.M. Bailey, M.A.R. Mian, E.R. Shipe, D.A. Ashley, W.A. Parrott, R.S. Hussey, and H.R. Boerma. 1996a. Identification of quantitative trait loci for plant height, lodging, and maturity in a soybean population segregating for growth habit. Theor Appl Genet 92:516-523.

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

Liu, B., T. Fujita, Z.-H. Yan, S. Sakamoto, D. Xu, and J. Abe. 2007. QTL mapping of domestication-related traits in soybean (*Glycine max*). Ann Bot-London 100:1027-1038.

Loewus, F.A., and M.W. Loewus. 1983. *myo*-Inositol: Its biosynthesis and metabolism. Ann Rev Plant Phys 34:137-161.

Loewus, F.A., and P.P.N. Murthy. 2000. *myo*-Inositol metabolism in plants. Plant Sci 150:1-19.

Lopez, H.W., F. Leenhardt, C. Coudray, and C. Remesy. 2002. Minerals and phytic acid interactions: is it a real problem for human nutrition? International Journal of Food Sci and Tech 37:727-739.

Mahajan, S., S.K. Konar, and D.G.B. Boocock. 2006. Standard biodiesel from soybean oil by a single chemical reaction. Journal of American Oil Chem Soc 83:641-644.

Mansur, L., J. Orf, K. Chase, T. Jarvik, P. Cregan, and K. Lark. 1996. Genetic mapping of agronomic traits using recombinant inbred lines of soybean. Crop Sci 36:1327-1663.

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

Maughan, P.J., M.A. Saghai Maroof, and G.R. Buss. 2000. Identification of quantitative trait loci controlling sucrose content in soybean (*Glycine max*). Mol Breed 6:105-111.

McBlain, B.A., and R.L. Bernard. 1987. A new gene affecting the time of flowering and maturity in soybeans. J Hered 78:160-162.

Nunes, A.C.S., G.R. Vianna, F. Cuneo, J. Amaya-Farfan, and G.d. Capdeville. 2006a. RNAi-mediated silencing of the *myo*-inositol-1-phosphate synthase gene (*GmMIPS1*) in transgenic soybean inhibited seed development and reduced phytate content. Planta 224:125-132.

Nunes, A.C.S., G.R. Vianna, F. Cuneo, J. Amaya-Farfan, G.d. Capdeville, E.L. Rech, and F.J.L. Aragao. 2006b. RNAi-mediated silencing of the *myo*-inositol-1-phosphate synthase gene (*GmMIPS1*) in transgenic soybean inhibited seed development and reduced phytate content. Planta 224:125-132.

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

Paetau, I., Chin-Zue Chen, Jay-lin Jane. 1994. Biodegradable Plastic Made from Soybean Products. 1. Effect of Preparation and Processing on Mechanical Properties and Water Absorption. Industrial Engineering and Chemistry Research 33:1821-1827.

Pharr, D.M., and H.N. Sox. 1984. Changes in carbohydrate and enzyme levels during the sink to source transition of leaves of *Cucumis sativus* L., a stachyose translocator. Plant Sci Lett 35:187-193.

Raboy, V. 1997. Accumulation and storage of phosphates and minerals, p. 441-477, *In* B. A. L. a. I. K. Vasil, ed. Advances in cellular and molecular biology of plants, Vol. 4. Kluwer Academic Publisher, Dordrecht, the Netherlands.

Raboy, V. 2002. Progress in breeding low phytate crops. Journal of Nutrition 132:503S-505S.

Rahman, S.M., and Y. Takagi. 1997. Inheritance of reduced linolenic acid content in soybean seed oil. Theor Appl Genet 94:299-302.

Ray, J.D., K. Hinson, J.E.B. Mankono, and M.F. Malo. 1995. Genetic Control of a Long-Juvenile Trait in Soybean. Crop Sci 35:1001-1006.

Rebetzke, G.J., J.W. Burton, J. T E Carter, and R.F. Wilson. 1998a. Genetic variation for modifiers controlling reduced saturated FA content in soybean. Crop Sci 38:303-308.

Rebetzke, G.J., J.W. Burton, J. Thomas E Carter, and R.F. Wilson. 1998b. Changes in agronomic and seed characteristics with selection for reduced palmitic acid content in soybean. Crop Sci 38:297-302.

Rockland, L.B.a.T.M.R. 1981. Legume protein quality. Food Technology 35:79-82.

Ross, A.J., W.R. Fehr, G.A. Welke, and S.R. Cianzio. 2000. Agronomic and seed traits of 1%-linolenate soybean genotypes. Crop Sci 40:383-386.

Saghai Maroof, M.A., N.M. Glover, R.M. Biyashev, G.R. Buss, and E.A. Grabau. 2009. Genetic Basis of the Low-Phytate Trait in the Soybean Line CX1834. Crop Sci 49:69-76.

Saravitz, D.M., D.M. Pharr, and J. Thomas E. Carter. 1987. Galactinol synthase activity and soluble sugars in developing seeds of four soybean genotypes. Plant Physiol 83:185-189.

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

Shi, J., H. Wang, K. Schellin, B. Li, M. Faller, J.M. Stoop, R.B. Meeley, D.S. Ertl, J.P. Ranch, and K. Glassman. 2007. Embryo-specific silencing of a transporter reduces phytic acid content of maize and soybean seeds. Nat Biotech 25:930-937.

Stevenson-Paulik, J., R.J. Bastidas, S.-T. Chiou, R.A. Frye, and J.D. York. 2005. Generation of phytate-free seeds in Arabidopsis through disruption of inositol polyphosphate kinases. Proc Natl Acad Sci (USA) 102:12612-12617.

Tasma, I.M., L.L. Lorenzen, D.E. Green, and R.C. Shoemaker. 2001. Mapping genetic loci for flowering time, maturity, and photoperiod insensitivity in soybean. Mol Breed 8:25-35.

Thorne, J.H. 1980. Kinetics of ^{14}C-photosynthate uptake by developing soybean fruit. Plant Physiol 65:975-979.

USB. 2007. Talksoy.com [Online] (verified 3/28/07).

USDA. 2007a. Soybean Meal: World Supply and Distribution [Online]. Available by United States Department of Agriculture http://www.fas.usda.gov/psdonline/psdReport.aspx (posted March 9, 2007; verified March 26).

USDA. 2007b. WASDE-444. World Agricultural Outlook Board, United States Department of Agriculture.

Kyazma. 2004. MapQTL 5, Software for the mapping of quantitative trait loci in experimental populations. Kyazma, Wageningen, Netherlands.

Kyazma B V. 2006. JoinMap 4, Software for the calculation of genetic linkage maps in experimental populations. Kyazma B V, Wageningen, Netherlands.

Walker, D.R., A.M. Scaboo, V.R. Pantalone, J.R. Wilcox, and 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., and J.F. Cavins. 1987. Gene symbol assigned for linolenic acid mutant in the soybean. Heredity 78:410.

Wilcox, J.R., J.W. Burton, C.J. Rebetzke, and R.F. Wilson. 1994. Transgressive segregation for palmitic acid in seed oil of soybean. Crop Sci 34:1248-1250.

Wilson, R.F. 2004. Seed composition, *In* H. R. Boerma and J. E. Specht, eds. Soybeans: Improvement, Production, and Uses, 3 ed. ASA, CSSA, SSA, Madison, WI.

Zhang, W.K., Y.J. Wang, G.Z. Luo, J.S. Zhang, C.Y. He, X.L. Wu, J.Y. Gai, and S.Y. Chen. 2004. QTL mapping of ten agronomic traits on the soybean (*Glycine max* L. Merr.) genetic map and their association with EST markers. Theor Appl Genet 108:1131-1139.

Chapter 2

Identification of a Candidate Gene Mutation Associated With the Low Stachyose Phenotype in Soybean Line PI200508

ABSTRACT

The recently released Williams 82 whole genome shotgun (WGS) sequence was exploited for candidate gene discovery to investigate the genetic factors underlying a major quantitative trait locus (QTL) contributing to low seed stachyose content in two separate populations derived from soybean line PI200508,. The physical interval containing a low stachyose QTL from PI200508 was identified in the WGS, and screened for areas that could be exploited for linkage mapping purposes. Microsatellite sequences designed in these areas were used to develop new markers, creating tighter linkage with the QTL. Examination of the Williams 82 sequence in this interval and the corresponding glyma0.1b gene model, available through the Phytozome website (http://www.phytozome.org), revealed one candidate gene with significant homology to previously characterized galactosyltransferase genes from *Arabidopsis thaliana*, pea (*Pisum sativum*), and cucumber (*Cucumis sativus*). Sequencing of the annotated coding region for this gene revealed a single unique sequence polymorphism between PI200508, and lines exhibiting wild type expression of stachyose content. The mutant phenotype appears to have arisen from a 3 bp deletion, and is easily distinguished from the wild type allele via a marker presented in this study. This mutation-specific marker explained 88% to 94% of the phenotypic variation for seed stachyose content, and 76% for seed sucrose content, traits that exhibited a strong negative correlation in this study ($p<0.01$).

INTRODUCTION

The transition of soybean [*Glycine max* (L.) Merrill] from an oilseed crop to a provider of renewable energy and as a protein source has increased the importance of breeding specialized soybean lines with unique seed composition traits. Seed composition is an extremely important factor to livestock production, as over 30 million tons of soybean meal are produced annually in the United States, mostly for use in animal feed. Of considerable interest to the animal feed industry is increasing the amount of metabolizable energy in soybean meal, which is largely controlled by the seed carbohydrate profile.

Carbohydrates found in soybean seed include the disaccharide sucrose, as well as the oligosaccharides raffinose and stachyose, which are raffinose family oligosaccharides (RFOs). The bulk of soybean seed carbohydrate content is comprised of sucrose, while RFOs make up the remaining fraction (Hymowitz and Collins, 1974). There is evidence that RFOs, and other non-digestible oligosaccharides can provide a beneficial probiotic function (reviewed by Roberfroid and Slavin, 2000), however, they are also responsible for off flavors, gastric discomfort, and flatulence. Most importantly, they are difficult to digest, and provide little metabolizable energy. Typical carbohydrate levels in soybean seeds are 5-7% sucrose, approximately 1% raffinose, and 3-4% stachyose. Decreasing seed stachyose levels below one percent would increase sucrose levels, and thus, provide a significant increase in metabolizable energy, creating a more efficient feed source.

RFOs have been reported to play roles in cold acclimation, desiccation tolerance, and seed viability (reviewed in Horbowicz and Obendorf, 1994). They are related in that

both carbohydrates are a series of galactosyl groups linked to sucrose via α-(1→6) glycosidic bonds. Raffinose and stachyose are tri- and tetra- saccharides, respectively. Biosynthesis of RFOs is presumed to be carried out by two types of enzymes, galactinol synthase, which generates galactinol from galactose and myo-inositol, and galactosyltransferases (e.g. raffinose and stachyose synthases), which are responsible for incorporating the galactosyl group from galactinol into the oligosaccharide (For a more thorough explanation, see Dey (1985)). Protein species from the latter group of enzymes have been well characterized in legumes such as pea (*Pisum sativum*) (Peterbauer et al., 2002; Peterbauer et al., 2003), adzuki bean (*Vigna angularis*) (Peterbauer et al., 1999), and lentil (*Lens culinaris*) (Hoch et al., 1999), as well as rice (*Oryza sativa*) (Li et al., 2007), cucumber (*Cucumis sativus*) (Osumi et al., 1998), and *Arabidopsis thaliana*, where they are also called seed imbibition proteins such as SIP1 and SIP2 (AT1G55740 and AT3G57520, respectively).

Although work has been done to identify genomic regions which affect sucrose content in soybeans (Kim et al., 2006; Maughan et al., 2000), there is little else publicly reported regarding the genomic locations affecting sucrose, raffinose, or stachyose accumulation in soybean seed. Bentsink et al. (2000) identified a number of seed carbohydrate quantitative trait loci (QTL) in an Arabidopsis RIL population and reported coincidence of QTL for sucrose, raffinose, and stachyose on Arabidopsis chromosome 1. Although it was not determined whether the observed QTL were caused by the same factor, the theory was presented that a single locus could have pleiotropic effects on the biosynthesis of these three carbohydrates. A similar phenomenon has been seen in

soybean, where a mutation at a single locus conferred a phenotype of low stachyose and low phytate (a cyclitol also derived from myo-inositol) (Hitz et al., 2002).

Considerable variation exists for these traits (Bianchi et al., 1984; Hymowitz and Collins, 1974), suggesting that allelic differences in the genes encoding these biosynthetic enzymes does exist, enabling breeders to manipulate these traits. However, Openshaw and Hadley (1981) reported heritability estimates ranging from only 0.14 to 0.69 for total seed sugar content through regression of F_3 progeny means onto individual F_2 parental plants. These low values of genetic variation seem to imply that this trait can be under significant environmental influence. This makes breeding for these traits more difficult, increasing the importance of identifying breeding stock that can contribute a stronger genetic effect. Recently, a low stachyose line, PI200508 was identified by Kerr and Sebastian (2000), with increased seed sucrose content and significantly reduced seed stachyose content. The allele responsible for this phenotype, *stc1a*, which confers a low stachyose phenotype to its offspring, is a recessive allele that likely arose from a mutation causing loss of function in a synthase enzyme involved in the RFO biosynthetic pathway. Loss of function soybean mutants have been exploited for a more desirable phenotype for traits such as low linolenic acid (Chapell and Bilyeu, 2006; Wilcox et al., 1984; Wilcox and Cavins, 1985).

The *stc1a* allele has been introgressed into cultivars used to produce soybean meal, however, many of these lines have lower yield than conventional cultivars. Since it has been shown in one study that *stc1a* has little effect on seedling emergence, yield, maturity, or lodging, (Neus et al., 2005), it is unclear whether decreased performance is a result of the *stc1a* allele, or whether the result of genetic background effects from a

recurrent parent with inferior agronomic characteristics. The objective of this study was to identify the location of the *stc1a* locus through interval mapping, and characterize the genetic cause of the locus through comparative genomics. Additionally, we present a mutation-specific marker that will aid the introgression of the locus into elite soybean cultivars, providing breeders the necessary tools to select from and develop soybean lines with lower seed RFO content.

MATERIALS AND METHODS

Genetic Materials

In this study, two F_2 populations were used: PI87013 × PI200508, and PI243545 × PI200508. In both populations, the low stachyose line PI200508 was crossed with lines expressing the normal stachyose phenotype. Populations generated from the crosses provided 60 and 82 individuals for this study, respectively, and will be referred to as population 1 and 2 in this report. All F_2 individuals were field grown in 2003 at Virginia Tech's Kentland Farm in Blacksburg, Virginia. Tissue collection for DNA extraction was done on F_2 plants, and phenotyping of seed composition traits was performed on F_3 seed, however, because of low F_3 seed inventory, only 44 individuals from population 1, and 58 individuals from population 2, could be phenotyped.

HPLC Analysis of Seed Oligosaccharide Content

F_3 seed used for determination of seed oligosaccharide content was harvested from individual F_2 plants grown during 2003. Extraction and quantitation of soluble seed oligosaccharides occurs simultaneously for sucrose, raffinose, and stachyose, and has been previously reported by Cicek et al. (2006). Seed from each sample were ground in a Cyclone sample mill (UDY Corporation) using a 1.0 mm mesh screen. One gram of ground soybean seed was used per sample for soluble oligosaccharide extraction. Each sample was thoroughly mixed with 10 mL of water (PureLab® Ultra) and shaken on a horizontal shaker at 200 rpm for 15 min. The sample was then centrifuged at 1800 g for

10 min. The soluble proteins from 5 mL supernatant were precipitated with 7 mL acetonitrile (HPLC-grade). The supernatant (1.5 mL) was then centrifuged at 12,200 g for 15 min. An aliquot of the supernatant (1 mL) was evaporated to dryness with compressed air using a Reactitherm heating/evaporation unit set at 98˚C. The resulting dried material was dissolved in 400 µl of water-acetonitrile (35:65, v/v) solution and loaded to the HPLC. Preparation of calibration standards, equipment used, and peak area calculations were as previously reported (Cicek et al., 2006).

DNA Extraction and Genotyping

Young first or second trifoliate leaves of field-grown plants were sampled for DNA extraction. DNA from parental lines and each F_2 individual, was isolated according to the protocol described previously by Saghai Maroof et al. (1984). DNA from each population was genotyped using publicly available simple sequence repeat (SSR, or microsatellite) markers (Cregan et al., 1999), as well as primers designed, for this study, from SSR sequences observed in the Williams 82 whole genome shotgun (WGS) sequence (DoE Joint Genome Institute, 2008). Because the WGS sequence is still in its draft release format (as of April, 2008), known as Glyma0, it has not been assembled into chromosomes, and exists as several hundred supercontigs (or scaffolds). The numerical identification of scaffolds used for primer design in this study will be reported. However, it is worth noting that these identifiers used in the Glyma0 draft release will be replaced with new nomenclature in the next release scheduled for the end of 2008, after which, actual sequences used will be more useful in locating regions of interest in the WGS

sequence. For markers designed in this study, SSR regions were identified with SSRIT (Temnykh et al., 2001), and primers were designed with the web-based Primer3 platform (Rozen and Skaletskyv, 2000). Molecular marker procedures were set up as described by Yu et al. (1994). Typical PCR reactions for SSR detection were denatured at 94ºC for 3 min, followed by 36 cycles at 94º C for 30 sec, 47ºC for 30 sec, 68ºC for 1 min, with a final extension at 68ºC for 8 min. To limit non-specific amplification, annealing temperatures were raised for primers with higher T_m, however, most annealing temperatures fell within the 47-55ºC range. Genotypes were visualized as previously reported (Saghai Maroof et al., 1994).

Mapping and Statistical Analysis

Linkage analysis and map generation was performed via JoinMap 4 (Van-Oiogen, 2006). Linkage groups were constructed using a logarithm of the odds (LOD) threshold of 4.0. Regression mapping of each linkage group was performed using the Kosambi mapping function, using linkages with recombination frequency smaller than 0.4, and a LOD larger than 1.0. Interval mapping for QTL analysis was performed with MapQTL 5 (Van-Oiogen, 2004). A default LOD threshold of 3.0 was used to establish significance. Pearson correlation coefficients between carbohydrate traits were calculated using SPSS statistical software package (release 15.0.0).

Nucleic Acid Sequencing and Comparison

Sequences from candidate model genes identified on Phytozome (http://www.phytozome.org) were used to BLAST against sequences in the GenBank EST database (Benson et al., 2008). Coding sequence annotation was derived from a combination of the Glyma0.1b gene model from the Glyma0 release of the soybean Whole Genome Shotgun (WGS) sequence (DoE Joint Genome Institute, 2008), and manual annotation of start and stop codon locations, which were not identified in the Glyma0.1b annotation for the candidate gene revealed in this study. Primers for sequencing were designed as outlined above to amplify the entire coding sequence of each predicted exon. PCR to generate fragments for sequencing was carried out under the following conditions; denatured at 94ºC for 3 min, followed by 36 cycles at 94º C for 30 sec, 58ºC for 30 sec, 68ºC for 1 min, with a final extension at 68ºC for 8 min. PCR fragments were amplified from PI200508, PI87013, PI243545, and Williams 82, and submitted for sequencing at the Virginia Bioinformatics Institute Core Laboratory Facility on a 3130xl Genetic Analyzer (ABI, Inc.). Sequence analysis and comparison between the four lines sequenced in this study was carried out via the Lasergene 7.1 software package (DNASTAR, Inc.), using SeqMan for contig assembly and comparison, and MegAlign for ClustalV multiple sequence alignment.

RESULTS

Statistical Analysis

Two populations derived from crossing PI200508 with two normal stachyose lines exhibit the bimodal distribution of a recessive trait whose genetic variance is under the control of a single gene, easily allowing separation of wild type and low stachyose individuals (Table 1). The seed stachyose content distribution seen in these populations closely fits ($p>0.1$) a 3:1 segregation ratio in both population 1 ($\chi^2=0.485$), and population 2 ($\chi^2=0.207$), matching the expected segregation ratio for a single recessive gene in a F_2 population. Seed RFO content values were plotted against seed sucrose content (Figure 1). Pearson correlation coefficients, determined through linear regression analysis, showed that seed stachyose content exhibited a strong negative correlation to seed sucrose content in both populations, while being significantly ($P<0.01$), negatively correlated with seed raffinose content in population 2 (Figure 1). Seed raffinose content was positively correlated with seed sucrose content in population 2 as well.

Linkage Analysis and Interval Mapping

Mapping in populations 1 and 2 revealed a major seed stachyose content QTL (LOD>20) in the interval between Sat213 and Satt643. The sequences from these markers were applied to BLAST through the Phytozome website to examine the interval, and identify a candidate gene in the Williams 82 whole genome shotgun (WGS) sequence. Three WGS supercontigs: 157, 158 and 631, were discovered that span this

region. To determine which of these supercontigs contained the PI200508 source of low stachyose, 3 primer pairs were designed from microsatellite sequences on the ends of these supercontigs. Polymorphism was high, with the exception of supercontig 631, on which suitable polymorphism was not found at the observed marker loci. Three of the new markers, 157MAT36, 157BAT21, and 158BAT36 (Table 2), were chosen for mapping because of their distal positions on their respective supercontigs. These markers were then incorporated into the existing linkage map (Figure 2). PI243545 and PI200508 exhibited lower levels of polymorphism in this region, and as a result, this population has fewer markers in its linkage map.

Candidate Gene Discovery

The highest interval mapping LOD scores were seen in the interval between 1157BAT21 and 157MAT36, making supercontig 157 the most likely interval to contain the gene affecting stachyose accumulation. That supercontig, and its complimentary gene model set (Glyma0.1b) were then visually observed through the phytozome Gbrowse interface. The Williams 82 sequence between 157BAT21 and 157MAT36 spans a physical distance of 724 kb, and contains 66 predicted genes, an average of one gene every 11 kb, based on the glyma0.1b gene model (DoE Joint Genome Institute,2008). One of the 66 predicted genes showed sequence homology to galactosyltransferase genes (seed imbibition proteins, raffinose synthases, stachyose synthases) from other plant species, and was considered the lone candidate gene in this region. The WGS gene model was further annotated to identify putative start and stop

codons, which were absent in the Phytozome annotation, and establish a predicted coding sequence for the gene. Final annotation of the putative gene included five exons (Figure 3), spanning a total coding length of 2346 bp. BLAST results for this putative coding sequence include a number of coding sequences from other plant species (Table 3). Queries using a translated sequence from our annotation against NCBI protein databases yielded similar results, with numerous quality hits for the previously mentioned, well-characterized raffinose/stachyose synthases, and galactosyltransferases. Primers designed to amplify each of the five predicted exons for sequencing are shown in Table 2.

Sequence Comparison and Mutation-Specific Marker Design

Candidate gene sequence comparison between three soybean lines exhibiting wild type seed stachyose levels, parental lines, PI87013, PI243545, and one check cultivar, Williams 82, and the low stachyose line PI200508, revealed sequence polymorphism unique to the low stachyose line at only one location, a 3 bp deletion from position 991-993 in the predicted coding sequence (Figure 3). Because a 3 bp difference can be readily observed via PCR fragment length analysis, primers were designed to exploit this deletion as a gene/mutation specific marker for screening purposes, (Forward 5'-GGACTTGAAGGAACAGTTTAGG-3', and Reverse 5'-CGTTACTGACGATCTTATCCAC-3') (Figure 4). The locus represented by this marker, termed *Rsm1* (raffinose synthase mutation) in this study (*rsm1* for the PI200508 allele as it segregates in a recessive manner) was mapped to confirm cosegregation with the low stachyose QTL observed on MLG C2. Interval mapping revealed that LOD

scores for seed stachyose content peak at the *Rsm1* locus (Figure 2), which in turn explains 88% and 94% of the phenotypic variation for the trait, respectively, in population 1 and population 2 (Table 4). In both populations, a major sucrose QTL cosegregated with the *Rsm1* locus, accounting for 76% and 60% of the phenotypic variation for that trait. A significant association between Rsm and seed raffinose content was observed in population 2.

In addition to the 3 bp deletion observed in PI200508, four additional single nucleotide polymorphisms (SNPs) were identified within the sequenced region, two of which would result in an amino acid change in the predicted reading frame (Figure 3). However, because these were unique to neither PI200508 nor wild type lines as a whole, they were not considered as candidates for causal mutation in PI200508.

DISCUSSION

In this study, we were able to detect the location of a major genetic factor contributing to the PI200508 trait of low stachyose, on MLG C2, through QTL analysis, and use this information in conjunction with the WGS to identify a candidate for the causal mutation responsible for this phenotype. The candidate gene containing this mutation appears to be a soybean galactosyltransferase gene, which shares close homology with characterized raffinose and stachyose synthase genes from other plant species. Comparison of this gene's sequence from four soybean lines identified a 3 bp deletion in PI200508, a locus containing a recessive mutation referred to here as *rsm1*, which is easily detected through PCR fragment-length analysis using primer sequences presented in the text.

Because the low stachyose trait from PI200508 segregates following a mendelian, dominant/recessive single gene pattern (Kerr and Sebastian, 2000; Neus et al., 2005), PI200508 was crossed with two normal stachyose lines, and the resultant F_2 progeny were examined. Distribution of seed stachyose content in this study supports reports of single gene inheritance of low stachyose from PI200508 (Table 1). If the PI200508 phenotype is the result of a loss of function mutation in a galactosyltransferase gene as the result of a frame shift, the expected segregation ratio in its F_2 progeny would be 3:1, wild type seed stachyose content to low. Segregation ratios in both populations did not significantly differ from this expected ratio ($P > 0.10$).

