CHARACTERIZATION OF A DIVERSE USDA COLLECTION OF WILD SOYBEAN (*Glycine soja* SIEBOLD & ZUCC.) ACCESSIONS AND SUBSEQUENT MAPPING FOR SEED COMPOSITION AND AGRONOMIC TRAITS IN A RIL POPULATION

A Dissertation
presented to
the Faculty of the Graduate School
at the University of Missouri-Columbia

In Partial Fulfillment
of the Requirements for the Degree
of Doctor of Philosophy in Plant Breeding, Genetics, and Genomics

---

by
THANG CAO LA
Dr. Andrew Scaboo, Dissertation Supervisor

JULY 2018

The undersigned, appointed by the Dean of the Graduate School,
have examined the Dissertation entitled

CHARACTERIZATION OF A DIVERSE USDA COLLECTION OF WILD SOYBEAN (*Glycine soja* SIEBOLD & ZUCC.) ACCESSIONS AND SUBSEQUENT MAPPING FOR SEED COMPOSITION AND AGRONOMIC TRAITS IN A RIL POPULATION

Presented by THANG CAO LA
A candidate for the degree of
Doctor of Philosophy in Plant Breeding, Genetics, and Genomics,
and hereby certify that, in their opinion, it is worthy of acceptance.

Dr. Andrew Scaboo, Chair

Dr. J. Grover Shannon, Committee member

Dr. D. Jason Gillman, Committee member

Dr. Henry T. Nguyen, Committee member

Dr. Dong Xu, Committee member

ACKNOWLEDGEMENTS

I would like to thank my advisor, Dr. Andrew Scaboo, for accepting me as one of his students and giving me the opportunity to learn and work in his lab, for his patient and supportive guidance of my progress, and for his financial support so that I could pursue the Ph.D. degree in Mizzou.

I would like to thank Dr. Jason Gillman for accepting to be one of my committee members. I am grateful for his patience and calmness to help me solve my problems and correct my mistakes in analyzing my data and writing my dissertation.

I would like to express my appreciation to Dr. Grover Shannon, Dr. Henry T. Nguyen, and Dr. Dong Xu for their valuable comments, advice, and critiques as members of my committee members.

I would like to send my thanks and appreciations to the current and past members of Dr. Andrew Scaboo's group for and assistance with field and lab work. They have taught me many things relating to not only my study but also my social life.

I would like to thank Cuu Long Delta Rice Research Institute, Vietnam for allowing me to take the opportunity to pursue my Ph.D. degree and for their continuous support since I started working there.

I would like to thank my family and my friends, especially my parents, for giving me continuous support and courage to finish my Ph.D. studies. Without their love and support, I doubt that I could accomplish my goals.

Last but not least, I would like to thank all people who have supported my so that I can go this far. Without your help, I could have not been able to accomplish this much.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

LIST OF TABLES

ABSTRACT

The relatively low genomic variation of current U.S. soybean [*Glycine max* (L.) Merill] cultivars constrains the improvement of grain yield, seed quality, and other agronomic traits within soybean breeding programs. Recently, a substantial effort has been undertaken to introduce novel genetic diversity present in wild soybean (*Glycine soja* Siebold & Zucc.) into new elite cultivars, in both public and private applied soybean breeding programs. The objectives of this research were to evaluate the phenotypic diversity within a core collection of 80 *G. soja* plant introductions (PIs) in the United States Department of Agriculture National Genetic Resources Program that were collected in China, Japan, Russia, and South Korea, and to analyze the correlations between agronomic and seed composition traits. Field tests were conducted in Missouri and North Carolina during three years, 2013, 2014, and 2015, in a randomized complete block design (n=3). The phenotypic data collected included plant maturity date, seed weight, and the seed concentration of protein, oil, essential amino acid, fatty acid, and soluble carbohydrates. Analyzing the data from six environments, we found genotype was a significant ($p < 0.0001$) source of variation for maturity date, seed weight, seed protein and amino acids, seed oil and fatty acids, and seed carbohydrates. Significant correlations were observed between numerous traits. The core collection had lower seed weight, higher seed content of protein, linolenic acid, raffinose and stachyose but lower seed content of oil and oleic acid than those of the cultivated soybean lines that were used as checks. The amino acid profile of the core collection was significantly different from that of the checks. An association analysis revealed 19 SNP that were significantly associated with maturity, seed weight, and seed contents of aspartic acid, glutamine, palmitic acid, oleic acid, and linoleic acid. The information and data collected in this study will be invaluable in guiding soybean breeders and geneticists in selecting promising *Glycine soja* plant introductions for research and cultivar improvement.