The galactosyltransferase candidate gene was initially investigated as a QTL between markers Sat213 and Satt643, on MLG C2. Segregation of low seed stachyose, and results of QTL mapping suggested that a single, recessive gene was responsible for low stachyose in PI200508, which was contained in the interval between public markers Sat213 and Satt634. Because the distance between these markers, 11.8-12.2 cM, spanning at least three supercontigs, was too large to clone and sequence, the interval was shortened through the targeted design of new marker using the newly released Williams 82 WGS (DoE Joint Genome Institute, 2008). This not only shortened the interval, but also revealed the WGS supercontig that contained the QTL. This approach of designing markers and mapping supercontig ends proved to be a rapid and straightforward method of navigating a genome that is fractionated and not assembled into chromosomes yet. Once the low stachyose QTL was mapped to a single supercontig, the gene models in this region were examined for homology to genes known to function in the biosynthesis of oligosaccharides. A candidate gene sequence was identified that shared coding sequence homologies of as much as 84%, 79%, and 76% to stachyose and raffinose synthase mRNA sequences from pea, cucumber, and Arabidopsis (Table 3).

Additionally, after completion of this study, a soybean raffinose synthase from Williams82 was released on GenBank (May, 2008). The predicted coding sequence from Williams82 sequenced in this study shares over 99.9% identity with this accession (EU651888) in aligned regions. Sequence alignment between EU651888 and the PI200508 sequence reported in this study clearly shows the 3 bp deletion in PI200508, which evidence suggests is the cause of the low seed stachyose phenotype.

In the current study, the identification of this candidate gene required only three new markers to be developed for mapping, however, this method of QTL entrapment via fine mapping facilitated by the WGS will be very useful when a gene of unknown or less obvious function is targeted, and could require the development of many more markers to isolate candidate genes. More comprehensive annotation to the WGS will eventually become available; however, the massive undertaking of finding empirical proof of function for each gene, or identifying a characterized ortholog from which to infer function, is an extremely laborious undertaking. Because of this, annotation in some genomic regions will likely remain vague for some time, increasing the usefulness of this approach to candidate gene discovery.

Although *Rsm1* shares closest homology with sequences from genes reported as raffinose synthases, and shares over 99.9% homology to a soybean raffinose synthase, EU651888, our mutation-specific marker showed association with seed raffinose content in only one population (Table 4). Some studies have shown that raffinose synthases do not transfer galactosyl groups to longer oligosaccarides such as raffinose and stachyose (Lehle and Tanner, 1973; Li et al., 2007). With these findings in mind, one theory for the strong effect of this mutation on seed stachyose content could be a severance in the early stages of the RFO biosynthetic pathway proposed in Dey et al. (1985), in which PI200508 has lost the ability to synthesize raffinose on any appreciable level, creating an environment deficient in galactosyl acceptors for enzymes responsible for stachyose synthesis. This sort of interruption in the stepwise biosynthesis of RFOs would presumably inhibit both raffinose and stachyose accumulation, and the negative relationship observed between these traits in this study, would not be expected (Figure 1).

Although the observations presented here suggest that the putative enzyme encoded by *Rsm1* is more involved with stachyose biosynthesis, gene expression and enzymatic activity assays will be necessary to clarify the role of *Rsm1* in seed RFO accumulation, as well as to determine whether the putative gene product exclusively targets raffinose for addition of galactosyl groups. Use of raffinose and stachyose as galactosyl acceptors for the synthesis of verbascose (a pentasaccharide RFO) has been reported in pea (Peterbauer et al., 2003). This is likely because raffinose and stachyose share structural similarities in the region where additional glycosidic bonds would be formed, but it is not beyond reason that a similar mechanism could be at work in the synthesis of stachyose. Additional steps to further characterize this candidate gene and its putative expression are underway.

As shown in this study, the recently released Williams 82 whole genome sequence is a powerful tool that can be applied towards gene discovery and identification, through comparative genomics. It facilitated the discovery of a candidate gene, strongly associated with seed stachyose content in soybean, an important industry trait, in separate mapping populations of only 60 and 82 F_2 individuals. As the progression from QTL to gene traditionally requires fine-mapping in very large populations, this achievement demonstrates the value of incorporating information from the WGS, and comparative genomics as a whole, into traditional mapping studies. In addition to its use towards further gene discovery, the WGS will greatly enhance breeding methods such as marker-assisted selection and detailed QTL analysis through fine mapping.

ACKNOWLEDGMENTS

This study was supported by the United Soybean Board. We thank Katie Haering for assistance with the seed saccharide content data collection.

REFERENCES

Benson, D.A., I. Karsch-Mizrachi, D.J. Lipman, J. Ostell, and D.L. Wheeler. 2008. GenBank. Nucleic Acids Res 36:D25-30.

Bentsink, L., C. Alonso-Blanco, D. Vreugdenhil, K. Tesnier, S.P.C. Groot, and M. Koornneef. 2000. Genetic analysis of seed-soluble oligosaccharides in relation to seed storability of Arabidopsis. Plant Physiol 124:1595-1604.

Bianchi, M.d.L.P., H.C. Silva, and G.L. Braga. 1984. Oligosaccharide content of ten varieties of dark-coated soybeans. J Agr Food Chem 32:355-357.

Chapell, A.S., and K.D. Bilyeu. 2006. A GmFAD3A mutation in the low linolenic acid soybean line C1640. Plant Breed 125:535-536.

Cicek, M.S., P. Chen, M.A.S. Maroof, and G.R. Buss. 2006. Interrelationships among agronomic and seed quality traits in an interspecific soybean recombinant inbred population. Crop Sci 46:1253-1259

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, and J.E. Specht. 1999. An integrated genetic linkage map of the soybean genome. Crop Sci 39:1464-1490.

Dey, P.M. 1985. D-Galactose Containing Oligosaccharides, p. 53-130, *In* P. M. Dey and R. A. Dixon, eds. Biochemistry of Storage Carbohydrates in Green Plants. Academic, New York.

DoE Joint Genome Institute. 2008. Soybean Genome Project [Online]. Available by DoE Joint Genome Institute http://www.phytozome.org (verified February 2008).

Hitz, W.D., T.J. Carlson, P.S. Kerr, and S.A. Sebastian. 2002. Biochemical and molecular characterization of a mutation that confers a decreased raffinosaccharide and phytic acid phenotype on soybean seeds. Plant Physiol 128:650-660.

Hoch, G., T. Peterbauer, and A. Richter. 1999. Purification and Characterization of Stachyose Synthase from Lentil (*Lens culinaris*) Seeds: Galactopinitol and Stachyose Synthesis. Arch Biochem Biophys 366:75-81.

Horbowicz, M., and R.L. Obendorf. 1994. Seed desiccation tolerance and storability: dependence on flatulence-producing oligosaccharides and cyclitols: review and survey. Seed Sci Res 4:385-405.

Hymowitz, T., and F.I. Collins. 1974. Variability of sugar content in seed of *Glycine max* (L.) Merrill and *G. soja* Sieb and Zucc. Agron J 66:239-240.

Kerr, P.S., and S.A. Sebastian. 2000. Soybean products with improved carbohydrate composition and soybean plants. USA Patent 6147193. Date Issued: November 14 2000.

Kim, H.K., S.T. Kang, and K.W. Oh. 2006. Mapping of putative quantitative trait loci controlling the total oligosaccharide and sucrose content of *Glycine max* seeds. J Plant Res 119:533-538.

Lehle, L., and W. Tanner. 1973. The function of *myo*-inositol in the biosynthesis of raffinose. European Journal of Biochemistry 38:103-110.

Li, S., T. Li, W.-D. Kim, M. Kitaoka, S. Yoshida, M. Nakajima, and H. Kobayashi. 2007. Characterization of raffinose synthase from rice (*Oryza sativa* L. var. Nipponbare). Biotechnology Letters 29:635-640.

Maughan, P.J., M.A. Saghai Maroof, and G.R. Buss. 2000. Identification of quantitative trait loci controlling sucrose content in soybean (*Glycine max*). Mol Breed 6:105-111.

Neus, J.D., W.R. Fehr, and S.R. Schnebly. 2005. Agronomic and seed characteristics of soybean with reduced raffinose and stachyose. Crop Sci 45:589-592.

Openshaw, S.J., and H.H. Hadley. 1981. Selection to modify sugar content of soybean seeds. Crop Sci 21:805-808.

Osumi, C., J. Nozaki, and T. Kida. 1998. Raffinose synthetase gene, method of producing raffinose and transgenic plant United States Patent 6166292 Date Issued: 12/26/2000 1998.

Peterbauer, T., J. Mucha, U. Mayer, M. Popp, J. Glössl, and A. Richter. 1999. Stachyose synthesis in seeds of adzuki bean (*Vigna angularis*): molecular cloning and functional expression of stachyose synthase. Plant J 20:509-518.

Peterbauer, T., L. Mach, J. Mucha, and A. Richter. 2002. Functional expression of a cDNA encoding pea (*Pisum sativum* L.) raffinose synthase, partial purification of the enzyme from maturing seeds, and steady-state kinetic analysis of raffinose synthesis. Planta 215:839-846.

Peterbauer, T., J. Mucha, L. Mach, D.A. Jones, C.L. Hedley, and A. Richter. 2003. Enzymatic control of the accumulation of verbascose in pea seeds. Plant Cell Environ 26:1385-1391.

Roberfroid, M., and J. Slavin. 2000. Nondigestible Oligosaccharides. Critical Reviews in Food Science and Nutrition 40:461-480.

Rozen, S., and H.J. Skaletskyv. 2000. Primer3 on the WWW for general users and for biologist programmers, p. 365-386, *In* S. Krawetz and S. Misener, eds. Bioinformatics Methods and Protocols: Methods in Molecular Biology. Humana Press, Totowa, NJ.

Saghai Maroof, M.A., R.M. Biyashev, G.P. Yang, Q. Zhang, and R.W. Allard. 1994. Extraordinarily polymorphic microsatellite DNA in barley: species diversity, chromosomal locations, and population dynamics. Proc Natl Acad Sci (USA) 91:5466-5470.

Saghai Maroof, M.A., K.M. Soliman, R.A. Jorgensen, and R.W. Allard. 1984. Ribosomal DNA spacer-length polymorphisms in barley: Mendelian inheritance, chromosomal location, and population dynamics. Proc Natl Acad Sci (USA) 81:8014-8018.

Temnykh, S., G. DeClerck, A. Lukashova, L. Lipovich, S. Cartinhour, and S. McCouch. 2001. Computational and experimental analysis of microsatellites in rice (*Oryza sativa* L.): frequency, length variation, transposon associations, and genetic marker potential. Genome Res 11:1441-1452.

Kyazma. 2004. MapQTL 5, Software for the mapping of quantitative trait loci in experimental populations. Kyazma, Wageningen, Netherlands.

Kyazma B V. 2006. JoinMap 4, Software for the calculation of genetic linkage maps in experimental populations. Kyazma B V, Wageningen, Netherlands.

Wilcox, J.R., J.F. Cavins, and N.C. Nielsen. 1984. Genetic alteration of soybean oil composition by a chemical mutagen. Journal of American Oil Chem Soc 61:97-100.

Wilcox, J.R., and J.F. Cavins. 1985. Inheritance of low linolenic acid content of the seed oil of a mutant in Glycine max. Theor Appl Genet 71:74-78.

Yu, Y.G., M.A.S. Maroof, G.R. Buss, P.J. Maughan, and S.A. Tolin. 1994. RFLP and microsatellite mapping of a gene for soybean mosaic virus resistance. Phytopathology 84:60-64.

Figure 1. Scatterplot of percent seed oligosaccharide content (Y axis) and percent seed sucrose content (X axis) in population 1 (left) and population 2 (right). Stachyose values are represented with square plots, and raffinose values by circles. Trend line equations are given at the top right of each plot, as are the correlation coefficients (R) between each trait pair.

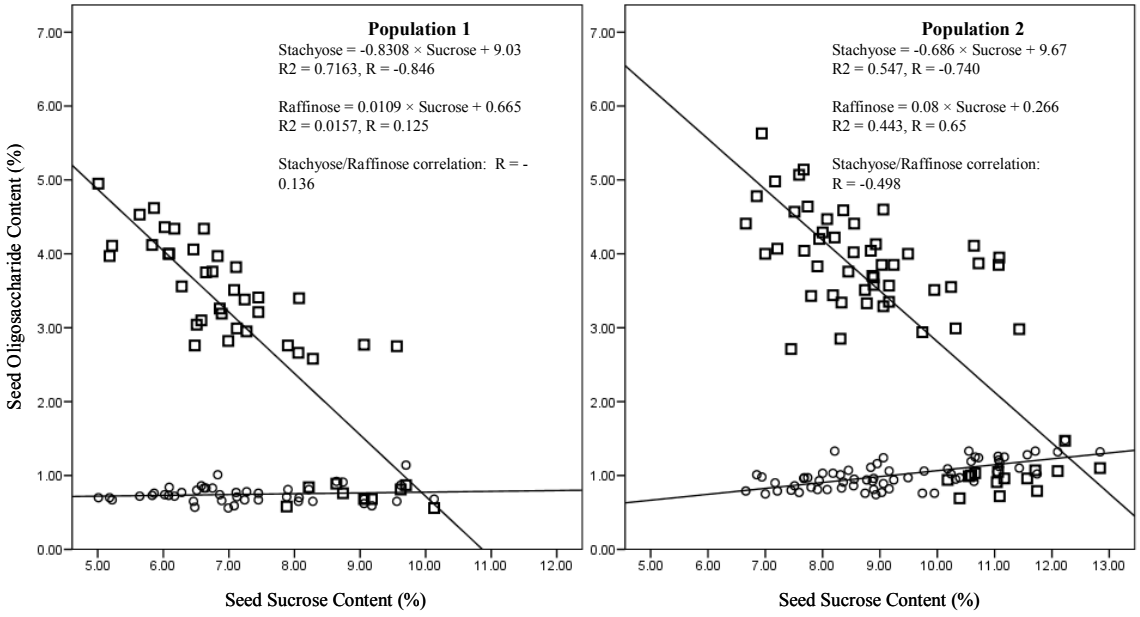

Figure 2. Linkage maps and corresponding LOD curves from the interval between Sat213 and Satt643 on MLG C2 from populations PI87013 × PI200508 (top) and PI243545 × PI200508 (bottom). Values on the left represent distance in centimorgans from the top marker. LOD curves for each trait are represented by (♦) Sucrose, (■) Raffinose, (▲) Stachyose.

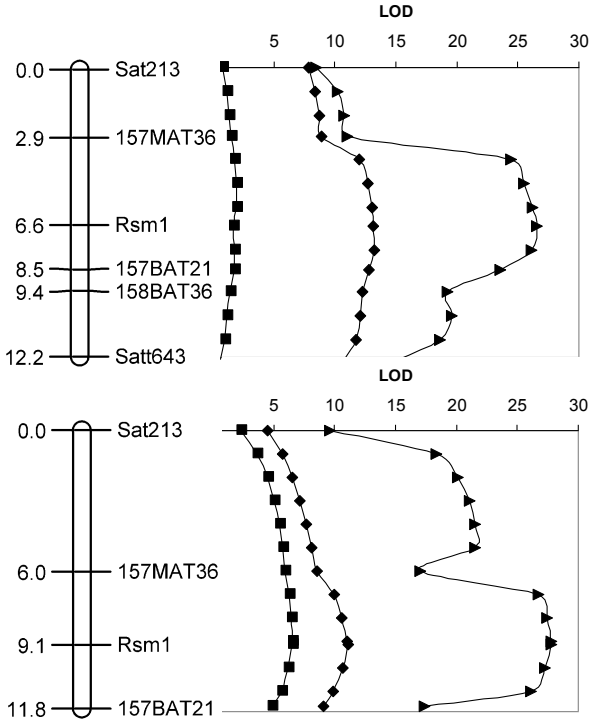

Figure 3. Alignment of 2346 bp wild type Rsm from Williams82, PI87013, PI243545, and PI200508. Sequences are in their predicted reading frame. SNPs that do not alter the translated amino acid sequence are designated with ▼, while polymorphisms that do are designated with ▲. The amino acid change from the Williams82 sequence from this study is shown in the latter case. Amplified region from mutation specific marker, Rsm1, is shown in bold, with horizontal bars representing primer annealing location. Partial translation is also displayed in the region of the 3 bp deletion in PI200508.

```
                                                                                                    85
Williams82  ATGGCTCCAAGCATAAGCAAAACTGTGGAACTAAATTCATTTGGTCTTGTCAACGGTAATTTGCCTTTGTCCATAACCCTAGAAG
PI87013     ATGGCTCCAAGCATAAGCAAAACTGTGGAACTAAATTCATTTGGTCTTGTCAACGGTAATTTGCCTTTGTCCATAACCCTAGAAG
PI243545    ATGGCTCCAAGCATAAGCAAAACTGTGGAACTAAATTCATTTGGTCTTGTCAACGGTAATTTGCCTTTGTCCATAACCCTAGAAG
PI200508    ATGGCTCCAAGCATAAGCAAAACTGTGGAACTAAATTCATTTGGTCTTGTCAACGGTAATTTGCCTTTGTCCATAACCCTAGAAG
                                                                                                    170
            GATCAAATTTCCTCGCCAACGGCCACCCTTTTCTCACGGAAGTTCCCGAAAACATAATAGTCACCCCTTCACCCATCGACGCCAA
            GATCAAATTTCCTCGCCAACGGCCACCCTTTTCTCACGGAAGTTCCCGAAAACATAATAGTCACCCCTTCACCCATCGACGCCAA
            GATCAAATTTCCTCGCCAACGGCCACCCTTTTCTCACGGAAGTTCCCGAAAACATAATAGTCACCCCTTCACCCATCGACGCCAA
            GATCAAATTTCCTCGCCAACGGCCACCCTTTTCTCACGGAAGTTCCCGAAAACATAATAGTCACCCCTTCACCCATCGACGCCAA
                                                                                                    255
            GAGTAGTAAGAACAACGAGGACGACGACGTCGTAGGTTGCTTCGTGGGCTTCCACGCGGACGAGCCCAGAAGCCGACACGTGGCT
            GAGTAGTAAGAACAACGAGGACGACGACGTCGTAGGTTGCTTCGTGGGCTTCCACGCGGACGAGCCCAGAAGCCGACACGTGGCT
            GAGTAGTAAGAACAACGAGGACGACGACGTCGTAGGTTGCTTCGTGGGCTTCCACGCGGACGAGCCCAGAAGCCGACACGTGGCT
            GAGTAGTAAGAACAACGAGGACGACGACGTCGTAGGTTGCTTCGTGGGCTTCCACGCGGACGAGCCCAGAAGCCGACACGTGGCT
                                                                                                    340
            TCCCTGGGGAAGCTCAGAGGAATAAAATTCATGAGCATATTCCGGTTTAAGGTGTGGTGGACCACTCACTGGGTCGGTAGCAACG
            TCCCTGGGGAAGCTCAGAGGAATAAAATTCATGAGCATATTCCGGTTTAAGGTGTGGTGGACCACTCACTGGGTCGGTAGCAACG
            TCCCTGGGGAAGCTCAGAGGAATAAAATTCATGAGCATATTCCGGTTTAAGGTGTGGTGGACCACTCACTGGGTCGGTAGCAACG
            TCCCTGGGGAAGCTCAGAGGAATAAAATTCATGAGCATATTCCGGTTTAAGGTGTGGTGGACCACTCACTGGGTCGGTAGCAACG
                                                                                                    425
            GACACGAACTGGAGCACGAGACACAGATGATGCTTCTCGACAAAAACGACCAGCTCGGACGCCCCTTTGTGTTGATTCTCCCGAT
            GACACGAACTGGAGCACGAGACACAGATGATGCTTCTCGACAAAAACGACCAGCTCGGACGCCCCTTTGTGTTGATTCTCCCGAT
            GACACGAACTGGAGCACGAGACACAGATGATGCTTCTCGACAAAAACGACCAGCTCGGACGCCCCTTTGTGTTGATTCTCCCGAT
            GACACGAACTGGAGCACGAGACACAGATGATGCTTCTCGACAAAAACGACCAGCTCGGACGCCCCTTTGTGTTGATTCTCCCGAT
                                                                                       ▼            510
            CCTCCAAGCCTCGTTCCGAGCCTCCCTGCAACCCGGTTTGGATGATTACGTGGACGTTTGCATGGAGAGCGGGTCGACACGTGTC
            CCTCCAAGCCTCGTTCCGAGCCTCCCTGCAACCCGGTTTGGATGATTACGTGGACGTTTGCATGGAGAGCGGCTCGACACGTGTC
            CCTCCAAGCCTCGTTCCGAGCCTCCCTGCAACCCGGTTTGGATGATTACGTGGACGTTTGCATGGAGAGCGGCTCGACACGTGTC
            CCTCCAAGCCTCGTTCCGAGCCTCCCTGCAACCCGGTTTGGATGATTACGTGGACGTTTGCATGGAGAGCGGCTCGACACGTGTC

         (Cys→Ser)
             ▲                                                                                      595
            TGTGGCTCCAGCTTCGGGAGCTGCTTATACGTCCACGTTGGCCATGACCCGTATCAGTTGCTTAGAGAAGCAACTAAAGTCGTTA
            TGTGGCTCCAGCTTCGGGAGCTGCTTATACGTCCACGTTGGCCATGACCCGTATCAGTTGCTTAGAGAAGCAACTAAAGTCGTTA
            TCTGGCTCCAGCTTCGGGAGCTGCTTATACGTCCACGTTGGCCATGACCCGTATCAGTTGCTTAGAGAAGCAACTAAAGTCGTTA
            TCTGGCTCCAGCTTCGGGAGCTGCTTATACGTCCACGTTGGCCATGACCCGTATCAGTTGCTTAGAGAAGCAACTAAAGTCGTTA
                             ▼                                                                      680
            GGATGCATTTGGGGACGTTCAAGCTTCTCGAGGAGAAAACCGCGCCAGTGATCATAGACAAGTTTGGTTGGTGTACATGGGACGC
            GGATGCATTTGGGGACTTTCAAGCTTCTCGAGGAGAAAACCGCGCCAGTGATCATAGACAAGTTTGGTTGGTGTACATGGGACGC
            GGATGCATTTGGGGACTTTCAAGCTTCTCGAGGAGAAAACCGCGCCAGTGATCATAGACAAGTTTGGTTGGTGTACATGGGACGC
            GGATGCATTTGGGGACTTTCAAGCTTCTCGAGGAGAAAACCGCGCCAGTGATCATAGACAAGTTTGGTTGGTGTACATGGGACGC
                                                                                                    765
            GTTTTACTTGAAGGTGCATCCCTCAGGTGTGTGGGAAGGGGTGAAAGGGTTGGTGGAGGGAGGGTGCCCTCCAGGGATGGTCCTA
            GTTTTACTTGAAGGTGCATCCCTCAGGTGTGTGGGAAGGGGTGAAAGGGTTGGTGGAGGGAGGGTGCCCTCCAGGGATGGTCCTA
            GTTTTACTTGAAGGTGCATCCCTCAGGTGTGTGGGAAGGGGTGAAAGGGTTGGTGGAGGGAGGGTGCCCTCCAGGGATGGTCCTA
```

GTTTTACTTGAAGGTGCATCCCTCAGGTGTGTGGGAAGGGGTGAAAGGGTTGGTGGAGGGAGGGTGCCCTCCAGGGATGGTCCTA

850

ATCGACGACGGGTGGCAAGCCATTTGTCACGACGAGGACCCCATAACGGACCAAGAGGGTATGAAGCGAACCTCCGCAGGGGAGC
ATCGACGACGGGTGGCAAGCCATTTGTCACGACGAGGACCCCATAACGGACCAAGAGGGTATGAAGCGAACCTCCGCAGGGGAGC
ATCGACGACGGGTGGCAAGCCATTTGTCACGACGAGGACCCCATAACGGACCAAGAGGGTATGAAGCGAACCTCCGCAGGGGAGC
ATCGACGACGGGTGGCAAGCCATTTGTCACGACGAGGACCCCATAACGGACCAAGAGGGTATGAAGCGAACCTCCGCAGGGGAGC

935

AAATGCCATGCAGGTTGGTGAAGTTGGAGGAAAATTACAAGTTCAGACAGTATTGTAGTGGAAAGGATTCTGAGAAGGGTATGGG
AAATGCCATGCAGGTTGGTGAAGTTGGAGGAAAATTACAAGTTCAGACAGTATTGTAGTGGAAAGGATTCTGAGAAGGGTATGGG
AAATGCCATGCAGGTTGGTGAAGTTGGAGGAAAATTACAAGTTCAGACAGTATTGTAGTGGAAAGGATTCTGAGAAGGGTATGGG
AAATGCCATGCAGGTTGGTGAAGTTGGAGGAAAATTACAAGTTCAGACAGTATTGTAGTGGAAAGGATTCTGAGAAGGGTATGGG

GlnValTyrValTrpHisAlaLeuCys

1020

TGCCTTTGTTAG**GGACTTGAAGGAACAGTTTAGGAGCGTGGAGCAGGTGTATGTGTGGCACGCGCTTTGTGGGTATTGGGGTGGG**
TGCCTTTGTTAG**GGACTTGAAGGAACAGTTTAGGAGCGTGGAGCAGGTGTATGTGTGGCACGCGCTTTGTGGGTATTGGGGTGGG**
TGCCTTTGTTAG**GGACTTGAAGGAACAGTTTAGGAGCGTGGAGCAGGTGTATGTGTGGCACGCGCTTTGTGGGTATTGGGGTGGG**
TGCCTTTGTTAG**GGACTTGAAGGAACAGTTTAGGAGCGTGGAGCAGGTGTATGTG---CACGCGCTTTGTGGGTATTGGGGTGGG**

1105

GTCAGACCCAAGGTTCCGGGCATGCCCCAGGCTAAGGTTGTCACTCCGAAGCTGTCCAATGGACTAAAATTGACAATGAAGGATT
GTCAGACCCAAGGTTCCGGGCATGCCCCAGGCTAAGGTTGTCACTCCGAAGCTGTCCAATGGACTAAAATTGACAATGAAGGATT
GTCAGACCCAAGGTTCCGGGCATGCCCCAGGCTAAGGTTGTCACTCCGAAGCTGTCCAATGGACTAAAATTGACAATGAAGGATT
GTCAGACCCAAGGTTCCGGGCATGCCCCAGGCTAAGGTTGTCACTCCGAAGCTGTCCAATGGACTAAAATTGACAATGAAGGATT

1190

TAGCGGTGGATAAGATCGTCAGTAACGGAGTTGGACTGGTGCCACCACACCTGGCTCACCTTTTGTACGAGGGGCTCCACTCCCG
TAGCGGTGGATAAGATCGTCAGTAACGGAGTTGGACTGGTGCCACCACACCTGGCTCACCTTTTGTACGAGGGGCTCCACTCCCG
TAGCGGTGGATAAGATCGTCAGTAACGGAGTTGGACTGGTGCCACCACACCTGGCTCACCTTTTGTACGAGGGGCTCCACTCCCG
TAGCGGTGGATAAGATCGTCAGTAACGGAGTTGGACTGGTGCCACCACACCTGGCTCACCTTTTGTACGAGGGGCTCCACTCCCG

1275

TTTGGAATCTGCGGGTATTGACGGTGTTAAGGTTGACGTTATACACTTGCTCGAGATGCTATCCGAGGAATACGGTGGCCGTGTT
TTTGGAATCTGCGGGTATTGACGGTGTTAAGGTTGACGTTATACACTTGCTCGAGATGCTATCCGAGGAATACGGTGGCCGTGTT
TTTGGAATCTGCGGGTATTGACGGTGTTAAGGTTGACGTTATACACTTGCTCGAGATGCTATCCGAGGAATACGGTGGCCGTGTT
TTTGGAATCTGCGGGTATTGACGGTGTTAAGGTTGACGTTATACACTTGCTCGAGATGCTATCCGAGGAATACGGTGGCCGTGTT

1360

GAGCTAGCCAAAGCTTATTACAAAGCGCTCACTGCTTCGGTGAAGAAGCATTTCAAAGGCAATGGGGTCATTGCGAGCATGGAGC
GAGCTAGCCAAAGCTTATTACAAAGCGCTCACTGCTTCGGTGAAGAAGCATTTCAAAGGCAATGGGGTCATTGCGAGCATGGAGC
GAGCTAGCCAAAGCTTATTACAAAGCGCTCACTGCTTCGGTGAAGAAGCATTTCAAAGGCAATGGGGTCATTGCGAGCATGGAGC
GAGCTAGCCAAAGCTTATTACAAAGCGCTCACTGCTTCGGTGAAGAAGCATTTCAAAGGCAATGGGGTCATTGCGAGCATGGAGC

1445

ATTGTAATGACTTCTTTCTCCTTGGTACCGAAGCCATAGCCCTTGGGCGCGTAGGAGATGATTTTTGGTGCACTGATCCCTCTGG
ATTGTAATGACTTCTTTCTCCTTGGTACCGAAGCCATAGCCCTTGGGCGCGTAGGAGATGATTTTTGGTGCACTGATCCCTCTGG
ATTGTAATGACTTCTTTCTCCTTGGTACCGAAGCCATAGCCCTTGGGCGCGTAGGAGATGATTTTTGGTGCACTGATCCCTCTGG
ATTGTAATGACTTCTTTCTCCTTGGTACCGAAGCCATAGCCCTTGGGCGCGTAGGAGATGATTTTTGGTGCACTGATCCCTCTGG

1530

AGATCCAAATGGCACGTATTGGCTCCAAGGGTGTCACATGGTGCACTGTGCCTACAACAGCTTGTGGATGGGGAATTTTATTCAG
AGATCCAAATGGCACGTATTGGCTCCAAGGGTGTCACATGGTGCACTGTGCCTACAACAGCTTGTGGATGGGGAATTTTATTCAG
AGATCCAAATGGCACGTATTGGCTCCAAGGGTGTCACATGGTGCACTGTGCCTACAACAGCTTGTGGATGGGGAATTTTATTCAG
AGATCCAAATGGCACGTATTGGCTCCAAGGGTGTCACATGGTGCACTGTGCCTACAACAGCTTGTGGATGGGGAATTTTATTCAG

(Val⟶Ile)

1615

CCGGATTGGGACATGTTCCAGTCCACTCACCCTTGTGCCGAATTCCATGCAGCCTCTAGGGCCATCTCTGGTGGACCAGTTTACG
CCGGATTGGGACATGTTCCAGTCCACTCACCCTTGTGCCGAATTCCATGCAGCCTCTAGGGCCATCTCTGGTGGACCAATTTACG
CCGGATTGGGACATGTTCCAGTCCACTCACCCTTGTGCCGAATTCCATGCAGCCTCTAGGGCCATCTCTGGTGGACCAATTTACG
CCGGATTGGGACATGTTCCAGTCCACTCACCCTTGTGCCGAATTCCATGCAGCCTCTAGGGCCATCTCTGGTGGACCAATTTACG

1700

TTAGTGATTGTGTTGGAAAGCACAACTTCAAGTTGCTCAAGAGCCTCGCTTTGCCTGATGGGACGATTTTGCGTTGTCAACACTA
TTAGTGATTGTGTTGGAAAGCACAACTTCAAGTTGCTCAAGAGCCTCGCTTTGCCTGATGGGACGATTTTGCGTTGTCAACACTA
TTAGTGATTGTGTTGGAAAGCACAACTTCAAGTTGCTCAAGAGCCTCGCTTTGCCTGATGGGACGATTTTGCGTTGTCAACACTA
TTAGTGATTGTGTTGGAAAGCACAACTTCAAGTTGCTCAAGAGCCTCGCTTTGCCTGATGGGACGATTTTGCGTTGTCAACACTA

1785

TGCACTCCCCACACGAGACTGTTTGTTTGAAGACCCCTTGCATGATGGGAAGACAATGCTCAAAATTTGGAATCTCAACAAATAT
TGCACTCCCCACACGAGACTGTTTGTTTGAAGACCCCTTACATGATGGGAAGACAATGCTCAAAATTTGGAATCTCAACAAATAT
TGCACTCCCCACACGAGACTGTTTGTTTGAAGACCCCTTGCATGATGGGAAGACAATGCTCAAAATTTGGAATCTCAACAAATAT
TGCACTCCCCACACGAGACTGTTTGTTTGAAGACCCCTTGCATGATGGGAAGACAATGCTCAAAATTTGGAATCTCAACAAATAT

1870

```
ACAGGTGTTTTGGGTCTATTTAATTGCCAAGGAGGTGGGTGGTGTCCCGTAACTAGGAGAAACAAGAGTGCCTCTGAATTTTCAC
ACAGGTGTTTTGGGTCTATTTAATTGCCAAGGAGGTGGGTGGTGTCCCGTAACTAGGAGAAACAAGAGTGCCTCTGAATTTTCAC
ACAGGTGTTTTGGGTCTATTTAATTGCCAAGGAGGTGGGTGGTGTCCCGTAACTAGGAGAAACAAGAGTGCCTCTGAATTTTCAC
ACAGGTGTTTTGGGTCTATTTAATTGCCAAGGAGGTGGGTGGTGTCCCGTAACTAGGAGAAACAAGAGTGCCTCTGAATTTTCAC
                                                                                    1955
AAACTGTGACATGCTTAGCGAGTCCTCAAGACATTGAATGGAGCAATGGGAAAAGCCCAATATGCATAAAAGGGATGAATGTGTT
AAACTGTGACATGCTTAGCGAGTCCTCAAGACATTGAATGGAGCAATGGGAAAAGCCCAATATGCATAAAAGGGATGAATGTGTT
AAACTGTGACATGCTTAGCGAGTCCTCAAGACATTGAATGGAGCAATGGGAAAAGCCCAATATGCATAAAAGGGATGAATGTGTT
AAACTGTGACATGCTTAGCGAGTCCTCAAGACATTGAATGGAGCAATGGGAAAAGCCCAATATGCATAAAAGGGATGAATGTGTT
                                                                                    2040
TGCTGTATATTTGTTCAAGGACCACAAACTAAAGCTCATGAAGGCATCAGAGAAATTGGAAGTTTCACTTGAGCCATTTACTTTT
TGCTGTATATTTGTTCAAGGACCACAAACTAAAGCTCATGAAGGCATCAGAGAAATTGGAAGTTTCACTTGAGCCATTTACTTTT
TGCTGTATATTTGTTCAAGGACCACAAACTAAAGCTCATGAAGGCATCAGAGAAATTGGAAGTTTCACTTGAGCCATTTACTTTT
TGCTGTATATTTGTTCAAGGACCACAAACTAAAGCTCATGAAGGCATCAGAGAAATTGGAAGTTTCACTTGAGCCATTTACTTTT
                                                                                    2125
GAGCTATTGACAGTGTCTCCAGTGATTGTGCTGTCAAAAAAGTTAATTCAATTTGCTCCAATTGGATTAGTGAACATGCTTAACA
GAGCTATTGACAGTGTCTCCAGTGATTGTGCTGTCAAAAAAGTTAATTCAATTTGCTCCAATTGGATTAGTGAACATGCTTAACA
GAGCTATTGACAGTGTCTCCAGTGATTGTGCTGTCAAAAAAGTTAATTCAATTTGCTCCAATTGGATTAGTGAACATGCTTAACA
GAGCTATTGACAGTGTCTCCAGTGATTGTGCTGTCAAAAAAGTTAATTCAATTTGCTCCAATTGGATTAGTGAACATGCTTAACA
                                                                                    2210
CTGGTGGTGCCATTCAGTCCATGGAGTTTGACAACCACATAGATGTGGTCAAAATTGGGGTTAGGGGTTGTGGGGAGATGAAGGT
CTGGTGGTGCCATTCAGTCCATGGAGTTTGACAACCACATAGATGTGGTCAAAATTGGGGTTAGGGGTTGTGGGGAGATGAAGGT
CTGGTGGTGCCATTCAGTCCATGGAGTTTGACAACCACATAGATGTGGTCAAAATTGGGGTTAGGGGTTGTGGGGAGATGAAGGT
CTGGTGGTGCCATTCAGTCCATGGAGTTTGACAACCACATAGATGTGGTCAAAATTGGGGTTAGGGGTTGTGGGGAGATGAAGGT
                                                                                    2295
GTTTGCATCAGAGAAACCAGTTAGTTGCAAACTAGATGGGGTAGTTGTAAAATTTGATTATGAGGATAAAATGCTGAGAGTGCAA
GTTTGCATCAGAGAAACCAGTTAGTTGCAAACTAGATGGGGTAGTTGTAAAATTTGATTATGAGGATAAAATGCTGAGAGTGCAA
GTTTGCATCAGAGAAACCAGTTAGTTGCAAACTAGATGGGGTAGTTGTAAAATTTGATTATGAGGATAAAATGCTGAGAGTGCAA
GTTTGCATCAGAGAAACCAGTTAGTTGCAAACTAGATGGGGTAGTTGTAAAATTTGATTATGAGGATAAAATGCTGAGAGTGCAA
                                             2346
GTTCCCTGGCCTAGTGCTTCAAAATTGTCAATGGTTGAGTTTTTATTTTGA
GTTCCCTGGCCTAGTGCTTCAAAATTGTCAATGGTTGAGTTTTTATTTTGA
GTTCCCTGGCCTAGTGCTTCAAAATTGTCAATGGTTGAGTTTTTATTTTGA
GTTCCCTGGCCTAGTGCTTCAAAATTGTCAATGGTTGAGTTTTTATTTTGA
```

Figure 4. Autoradiograph of Rsm PCR fragments amplified from PI200508 and four lines with wild type seed stachyose levels. V71-370 is a *G. max* breeding line with wild type seed RFO content.

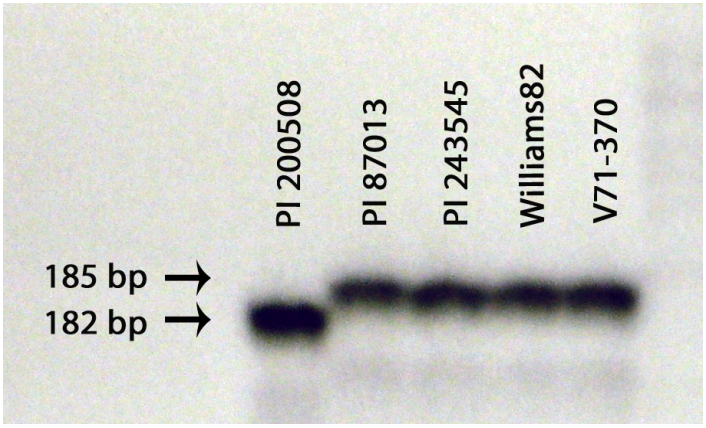

Table 1. Seed stachyose content (%) means and chi-square values in two soybean populations segregating for low seed stachyose. Neither population shows significant deviation from the expected 3:1 wild type:low stachyose ratio (P>0.1)

	Population 1							
	Seed Stachyose Content			N				
	μ (%)	σ	Range	Observed	Expected	χ2	df	P
Wild type	3.57	0.643	2.58-4.95	35	33	0.485	1	0.486
Low stachyose	0.74	0.122	0.56-0.89	9	11			

	Population 2							
	Seed Stachyose Content			N				
	μ (%)	σ	Range	Observed	Expected	χ2	df	P
Wild type	3.96	0.641	2.71-5.63	45	43.5	0.207	1	0.649
Low stachyose	0.98	0.185	0.69-1.47	13	14.5			

Table 2. Primer sequences from markers used for linkage analysis, and amplification of fragments for sequencing.

Linkage Analysis

Primer ID	Sequence	Target†
157MAT36 F	AACTCTTGACAATTTATTCTCTTTTT	(AT)36
157MAT36 R	TTTTTGTGTTAAATGCTAATATACA	
157BAT21 F	ACCATTGATCTCACCAAATAAC	(AT)21
157BAT21 R	GTGACAGCAGTTACATTTTGAA	
158BAT36 F	AACTGGTTGCTTTCATCATAAC	(AT)36
158BAT36 R	TATGTTAGCCAGCATACAAAAA	
Rsm1 F	GGACTTGAAGGAACAGTTTAGG	3 bp deletion in Rsm
Rsm1 R	CGTTACTGACGATCTTATCCAC	

Sequencing

Primer ID	Sequence	Target
GatX1a F	ATAGCAAACCTAAGCACCAAAC	Exon 1
GatX1a R	TCTCCTCGAGAAGCTTGAAC	
GatX1b F	GCTCCAAGCATAAGCAAAAC	Exon 1
GatX1b R	TCTCCTCGAGAAGCTTGAAC	
RsmX5-1 F	AAAACATAATAGTCACCCCTTCA	Exon 1
RsmX5-1 R	GTCCACGTAATCATCCAAACC	
RsmX5-2 F	ATGATGCTTCTCGACAAAAAC	Exon 1
RsmX5-2 R	AAACTCCAACAAAACAAAACAGT	
RsmX4-1 F	AAAATTAGGCAGTAGGTCTAGGAA	Exon 2
RsmX4-1 R	TCATTGTCAATTTTAGTCCATTG	
RsmX4-2 F	TAGGGACTTGAAGGAACAGTTTA	Exon 2
RsmX4-2 R	AAAAAGAAAACAAAAATGCAGAG	
RsmX3-1 F	CACAACGGATTTTTCAATATTTC	Exon 3
RsmX3-1 R	ATTGTTTATTATAAGGGCTGCAA	
RsmX3-2 F	GCTATACACAACGGATTTTTCA	Exon 3
RsmX3-2 R	TTGTTTATTATAAGGGCTGCAAT	
RsmX2 F	CCATCAAATTTTCTAACCCATAA	Exon 4
RsmX2 R	TGATTAGTCCCTAACACTGACAA	
RsmX1-1 F	CAGGGATCAAAATAAAAACTCAA	Exon 5
RsmX1-1 R	ATATTTGTTCAAGGACCACAAAC	
RsmX1-2 F	TGTTAAGCATGTTCACTAATCCA	Exon 5
RsmX1-2 R	CACACCATAATGACATTGGAA	
GatX5a F	TGAGTGAAAATTAGTTGAAACCAT	Exon 5
GatX5a R	ACCTTCAGGGATCAAAATAAAA	
GatX5b F	TCTATTTAATTGCCAAGGAGGT	Exon 5
GatX5b R	AATACGCATGGGAAACTAAAAC	

† Target for primers designed for mapping or screening are microsatellite motifs or in the case of Rsm1, a three bp deletion, while those for sequencing are their targeted predicted exon

Table 3. BLAST output of homologous mRNA sequences containing the descriptors raffinose synthase, stachyose synthase, and galactosyltransferase that share homology to candidate gene. The percent identity between represents similarity in aligned regions.

Organism	Accession	Description	Max ident	E value
A. thaliana	AY062781	Raffinose synthase	76%	2.00E-165
	AY081645	Raffinose synthase	76%	2.00E-165
	NM_123403	Raffinose synthase	76%	2.00E-165
	AK229121	Raffinose synthase	71%	2.00E-44
	NM_116428	Stachyose synthase	71%	2.00E-44
Alonsoa meridionalis (mask flower)	AJ487030	Stachyose synthase	70%	2.00E-56
Cucumis sativus (Cucumber)	DQ641072	Raffinose synthase	79%	6.00E-108
	AF073744	Raffinose synthase	70%	8.00E-164
	EU096496	Stachyose synthase	68%	4.00E-47
	EU096497	Stachyose synthase	68%	4.00E-47
	EU096498	Stachyose synthase	68%	4.00E-47
Pisum sativum (Pea)	AJ311087	Stachyose synthase	84%	8.00E-37
	AJ512932	Stachyose synthase	84%	7.00E-38
	AJ426475	Raffinose synthase	77%	0
Stachys sieboldii (Chinese artichoke)	AJ344091	Stachyose synthase	71%	5.00E-52
Vigna angularis (Azuki bean)	Y19024	Stachyose synthase	69%	1.00E-46

Table 4. Results from interval mapping at the Rsm1 locus on MLG C2. Additive effects indicate the percent change in seed carbohydrate content due to the wild type allele of Rsm1.

	Population 1			Population 2		
	Sucrose	Raffinose	Stachyose	Sucrose	Raffinose	Stachyose
LOD	13.2	2.0 NS	26.6	11.1	6.6	27.8
% Var	75.6	19.7	93.8	59.6	42.3	88.3
Additive	-1.6	-0.02	1.7	-1.9	-0.1	1.8

NS not significant

Chapter 3

QTL mapping of seed traits in an interspecific soybean population

ABSTRACT

QTL analysis was utilized to identify genomic regions contributing to soybean traits of seed sucrose, raffinose, stachyose, and phytate content, as well as seed weight, in an interspecific cross of soybean (*Glycine max* (L.) Merr.). This study utilized a linkage map, generated from a F_8 RIL population, in conjunction with trait data collected from subsequent generations, to perform interval mapping. The linkage map consists of 257 microsatellite, RFLP, and EST-derived markers, spanning all 20 soybean chromosomes/molecular linkage groups, covering a total distance of 1617.3 cM. A total of 25 putative QTL were identified from this population, representing 12 distinct areas on the linkage map. Of these 12 regions, nine are supported in multiple environments, with at least one supported QTL for each trait. As reports of QTL for seed raffinose, stachyose, and phytate content QTL are limited in published literature, this information could be useful to breeders executing a marker-assisted breeding strategy, as well as aiding future studies to identify and characterize specific genes involved in the expression of these traits.

INTRODUCTION

Along with its uses in the processing industry as a protein and oil source, soybean (*Glycine max* (L.) Merr.) is an important food source for humans and livestock. Accumulation of carbohydrates, mainly the saccharides sucrose, raffinose, and stachyose, is an important seed quality trait as it affects the applications of soy in the global market as human and animal feed. Likewise, seed phytate content is an important aspect for nutritional and environmental reasons. Seed weight, as a component of yield, is important as growers attempt to maximize production on available space. Consumption of high levels of raffinose family oligosaccharides (RFOs, notably raffinose and stachyose) has been shown to cause various negative side-effects such as, increased flatulence, diarrhea, and reduced metabolizable energy when consumed by monogastric animals (Coon, 1990; Yamka R. M. et al., 2003). A recently characterized recessive soybean allele, *stc1a,* has been shown to contribute to low raffinose and stachyose, and increased sucrose levels (Kerr and Sebastian, 2000). Neus et al. (2005) found no negative associations between the *stc1a* allele and agronomic or seed quality traits. Additional information regarding the genetic basis of saccharide-related traits could accelerate the discovery of new alleles contributing to RFO production, and development of specialized cultivars.

Physiological trait Quantitative Trait Loci (QTL) have been identified in many crops, including *Zea mays* L. (Hund et al., 2004), *Sorghum bicolor* (L.) Moench (Pereira and Lee, 1995), *Brassica rapa* L. (Song et al., 1995), and *Hordeum vulgare* L. (Li et al., 2005). Soybean reproductive, morphological, and seed quality QTL have also been

studied. Funatsuki et al. (2005) mapped QTL related to seed set and yield under chilling stress, and in another study, investigated QTL controlling pod shattering (2006). QTL affecting protein and oil content have been reported (Brummer et al., 1997; Diers et al., 1992; Mansur et al., 1993; Panthee et al., 2005), as well as QTL for isoflavone content (Primomo et al., 2005). QTL associated with seed sucrose content were identified by Maughan (2000), while seed weight QTL were reported by Maughan et al. (1996) and Mian et al (1996).

This study utilized a recombinant inbred line (RIL) population developed from an interspecific cross to identify putative seed trait QTL. A previous study involving this population investigated relationships between seed carbohydrates and agronomic traits, and revealed no significant genotype × environment interaction for seed sucrose, raffinose, or stachyose content, suggesting that these traits can be effectively modified through selection (Cicek et al., 2006). A study involving QTL analysis of RFLP markers in an F_2 population derived from identical parental lines revealed a number of regions associated with seed sucrose content (Maughan et al., 2000). However, association between molecular markers and raffinose and stachyose content was not investigated, nor were the additional traits included in this study. The objective of this study was to investigate the genetic basis of sucrose, RFO, and phytate accumulation, and seed weight, in soybean seed through the generation of a dense linkage map, and identification of QTL related to these traits.

MATERIALS AND METHODS

Mapping Population

The development and advancement of the RIL population used in this study has been described elsewhere through the F_8 generation (Cicek et al., 2006). In brief, a 308 line F_{12} RIL population was developed from the interspecific cross between the adapted large-seeded *G. max* breeding line, V71-370, with yellow seed coat and high sucrose and a small-seeded *G. soja* plant introduction, PI407162, with black seed coat and low sucrose, will be further denoted as VPRI. The VPRI RILs were developed from individual F_2 plants by a single seed descent method to advance the lines beyond the F_2 generation. The entire RIL population was planted at Whitethorne, Virginia in 2005, 2006, and 2007, representing the F_{10}, $F_{11,}$ and F_{12} generations, respectively. An additional planting of F_{12} RILs occurred at the Eastern Virginia Agricultural Research and Extension Center (EVAREC) in Warsaw, Va, during 2007, making a total of four possible datasets for this population, representing different environments. Because of the viny, prostrate growth habit that most lines inherited from the *G. soja* parent, row spacing was set at 2.3 m. Each entry was planted in a single 3.0 m long row, at 30 seeds per row. Each parent and the check cultivar were entered three times in each block at each location. To continue the single seed descent advancement, individual plants per row were randomly selected as progenitors for the next generation, and trained to a stake to avoid seed contamination from adjacent plants. Seed from these single plants was harvested separately from the remaining plants in each row, and was not used for trait

evaluation. The result was two sets of seed per RIL, with seed harvested from a single plant used to propagate the next generation, and bulk seed used for phenotypic screening. Additionally, tissue was sampled from these single plant selections to extract DNA.

Quantitative Data Analysis of Sugar Content

Seed sugar content data representing three environments were used in this study. Environments 1 and 2 were from a previously collected data (Cicek et al., 2006), representing 1999 plantings in Blacksburg and Warsaw. Environment 3 was Blacksburg, 2005. In each case, bulk harvested seed from each RIL and parental line was ground in a Cyclone sample mill (UDY Corporation) using a 0.08 mm mesh sieve. One gram of ground soybean seed was mixed with 10 mL of double distilled water and shaken on a horizontal shaker at 200 rpm for 15 min. The sample was then centrifuged at 1800 g for 10 min. Soluble proteins from 5 mL of supernatant were precipitated by adding 7 mL acetonitrile (100% HPLC-grade), and the supernatant (1.5 mL) was removed and centrifuged at 12,200 g for 10 min. One milliliter of supernatant was evaporated to dryness with compressed air using a Reactitherm heating/evaporation unit set at 98˚C. The resulting dried material was dissolved in 400 µL water-acetonitrile (35:65, v/v) solution and loaded to the HPLC. Preparation of calibration standards, equipment used, and peak area calculations were as previously reported (Cicek et al., 2006).

Seed Phytate Analysis

Seed phytate content was determined using a modified colorometric method (Gao et al., 2007), and was observed in two environments, Blacksburg 2005 and 2006. For each sample, 0.5 g of ground soybean powder (see sugar extraction above) was placed into a 15 mL Falcon tube. Ten milliliter HCl (0.65 M, 2.4%) was added, vortexed, and shaken overnight (16 hours) at 220 rpm. The suspended solids were separated from the crude extract in a Sorvall RT6000B centrifuge at 3500 rpm, and 10ºC for 20 min. The supernatant was poured into another 15 mL Falcon tube containing 1 g NaCl. The tube was vortexed, and the solution held at 4º C for 1 hr to allow precipitation of soluble protein. Each sample was then centrifuged at 3500 rpm, and 10ºC for 20 min. One milliliter of the solution phase was diluted 25 times with double distilled water in a 50 mL Falcon tube, and 3 mL was transferred to a 15 mL Falcon tube. One milliliter of Wade Reagent (0.03% $FeCl_3.6H_2O$ + 0.3% sulfosalicylic acid) was added, and each tube vortexed and centrifuged at 3500 rpm, at 10ºC for 10 min.