In addition the identification of quantitative trait loci (QTLs) associated with the contents of seed protein and oil, maturity, branching traits, height, lodging, and yield in a recombinant inbred line (RIL) population developed from one single F$_2$ plant from the cross between Osage and PI593983 was carried out. The mapping population in this study included 164 F$_{4:6}$ recombinant inbred lines (RILs) derived from a cross between Osage, a cultivated soybean variety, and PI593983, a wild soybean accession. Field tests were carried out in Missouri for two years during 2016 and 2017, in a randomized complete block design (n=2). Both protein and oil contents showed high heritabilities. Seed protein and seed oil were negatively correlated (–0.77). A total of 4,374 polymorphic markers were used to construct a genetic linkage map, and nine QTLs for protein content, explained 7.6 to 36.7% of variance, and seven QTLs for oil content, explained for 7.8 to 29.7% of variance, were detected using composite interval mapping. addition we identified eight novel QTLs and confirmed sixteen QTLs associated with maturity (R^2 = 6.4 to 26.3%), plant height (R^2 = 7.4 to 15.5%), and total branch length (R^2 = 9.3% and 14.5%) in individual and across environments, and the ratio of total branch length to plant height (R^2 = 11.8%), yield (R^2 =12.8 and 15.7), and lodging (R^2 = 12.1 and 13.4) in individual studied environments. Sixteen QTLs for maturity, yield, and plant height confirmed previously reported QTLs, and eight QTLs have not been reported before. The results of this study will facilitate the identification of the causative genes for seed protein and oil, maturity, height, lodging, and branching traits, and will help soybean breeder improve soybean performance by developing markers for marker-assisted selection.

Chapter I:

INTRODUCTION

Soybean, soybean production and its increasing demand

Soybean [*Glycine max* (L.) Merr.] is one of the most valuable grain legumes in the world. Cultivated soybean has 349-396g kg^{-1} protein and 190-235g kg^{-1} oil (Liu, 1997; Hildebrand et al., 2008). Some soyfood such as tofu, soy sauce, miso, and natto are traditional food made of soybean. Generally, soybean seed is used to produce protein-rich food and seed oil for animal consumption, industrial purposes, or food processing. Soybean oil is used not only as cooking oil but also as an ingredient in salad oil, mayonnaise, and margarine. Soy meal is the high-protein remains of soybean seed after taking away the oil Liu (1997).In the US, soybean meal provides most of the protein required by livestock or it can be used for human consumption.

It is the highest produced oilseed in the world, accounting for more than 50% of the total production (USDA-FAS 2015). The largest soybean producer in the world is the U.S., followed by Brazil, Argentina, China, and India (USDA-FAS 2015). In the U.S., soybean is also the dominant oilseed crop, with the harvested area increased continuously from 9,572,839 hectares in 1960 to 30,858,494 hectares in 2013, and soybean production increased from 15,106,820 metric tons in 1960 to 91,388,634 metric tons in 2013 (USDA-FAS 2015). US is also the biggest soybean exporter (USDA-FAS 2015), although China is the fourth largest soybean producer, it is also the largest importer of soybean (USDA-FAS 2015).

Soybean seed has about 20% oil and 40% protein based on dry weight (Liu, 1997). Generally, soybean seed is used to produce oil for industrial purpose or food processing.