Seven prepared phytate standards were used to generate a standard curve from which sample concentration was determined. Calibration standards of sodium phytate (Sigma, 560 ppm phytate-P) were prepared in 25% NaCl solution. The seven standard concentrations were 0, 1.12, 2.24, 3.36, 5.6, 7.84, and 11.2 parts per million (ppm) phytate-P. These samples were combined with Wade reagent (3 mL standard to 1 mL Wade reagent) and absorbance was read with the samples on a Beakman Coulter DU 640 spectrophotometer at 500 nm.

Observations of Seed Weight

Seed size/weight analysis was determined as 200 seed weight. An automatic seed counter was used to rapidly count 200 seed, which were then weighed. Data was collected from three environments, Blacksburg 2005, Blacksburg 2007, and Warsaw 2007.

Data Analysis

Trait means, distribution, and analysis of variance (ANOVA) were determined using the SPSS statistical software package (release 15.0.0) (SPSS, Inc., 2006). ANOVA was performed using the general linear model Univariate command, with trait values from parental lines and check cultivar removed from the dataset.

DNA Extraction and Genotyping

Young leaves of field-grown 6-week old F_8 plants were cut for DNA extraction. An equal amount of tissue was taken from each single plant and bulked by rows. DNA from each parent and RIL was isolated according to the protocol described previously by Saghai Maroof et al. (1984). Microsatellite markers used in this study include those from previously reported sequences (Cregan et al., 1999), as well as SSR primers designed to amplify EST sequences. SSR procedures were as described previously (Saghai Maroof et al., 1994). Briefly, a 10 µl PCR reaction contained 50 ng genomic DNA, 0.01 mM each primer, 10x reaction buffer, 3 mM $MgCl_2$, 200 mM each dATP, dGTP, and dTTP, 5 mM dCTP. 1.0 U of AmpliTaq DNA polymerase (Perkin-Elmer Cetus, Norwalk, Ct.), and

0.05 μCi [α-^{32}P] dCTP. The reaction was denatured at 94˚C for 3 min, followed by 32 cycles at 94˚C for 1 min, 47˚C for 30 sec, and 68˚C for 1 min, with a final extension step at 68˚C for 7 min in a 96-well Perkin Elmer GeneAmp 9700 thermocycler. PCR products were denatured at 94˚C for 5 min after loading buffer (98% deionized formamide, 10 mM EDTA (pH 8.0), 0.25% xylene cyanol FF, 0.25% bromophenol blue) was added. Five microliters of each sample were loaded on a polyacrylamide denaturing gel and separated at 1500-V constant power in 1x TBE (Tris-Borate-EDTA) running buffer, using a DNA sequencing unit (Model STS-45, IBI, New-Haven, Ct.). Gels were immediately covered with plastic wrap and exposed to X-ray film.

Primer design from sequence information

Primers designed specifically for this study utilize sequence information from amplified sequences putatively located in genomic regions thought to contribute to manifestation of traits of interest. The process involves the use of the NCBI-based cross-database search engine, Entrez (NCBI, 2008), as well as the recently publicized Williams 82 soybean Whole Genome Shotgun (WGS) sequence, available online (www.phytozome.org) (DoE JGI, 2008). Sequence hits from searches for raffinose synthase were considered candidates for primer design, as well as a soybean galactosyltransferase gene homolog identified in the Williams 82 WGS sequence. These sequences were examined for SSR regions using the web based platform, SSRIT (Temnykh et al., 2001), and primers were designed around repeat sequences with the web-based program, Primer3 (Rozen and Skaletskyv, 2000)

Construction of a Genetic Linkage Map and QTL Mapping

A 187 loci linkage map had already been generated from this population for a previous study. To expand on this existing map, 194 molecular markers were screened for polymorphism between parental lines V71-370 and PI407162. Of these, 157 were publicly available SSR markers, reported by Cregan et al. (1999), an additional 31 were designed for this study from raffinose synthase related EST sequences obtained from GenBank (NCBI, 2008), two primer pairs were designed from WGS sequence information from a putative soybean galactosyltransferase gene homolog, and four were chosen from a previous study due to their proximity to a putative soybean galactosyltransferase gene, reported to contribute to soybean seed carbohydrate content (Skoneczka et al., 2009). SSR markers were chosen based on their position reported in the genetic linkage map of the soybean genome reported by Cregan et al. (1999). This previously published map was used to select markers that would most likely extend or fill gaps in our linkage maps. Additionally, the linkage map contained markers reported in Godwin (2003). Classical markers, determinate stem (*dt1*), pod color (*L1*), and pubescence color (*t*) were also scored and incorporated into the existing map. Parental line V71-370 was a large seeded, determinate, tan pod, grey pubescent soybean line, while PI407162 showed indeterminate growth, and had black pods with tawny (brown) pubescence.

Linkage analysis and map generation were performed using JoinMap 4 (Van-Oiogen, 2006). A logarithm of the odds (LOD) threshold of 4.0 was used to establish

significant linkage between markers. Regression mapping of each linkage group was performed using the Kosambi mapping function, using linkages with recombination frequency smaller than 0.4, and a LOD larger than 1.0. The carbohydrate quantitative trait data collected from the RIL population at Blacksburg in 2005 were subjected to QTL analysis, along with two previously gathered carbohydrate datasets (Cicek et al., 2006). These previous data were gathered from $F_{7:8}$ RILs, planted at two locations; Kentland Farm, Blacksburg; and the Eastern Virginia Agricultural Research Center (EVAREC), Warsaw, VA in 1999. Inclusion of these additional datasets was done to validate any observed QTL, and also to exploit the larger, greater density map generated for this study to discover any previously undetected QTL. Interval mapping for QTL analysis of seed sucrose, raffinose, stachyose, and phytate content, as well as 200 seed weight was performed with MapQTL 5 (Van-Oiogen, 2004). A default LOD score of 3.0 was used as the significance threshold for observed QTL.

RESULTS

Trait Distribution and Analysis of Variance

Continuous distribution was observed for seed composition traits and 200 seed weight (Figures 1-5). Examination of seed sucrose, stachyose, and phytate content, and 200 seed weight showed similar distributions across all years, while seed raffinose content appeared to be slightly higher in the Blacksburg 2005 dataset. Although the mean trait values between environments appeared quite similar (Table 1), ANOVA showed significant differences in observed variation among the RILs for each trait ($p < 0.0001$). Environments differed significantly for all traits except phytate (Table 2).

Linkage Analysis

Of 194 markers (157 publicly available, 37 designed for this study) screened for polymorphism between V71-370 and PI407162, 146 were polymorphic between the parents of the RILs. A high rate of polymorphism was observed for the 157 publicly available markers, with polymorphism observed at 139 of the screened loci (Table 3). Lower polymorphism was observed at marker loci developed from EST for this study, with only seven markers of 37 showing polymorphism between parents (Table 4). Of the 146 markers exhibiting polymorphism between parents, 91 were selected for mapping in this study due to their reported linkage group position, or in the case of those generated in this study, for their putative relationship with seed carbohydrate accumulation.

Genotyping data from these 91 markers were combined with the existing marker data for this mapping population, resulting in a 278 marker dataset for linkage analysis.

The resultant linkage map included 257 linked loci, spanning 1617.3 cM across all 20 soybean molecular linkage groups (MLG). Twenty one loci were unlinked. Two markers designed from raffinose synthase-related EST sequences, SRS3-1 and AF126550b, mapped to MLG L and A2. Four markers, selected from a separate study due to their tight linkage to a putative soybean galactosyltransferase gene were found to be polymorphic between V71-370 and PI407162. One of these markers, 157MTA25, mapped to MLG C2. One of two markers designed from WGS sequence information from a potential homolog of that putative galactosyltransferase gene, was polymorphic, SRS35-1, which mapped to MLG A1.

QTL analysis

Interval mapping revealed a number of QTL for seed composition traits, as well as 200 seed weight (Table 5). Seed carbohydrate data collected from F_8 plants, across two locations, showed consistent agreement between QTL observed for seed raffinose, and stachyose content. However, significant sucrose content QTL were only detected in the Blacksburg 1999 and Warsaw 1999 environments. QTL from these two sets of data were observed on MLG A1, A2, F, and M (Figure 6). In the case of the QTL on MLG F and M, the QTL were seen at identical positions across environments, while the QTL on A2 were within 2 cM of each other. The putative QTL on MLG A2 showed the greatest magnitude (LOD = 4.47, 4.15) and explained as much as 10.5% of the phenotypic

variation for seed sucrose content in the 1999 Blacksburg dataset. Interval mapping of seed sucrose content data from 2005 revealed only minor-effect QTL ($2.5 < LOD < 3.0$) on MLG A1 and I. This QTL on MLG A1 is supported by an observed sucrose QTL on A1 in the Blacksburg 1999 dataset, which existed within 1.0 cM of each other on that linkage group.

QTL were observed for seed raffinose content on three linkage groups (Figure 7). Raffinose QTL were observed in multiple locations on MLG L, in the Blacksburg and Warsaw datasets from 1999, located within 7.5 cM of each other. Significant LOD scores were also seen on MLG D1b in the Blacksburg 1999 data, and on MLG C2 in Blacksburg 2005. LOD scores for seed raffinose content ranged from 3.0 on MLG D1b in 1999 Blacksburg, to 6.78 on MLG L from Warsaw 1999. Each observed QTL individually contribute to over 10 percent of the phenotypic variation observed in their respective datasets.

QTL for seed stachyose content were, like sucrose content QTL, highly conserved between the Blacksburg and Warsaw datasets gathered during 1999 (Figure 8). QTL were present on MLG A1, D2, and O in both of these datasets, while only one QTL was observed in Blacksburg 2005, on MLG O. The QTL observed in the first two datasets were highly localized, occurring within 2 cM of each other, while the QTL observed on MLG O in the third dataset, Blacksburg 2005, was detected 5.5 and 7.5 cM upstream of the QTL detected in the Blacksburg 1999 and Warsaw 1999 environments, in a different marker interval.

Three putative QTL for seed phytate content were identified on MLG I and G (Figure 9). The two QTL on I were present in data gathered in Blacksburg during 2005

and 2006, at positions 1 and 23.8 cM on that linkage group. This putative QTL was responsible for 13.5% of the phenotypic variation observed for seed phytate content in Blacksburg 2005. The third putative QTL on MLG G, explaining 6.6% of the phenotypic variation was observed only in Blacksburg 2005.

Four QTL were identified for 200 seed weight on MLG M and H (Figure 10). QTL were detected between 60.0 and 65.0 cM on MLG M in each of the three environments in this study. The maximum LOD score for this region on MLG M was 6.51 in Warsaw 2007, and accounted for 14.2% of the phenotypic variation. This translates to a mean additive effect of 1.18 g in individuals harboring the V71-370 allele at this QTL. A single QTL was observed on MLG H in the Blacksburg 2007 dataset, explaining 7.4% of the phenotypic variance.

Three classical trait loci, pod color (*L1*), pubescence color (*T*), and growth habit, or determinate stem (*Dt1*), were also mapped in this study. *L1* and *Dt1* mapped to MLG L, while *T* mapped to MLG C2 (Figure 7).

DISCUSSION

Putative QTL for traits of soybean seed carbohydrate content, phytate content, and seed weight, were identified in an interspecific RIL population. These 25 putative QTL represent 12 distinct genomic regions on the linkage map generated from an interspecific RIL population. Nine of these QTL-associated regions were supported by significant LOD scores across more than one environment. Raffinose and stachyose seed content QTL have not been previously reported, making those identified here the first. As the development of soybean lines with modified levels of sucrose and RFOs becomes more important to the human and animal food industries, this information could be useful to breeders using marker-assisted selection.

In the present study we detected genomic regions on MLG A1, A2, F and M that putatively harbor QTL for seed sucrose content (Table 5). Sucrose content QTL were compared to previous results from a study involving an F_2 population developed from the same parental lines (Maughan et al., 2000). The genomic region on MLG A2 was reported to have the second greatest effect in that study, explaining 12.1% of the phenotypic variance, compared to the 10.5 and 9.7% seen in this study. Additionally, the seed sucrose content QTL identified in this study on MLG M and F were also detected by Maughan et al. (2000). Molecular marker GMSC514, on MLG M, was reported to be associated with seed sucrose content by Maughan et al. (2000); however, the QTL identified in the present study on MLG M, between Satt463 and Satt551, appeared to be

42 cM away from GMSC514, bringing into question whether both studies were referring to the same genomic region.

As mentioned, soybean QTL for specific RFOs have not been reported previously, making comparison to previous literature difficult. In this study, seed raffinose content QTL were detected on MLG L in multiple environments, and in one of the three environments on two additional linkage groups, C2 and D1b (Figure 7). Seed stachyose content QTL mapped to MLG A1 and D2 in two environments, while QTL observed on MLG O were observed at all three (Figure 8). In a study investigating seed sucrose and total oligosaccharide content, Kim et al. (2006) also identified a region on MLG C2, near the *T* locus, that was responsible for as much as 14.8% of the phenotypic variance for total seed oligosaccharide content. Additionally, they report a region on MLG L, near Satt166, that was significantly associated with seed oligosaccharide content. These QTL were significantly associated with total seed oligosaccharide content and not seed sucrose content, suggesting that the region involved could affect seed RFO accumulation. Although these markers previously reported to be associated with seed RFO content were not used in this study, public mapping efforts such as those by Cregan et al. (1999), indicated that the region identified by Kim et al. (2006) coincided with the putative seed raffinose content QTL identified in this study on MLG C2 and L.

A recent study involving the low stachyose soybean line PI200508 reported a mutation-specific marker, mapping to MLG C2 that accounted for 88.3%-93.8% of the phenotypic variation for seed stachyose content in two populations derived from low stachyose/normal stachyose parental crosses, and also showed strong association with seed raffinose content in one population (Skoneczka et al., 2009). The mutation

exploited for mapping in this case was in a putative soybean galactosyltransferase gene with distinct homology to well characterized raffinose and stachyose synthase genes from other plant species.

To further investigate the accumulation of seed raffinose and stachyose content, a molecular marker used to identify the mutation in PI200508 was mapped in the VPRI population. The markers 157MTA25, 157MAT28, 157BAT37, and BAT21 were designed from the soybean WGS sequence by Skoneczka et al. (2009), and were found to be polymorphic between V71-370 and PI407162. One of these markers was incorporated into the linkage map, where it mapped 8.5 cM away from the seed raffinose content QTL on MLG C2 observed in 2005 Blacksburg. Because of this observed distance, further investigation of this region would be necessary to determine whether the seed raffinose content QTL identified in this study on MLG C2 is related to the presence of a putative galactosyltransferase gene on that chromosome. However, the additional report of QTL for total seed oligosaccharide content on MLG C2 and L, by Kim et al. (2006), suggests that these regions could be of interest to plant breeders. Other seed RFO content QTL identified in this study on MLG A2, D2, O, have no precedence in existing literature.

A homologous sequence from the putative galactosyltransferase gene was investigated for potential involvement in seed raffinose and/or stachyose accumulation. This marker, SRS35-1, showed no significant association with sucrose, raffinose or stachyose content in this study, despite its 8 cM proximity to the sucrose QTL on MLG A1. Additionally, two mapped markers designed from raffinose synthase-related EST sequences, SRS3-1 (MLG L) and AF126550b (MLG A2), showed no significant trait-marker association. This does not imply that the genes represented by these markers

have no role in oligosaccharide biosynthesis and accumulation, but that there was no detectable effect in the experimental population used in this study. It is possible that the detected polymorphism between parents at these marker loci which allowed them to be used in linkage analysis did not confer detectible phenotypic differences between V71-370 and PI470162 alleles in the genes they represent.

Three putative QTL for seed phytate content were identified on MLG I and G in this study (Figure 9). From two possible sets of data for this trait, QTL were identified on MLG I in both. However, peak LOD scores on this linkage group were observed 22.8 cM apart from each other (Figure 9). Having only two datasets to compare, makes it difficult to discern whether these results represent two distinct genetic factors affecting this trait, or whether the QTL analysis simply lacked the resolution necessary to definitively identify a single marker interval which contained the QTL. Inspection of the LOD curves for this linkage group would suggest the latter, as curves generated from both environments exhibited multiple peaks across the front half of MLG I. Further resolution of this region may be obtained through the incorporation of additional datasets for comparison.

Phytate QTL, like those for seed raffinose and stachyose content, are limited in published literature. Using a population developed from a soybean line containing a low phytate mutation, CX1834, low phytate QTL have been identified on MLG L and N (Gao et al., 2008; Walker et al., 2006). A later study putatively identified the genetic factors responsible for these QTL as soybean *MRP* genes (Saghai Maroof et al., 2009). Another class of genes involved in phytate biosynthesis, the *MIPS* class of genes, were mapped in the study by Maroof et al. (2009). Primers designed from the *MIPS2* sequence mapped to

MLG G; however, the mapped location of that gene was on the proximal end of the linkage group, while the QTL observed in this study was in the middle, making it unlikely that they represent the same entity.

Two hundred seed weight QTL were identified on MLG M in all three environments, while an additional one was identified on MLG H in the 2007 Blacksburg dataset (Table 5). Because the peak LOD scores on MLG M, from each of the three datasets, existed within 5.0 cM of each other, they likely represent the same genetic entity. Among other QTL, Teng et al. (2008) identified a QTL on MLG M between Satt150 and Satt220 that accounted for as much as 25.23% of the phenotypic variance for 100 seed weight. Although these markers were not mapped in the RIL population investigated here, they have been reported to map in the vicinity of the QTL observed in this study on MLG M. Additional studies have also reported QTL for seed size on this linkage group (Chapman et al., 2003; Csanadi et al., 2001; Hyten et al., 2004; Liu et al., 2007). Another study involving three experimental populations across three locations identified QTL for seed weight on MLG M, as well as on MLG H (Hoeck et al., 2003).

Although the detection of QTL on the same linkage groups in multiple environments, representing multiple generations of this RIL population, would suggest the presence of stable QTL that could be integrated into a breeding program, there seems to be significant environmental influence on the expression of seed traits as many of these QTL were not identified in all of the observed environments. It is noteworthy that the most recent seed sucrose content dataset, for example, collected in Blacksburg during 2005 from the F_{10} generation, yielded no sucrose content QTL with LOD scores greater than the designated 3.0 threshold for significance. Two minor-effect QTL ($2.5 < LOD <$

3.0) were detected in this dataset. One explanation for this could be the effect of environment on quantitative traits such as seed sucrose content. Collectively, the QTL detected in the 1999 Blacksburg and Warsaw datasets accounted for only 34.0% and 24.8% of the phenotypic variance, leaving the majority of the variance for this trait unexplained. This unexplained variance could be contributed by undetected QTL, but also environment. It is possible that environmental effect was responsible for masking QTL in the 2005 Blacksburg dataset. An additional contributor to variation observed in later datasets, could be gene flow between RILs in this population between the F_8 generation, from which DNA was isolated for linkage analysis, and the F_{10} generation. Although outcrossing rates in cultivated soybean have been reported to be below 3%, (Garber and Odland, 1926; Kiang et al., 1992), observed multi-locus outcrossing rates from 9.3- 19% were reported in one study involving four wild populations of *G. soja* (Fujita et al., 1997). Outcrossing rates within the VPRI RIL population are unknown, raising the possibility that increased hybridization rates could occur in this population due to its *G. max*/*G. soja* lineage. In this study, seed sucrose data from F_{10} plants was subjected to QTL analysis with marker data generated from DNA isolated from F_8 plants. The two generations separating the sources of phenotypic and genetic data could have allowed the transfer of genetic material between RILs, if heightened outcrossing rates exist in this population. Transfer of this kind, could be limiting the detection of QTL in later RIL generations.

This study identified 25 putative QTL for traits of soybean seed carbohydrate content, phytate content, and seed weight in an interspecific RIL population, located on 12 MLG on the linkage map generated in this study. Nine of these QTL were supported

by significant trait-marker associations in multiple environments. Since there is no precedence for raffinose and stachyose seed content-specific QTL in the present literature, those reported here are among the first. Further investigation of these QTL would be needed to determine whether those detected in only one environment are false positives, or merely subject to environmental influence. It is likely that the QTL reported in this study are not all of the QTL for these traits. Expansion of our existing linkage map would be needed for more comprehensive examination of the genetic basis for these traits. This would facilitate the investigation of genes on more distal chromosomal regions further identifying and clarifying QTL affecting traits valuable to plant breeders.

REFERENCES

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

Chapman, A., V.R. Pantalone, A. Ustun, F.L. Allen, D. Landau-Ellis, R.N. Trigiano, and P.M. Gresshoff. 2003. Quantitative trait loci for agronomic and seed quality traits in an F_2 and $F_{4:6}$ soybean population. Euphytica 129:387-393.

Cicek, M.S., P. Chen, M.A.S. Maroof, and G.R. Buss. 2006. Interrelationships among agronomic and seed quality traits in an interspecific soybean recombinant inbred population. Crop Sci 46:1253-1259

Coon, C.N., K. L. Leske, O. Akavanichan, T. K. Cheng. 1990. Effect of oligosaccharide-free soybean meal on true metabolizable energy and fiber digestion in adult roosters. Poul Sci 69:787-793.

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, and J.E. Specht. 1999. An integrated genetic linkage map of the soybean genome. Crop Sci 39:1464-1490.

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

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

Doe JGI. 2008. Soybean Genome Project [Online]. Available by DoE Joint Genome Institute http://www.phytozome.org (verified February 2008).

Fujita, R., M. Ohara, K. Okazaki, and Y. Shimamoto. 1997. The extent of natural cross-pollination in wild soybean (*Glycine soja*). J Hered 88:124-128.

Funatsuki, H., K. Kawaguchi, S. Matsuba, Y. Sato, and M. Ishimoto. 2005. Mapping of QTL associated with chilling tolerance during reproductive growth in soybean. Theor Appl Genet 111:851-861.

Funatsuki, H., M. Ishimoto, H. Tsuji, K. Kawaguchi, M. Hajika, and K. Fujino. 2006. Simple sequence repeat markers linked to a major QTL controlling pod shattering in soybean. Plant Breed 125:195-197.

Gao, Y., C. Shang, M.A.S. Maroof, R.M. Biyashev, E.A. Grabau, P. Kwanyuen, J.W. Burton, and G.R. Buss. 2007. A modified colorimetric method for phytic acid analysis in soybean. Crop Sci 47:1797-1803.

Gao, Y., R.M. Biyashev, M.A.S. Maroof, N.M. Glover, D.M. Tucker, and G.R. Buss. 2008. Validation of Low-Phytate QTLs and evaluation of seedling emergence of Low-Phytate Soybeans. Crop Sci 48:1355-1364.

Garber, R.J., and T.E. Odland. 1926. Natural crossing in soybeans. Agron J 18:967-970.

Godwin, M.J. 2003. Molecular Mapping of Disease-Related Expressed Sequence Tags and Resistance Gene Analogues in Soybean, Virginia Polytechnic Institute and State University.

Hoeck, J.A., W.R. Fehr, R.C. Shoemaker, G.A. Welke, S.L. Johnson, and S.R. Cianzio. 2003. Molecular Marker Analysis of seed size in soybean. Crop Sci 43:68-74.
Hund, A., Y. Fracheboud, A. Soldati, E. Frascaroli, S. Salvi, and P. Stamp. 2004. QTL controlling root and shoot traits of maize seedlings under cold stress. Theor Appl Genet 109:618-629.
Hyten, D.L., V.R. Pantalone, C.E. Sams, A.M. Saxton, D. Landau-Ellis, T.R. Stefaniak, and M.E. Schmidt. 2004. Seed quality QTL in a prominent soybean population. Theor Appl Genet 109:552-561.
Kerr, P.S., and S.A. Sebastian. 2000. Soybean products with improved carbohydrate composition and soybean plants. USA Patent 6147193. Date Issued: November 14 2000.
Kiang, Y.T., Y.C. Chiang, and N. Kaizuma. 1992. Genetic Diversity in Natural Populations of Wild Soybean in Iwate Prefecture, Japan. J Hered 83:325-329.
Kim, H.K., S.T. Kang, and K.W. Oh. 2006. Mapping of putative quantitative trait loci controlling the total oligosaccharide and sucrose content of *Glycine max* seeds. J Plant Res 119:533-538.
Li, J.Z., X.Q. Huang, F. Heinrichs, and M.W. Ganal. 2005. Analysis of QTLs for yield, yield components, and malting quality in a BC3-DH population of spring barley. Theor Appl Genet 110:356-363.
Liu, B., T. Fujita, Z.-H. Yan, S. Sakamoto, D. Xu, and J. Abe. 2007. QTL mapping of domestication-related traits in soybean (*Glycine max*). Ann Bot-London 100:1027-1038.
Mansur, L.M., K.G. Lark, H. Kross, and A. Oliveria. 1993. Interval mapping of quantitative trait loci for reproductive, morphological, and seed traits of soybean (*Glycine max* L.). Theor Appl Genet 86:907-913.
Maughan, P.J., M.A. Saghai Maroof, and G.R. Buss. 1996. Molecular-marker analysis of seed-weight: genomic locations, gene action, and evidence for orthologous evolution among three legume species. Theor Appl Genet 93:574-579.
Maughan, P.J., M.A. Saghai Maroof, and G.R. Buss. 2000. Identification of quantitative trait loci controlling sucrose content in soybean (*Glycine max*). Mol Breed 6:105-111.
Mian, M.A.R., M.A. Bailey, J.P. Tamulonis, E.R. Shipe, J. T. E. Carter, W.A. Parrott, D.A. Ashley, R.S. Hussey, and H.R. Boerma. 1996. Molecular markers associated with seed weight in two soybean populations. Theor Appl Genet 93:1011-1016.
NCBI. 2008. NCBI Homepage [Online] http://www.ncbi.nlm.nih.gov (verified February 2008).
Neus, J.D., W.R. Fehr, and S.R. Schnebly. 2005. Agronomic and seed characteristics of soybean with reduced raffinose and stachyose. Crop Sci 45:589-592.
Panthee, D.R., V.R. Pantalone, D.R. West, A.M. Saxton, and C.E. Sams. 2005. Quantitative trait loci for seed protein and oil concentration, and seed size in soybean. Crop Sci 45:2015-2022.
Pereira, M.G., and M. Lee. 1995. Identification of genomic regions affecting plant height in sorghum and maize. Theor Appl Genet 90:380-388.
Primomo, V.S., V. Poysa, G.R. Ablett, C. Jackson, M. Gijzen, and I. Rajcan. 2005. Mapping QTL for individual and total isoflavone content in soybean seeds. Crop Sci 45:2454-2464.

Rozen, S., and H.J. Skaletskyv. 2000. Primer3 on the WWW for general users and for biologist programmers, p. 365-386, *In* S. Krawetz and S. Misener, eds. Bioinformatics Methods and Protocols: Methods in Molecular Biology. Humana Press, Totowa, NJ.

Saghai Maroof, M.A., R.M. Biyashev, G.P. Yang, Q. Zhang, and R.W. Allard. 1994. Extraordinarily polymorphic microsatellite DNA in barley: species diversity, chromosomal locations, and population dynamics. Proc Natl Acad Sci (USA) 91:5466-5470.

Saghai Maroof, M.A., N.M. Glover, R.M. Biyashev, G.R. Buss, and E.A. Grabau. 2009. Genetic Basis of the Low-Phytate Trait in the Soybean Line CX1834. Crop Sci 49:69-76.

Saghai Maroof, M.A., K.M. Soliman, R.A. Jorgensen, and R.W. Allard. 1984. Ribosomal DNA spacer-length polymorphisms in barley: Mendelian inheritance, chromosomal location, and population dynamics. Proc Natl Acad Sci (USA) 81:8014-8018.

Skoneczka, J.A., M.A. Saghai Maroof, C. Shang, and G.R. Buss. 2009. Identification of candidate gene mutation associated with low stachyose phenotype in soybean line PI200508. Crop Sci 49:247-255.

Song, K., M.K. Slocum, and T.C. Osborn. 1995. Molecular-marker analysis of genes controlling morphological variation in *Brassica rapa* (syn. Campestris). Theor Appl Genet 90:1-10.

Temnykh, S., G. DeClerck, A. Lukashova, L. Lipovich, S. Cartinhour, and S. McCouch. 2001. Computational and experimental analysis of microsatellites in rice (*Oryza sativa* L.): frequency, length variation, transposon associations, and genetic marker potential. Genome research 11:1441-1452.

Teng, W., Y. Han, Y. Du, D. Sun, Z. Zhang, L. Qiu, G. Sun, and W. Li. 2008. QTL analyses of seed weight during the development of soybean (*Glycine max* L. Merr.). Heredity 102:372-380.

Kyazma. 2004. MapQTL 5, Software for the mapping of quantitative trait loci in experimental populations. Kyazma, Wageningen, Netherlands.

Kyazma B V. 2006. JoinMap 4, Software for the calculation of genetic linkage maps in experimental populations. Kyazma B V, Wageningen, Netherlands.

Walker, D.R., A.M. Scaboo, V.R. Pantalone, J.R. Wilcox, and H.R. Boerma. 2006. Genetic Mapping of Loci Associated with Seed Phytic Acid Content in CX1834-1-2 Soybean. Crop Sci 46:390-397.

Yamka R. M., U. Jamikorn, A. D. True, and D.L. Harmon. 2003. Evaluation of soybean meal in canine foods. Anim Feed Sci Tech 109:121-132.

Figure 1. Frequency distribution of seed sucrose content in three sets of data for the 308 member VPRI RIL population. Means for each parent, planted in three plots within each year, are given along with the standard deviation.

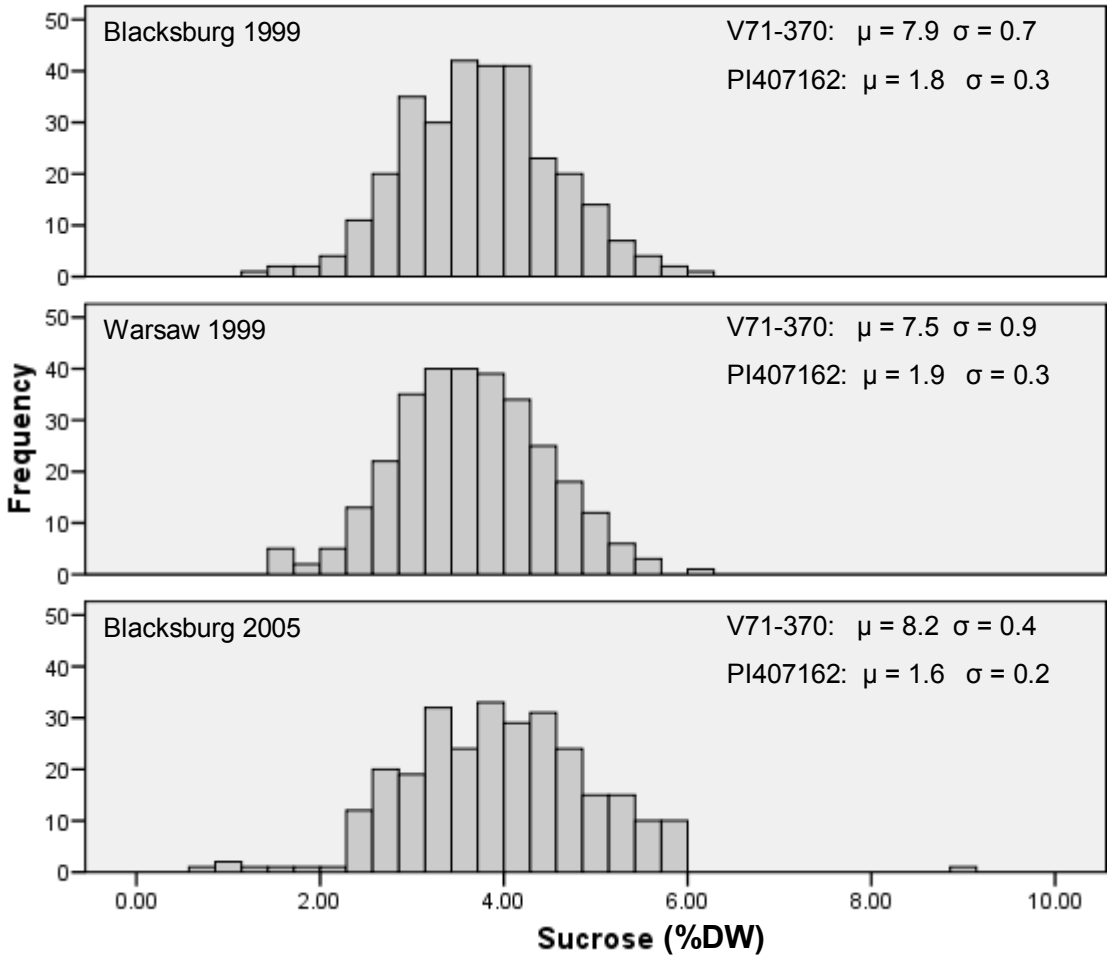

Figure 2. Frequency distribution of seed raffinose content in three sets of data for the 308 member VPRI RIL population. Means for each parent, planted in three plots within each year, are given along with the standard deviation.

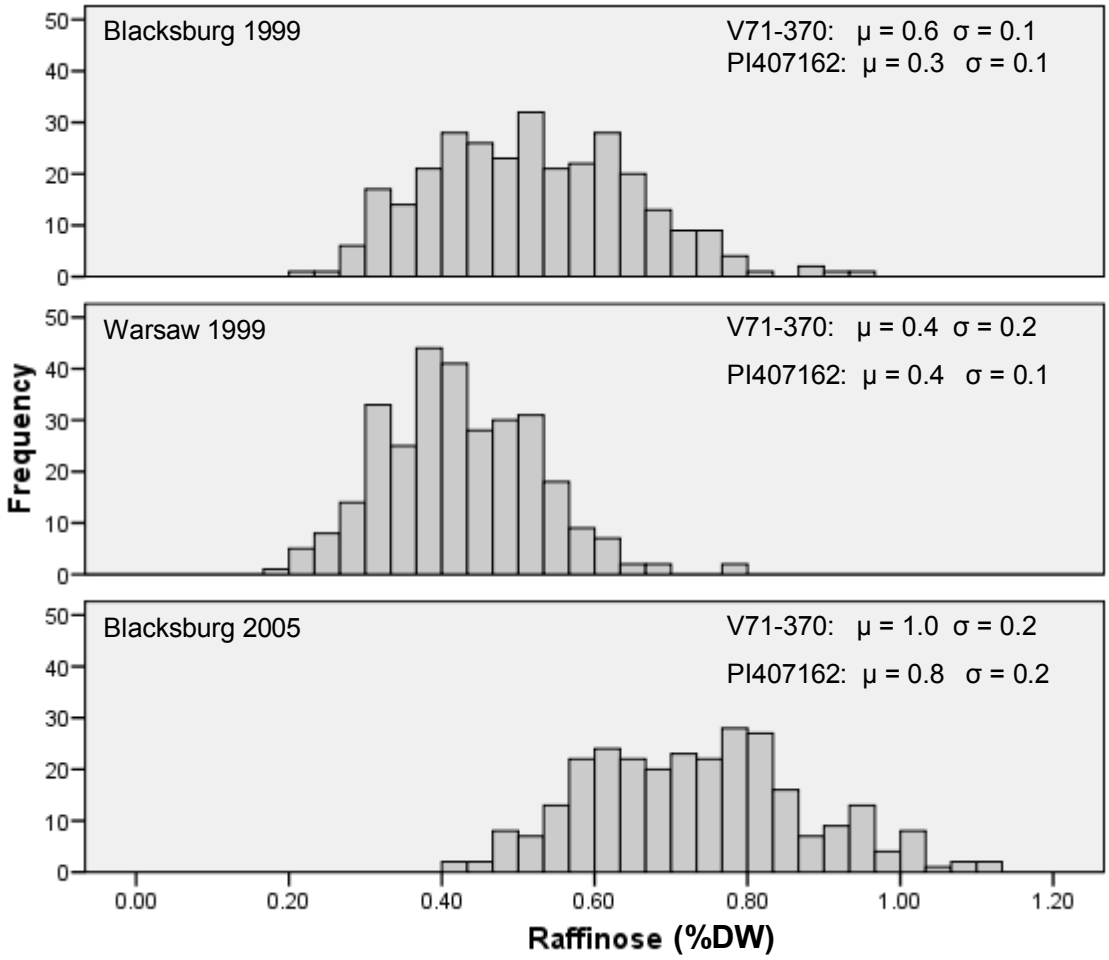

Figure 3. Frequency distribution of seed stachyose content in three sets of data for the 308 member VPRI RIL population. Means for each parent, planted in three plots within each year, are given along with the standard deviation.

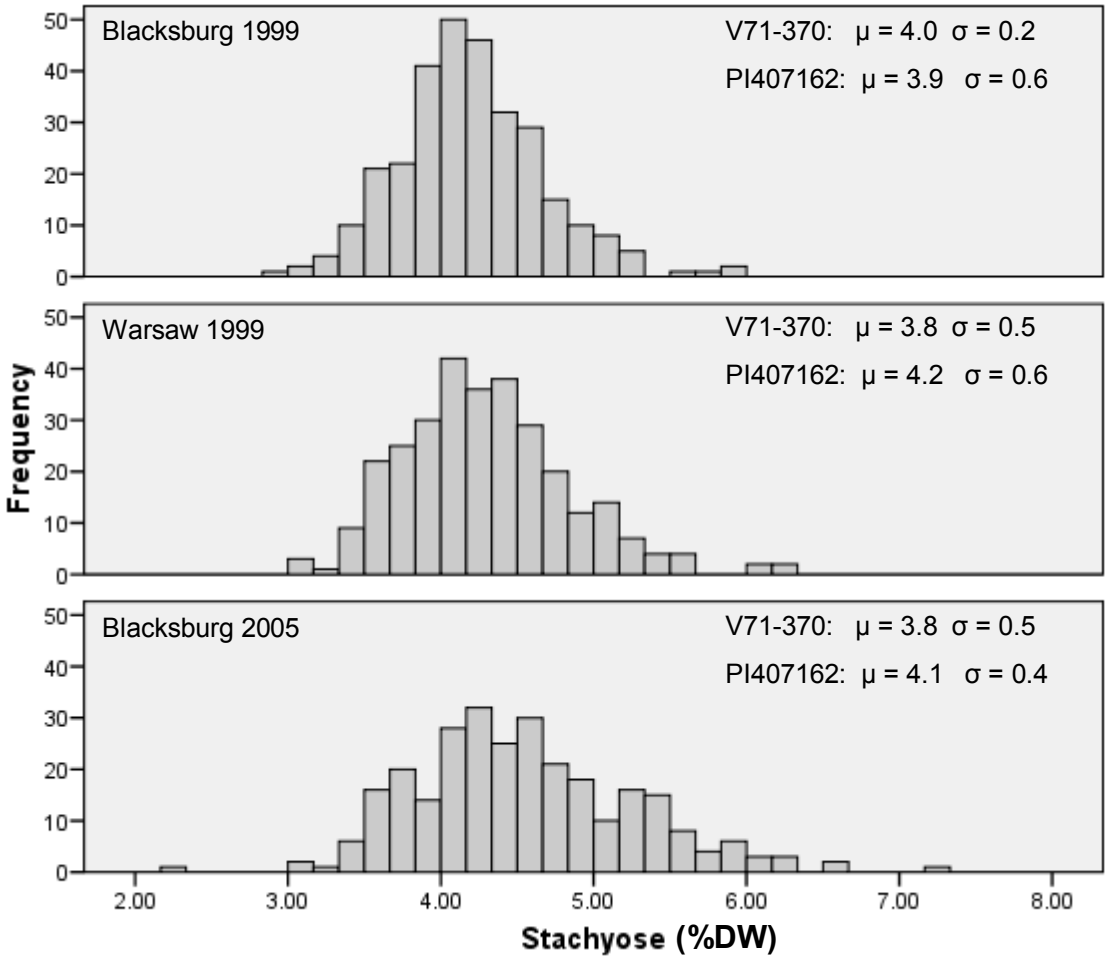

Figure 4. Frequency distribution of seed phytate content in two sets of data for the 308 member VPRI RIL population. Means for each parent, planted in three plots within each year, are given along with the standard deviation.

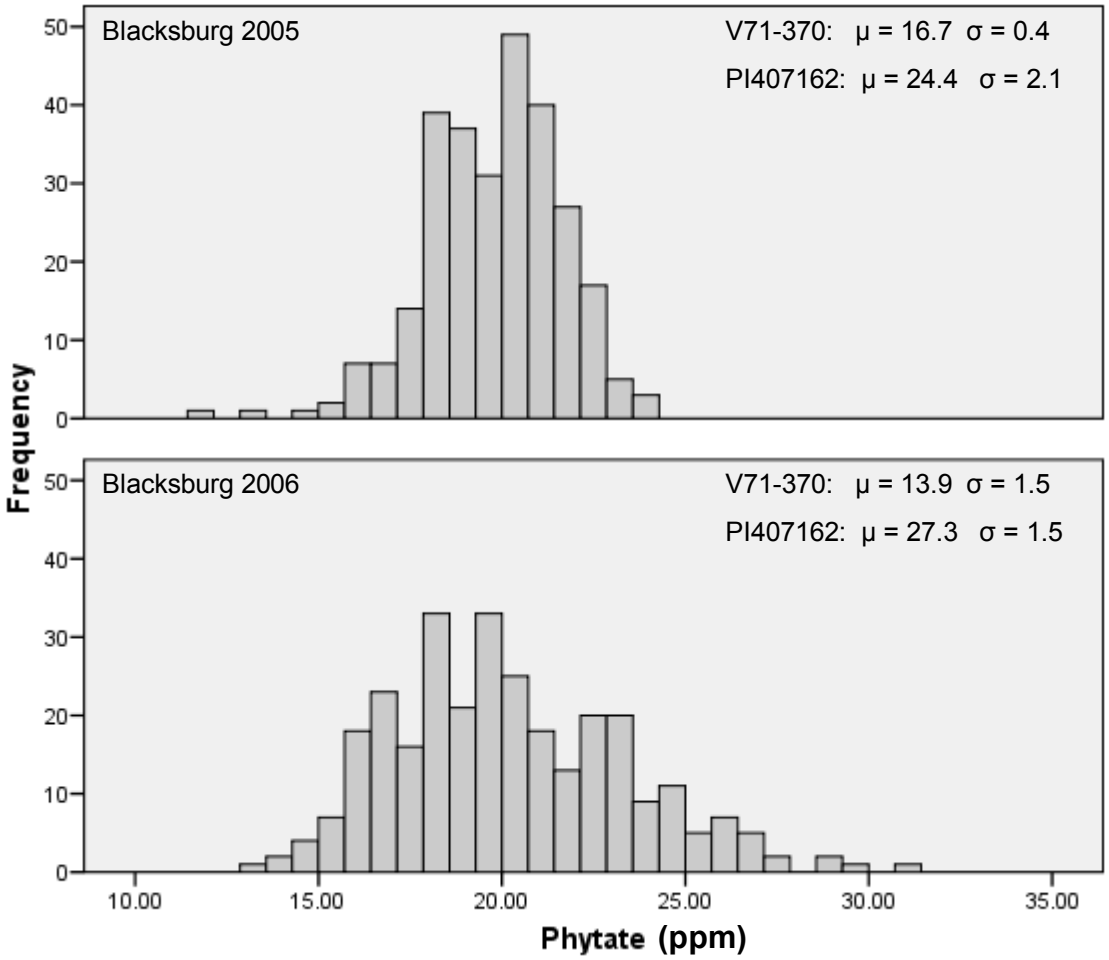

Figure 5. Frequency distribution of 200 seed weight in three sets of data for the 308 member VPRI RIL population. Means for each parent, planted in three plots within each year, are given along with the standard deviation.

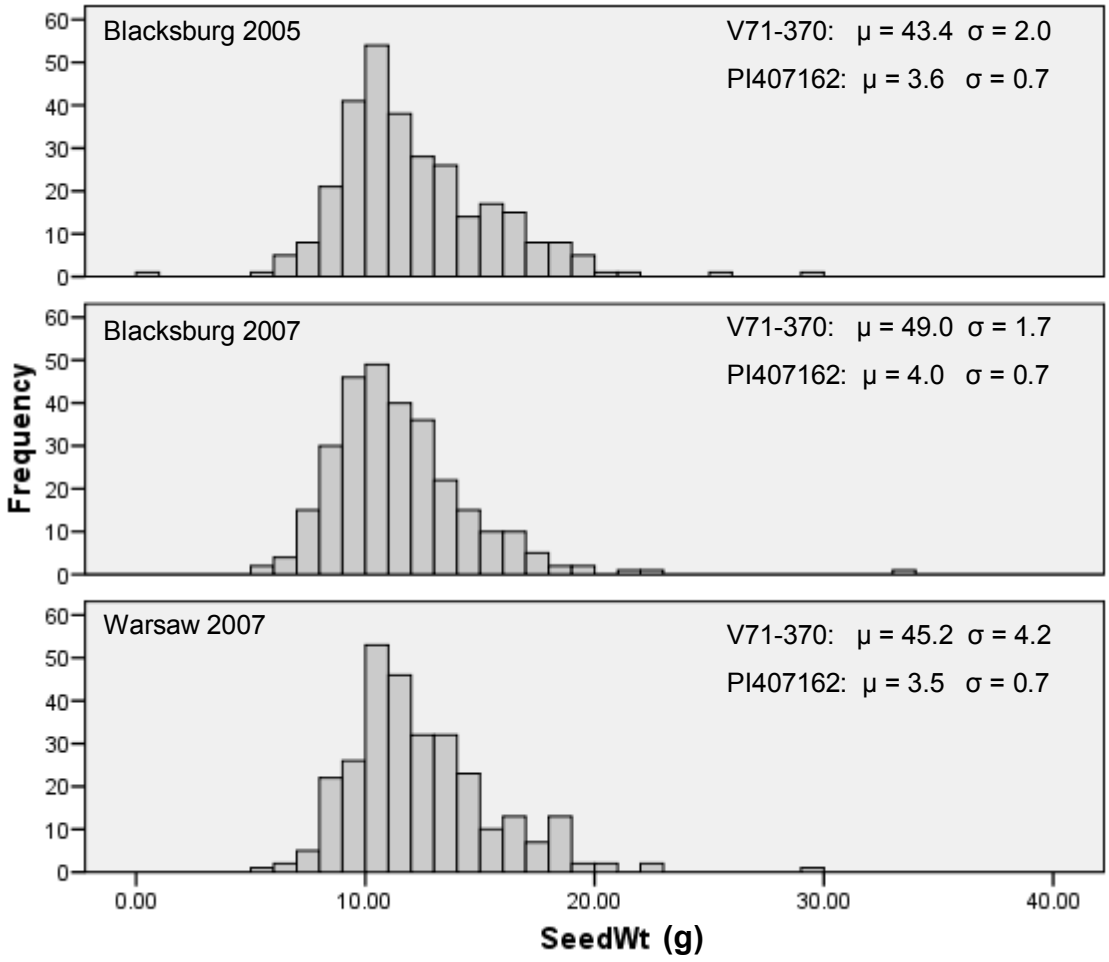

Figure 6. Linkage maps for molecular linkage groups (MLG) containing significant QTL for seed sucrose content in the VPRI RIL population. Distance in cM from the top is given to the left of each marker. LOD curves are aligned with their corresponding linkage map.

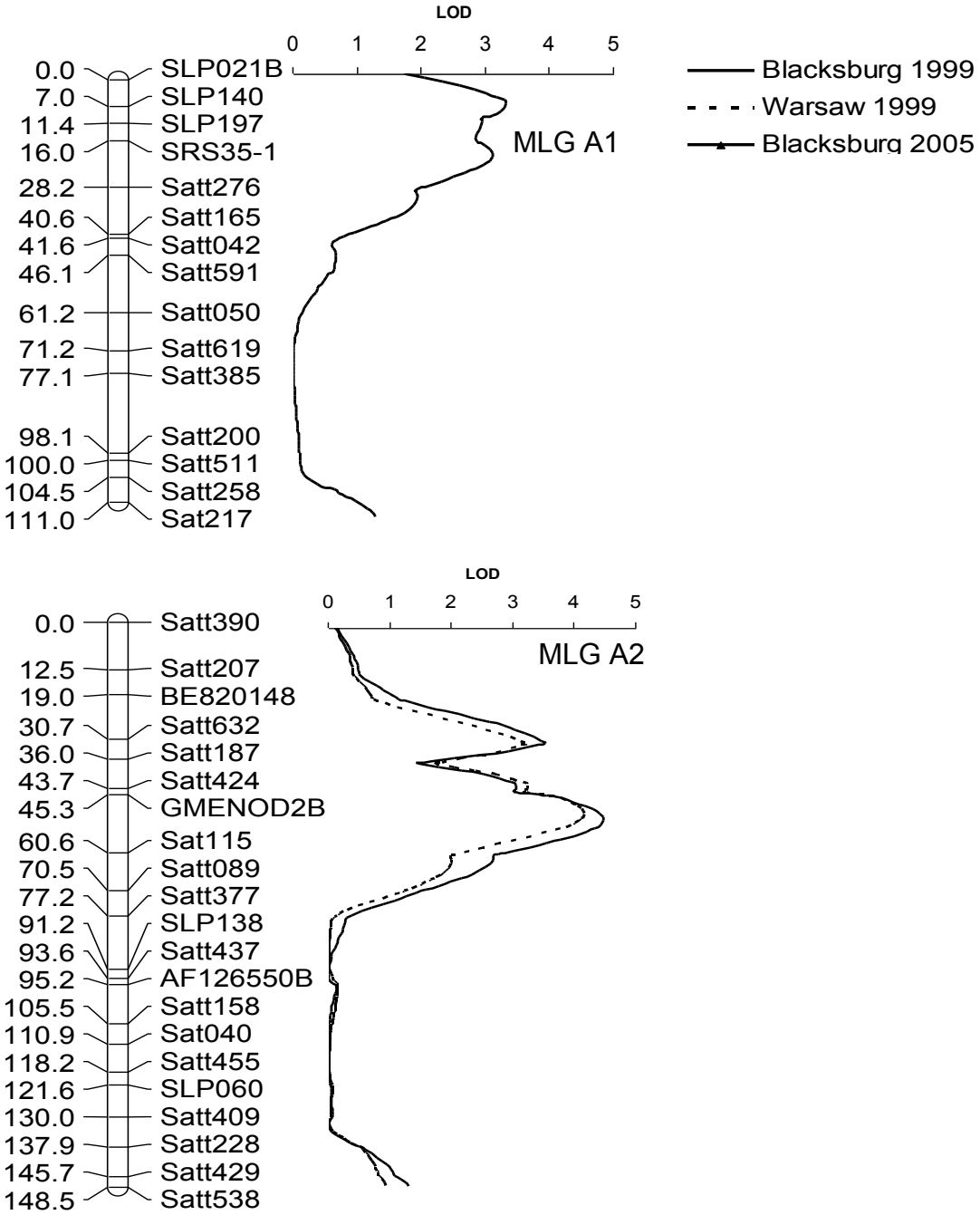

Figure 6 (cont.).