Soybean oil is used not only as cooking oil but also as an ingredient in salad oil, mayonnaise, and margarine. The byproducts are widely used for livestock and aquaculture feed, and in industrial products such as biodiesel. In the US, soybean meal, the high protein containing byproduct of the oil removal process, is the major ingredient of livestock feed since soy protein is the cheapest protein that provides a complete panel of amino acids that are required by livestock. A small amount of soybean is made into human foods such as such as tofu, soymilk, soy sauce, infant formula, miso, and natto (Liu, 1997).

Branching trait in soybean

The growth and development of branches determine the canopy architecture of a plant and lodging resistance, and, hence, affect yield. Narrow-row/high-density planting (>25 plants m^{-2} and 40-50cm interrow spacing) is widely applied in the USA (Heatherly and Elmore, 2004); however, lower plant densities (<20 plants m^{-2}) are used in Korea to avoid lodging and disease, and lower the costs of seed and labor (Cho and Kim, 2010). Agudamu et al. (2016) and Cox et al. (2010) stated that the decrease in plant density could be compensated by the increased number of branches on the main stems.

The branching trait is predominantly influenced by environmental conditions including population density, nutrient availability, row spacing, and planting date (Acock and Acock, 1987; Weaver et al., 1991; Asanome and Ikeda, 1998; Foroutan-pour et al., 1999). Acock and Acock (1987) stated that the number of branches and branch length reduced in response to an increased number of shading weeks. Settimi and Board (1988) found that an early planting date with optimal photoperiod improved the branch distribution of soybean. Schon and Blevins (1990) reported that soybean had more branches when boron was applied. Settimi and Board (1988) found that planting date with optimal photoperiod resulted in an improvement of the branch distribution in soybean. Board and Kahlon (2013) stated that the variation in branch development under subnormal density resulted in differences in yield. Shim et al. (2017) observed differences in branch development among soybean varieties in which American/Chinese soybean varieties had fewer branches than Korean/Japanese soybean varieties. Because of the dominant influences of environment on soybean branching phenotype, genomic studies about this trait are limited.

Chemical composition of soybean seed

Protein and oil

Typically, soybean seed has 200 g kg^{-1} oil and 400 g kg^{-1} protein based on dry weight (Liu, 1997; Wilson, 2004). The soybean seed content of protein and oil affect the price of soybean, and soybean with high content of protein and oil is preferable by soybean processors and consumers (Brumm and Hurburgh, 1990; Orf and Helms, 1994).

Soybean oil is used not only as cooking oil but also as ingredient in salad dressings, mayonnaise, and margarine (Wilson, 2004; Wilson, 2008). Soybean oil can also be used in industrial products such as biodiesel fuel and plastics (http://www.soystats.com). A small amount of soybean is made into human foods such as tofu, soymilk, soy sauce, infant formula, miso, and natto (http://www.soyfoods.org/). In the US, soybean meal, the high protein containing soybean product after the oil removal process, is the major ingredient of livestock feed since soy protein is the cheapest protein that provides a complete panel of amino acids required by livestock (Chiari et al., 2004; Yesudas et al., 2013). Pettersson and Pontoppidan (2013) reported that soybean meal and oil accounted for approximately 60% and 40% of soybean’s value, respectively. Willis (2003) stated that soybean seed with

≥415 g kg^{-1} protein content and 220 g kg^{-1} based on dry weight was the standard to achieve soybean meal with ≥475 g kg^{-1} protein.

The contents of oil and protein in soybean seed range from 81 g kg^{-1} to 279 g kg^{-1} and from 341 g kg^{-1} to 568 g kg^{-1}, respectively; however, there are only a few accessions with high contents of both protein and oil in the USDA Soybean Germplasm Collection (Chung et al., 2003; Wilson, 2004). These traits are quantitative traits, controlled by many genes, and strongly affected by environment (Lee et al., 2007). Chung et al. (2003) stated that the contents of soybean protein and oil varied significantly in different regions of the USA. Specht et al. (2001) reported the depression of seed protein due to drought stress. Bellaloui et al. (2015) stated that cool temperature had negative effect on seed storage protein. And oil content increases when the temperature is warm (Carrera et al., 2011).

Fatty acids in soybean seed

There are five main types of fatty acid in soybean seed oil. The general fatty acid profile is of 54% linoleic acid, 22% oleic acid, 10% palmitic acid, 10% linolenic acid, and 4% stearic acid (Wilson, 2004). Palmitic acid, stearic acid, oleic acid, linoleic acid, and linolenic acid can be written as 16:0, 18:0, 18:1, 18:2, and 18:3, respectively. The first and second numbers represent the number of carbon atoms and double-bonds, respectively, in the molecule of those acids. Based on the number of double bonds in their molecule, these fatty acids are either categorized in to; saturated acids without a double bond (palmitic acid and stearic acid), monosaturated acid with one double bond (oleic acid), or polyunsaturated acid with more than one double-bond (linoleic acid and linolenic acid).

Unsaturated fatty acids are less oxidatively stable than saturated fatty acids. Under cold weather, saturated fatty acid may inhibit the flow of biodiesel because of its low melting temperature. Consuming foods containing saturated fatty acid may increase the cholesterol level in blood and cause heart problems. In 2004, Wilson stated that soybean oil with lower saturated fat would be more desired to improve cold flow.

Both linoleic acid and linolenic acid are essential for human body. Linoleic acid, or omega-6, plays important role in brain and heart function, and regulating normal growth and development of human (Uauy et al., 2000). Linolenic, or omega-3, is essential for development the eyes and brain of a fetus, and for the improvement of health (Uauy et al., 2000). Because of these health benefits, soybean seed with high content of polyunsaturated acids is preferable when the seed is consumed directly without cooking process (Brouwer et al., 2004) . White (2007) reported that fatty acid with more double bonds would be oxidized easier. Low oxidative stability of polyunsaturated fatty acid is desirable for drying oils (Cahoon, 2003). To develop soybean line with high linolenic for this purpose, Eckert et al. (2006) overexpressed the fatty acid desaturase 3 (FAD3) and obtained soybean lines with up to 50% linolenic acid content. However, because of low stability under high temperature, oil with high content of polyunsaturated fatty acid for cooking purpose becomes oxidized easily and induces off-flavors (Lee et al., 2009). Warner and Fehr (2008) also reported that the processing food products using this kind of oil had shorter storage time. In addition, using oil with a high content of polyunsaturated fatty acid as biofuel usually also has problems with the flow of the fuel and oil filter because it causes viscous materials (Yadav, 1996).

Lee et al. (2007), and Warner and Gupta (2005) stated that oil with high content of oleic acid would be desirable because it would be more stable for greater application in food processing and industrial products. Therefore, Fehr (2007), Lee et al. (2007), and Wilson (2004) suggested that soybean lines with high oleic acid would be desirable for wider range of applications.

Soluble carbohydrates

Typically, soluble carbohydrates in soybean seed consist of 41.3-67.5% sucrose, 5.2-15.8% raffinose, 12.1-35.2% stachyose, glucose, and fructose (Yazdi-Samadi, 1977; Eldridge et al., 1979). Pinitol, myo-inositol, verbascose, galactose, arabinose, and mannose have also been found in soybean seed, though on in very low amount (Yazdi-Samadi, 1977; Schweizer, 1978; Eldridge et al., 1979). Glucose, fructose, and sucrose are desirable carbohydrates in soybean seed because they induce sweet taste and are easily digestible (Joseph, 1975). Raffinose and stachyose belong to raffinose family of oligosaccharides. Monogastric animals cannot digest RFOs because of the lacking of enzyme that breaks α-galactosidic linkages in these oligosaccharides. RFOs are classified as anti-nutritional factors, and their presence in soybean meal may cause flatulence, diarrhea, and other problems relating to digestion (Schweizer, 1978; Sebastian et al., 2000). When soybean oligosaccharides were added to swine diets, the nutrient digestibility decreased from 1.1 to 7.4 units (Leske and Coon, 1999). One of the methods to reduce the amount of RFOs in soybean meal is using ethanol extraction; however, this method is not economically feasible because it also reduces the amount of sucrose (Leske and Coon, 1999). Therefore, reducing RFO content in soybean seed through plant breeding is more approachable and efficient. Quantitative trait loci for soybean sucrose and oligosaccharide content have been reported (Maughan et al., 2000; Kim et al., 2005).