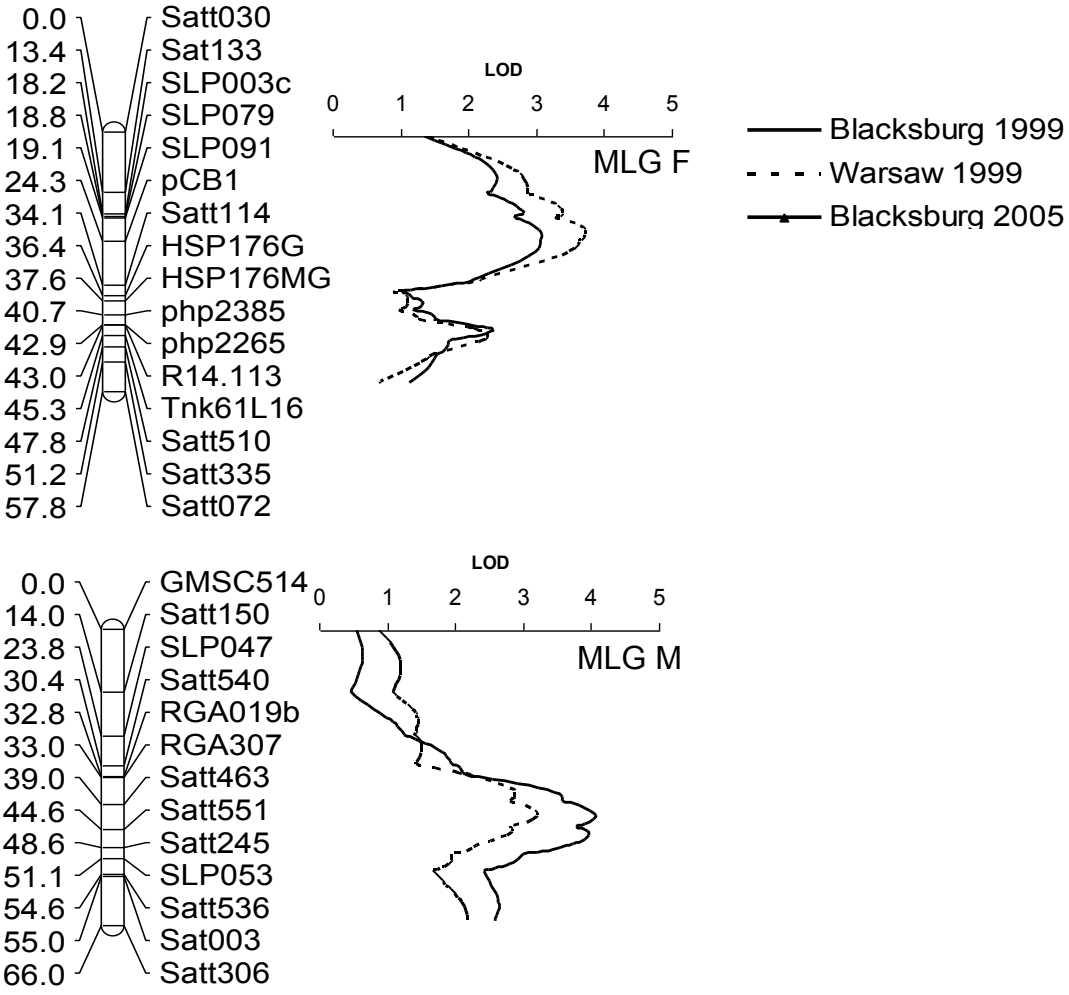

Figure 7. Linkage maps for molecular linkage groups (MLG) containing significant QTL for seed raffinose content in the VPRI RIL population, as well as three classical trait loci. Distance in cM from the top is given to the left of each marker. LOD curves are aligned with their corresponding linkage map. Pod color (*L1*) and determinate stem (*Dt1*) mapped to MLG L, while pubescence color (*T*) mapped to MLG C2.

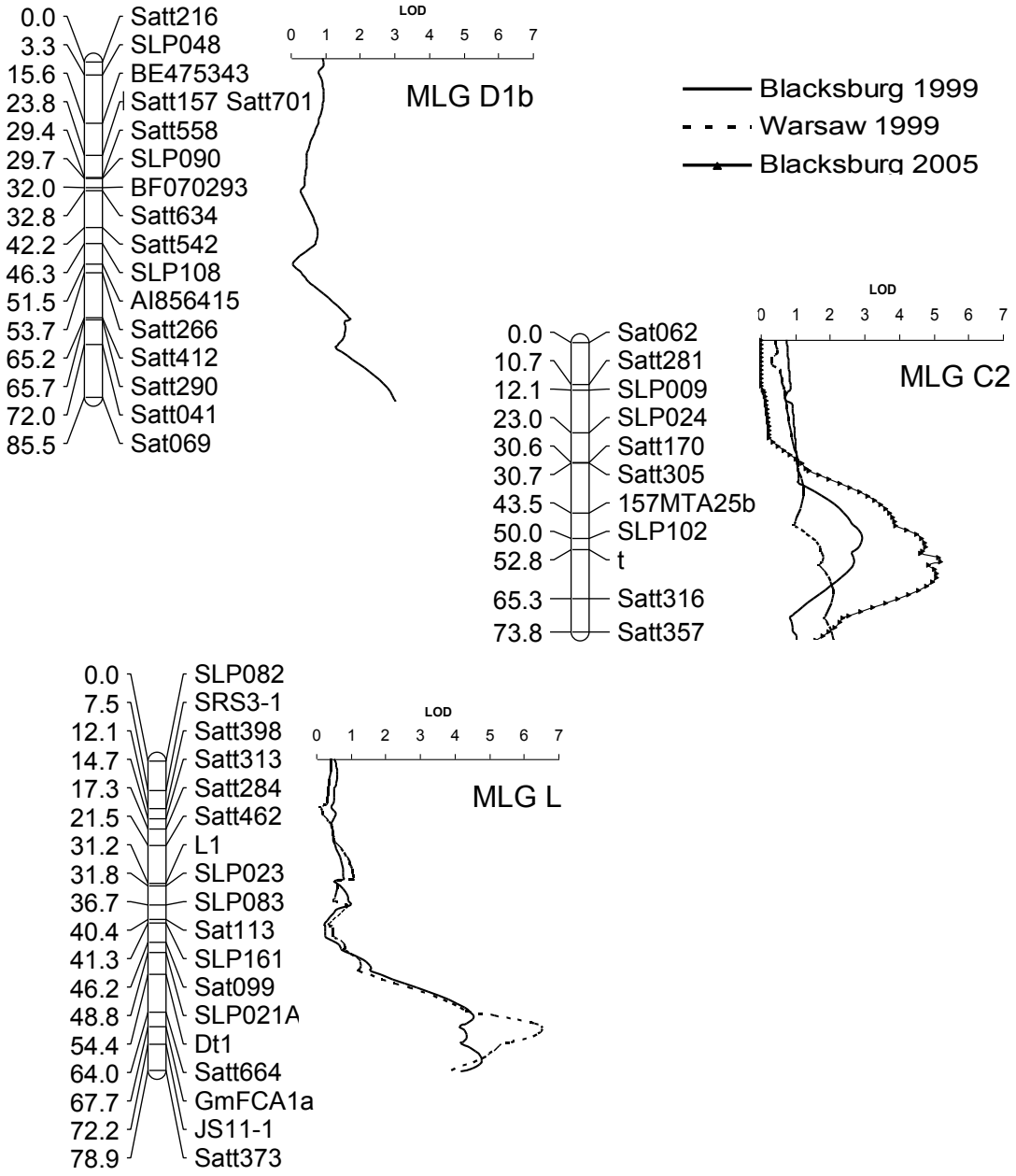

Figure 8. Linkage maps for molecular linkage groups (MLG) containing significant QTL for seed stachyose content in the VPRI RIL population. Distance in cM from the top is given to the left of each marker. LOD curves are aligned with their corresponding linkage map.

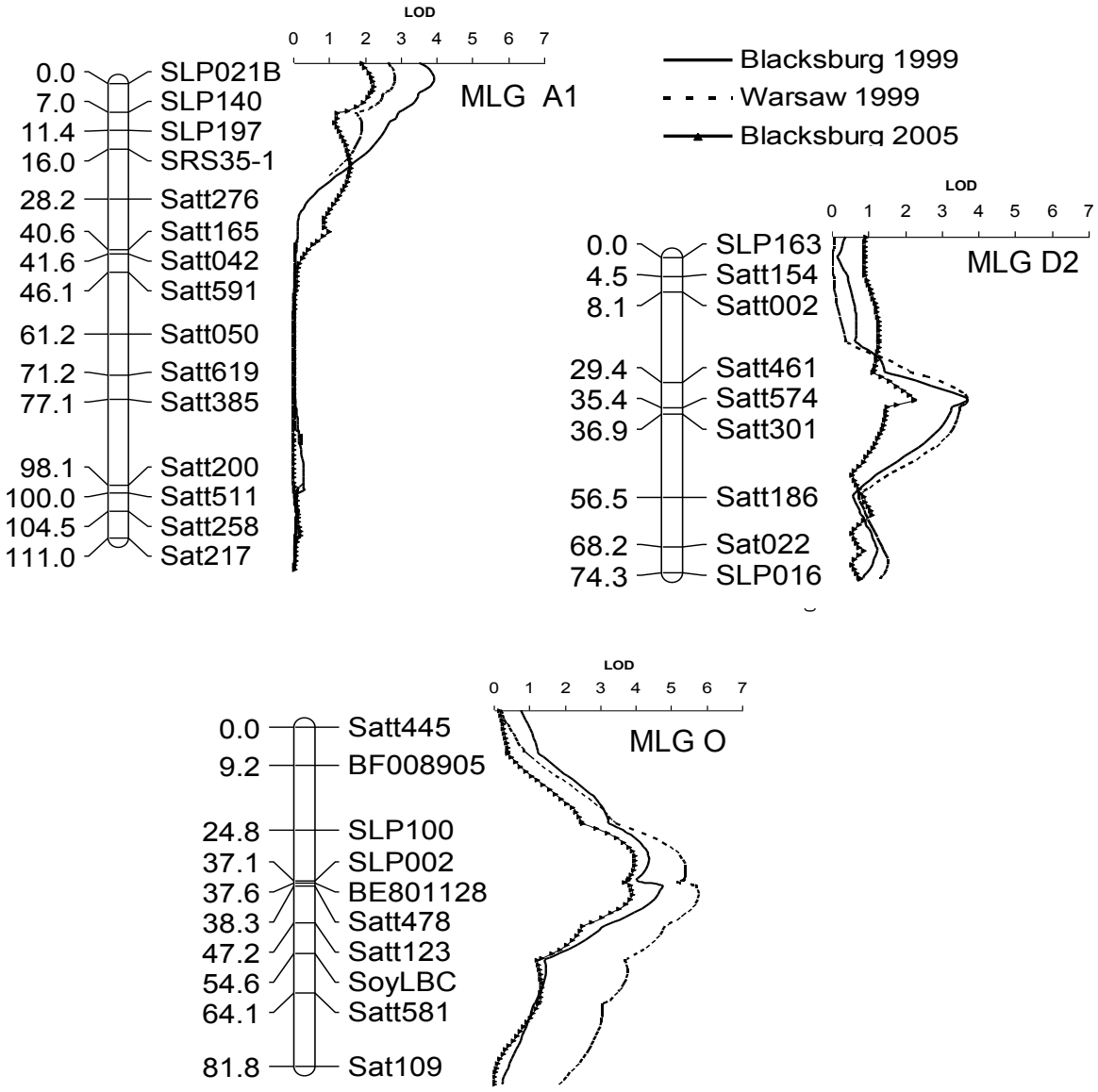

Figure 9. Linkage maps for molecular linkage groups (MLG) containing significant QTL for seed phytate content in the VPRI RIL population. Distance in cM from the top is given to the left of each marker. LOD curves are aligned with their corresponding linkage map.

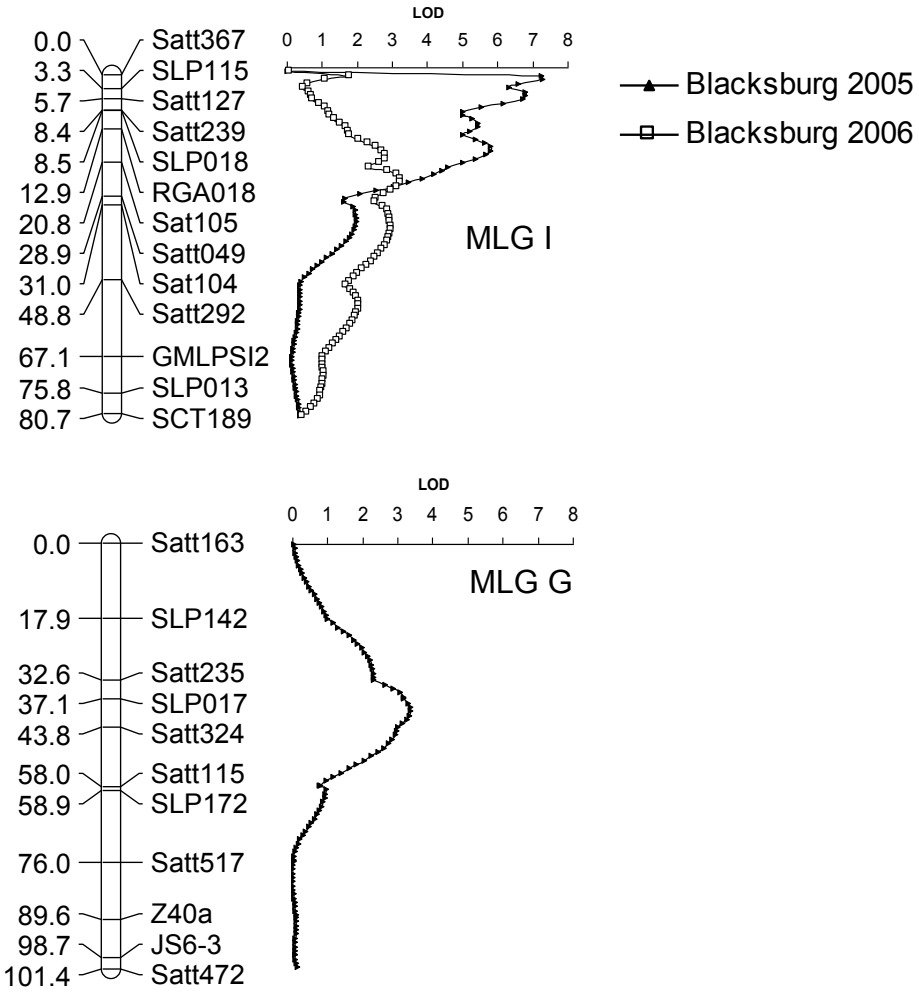

Figure 10. Linkage maps for molecular linkage groups (MLG) containing significant QTL for 200 seed weight in the VPRI RIL population. Distance in cM from the top is given to the left of each marker. LOD curves are aligned with their corresponding linkage map.

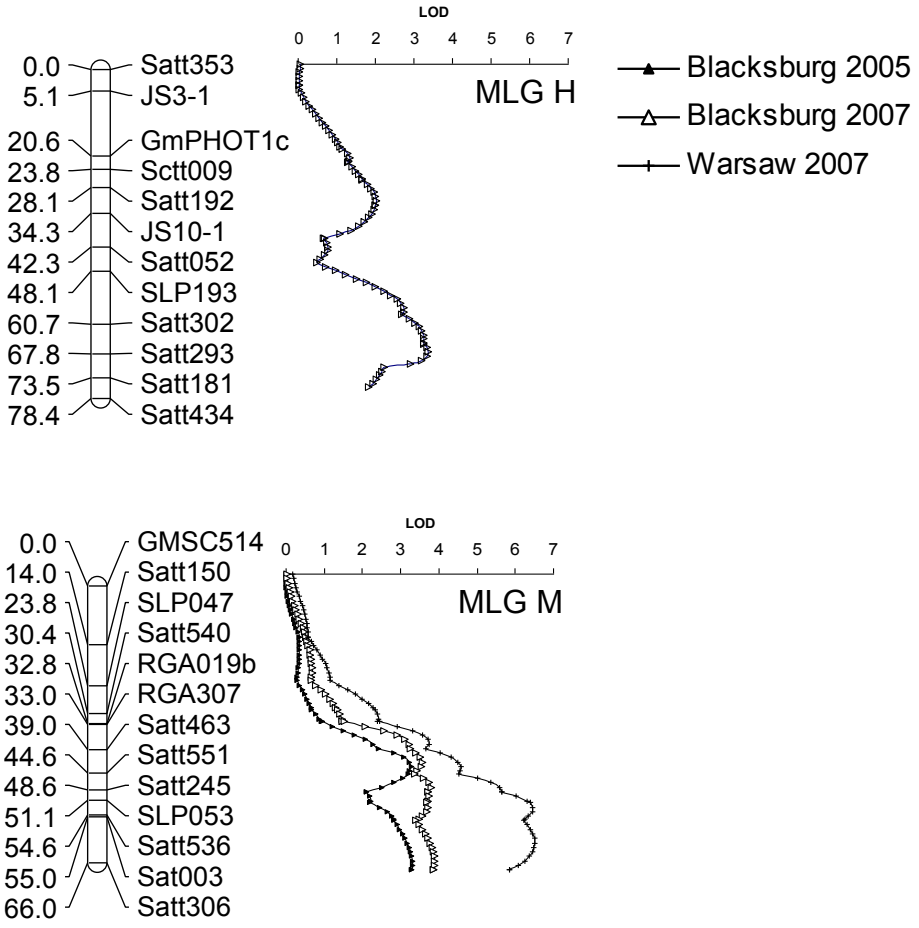

Table 1. Replicate means and standard deviations for traits investigated in the VPRI RIL population across each environment and combined. Environments for seed sucrose, raffinose, and stachyose content represent plantings in Blacksburg 1999, Warsaw 1999, and Blacksburg 2005. Environments for seed phytate content represent Blacksburg 2005 and 2006 plantings, and 200 seed weight was observed from Blacksburg 2005, Blacksburg 2007, and Warsaw 2007 plantings.

	Environment 1			Environment 2			Environment 3			Combined		
	μ	N	σ	μ	N	σ	μ	N	σ	μ	N	σ
Sucrose (%DW)	3.74	300	0.82	3.64	300	0.81	3.96	282	1.02	3.77	882	0.90
Raffinose (%DW)	0.52	300	0.13	0.42	300	0.10	0.73	282	0.14	0.55	882	0.18
Stachyose (%DW)	4.20	300	0.47	4.29	300	0.55	4.55	282	0.71	4.34	882	0.60
Phytate (ppm)	19.81	281	1.84	20.20	299	3.75	-	-	-	20.01	580	2.99
SeedWt (g)	12.14	294	3.37	11.46	291	3.05	12.41	292	3.14	12.00	877	3.21

Table 2. Analysis of variance (ANOVA) for seed sucrose, raffinose, stachyose, and phytate content, as well as seed weight, in the VPRI RIL population developed from a cross between V71-370 and PI407162 grown in multiple environments.

	Source	df	MS	F	P
Sucrose	RIL	300	1.865	7.965	<0.0001
	Environment	2	5.942	25.370	<0.0001
	Error	579	0.234		
	Total	881			
Raffinose	RIL	300	0.025	2.270	<0.0001
	Environment	2	7.244	654.045	<0.0001
	Error	579	0.011		
	Total	881			
Stachyose	RIL	300	0.609	2.998	<0.0001
	Environment	2	9.473	46.607	<0.0001
	Error	579	0.203		
	Total	881			
Phytate	RIL	298	9.382	2.276	<0.0001
	Environment	1	10.116	2.454	0.1184 NS
	Error	278	4.122		
	Total	577			
SeedWt	RIL	299	24.751	10.552	<0.0001
	Environment	2	70.996	30.269	<0.0001
	Error	574	2.346		
	Total	875			

NS not significant (P > 0.05)

Table 3. Results from diagnostic screen of 157 soybean SSRs to determine marker polymorphism between parental lines V71-370 and PI407162. Primers are arranged by linkage group, and their position on their respective linkage groups as reported by Cregan et al. (1999). Selection of polymorphic markers for linkage mapping is also indicated.

Locus	LG	Map position	Polymorphic	Mapped
Satt572	A1	14.7	Yes	Yes
Satt593	A1	25.6	Yes	No
Satt449	A1	27.8	Yes	No
Satt526	A1	27.9	No	--
Satt591	A1	31.1	Yes	Yes
Satt717	A1	52.0	Yes	No
Sat407	A1	52.3	Yes	No
Satt619	A1	69.2	Yes	No
Satt599	A1	85.6	No	--
Satt225	A1	95.2	No	--
Satt258	A1	95.5	Yes	Yes
Satt211	A1	96.0	Yes	No
Sat271	A1	97.8	No	--
Sat217	A1	101.6	Yes	No
BE820148	A2	35.9	Yes	Yes
Satt632	A2	51.5	Yes	Yes
GMENOD2B	A2	58.4	Yes	Yes
Satt228	A2	154.1	Yes	Yes
Satt538	A2	159.6	Yes	Yes
Satt378	A2	165.7	Yes	Yes
BE806308	B1	0.0	Yes	No
Sat272	B1	14.3	Yes	No
Sat270	B1	22.0	Yes	No
Satt638	B1	37.8	Yes	Yes
Satt519	B1	57.9	Yes	Yes
Satt298	B1	64.9	Yes	No
Sat348	B1	72.0	No	--
Satt597	B1	73.8	Yes	No
Satt484	B1	118.5	No	--
Satt453	B1	124.0	Yes	Yes
Sat331	B1	125.7	Yes	No
BE801538	B1	126.5	No	--
AQ851479	B1	128.7	No	--
Sat177	B2	7.8	Yes	No
Sat264	B2	12.6	Yes	No
Sat287	B2	31.9	Yes	No

Satt601	B2	67.7	Yes	Yes
Sat424	B2	100.1	Yes	No
Satt690	C1	5.4	Yes	No
Sat337	C1	32.1	Yes	No
Sat140	C1	41.4	Yes	No
Satt190	C1	73.3	Yes	No
Satt718	C1	73.8	No	--
Satt661	C1	74.4	No	--
Satt670	C1	85.4	Yes	No
Satt713	C1	89.0	Yes	Yes
Sat235	C1	94.6	Yes	No
Satt681	C2	3.2	Yes	No
Satt281	C2	40.3	Yes	No
Satt520	C2	42.4	Yes	No
Satt291	C2	45.8	Yes	No
sat 336	C2	51.8	Yes	No
Sat153	C2	62.0	Yes	No
Satt170	C2	70.6	Yes	No
Satt450	C2	89.3	Yes	No
Sat 213	C2	90.9	Yes	No
sat 246	C2	91.8	Yes	No
Satt643	C2	94.7	Yes	No
Satt376	C2	97.8	Yes	No
Satt363	C2	98.1	Yes	No
sat 076	C2	99.2	Yes	No
Satt286	C2	101.8	Yes	No
sat 402	C2	103.3	Yes	No
Satt557	C2	112.2	Yes	No
Satt289	C2	112.4	Yes	No
Satt134	C2	112.8	Yes	No
Satt658	C2	113.6	Yes	No
Satt708	C2	115.5	Yes	No
Satt368	D1a	43.8	Yes	No
Satt580	D1a	58.5	Yes	Yes
Satt370	D1a	59.9	Yes	No
Satt203	D1a	61.9	Yes	No
Satt383	D1a	64.3	Yes	Yes
sat 106	D1a	66.4	Yes	No
Satt603	D1a	66.4	Yes	Yes
Satt407	D1a	99.6	Yes	Yes
BE021153	D1b	30.2	No	--
BE475343	D1b	30.7	Yes	Yes
Satt701	D1b	40.0	Yes	Yes
Satt558	D1b	43.9	Yes	Yes
Satt634	D1b	46.6	Yes	Yes
BF070293	D1b	47.3	Yes	Yes
AI856415	D1b	50.1	Yes	Yes
Satt296	D1b	52.6	Yes	No
Satt542	D1b	53.0	Yes	No

Satt412	D1b	72.6	Yes	Yes
Sctt008	D2	3.2	Yes	No
Sct 192	D2	11.8	Yes	Yes
Satt014	D2	29.6	Yes	Yes
Satt002	D2	47.7	Yes	Yes
Satt443	D2	51.4	Yes	No
Satt582	D2	53.9	Yes	No
Satt154	D2	57.1	Yes	Yes
Satt461	D2	80.2	Yes	No
Satt574	D2	87.7	Yes	Yes
Satt186	D2	105.5	Yes	Yes
Satt031	D2	115.7	Yes	No
BE347343	E	65.7	Yes	No
GMRUBP	F	0.0	No	--
BE806387	F	23.0	No	--
Satt334	F	78.1	Yes	No
Satt288	G	76.8	No	--
Satt353	H	8.5	Yes	No
Satt568	H	27.6	Yes	No
Satt192	H	44.0	Yes	No
Satt442	H	47.0	Yes	No
Satt052	H	64.1	Yes	No
Satt314	H	69.1	Yes	No
Satt181	H	91.1	Yes	Yes
Satt434	H	105.8	Yes	Yes
GMLPSI2	I	97.0	Yes	Yes
Satt405	J	12.4	Yes	No
Satt285	J	25.5	Yes	No
sct 065	J	32.1	Yes	No
Satt693	J	33.9	Yes	No
Satt414	J	37.0	Yes	Yes
Satt596	J	39.6	Yes	No
Satt456	J	40.8	Yes	No
Sat366	J	52.7	Yes	No
Satt244	J	65.0	Yes	No
Satt547	J	67.8	Yes	No
Satt431	J	78.6	Yes	No
Satt715	K	0.9	Yes	No
Satt539	K	1.8	Yes	No
Satt555	K	42.7	Yes	No
Satt544	K	43.4	Yes	No
Satt417	K	46.2	Yes	No
Satt552	K	46.4	Yes	No
Satt518	K	46.6	Yes	No
Satt628	K	49.6	No	--
Satt617	K	50.9	Yes	No
Satt710	K	51.0	Yes	No
Satt559	K	55.0	Yes	No
Satt725	K	56.9	Yes	No

Sat 044	K	58.0	Yes	No
Sat 043	K	61.7	Yes	Yes
Sct 190	K	77.4	No	--
Satt475	K	78.7	Yes	No
Sat 167	K	85.2	Yes	No
Sat 243	K	86.8	Yes	No
Sat 352	K	93.6	Yes	No
Sat 293	K	99.1	Yes	No
Satt588	K	117.0	Yes	No
Satt448	L	64.7	Yes	No
Satt527	L	70.4	Yes	Yes
Satt561	L	71.4	No	--
Satt006	L	92.0	Yes	No
Satt664	L	92.7	Yes	Yes
Satt404	M	0.8	Yes	No
Satt435	M	38.9	Yes	No
BE823543	M	78.4	No	--
Satt237	N	75.0	Yes	No
Satt339	N	75.9	Yes	No
BF008905	O	29.0	Yes	Yes
BE801128	O	69.0	Yes	Yes
Satt581	O	106.0	Yes	Yes
Sat 307	O	123.4	Yes	No

Table 4. Results of diagnostic screen of markers designed from EST and whole genome shotgun (WGS) sequence information for this study. For mapped markers, their molecular linkage group and cM position as determined by linkage mapping are given.