Amino acids

Soybean meal is the major source of essential amino acids of poultry and swine. Friedman and Brandon (2001) reported that the amino acid profile of soybean meal played more important role than the protein content of it did.

The amino acid composition in soybean seed is not ideal for monogastric animal feed. Soybean seed protein is deficient in sulfur-containing amino acis, methionine (Met), threonine (Thr), and lysine (Lys), when it is used as main source of protein for animal consumption. Monogastric animals cannot synthesize Met, Thr, and Lys, so these essential amino acids must be provided from the animals' diet. Supplementary ingredients are utilized to overcome this limitation. Imsande (2001) reported that the annual costs of supplementing methionine by pork nurseries and poultry growers was about $100 million. In addition, the supplementation of Met, a sulfur-containing amino acid, may lead to the production of undesired volatile sulfides (George and De Lumen, 1991). Therefore, the goals of soybean breeding are not only in improving amino-acid assimilation but also increasing the relative percentage of methionine (Met), Lysine (Lys), and threonine (Thr) (Durham, 2003). Clarke and Wiseman (2000) calculated that if Lys, Met, and Thr concentrations increase by 10%, the commercial meal value will increase by $5.1 to $10.6, $3.0, and $8.8/T, respectively. To increase soybean value and decrease negative effects from using supplements, it is necessary to develop soybean cultivar with an improved amino acid composition.

Correlation between oil fatty acid profile, oil content, soluble carbohydrates, and protein content

The contents of oil and protein in soybean seed vary greatly, from 8.1% to 27.9% and from 34.1% to 56.8%, respectively (Wilson, 2004). These traits are quantitative, controlled by many genes, and strongly affected by environment (Lee et al., 2007). Although some QTLs relating to higher protein without decreasing oil content have been reported (Lee et al., 1996; Eskandari et al., 2013), a strong negative correlation between these traits has also been reported (Shannon et al., 1972; Cober and D Voldeng, 2000; Lee et al., 2007; Ramteke et al., 2010). Because of this negative correlation, it is very challenging to increase the content of either seed oil or protein without decreasing the other. Chung et al. (2003) suggested that this negative correlation might be due to either a single pleiotropic QTL or tightly association between high protein allele(s) and low oil allele(s). Until now, 186 QTLs and 154 QTLs have been found that they are associated with the content of seed oil and protein, respectively (www.SoyBase.org, "SoyBase browser", accessed 05/20/2018).

In seed oil, oleic acid content is negatively correlated with other fatty acids (Brace et al., 2011; La et al., 2014). These negative correlation could be explained by the flux in the lipid biosynthesis pathway (Brace et al., 2011).

Hymowitz (1972) reported a positive correlation between the total sugar and oil content in soybean seed; however, both of them were negatively correlated with protein content.

Genetic diversity

Both wild soybean (*Glycine soja* Sieb. and Zucc.) and cultivated soybean [*Glycine max* (L.) Merr.] are members of the subgenus Soja, which belongs to the genus Glycine within the family Leguminosae. Wild soybean has been considered the closest relative of cultivated soybean (Carter et al., 2004). Genetically, both have 20 chromosomes (2n = 40), can be sexually crossed, carry out normal meiotic chromosome pairing, and produce viable, fertile hybrids (Carter et al., 2004). Morphologically, wild soybean differs from cultivated soybean in that wild soybean flowers later, lodges more, has more lateral branches, produces small black seed that are dispersed by shattering (Funatsuki et al., 2006; Liu et al., 2007). Kim et al. (2010) reported that there was a 0.31% difference in areas where the SNPs and indels were aligned precisely when they aligned the 48.8-Gb Illumina Genome Analyzer short DNA reads of *Glycine soja* var. IT182932 to the *Glycine max* reference genome. They also stated that about 21% of the expanded area in cultivated soybean comparing to wild soybean contained transposable elements.