ID	Source**	Polymorphic	LG	VPRI F_8
SRS1-1	E24424	No	--	--
SRS1-2	E24424	No	--	--
SRS1-3	E24424	No	--	--
SRS1-4	E24424	No	--	--
SRS2-1	E25448	No	--	--
SRS2-2	E25448	No	--	--
SRS2-3	E25448	No	--	--
SRS3-1	AC185960	Yes	L	7.5
SRS4-1	AC186737	No	--	--
SRS4-2	AC186737	No	--	--
SRS4-3	AC186737	No	--	--
SRS5-1	AC185960	No	--	--
SRS5-2	AC185960	No	--	--
SRS5-3	AC185960	No	--	--
SRS6-1	AC166642	No	--	--
SRS6-2	AC166642	No	--	--
SRS6-3	AC166642	No	--	--
SRS7-1	AC186737	No	--	--
SRS7-2	AC186737	No	--	--
SRS8-1	AC185960	No	--	--
SRS8-2	AC185960	No	--	--
SRS9-1	AC166642	No	--	--
SRS9-2	AC166642	No	--	--
SRS9-3	AC166642	No	--	--
SRS10-1	AC185960	No	--	--
SRS10-2	AC185960	No	--	--
SRS35-1	WGS	Yes	A1	16.0
35SRS-AT20	WGS	No	--	--
AF126550a	AF126550	No	--	--
AF126550b	AF126550	Yes	A2	95.2
AF126550c	AF126550	No	--	--
AF126550d	AF126550	No	--	--
AF126550e	AF126550	No	--	--
157MTA25	WGS	Yes	C2	43.5
157MAT28	WGS	Yes	Not Mapped	--
157BAT37	WGS	Yes	Not Mapped	--
157BAT21	WGS	Yes	Not Mapped	--

**Soy Raffinose Synthase (SRS) markers were designed from EST sequences. The accession number of their source sequence is indicated. WGS sequence markers were developed around a reported soybean galactosyltransferase gene sequence

Table 5. Detected QTL in the VPRI RIL population, generated by interval mapping. Linkage group and position is given in centimorgans. The logarithm of odds (LOD) score and the percent of the phenotypic variance that is explained at each QTL is also provided. Additive values are the mean phenotypic difference between QTL alleles expressed as percent seed dry matter for carbohydrate traits, ppm for phytate content, and grams for seed weight.

Sucrose	**Group**	**Position**	**LOD**	**% Expl.**	**Additive**
Blacksburg 1999	A1	8.0	3.3	8.4	0.24
	A2	51.3	4.5	10.5	0.27
	F	23.1	3.1	7.0	0.22
	M	42.0	4.1	8.1	0.24
Warsaw 1999	A2	49.3	4.2	9.7	0.25
	F	23.1	3.7	8.4	0.23
	M	42.0	3.2	6.7	0.21
Blacksburg 2005	I	17.9	2.9	6.1	0.25
	A1	7.0	2.7	6.6	0.26
Raffinose	**Group**	**Position**	**LOD**	**% Expl.**	**Additive**
Blacksburg 1999	L	76.2	5.0	12.9	0.05
	D1b	85.5	3.0	11.1	0.04
Warsaw 1999	L	68.7	6.8	11.9	0.03
Blacksburg 2005	C2	52.0	5.2	11.2	-0.05
Stachyose	**Group**	**Position**	**LOD**	**% Expl.**	**Additive**
Blacksburg 1999	O	38.3	4.7	7.3	-0.13
	A1	107.5	3.9	7.7	0.13
	D2	35.4	3.7	5.5	-0.11
Warsaw 1999	O	40.3	5.7	9.5	-0.17
	A1	107.5	2.8	5.7	0.13
	D2	35.4	3.7	5.6	-0.13
Blacksburg 2005	O	32.8	4.0	10.0	-0.23
Phytate	**Group**	**Position**	**LOD**	**% Expl.**	**Additive**
Blacksburg 2005	I	1.0	7.6	13.5	-0.68
	G	40.1	3.4	6.6	0.47
Blacksburg 2006	I	23.8	3.2	6.4	-0.97
Seed Weight	**Group**	**Position**	**LOD**	**% Expl.**	**Additive**
Blacksburg 2005	M	65.0	3.3	8.2	0.50
Blacksburg 2007	H	69.8	3.4	7.4	0.83
	M	64.0	3.9	8.6	0.89
Warsaw 2007	M	60.0	6.5	14.2	1.18

Chapter 4

Genetic mapping of soybean maturity genes in an interspecific cross

ABSTRACT

This study involves a recombinant inbred line (RIL) population derived from a cross between V71-370 [*Glycine max* (L.) Merr] and PI 407162 (*Glycine soja* Seib. & Zucc.). The parents differed in time to maturity when grown in Virginia, and the RILs that arose from this cross exhibited a wide range of values for the trait, with many lines exhibiting transgressive segregation. Maturity was scored as days to R8, full maturity. Trait data was used to perform quantitative trait loci (QTL) analysis, using an existing molecular linkage map developed from this population, spanning 1618 cM over all 20 soybean molecular linkage groups (MLG). Large effect QTL were observed on MLG L and H, closely associated with microsatellite markers Satt 664 and Sctt009. Investigation of the genomic regions surrounding these markers in the soybean WGS sequence (www.phytozome.net), revealed two likely candidate genes that could be responsible for the detected QTL. These genes showed significant homology to genes known to affect flowering time and photoperiodism in *Arabidopsis thaliana*. Markers designed from these candidate gene sequences were closely linked to the previously observed QTL. QTL analysis after incorporation of these gene specific markers on MLG L and H indicated that the observed QTL were responsible for as much as 27.3 and 12.9 percent of the phenotypic variation observed for time to maturity, indicating they play a significant role in that trait. This report is among the first to directly link flowering time and photoperiodism gene orthologs with time to maturity data in an economically important crop through QTL analysis.

INTRODUCTION

In the interest of developing soybean lines that can be grown over a wide range of locations, plant breeders have devoted considerable effort to investigating the means by which different cultivars interact with their environment. Time to maturity, defined as the length of time between planting and harvest, is a critical trait in determining where a given soybean cultivar can be grown. Although temperature has been shown to play a role in onset of maturity, the predominant factor affecting this trait is photoperiod, or daylength.

Like many flowering plants, soybeans depend on daylength as an environmental cue to enter the reproductive phase and initiate flower development. They are a short day plant, meaning that the daylength must consistently be below a critical length for this to happen. This critical length can vary greatly between cultivars and will affect whether a soybean line will flower and mature within the growing season at a given location. Certain soybean genetic factors, termed photoperiod-sensitivity, or maturity genes, have been described to affect the plant's response to daylength. These genes have been reported to influence this trait in soybean; *E1* and *E2* (Bernard, 1971), *E3* (Buzzell, 1971), *E4* (Buzzell and Voldeng, 1980), *E5* (McBlain and Bernard, 1987), *E6* (Bonato and Vello, 1999), *E7* (Cober and Voldeng, 2001), and *J* (Ray et al., 1995). The genes *E1, E3, E4,* and *E7* were found to influence the onset of soybean reproductive stages in response to artificially induced long daylength (Bernard, 1971; Buzzell, 1971; Buzzell and Voldeng, 1980; Cober and Voldeng, 2001). These studies observing maturity under artificial conditions show that alleles responsible for late flowering and maturity appear

dominant to those conferring early flowering and maturity, while the dominant alleles from genes *E6* and *J* promote earlier flowering (Bonato and Vello, 1999; Ray et al., 1995).

Although the genomic locations of all of these maturity genes have not been established, many of them have been placed on soybean MLG through linkage analysis. Weiss (1970) established linkage between the classical locus *t* (pubescence color) and *E1* on what is now known as MLG C2 (Cregan et al., 1999). It was later discovered that another factor existed in that region, called *E7*, that showed tight linkage with previously reported maturity-linked markers on MLG C2 (Cober and Voldeng, 2001). An earlier study by Cober and Voldeng (1996) established linkage between maturity locus *E3* and the determinate stem locus, *dt1*, which lies on MLG L (Cregan et al., 1999). One of the earlier studies investigating association between a maturity gene and molecular markers reported that *E4* is closely associated with microsatellite markers on MLG I (Molnar et al., 2003). A recent study also mapped *E4* to MLG I between markers Satt496 and Satt354 using near-isogenic lines (Liu et al., 2008). They also presented a gene-specific marker, GmphyA2 that cosegregated with *E4*, suggesting that it encodes a phytochrome gene. E2 is reported to reside on MLG O by Cregan et al. (1999), but the locations of the remaining three soybean maturity genes remain unknown, or unreported.

Time to flowering and maturity QTL have been reported in many studies. Keim et al. (1990) reported five RFLP markers that were significantly associated with pod maturity on what are now known to be MLG C1, C2, and D1, explaining as much as 21% of the observed variation for the trait. A major QTL on MLG C2 was also reported by Mansur et al. (1996), in addition to minor effect QTL on MLG L and M. Tasma et al.

(2001) reported a QTL on MLG C2 as well, that was significantly associated with maturity traits, and explained as much as 47% of the phenotypic variance for days to R1 (days to first flower). Minor QTL were also reported on MLG A1, G, J, and L. Another study investigating the association between maturity and RFLP markers identified significant markers on MLG B1, C1, and L (Lee et al., 1996). Microsatellite markers associated with maturity have been reported on MLG A1, E, H, and I (Panthee et al., 2007).

Despite the knowledge of the locations of many soybean maturity genes, little is known about the precise nature of these genes, and their mode of action. A solution to the task of characterizing soybean *E* genes could lie in comparative studies with organisms in which genes responsible for the transition to reproductive stages have been more extensively studied. One example is *Arabidopsis thaliana*, which has been used extensively as a model to investigate flowering and maturity in plants. Using induced mutant lines, and investigating natural variation, at least 80 loci have been identified that affect flowering time (Koornneef et al., 1998; Levy and Dean, 1998). This information could be useful in a comparative sense, allowing orthologous genes to be identified in soybean. Tasma and Shoemaker (2003) mapped 10 Arabidopsis flowering time gene homologs in three soybean populations, which were distributed across nine MLG. Using near-isogenic lines, they were able to associate one mapped homolog, *FCA*, with the soybean maturity gene *E3*. Incorporating this knowledge into future studies, especially those involving trait data could significantly elucidate the role these gene homologs play in determining time to flower and maturity in soybean.

This study utilized an existing 257 locus genetic linkage map generated from an interspecific cross to identify QTL related to time to maturity in soybean. This linkage map was then used to anchor the soybean WGS sequence (www.phytozome.net) in an effort to identify candidate gene sequences that could be responsible for observed QTL. Primer sequences designed from candidate genes that cosegregate with observed QTL will be presented in the hope that they can be used in breeding programs to select for early or late maturing soybean lines, thereby accelerating new cultivar development.

MATERIALS AND METHODS

Plant Material

The population used in this investigation has been described elsewhere (Cicek et al., 2006). In brief, a F_{12} RIL population (N=300) was developed from the interspecific cross between the adapted large-seeded *G. max* breeding line, V71-370, and a small-seeded, *G. soja* plant introduction, PI 407162, with black seed coat. V71-370 is a maturity group 5 line, while PI 407162 is in maturity group 4, and the offspring derived from crosses between these lines exhibited a wide range of segregation for time to maturity. The RILs were developed from different F_2 plants by the single seed descent method, advancing the plant population from F_2 to F_{12}.

Field Observation

The 300 RILs, three plots of each parent, and three plots of one check cultivar, Chesapeake, were planted in an identical serpentine fashion at the Kentland Research Farm near Blacksburg, Virginia during the summers of 2006 and 2007. The planting dates of these experiments, which were performed in different fields, were May 24, 2006 and June 4, 2007. These plantings consisted of the F_{11} and F_{12} generations of this population, respectively. Each entry was planted in a single 1.2 m long row, at 30 seeds per row with 2.3 m between rows. Each set of 30 seeds was lightly scarified with sandpaper prior to planting.

Time to maturity data were observed in the field for each plot in both environments. Observations were made every two days. Maturity scores were given based on days to maturity after the first RIL was considered mature. Individual RILs were regarded as mature when they reached R8, or full maturity (95% of pods were mature based on color change). Because of the indeterminate growth habit of the *G. soja* parent, pod ripening was not uniform in many RILs. To allow these RILs to reach R8, pod shattering at the earliest developing nodes was allowed to occur before they were harvested.

DNA Extraction and Genotyping

Young first or second trifoliate leaves of field-grown plants were sampled for DNA extraction. DNA from each RIL, was isolated according to the protocol described previously by Saghai Maroof et al. (1984). DNA from each RIL was genotyped using publicly available simple sequence repeat (SSR, or microsatellite) markers (Cregan et al., 1999), as well as primers designed for this study, from SSR sequences observed in the Williams 82 whole genome shotgun (WGS) sequence (DoE Joint Genome Institute, 2008). The web based platform, SSRIT, was used to identify SSR regions from sequence information (Temnykh et al., 2001), and primers were designed with the web-based program, Primer3 (Rozen and Skaletskyv, 2000). Molecular marker procedures were set up as described by Yu et al. (1994). Typical PCR reactions for SSR detection were denatured at 94ºC for 3 min, followed by 36 cycles at 94º C for 30 sec, 47ºC for 30 sec, 68ºC for 1 min, with a final extension at 68ºC for 8 min. To limit non-specific

amplification, annealing temperatures were raised for primers with higher T_m, however, most annealing temperatures fell within the 47-55ºC range. Genotypes were visualized as previously reported (Saghai Maroof et al., 1994).

Mapping and Statistical Analysis

This study utilized existing marker data from a 257 locus linkage map derived from an interspecific cross. Linkage analysis and map generation were performed via JoinMap 4 (Van-Oiogen, 2006). Linkage groups were constructed using a logarithm of the odds (LOD) threshold of 4.0. Regression mapping of each linkage group was performed using the Kosambi mapping function, using linkages with recombination frequency smaller than 0.4, and a LOD greater than 1.0. Interval mapping for QTL analysis was performed with MapQTL 5 (Van-Oiogen, 2004). A default LOD threshold of 3.0 was used to establish significance. Trait means, distribution, and analysis of variance (ANOVA) were determined using the SPSS statistical software package (release 15.0.0) (SPSS, Inc., 2006). ANOVA was performed using the general linear model Univariate command, with trait values from parental lines and check cultivar removed from the dataset.

Identification of Candidate Gene Sequences

Markers identified as flanking a QTL interval identified in this study were used to anchor the soybean WGS sequence (DoE JGI, 2008). The sequences of flanking markers and their amplified regions were obtained from the NCBI UniSTS database (http://www.ncbi.nlm.nih.gov). These sequences were then used to BLAST against the

soybean WGS sequence, identifying the chromosome and physical interval that coincided with the linkage map position of the observed QTL. The physical interval identified through this method, and its corresponding gene model, Glyma1, was then visually examined using the Phytozome graphical genome browser interface (www.phytozome.net). Sequences from observed candidate gene models were then used to BLAST against the NCBI EST database to establish homology to genes from other plant species that are known to affect flowering time and/or maturity, and confirm them as candidates for the causal genetic factors of observed QTL. Sequences from candidate gene models were examined for likely regions of polymorphism, such as microsatellites, that could be exploited for mapping purposes. This included repeat motifs in putative intron sequences, as well as from the genomic regions surrounding the modeled gene. In the latter case, primers were designed to amplify regions no farther than 5 kb from the gene model of interest. These markers were then mapped to confirm their location in the existing linkage map and determine their relationship to observed QTL.

RESULTS

Distribution and Analysis of Variance

The observation dates of the first mature RILs were September 18 (123 days after planting) in 2006 and September 22 (110 days after planting) in 2007. Distribution of the maturity trait across both years appeared bimodal in nature (Figure 1). The manifestation of distinct peaks suggests that time to maturity in this experimental population is under the control of two or more genes. In both years of this study, many transgressive segregants were observed for time to maturity. RILs were observed to mature much earlier and later than the parents used to generate this cross, indicating a complex interaction of genes contributing to time to maturity. Trait means for RILs and parental lines are given in Table 1. Significant variation among RILs, as well as between environments, was observed (Table 2).

Linkage and QTL analysis

QTL analysis through interval mapping identified markers with significant association with time to maturity on MLG H and L. The results of this analysis suggested that the gene(s) contributing to these QTL could be found in the marker intervals between Satt353 and Sctt009 on MLG H, and Satt664 and Satt373 on MLG L (Figure 2). In both cases, these QTL locations were supported across two environments, representing two growing seasons, and were responsible for explaining as much as 12.9%

and 27.3% of the phenotypic variation for time to maturity, respectively (Table 3). Calculation of the additive effect of these QTL revealed that time to maturity was influenced by as many as four days in the case of the QTL on MLG L.

Candidate gene discovery

Examination of the observed QTL intervals in the soybean WGS sequence revealed the presence of two candidate genes that could be responsible for the manifestation of QTL in these regions. The sequences from molecular markers Satt353 (BH126542) and Sctt192 (BH126785) were subjected to BLAST via the phytozome website. This revealed a physical interval of 5,455,321 bp between these markers on soybean chromosome 12 (MLG H) of the Glyma1 gene model set. The physical distance between Satt664 (CC454019) and Satt373 (BH126560) was revealed to be 3,081,892 bp on chromosome 19 (MLG L). Observation of the modeled genes in these intervals revealed two soybean sequences that shared homology with Arabidopsis genes known to play a role in flowering. The modeled gene Glyma12g07890 on chromosome 12 (MLG H) shares homology (E-value = 0) to the Arabidopsis phototropin gene *PHOT1* (AT3G45780). The candidate gene model observed on chromosome 19 (MLG L), Glyma19g40570, shares significant homology (E-value = $3e^{-18}$) to the Arabidopsis *FCA* gene (AT4G16280). Subjecting the transcribed predicted coding sequence of the model Glyma12g07890 to protein BLAST via the NCBI database revealed further orthologous phototropin gene products from *Phaseolus vulgaris*, *Vicia faba*, *Pisum sativum*, *Solanum lycopersicum*, and *Lycopersicon esculentum* in addition to Arabidopsis (Table 4).

Likewise, the Glyma19g40570 putative protein has flowering time related orthologs in *Oryza sativa* cv. *Indica*, *Triticum aestivum*, *Hordeum vulgare*, and *Lolium perenne*, in addition to multiple isoforms of *FCA* in Arabidopsis. The loci representing putative soybean genes, Glyma12g07890 and Glyma19g40570 were referred to as GmPHOT1 and GmFCA1 in this study, due to their homology with phototropin and *FCA* genes, respectively.

Four primer pairs were designed from each of the model gene sequences of interest for use in linkage analysis (Table 5). Primers for each putative gene were designed to amplify possible sources of polymorphism between parental lines, such as microsatellites. When possible, target sequences were chosen from predicted exons in the hope of developing gene-specific markers. One primer pair designed from GmPHOT1 was polymorphic, GmPHOT1c, while two primers designed from GmFCA1 were able to be used for linkage analysis, GmFCA1a, and GmFCA1b. Linkage analysis including these markers showed that GmPHOT1c and GmFCA1a cosegregated with the two major QTL for time to maturity on their respective MLG, H, and L. GmFCA1b mapped to MLG N, 17.1 cM from Sat091 and 2.9 cM from Satt022 (results not shown). GmPHOT1c cosegregated with the greatest observed LOD score on MLG H from both sets of maturity data, yielding an LOD of 8.8 and 7.9 in 2006 and 2007 (Table 3). These QTL were responsible for 12.9% and 12.2% of the observed phenotypic variation. GmFCA1a mapped within 4 cM of the observed QTL on MLG L that was responsible for as much as 27.3% of the phenotypic variation for time to maturity.

DISCUSSION

This study compared time to maturity in an interspecific RIL population. Two genomic regions were identified in this population through QTL analysis to affect this trait. Additionally, two candidate genes were presented on MLG H and L that are potentially responsible for detected QTL. In the case of MLG H, a gene specific marker, GmPHOT1c, designed from a putative soybean phototropin gene cosegregated with the observed QTL on that linkage group. To date there have been no reports of soybean maturity QTL on MLG H, with the exception of one association between microsatellite locus Satt293 and maturity in a *Glycine max* RIL population (Panthee et al., 2007). However, the linkage map position of Satt293 is 42.7 cM away from GmPHOT1c according the linkage map generated in this study, and over 50 cM away according to Cregan et al. (1999), suggesting it is a separate genetic entity from the one detected in this study.

Phototropins are blue light receptor proteins that were previously known as non-phototropic hypocotyl (NPH) proteins (Briggs et al., 2001). They not only influence plant phototropic response, or the orientation of shoot and leaf surfaces for optimal light exposure, but also chloroplast migration in response to changing light intensity (Jarillo et al., 2001; Kagawa et al., 2001; Sakai et al., 2001). Intracellular chloroplast movement not only optimizes light exposure, but also protects photosynthetic machinery from excessive light intensity. Although there have been no reports of photropins influencing maturity, they have been shown to control light-induced shoot growth inhibition in

conjunction with another plant blue light receptor, cryptochrome (Parks et al., 2001). Although speculative, it warrants further investigation into whether phototropins could be part of a signaling pathway responsible for late-season shoot dormancy in response to daylength.

A possible candidate gene contributing the QTL observed on MLG L was identified in this study as a putative soybean gene in the WGS sequence with high similarity to *FCA* genes from multiple organisms. The linkage map position of gmFCA1a, the marker developed from this putative gene, segregated closely with the observed QTL region on MLG L in both environments. QTL related to maturity have been reported on MLG L (Cober and Voldeng, 1996; Lee et al., 1996). Maturity gene *E3* was found to be associated with Satt229 on MLG L using near-isogenic lines (Molnar et al., 2003). A study involving markers derived from soybean clones sharing homology to Arabidopsis *FCA* gene mapped a soybean *FCA* ortholog, GenBank accession AI930903, to MLG L, 29.4, 10.9, and 19.6 cM upstream of Satt373 in three separate mapping populations (Tasma and Shoemaker, 2003). This is comparable to the results of this study, in which GmFCA1a mapped 11.2 cM upstream from Satt373. Subjecting AI930903 to BLAST revealed two possible locations in the WGS sequence. The first was on chromosome Gm03 (MLG N), sharing homology with model gene Glyma03g37950 (expect value: $3.7e^{-81}$), while the most significant expect value was observed on Gm19 (MLG L) with shared homology with Glyma19g40570 (expect value: $2e^{-107}$). This strong homology suggests that the marker developed in this study represents the same genetic entity that was mapped by Tasma and Shoemaker, who claim that

comparison of this loci's map position to classical linkage maps make it a likely candidate for soybean maturity gene *E3* (2003).

FCA has been cloned and characterized in Arabidopsis. The protein encoded by this gene contains two RNA-binding domains, and a tryptophan-tryptophan protein interaction domain, suggesting it plays a role in transcriptional regulation of genes involved in flowering (Macknight et al., 1997). *FCA* mutants of the early flowering Arabidopsis ecotype, Landsberg erecta, were found to flower late under long day conditions, yet were hypersensitive to vernalization (Koornneef et al., 1991).

The observation of an *FCA* ortholog on Gm03 (MLG N) is interesting, as one of the markers developed for this study, GmFCA1b mapped to MLG N. GmFCA1b was the only other marker, developed from the putative soybean *FCA* sequence that was polymorphic in the RIL mapping population used in this study, and seems to represent another soybean *FCA* ortholog, Glyma03g37950. However, it will not be discussed further here as it showed no significant effect on time to maturity in this study. In addition to Glyma03g37950, subjecting the model nucleotide sequences of Glyma19g40570 and the putative soybean phototropin gene, Glyma12g07890, to BLAST via phytozome reveals the presence of 11 paralogous genes in the WGS sequence (Table 6). These possible paralogs exist on eight chromosomes, and although these chromosomes did not contain any markers that were significantly associated with time to maturity in this study, the strong association of GmPHOT1 and GmFCA1 with time to maturity suggest that these paralogous sequences could be useful resources for the further identification of genomic regions associated with that trait.

It is interesting to note that while two candidates for genes controlling time to maturity in soybean have been identified in this study, the QTL that they are associated with, although highly significant, together account for only 40.2% and 31.2% of the observed phenotypic variance in 2006 and 2007 (Table 3). Although time to maturity is a trait that is heavily influenced by environment, and can be somewhat subjective when quantified in the manner employed in this study, this suggests that there could be additional genetic components affecting its manifestation, which were not identified in this study. It was unexpected that the QTL alleles of the later maturing parent V71-370 was responsible for earlier maturity, as was indicated by the negative additive effects for each QTL. It has been shown that the effect of *E3*, which increases sensitivity to long daylength (retards flowering) (Buzzell, 1971), can be reversed by the allelic state of other maturity genes. The recessive allele *e4* was shown to contribute to artificial long day insensitivity (promotes flowering) only in the presence of *e3*. Also, studies investigating maturity QTL *FT1* (*E1*) and *FT3* (*E3*) showed that *ft1* reduced the effect of *Ft3* (Watanabe et al., 2004; Yamanaka et al., 2000). It is unknown whether epistasis was a factor in the V71-370 × PI407162 RIL population, yet it further suggests that undetected genetic factors in this population could be affecting maturity.