Soybean predominantly self-pollinates, leading to increased homozygosity and overall decreased genomic variation as compared to outcrossing species (Lam et al., 2010). The genomic variation of cultivated soybean is even less due to the strong pressure of domestication and modern plant breeding (Tanksley and McCouch, 1997). Founding events are activities that a few individuals represent a crop in a new region or a few cultivars are used to develop a crop. According to Halliburton (2004), domestication and founding events may cause a decrease in genetic diversity, an increase in linkage disequilibrium, changes in allele frequencies, and elimination of rare alleles in the targeted population. The reduction of genetic diversity can have negative effects on future genomic gain, makes the crop more susceptible to emerging pests and disease, and endanger future food-safety (Council, 1972; Esquinas-Alcázar, 2005).

According to Gizlice et al. (1994) only 80 out of at least 45,000 unique Asian landraces account for more than 99% of 258 North American public soybean cultivars released during the period from 1947 to 1988. And 17 out of these 80 can define up to 84% of the genetic base of those 258 cultivars (Gizlice et al., 1994; Hyten et al., 2006). Thorne and Fehr (1970) stated that the tremendous increase in soybean yield in the U.S. was mostly due to using elite soybean lines as parents to make crosses. However, this progress accelerates the erosion of cultivated soybean's gene pool (Concibido et al., 2003). According to Zhou et al. (2015), the landraces and improved cultivars in their study lost more than 50% sequences relating to nematode resistance in *G. soja*. Zhou et al. (2015) also reported that the studied landraces and cultivars had decreased genetic diversity and increased linkage disequilibrium. Therefore, the tendency of soybean to self-pollinate combined with the narrow genetic base of North American public soybean varieties creates a need for genetically diverse germplasm to allow for improvement of agronomic and seed quality traits.

To address the problem of narrow genetic diversity in soybean, the United States Department of Agriculture (USDA) maintains a soybean germplasm collection of 1,168 wild soybean accessions and 18,480 *G. max* accessions (Song et al., 2015). Crop relatives and exotic germplasm are important genetic resources for improving agriculture productivity, yet wild soybean (*Glycine soja*) has been largely under-utilized in breeding efforts focused on broadening the narrow genetic background of cultivated soybeans (Jin et al., 2003; Lee et al., 2008). Hymowitz (1970) indicated that wild soybean (*Glycine soja*

Sieb. et Zucc) would most possibly be the ancestor of cultivated soybean, although there could be some cytogenic barriers. Using wild germplasm in a breeding program is limited because the beneficial genes relating to yield and seed quality are usually in genetic linkage with undesired traits such as logging and seed shattering (Concibido et al., 2003). To break the genetic linkages, backcrossing to the adapted parent is one solution, and this process makes the breeding program more costly and time-consuming. Using molecular markers in breeding (Concibido et al., 2003; Palomeque et al., 2009) programs makes the process faster and more accurate because the selection is based on genotype rather than on phenotype, and allows the transfer of beneficial genes into elite lines with none or fewer unfavorable genes (Tanksley and McCouch, 1997). Molecular markers have been used to utilize many genes from exotic or unadapted germplasm. Many of these markers relating to pest and disease resistance include soybean cyst nematode (Webb et al., 1995; Concibido et al., 1997), corn earworm (Rector et al., 2000; Terry et al., 2000), brown stem rot (Klos et al., 2000), soybean mosaic virus (Hayes et al., 2000). Some studies have discovered markers relating to seed quality traits including seed oil and seed protein, and yield (Qiu et al., 2011; Kim et al., 2012).