Time to maturity has been attributed to maturity genes, which control the trait by influencing two factors, time to flowering, and duration of the reproductive phase (Kumudini et al., 2007). With this in mind, it is possible that measuring maturity as time to reproductive stage R8, or full maturity, resulted in detected QTL for either of the above factors, or a combination of both. Additionally, measuring maturity in this manner could have caused QTL for time to flowering, or those controlling length of reproductive

stage, to be masked by the other. For example, QTL for delayed flowering may be masked if the prolonged vegetative phase is followed by rapid seed set and fill. In order to reveal all QTL contributing to soybean maturity in this population, further studies observing these two aspects separately, and in greater detail, may be necessary.

Using observations made across two environments, maturity QTL in an interspecific soybean RIL population have been mapped to MLG L and H, explaining as much as 27.3%and 12.9% of the phenotypic variation respectively. Additionally, the use of the Williams82 WGS sequence as a resource for gene discovery has identified two gene models as putative maturity genes.

The location of the model gene GmFCA1, a putative ortholog of the Arabidopsis flowering gene *FCA*, on MLG L, makes it a candidate for *E3*, as suggested by Tasma and Shoemaker (2003). However, a recent map-based cloning study employing *E3* residual heterozygous lines has reported association between this maturity gene and the soybean gene *Phytochrome A3* (*GmphyA3*) on MLG L (Watanabe et al., 2009). BLAST results indicate that *GmphyA3* is approximately 571 kb upstream of GmFCA1, indicating that further work is necessary to elucidate the cause of the maturity QTL observed on MLG in this study. To date, there have been no QTL or maturity genes mapped to the region on MLG H that was observed to affect time to maturity in this study. Because of this, it cannot be determined whether GmPHOT1 could be a previously uncharacterized maturity gene, or a known gene for which the map location has not been reported, such as *E5* or *E6*.

For both QTL, the strong association between candidate gene specific markers and QTL indicates that markers developed in this study could be useful in a molecular

breeding program. Existing cultivars could be adapted to grow in a wider range of environments, and developing cultivars could be tailored to grow in any desired region. Additionally, maturity related markers could be used to facilitate cultivar development by allowing the selection of early or late maturity alleles in winter nurseries, where phenotypic selection for maturity is not feasible. This could become applicable, as fluctuating environmental conditions on the global scale turn from theory to empirical truth. Further investigation of these loci is necessary to confirm their association with time to maturity in soybean. Additional studies of these, and other QTL identified in future studies, could facilitate the cloning of soybean maturity genes, allowing the time to maturity to be altered, and increase our knowledge of the genes controlling time to maturity in soybean.

REFERENCES

2008. Soybean Genome Project [Online]. Available by DoE Joint Genome Institute http://www.phytozome.org (verified February 2008).

Bernard, R.L. 1971. Two major genes for time of flowering and maturity in soybeans. Crop Sci 11:242-244.

Bonato, E.R., and N.A. Vello. 1999. *E6*, a dominant gene conditioning early flowering and maturity in soybeans. Genet Mol Biol 22:229-232.

Briggs, W.R., C.F. Beck, A.R. Cashmore, J.M. Christie, J. Hughes, J.A. Jarillo, T. Kagawa, H. Kanegae, E. Liscum, A. Nagatani, K. Okada, M. Salomon, W. Rudiger, T. Sakai, M. Takano, M. Wada, and J.C. Watson. 2001. The Phototropin Family of Photoreceptors. Plant Cell 13:993-997.

Buzzell, R.I. 1971. Inheretance of a soybean flowering response to fluorescent-daylength conditions Can J Genet Cytol 13:703-707.

Buzzell, R.I., and H.D. Voldeng. 1980. Inheritance of insensitivity to long daylength. Soybean Genetics Newsletter:26-29.

Cicek, M.S., P. Chen, M.A.S. Maroof, and G.R. Buss. 2006. Interrelationships among agronomic and seed quality traits in an interspecific soybean recombinant inbred population. Crop Sci 46:1253-1259

Cober, E.R., and H.D. Voldeng. 1996. *E3* and *Dt1* linkage. Soybean Genetics Newsletter 23:56-57.

Cober, E.R., and H.D. Voldeng. 2001. A new soybean maturity and photoperiod-sensitivity locus linked to *E1* and *T*. Crop Sci 41:698-701.

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, and J.E. Specht. 1999. An integrated genetic linkage map of the soybean genome. Crop Sci 39:1464-1490.

Jarillo, J.A., H. Gabrys, J. Capel, J.M. Alonso, J.R. Ecker, and A.R. Cashmore. 2001. Phototropin-related NPL1 controls chloroplast relocation induced by blue light. Nature 410:952-954.

Kagawa, T., T. Sakai, N. Suetsugu, K. Oikawa, S. Ishiguro, T. Kato, S. Tabata, K. Okada, and M. Wada. 2001. Arabidopsis NPL1: A Phototropin Homolog Controlling the Chloroplast High-Light Avoidance Response. Science 291:2138-2141.

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

Koornneef, M., C.J. Hanhart, and J.H.v.d. Veen. 1991. A genetic and physiological analysis of late flowering mutants in *Arabidopsis thaliana*. Mol Gen Genet 229:57-66.

Koornneef, M., C. Alonso-Blanco, H.B.-d. Vries, C.J. Hanhart, and A.J. Peeters. 1998. Genetic control of flowering time in *Arabidopsis*. Annual Review of Plant Physiology 49:345-370.

Kumudini, S.V., P.K. Pallikonda, and C. Steele. 2007. Photoperiod and E-genes Influence the Duration of the Reproductive Phase in Soybean. Crop Sci 47:1510-1517.

Lee, S.H., M.A. Bailey, M.A.R. Mian, J. T. E. Carter, D.A. Ashley, R.S. Hussey, W.A. Parrott, and H.R. Boerma. 1996. Molecular Markers Associated with Soybean Plant Height, Lodging, and Maturity across Locations. Crop Sci 36:728-735.

Levy, Y.Y., and C. Dean. 1998. The Transition to Flowering. Plant Cell 10:1973-1989.

Liu, B., A. Kanazawa, H. Matsumura, R. Takahashi, K. Harada, and J. Abe. 2008. Genetic redundancy in soybean photoresponses associated with duplication of the Phytochrome A gene. Genetics 180:995-1007.

Macknight, R., I. Bancroft, T. Page, C. Lister, R. Schmidt, K. Love, L. Westphal, G. Murphy, S. Sherson, C. Cobbett, and C. Dean. 1997. FCA, a gene controlling flowering time in Arabidopsis, encodes a protein containing RNA-binding domains. Cell 89:737-745.

Mansur, L., J. Orf, K. Chase, T. Jarvik, P. Cregan, and K. Lark. 1996. Genetic mapping of agronomic traits using recombinant inbred lines of soybean. Crop Sci 36:1327-1663.

McBlain, B.A., and R.L. Bernard. 1987. A new gene affecting the time of flowering and maturity in soybeans. J Hered 78:160-162.

Molnar, S.J., S. Rai, M. Charette, and E.R. Cober. 2003. Simple sequence repeat (SSR) markers linked to *E1*, *E3*, *E4*, and *E7* maturity genes in soybean Genome 46:1024-1036.

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

Parks, B.M., K.M. Folta, and E.P. Spalding. 2001. Photocontrol of stem growth. Current Opinion in Plant Biology 4:436-440.

Ray, J.D., K. Hinson, J.E.B. Mankono, and M.F. Malo. 1995. Genetic Control of a Long-Juvenile Trait in Soybean. Crop Sci 35:1001-1006.

Rozen, S., and H.J. Skaletskyv. 2000. Primer3 on the WWW for general users and for biologist programmers, p. 365-386, *In* S. Krawetz and S. Misener, eds. Bioinformatics Methods and Protocols: Methods in Molecular Biology. Humana Press, Totowa, NJ.

Saghai Maroof, M.A., R.M. Biyashev, G.P. Yang, Q. Zhang, and R.W. Allard. 1994. Extraordinarily polymorphic microsatellite DNA in barley: species diversity, chromosomal locations, and population dynamics. Proc Natl Acad Sci (USA) 91:5466-5470.

Saghai Maroof, M.A., K.M. Soliman, R.A. Jorgensen, and R.W. Allard. 1984. Ribosomal DNA spacer-length polymorphisms in barley: Mendelian inheritance, chromosomal location, and population dynamics. Proc Natl Acad Sci (USA) 81:8014-8018.

Sakai, T., T. Kagawa, M. Kasahara, T.E. Swartz, J.M. Christie, W.R. Briggs, M. Wada, and K. Okada. 2001. Arabidopsis nph1 and npl1: Blue light receptors that mediate both phototropism and chloroplast relocation. Proc Natl Acad Sci (USA) 98:6969-6974.

Tasma, I.M., L.L. Lorenzen, D.E. Green, and R.C. Shoemaker. 2001. Mapping genetic loci for flowering time, maturity, and photoperiod insensitivity in soybean. Mol Breed 8:25-35.

Tasma, I.M., and R.C. Shoemaker. 2003. Mapping flowering time gene homologs in soybean and their association with maturity (E) loci. Crop Sci 43:319-328.

Temnykh, S., G. DeClerck, A. Lukashova, L. Lipovich, S. Cartinhour, and S. McCouch. 2001. Computational and experimental analysis of microsatellites in rice (*Oryza sativa* L.): frequency, length variation, transposon associations, and genetic marker potential. Genome research 11:1441-1452.

Kyazma. 2004. MapQTL 5, Software for the mapping of quantitative trait loci in experimental populations. Kyazma, Wageningen, Netherlands.

Kyazma B V. 2006. JoinMap 4, Software for the calculation of genetic linkage maps in experimental populations. Kyazma B V, Wageningen, Netherlands.

Watanabe, S., T. Tadjuddin, N. Yamanaka, M. Hayashi, and K. Harada. 2004. Analysis of QTLs for reproductive development and seed quality traits in soybean using recombinant inbred lines. Breed Sci 54:399-407.

Watanabe, S., R. Hideshima, Z. Xia, Y. Tsubokura, S. Sato, Y. Nakamoto, N. Yamanaka, R. Takahashi, M. Ishimoto, T. Anai, S. Tabata, and K. Harada. 2009. Map-Based Cloning of the Gene Associated With the Soybean Maturity Locus E3. Genetics 182:1251-1262.

Weiss, M.G. 1970. Genetic linkage in soybeans: Linkage group I. Crop Sci 10:69-72.

Yamanaka, N., Y. Nagamura, Y. Tsubokura, K. Yamamoto, R. Takahashi, H. Kouchi, M. Yano, T. Sasaki, and K. Harada. 2000. Quantitative trait locus analysis of flowering time in soybean using a RFLP linkage map. Breed Sci 50:109-115.

Yu, Y.G., M.A.S. Maroof, G.R. Buss, P.J. Maughan, and S.A. Tolin. 1994. RFLP and microsatellite mapping of a gene for soybean mosaic virus resistance. Phytopathology 84:60-64.

Figure 1. Frequency distribution of maturity scores in an advanced RIL population derived from an interspecific cross between *G. soja* and *G. max* from two plantings at the Kentland Research Farm (Blacksburg, Virginia) during 2006 (N=300) and 2007 (N=300). Scoring for each RIL was based on days to maturity (R8) after the first mature RIL was observed.

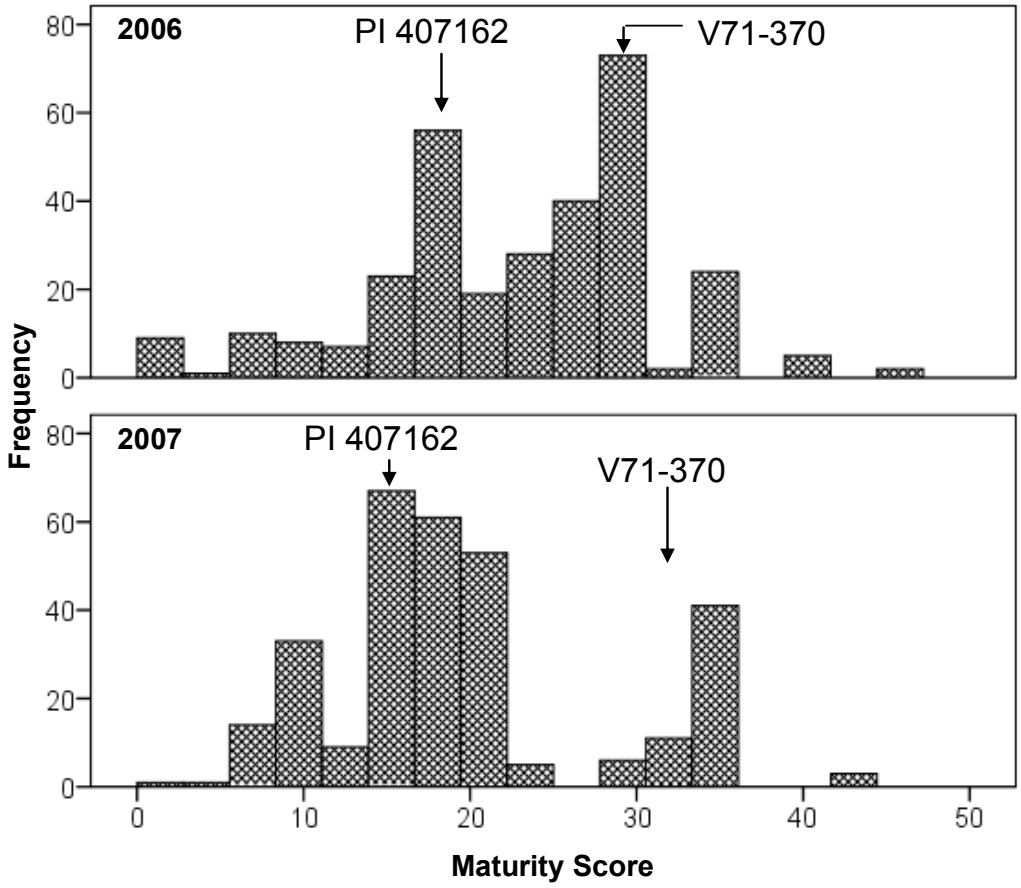

Figure 2. LOD plots generated from interval mapping of time to maturity trait data on molecular linkage groups (MLG) L and H. Solid lines represent data collected from plants grown at Kentland Research Farm (Blacksburg, Virginia) during 2006, while dashed lines represent data collected in 2007.

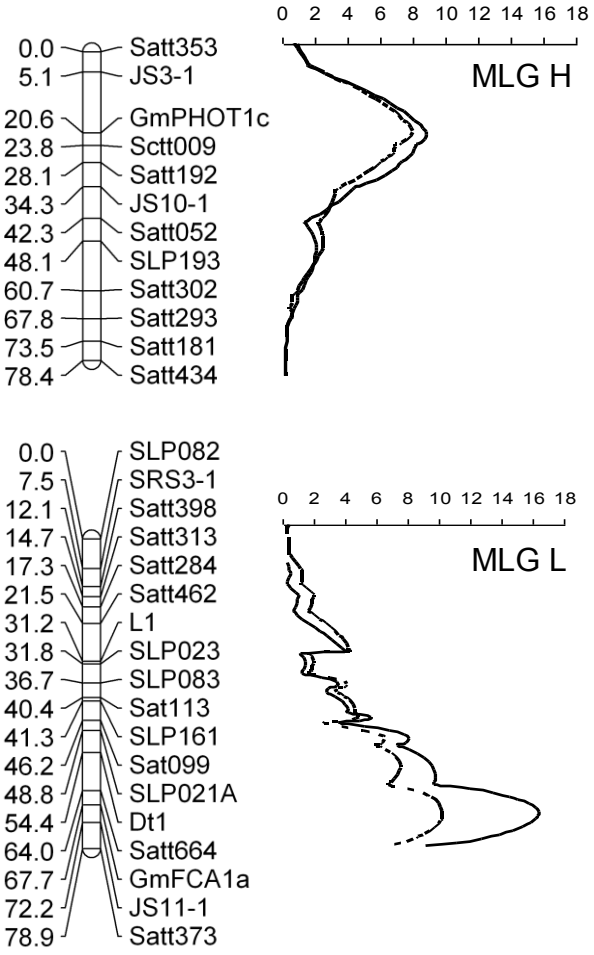

Table 1. Time to maturity means, standard deviation, and range for RILs and parents, V71-370 and PI407162 in 2006 and 2007. Trait is expressed as days after the earliest maturing RIL has reached R8 (95% mature pods based on color change).

	Blacksburg 2006				Blacksburg 2007			
	N	μ	σ	Range	N	μ	σ	Range
RILs	300	23.0	8.4	0-46	300	19.9	8.4	0-42
V71-370	3	25.7	3.8	23-30	3	31.7	0.6	31-32
PI407162	3	19.7	0.6	19-20	3	18.5	4.9	15-22

Table 2. Analysis of variance (ANOVA) for time to maturity trait in an interspecific RIL population developed from a cross between V71-370 and PI407162, grown at Kentland Research Farm (Blacksburg, Virginia), in 2006 and 2007.

	Source	df	MS	F	P
Maturity	RIL	299	121.809	6.485	<0.0001
	Environment	1	1454.297	77.431	<0.0001
	Error	297	18.782		
	Total	598			

Table 3. Significant interval mapping results (LOD>3.0) for observed time to maturity trait in an interspecific RIL population developed from a cross between V71-370 and PI407162, grown at Kentland Research Farm (Blacksburg, Virginia), in 2006 and 2007. Maximum LOD scores were observed at the gene-specific marker on MLG H, while they were observed within 4 cM of the gene-specific marker on MLG L. Additive values represent the mean phenotypic effect of the V71-370 allele in days.

MLG	Year	Position	LOD	% Expl.	Additive
H	2006	20.6	8.7	12.9	-3.0
	2007	20.6	7.9	12.2	-2.9
L	2006	70.7	16.3	27.3	-4.4
	2007	71.7	10.1	19.0	-3.7

Table 4. BLAST results displaying orthologs for two gene models, Glyma12g07890 and Glyma19g40570, that are candidates for genes responsible for the time to maturity trait in soybean.

GmPHOT1 -- MLG H

Organism	Accession	Locus Descriptor	Score	E value
Adiantum capillus-veneris	BAD16729	phototropin 2	1067	0
	BAA95669	phototropin	1041	0
Arabidopsis thaliana	NP_190164	PHOT1	1311	0
	BAH19411	AT3G45780	1078	0
	NP_851210	PHOT2	1055	0
	XP_001766409	PHOTB1	1047	0
	NP_851212.2	PHOT2	1041	0
	BAD94575	Nonphototropic hypocotyl	913	0
Avena sativa	AAC05083	NPH1-1	1227	0
	AAC05084	NPH1-2	1222	0
Lycopersicon esculentum	ABI48276	phototropin-2	1060	0
Oryza sativa (japonica)	BAA84780	Nonphototropic hypocotyl	1211	0
	CAB65325	Nonphototropic hypocotyl	1207	0
	ABG21842	Nonphototropic hypocotyl	1130	0
	BAA84779	Nonphototropic hypocotyl	1031	0
Phaseolus vulgaris	BAD89966	phototropin	1794	0
	BAD89967	phototropin	1424	0
	BAD89968	phototropin	1055	0
Physcomitrella patens subsp. patens	XP_001785726	PHOTB2	1029	0
	XP_001774614	PHOTA2	996	0
	XP_001774256	PHOTA1	990	0
	XP_001763104	PHOTA4	982	0
	XP_001765408	PHOTA3	895	0
Pisum sativum	AAM15725	phototropin 1	1439	0
	AAB41023.2	phototropin-like	1437	0
Solanum lycopersicon	ABN42185.2	phototropin-1	1378	0
	ABS57001	phototropin-2	1060	0
Vicia faba	BAC23098	phototropin	1466	0
	BAC23099	phototropin	1415	0
Zea mays	NP_001104886	blue-light	1218	0
	NP_001147477	phototropin-1	1020	0

GmFCA1 -- MLG L

Organism	Accession	Locus Descriptor	Score	E value
Arabidopsis thaliana	NP_850472	Flowering time control protein	251	1.00E-64
	CAB05388	FCA gamma	227	2.00E-57
	CAB05389	FCA delta	227	2.00E-57
	NP_193363.4	FCA	227	2.00E-57
	NP_849543.2	FCA	227	2.00E-57
	CAB05392	FCA delta	226	3.00E-57
	CAB10407	FCA gamma protein	226	3.00E-57
Pisum sativum	AAX20016	FCA gamma	224	2.00E-56
Brassica napus	AAL61622	FCA gamma	214	1.00E-53
Oryza sativa (indica)	AAQ74973	Flowering time control protein	211	1.00E-52
	AAQ74972	Flowering time control protein	211	1.00E-52
	AAQ17123	Flowering time control protein	211	1.00E-52
Triticum aestivum	AAP84398	FCA protein	206	4.00E-51
	AAP84401	FCA protein	206	4.00E-51
	AAP84409	FCA protein	206	6.00E-51
	AAP84389	FCA protein	206	6.00E-51
	AAP84393	FCA protein	206	6.00E-51
	AAP84400	FCA protein	206	6.00E-51
	AAP84417	FCA-A1	206	6.00E-51
	AAP84396	FCA protein	206	6.00E-51
	AAP84406	FCA protein	206	6.00E-51
	AAP84412	FCA protein	206	6.00E-51
	AAP84395	FCA protein	205	7.00E-51
	AAP84419	FCA-B2	205	7.00E-51
	AAP84399	FCA protein	205	1.00E-50
	AAP84390	FCA protein	205	1.00E-50
	AAP84391	FCA protein	205	1.00E-50
	AAP84405	FCA protein	204	1.00E-50
	AAP84384	FCA protein	204	1.00E-50
	AAP84402	FCA protein	204	1.00E-50
	AAP84410	FCA protein	204	1.00E-50
	AAP84420	FCA-D1	204	2.00E-50
	AAP84381	FCA protein	204	2.00E-50
	AAP84404	FCA protein	204	2.00E-50
	AAP84379	FCA protein	204	2.00E-50
	AAP84397	FCA protein	204	2.00E-50
	AAP84377	FCA protein	204	2.00E-50

	AAP84416	FCA protein	204	2.00E-50
	AAP84383	FCA protein	203	3.00E-50
	AAP84413	FCA protein	203	3.00E-50
	AAP84388	FCA protein	203	3.00E-50
	AAP84415	FCA protein	203	4.00E-50
	AAP84403	FCA protein	202	5.00E-50
	AAP84387	FCA protein	202	5.00E-50
	AAP84407	FCA protein	202	6.00E-50
	AAP84380	FCA protein	202	6.00E-50
	AAP84392	FCA protein	202	8.00E-50
	AAP84408	FCA protein	202	8.00E-50
	AAP84382	FCA protein	201	1.00E-49
	AAP84386	FCA protein	201	1.00E-49
	AAP84378	FCA protein	201	2.00E-49
	AAP84414	FCA protein	201	2.00E-49
	AAP84411	FCA protein	200	2.00E-49
	AAP84376	FCA protein	200	3.00E-49
Hordeum vulgare	ACI16484.2	Flowering time control protein	199	5.00E-49
Lolium perenne	AAT72460	FCA gamma	195	8.00E-48

Table 5: Primers designed for mapping two putative soybean genes, Glyma19g40570 (GmFCA1) and Glyma12g07890 (GmPHOT1), suspected to play a role in time to maturity based on their proximity to maturity QTL observed in a RIL population developed from a cross between V71-370 and PI407162.

ID	Forward	Reverse	Location
GmFCA1a	TATGTTTGGATTGGGTGATTT	GCTGGTTTAGATGAGTGGAAA	Intron 5 of Gm19:46942035
GmFCA1b	TGTACTGCAACCCCTAATTGT	TTTGTAGCAATTTGAAGGTCA	Intron 12 of Gm19:46942035
GmFCA1c	AGTGAAAACTATTTGGTCTTTGC	CAGCACTTACAAATTGGAACC	~10,000 bp upstream of Gm19:46942035
GmFCA1d	GTTGCTGCACTGGGATATTTA	TATTGGATGGAAACAGGAAAC	~7,000 bp upstream of Gm19:46942035
GmPHOT1a	CTCTCCTCGAGACAACTGAAG	TGAAATTTTCTTTTCAGCTC	In predicted exon 22 of Gm12:5532960
GmPHOT1b	TACAGAGAGTGAGAGGGATGG	GGAGGAGCGGAGTCCTTA	In predicted exon 1 of Gm12:5532960
GmPHOT1c	AATAAGCAGTGCCACAAGTTC	GAATTGTGAGGAAGGTTCAGA	In predicted 5'-UTR of Gm12:5532960
GmPHOT1d	TTTTATCCAACATCCCAAATC	TGCTAACAATATTCTCGTCAAAG	~1k upstream of Gm12:5532960

Table 6. Potential soybean paralogs of Glyma19g40570 (GmFCA1), and Glyma12g07890 (GmPHOT1), obtained from subjecting putative coding sequences from the Glyma1 gene model set, available at www.phytozome.net.

ID	Model Id	Chromosome (MLG)	E value
GmFCA1	Glyma03g37950	Gm03 (N)	6E-129
	Glyma20g02640	Gm20 (I)	4E-28
GmPHOT1	Glyma18g17300	Gm18 (G)	4E-79
	Glyma17g31210	Gm17 (D2)	2E-44
	Glyma20g11630	Gm20 (I)	2E-37
	Glyma15g04850	Gm15 (E)	2E-34
	Glyma13g40550	Gm13 (F)	6E-32
	Glyma03g01340	Gm03 (N)	1E-23
	Glyma08g33560	Gm08 (A2)	1E-05
	Glyma08g42040	Gm08 (A2)	7E-13
	Glyma16g19560	Gm16 (J)	1E-05