Soybean germplasm core collection

Utilization of over 1,100 accessions in the wild soybean germplasm collection for soybean breeding is unmanageable and impractical for public and private breeders using conventional breeding techniques and marker assisted selection (MAS). While the use of MAS has increased the utility of wild soybean to breeders, the agronomically undesirable traits of wild germplasm can be avoided during population development by backcrossing with elite varieties and by evaluating large segregating populations (Ertl and Fehr, 1985;

Carpenter and Fehr, 1986; LeRoy et al., 1991; Sebolt et al., 2000; Kabelka et al., 2006; Zhang and Huang, 2011; Akpertey et al., 2014; Shivakumar et al., 2016). To address the problem, Frankel and Brown (1984) suggested the establishment of a core collection with a limited number of accessions derived from the original collection; representing about 10% of the full collection. This core collection should represent the genetic diversity of the original collection with the lowest number of redundant accessions. A core collection is easier to evaluate and more efficient to utilize. Core collections were successfully developed with multiple crops including maize (*Zea mays*), rice (*Oryza sativa*), wheat (*Triticum aestivum*), and peanut (*Arachis hypogaea*) (Holbrook et al., 2000; Coimbra et al., 2009; Bordes et al., 2011; Liu et al., 2016). Soybean core collections exist in East Asia (Qiu et al., 2013) and Brazil (Priolli et al., 2013). Domesticated soybean core collections composed of a portion of the 18,480 USDA *G. max* accessions have also been developed (Oliveira et al., 2010). Even smaller mini-core collections that represent the most diverse 1% of the accessions have been developed for multiple crops including maize (*Zea mays*), rice (*Oryza sativa*), wheat (Triticum aestivum), and peanut (*Arachis hypogaea*) (Holbrook et al., 2000; Coimbra et al., 2009; Bordes et al., 2011; Liu et al., 2016).

QTL mapping

Seed content of protein and oil

There have been 322 QTLs and 240 QTLs associated with the content of seed oil and protein, respectively (http://www.soybase.org, "SoyBase browser", accessed 05/20/2018). These QTLs have been located on all chromosomes (http://www.soybase.org, "SoyBase browser", accessed 05/20/2018). Brummer et al. (1997) studied eight different soybean populations developed in Mid-west USA in multiple environments. They reported nine QTLs significantly associated with seed protein content and seven QTLs significantly associated with seed oil content. Hyten et al. (2004) evaluated 131 F_6-derived recombinant inbred lines developed from a cross between Essex and Williams in six different environments. They identified four QTLs for protein and six QTLs for oil. Wang et al. (2014) studied two different populations (SD02-4-59 × A02-381100 and SD02-911 × SD00-1501) in multiple environments and detected 11 QTLs for protein content and 16 QTLs for oil content. Most of reported QTLs have not been confirmed (Panthee et al., 2005; Pathan et al., 2013; Phansak et al., 2016).

Diers et al. (1992) studied a population developed from a cross between an experimental line of Iowa State University [*Glycine max* (L.) Merr.] and a wild soybean plant introduction PI468916 (*G. soja* Sieb. and Zucc.). They reported two QTLs, one on Chr. 15 and one on Chr. 20, associated with high protein. Nichols et al. (2006) and Fasoula et al. (2004) confirmed these QTLs by following the guidelines that was suggested by Soybean Genetics Committee (http://soybase.org/). The QTLs on Chr. 15 and 20 were named cqSeed protein-001 (Fasoula et al., 2004) and cqSeed protein-003, respectively (Nichols et al., 2006). Numerous QTLs with large effect on protein content have been located in these two locations with different significance levels (Sebolt et al., 2000; Chung et al., 2003; Wang et al., 2014; Warrington et al., 2015; Kim et al., 2016; Phansak et al., 2016; Qi et al., 2016). The difference in significance levels could be explained by the differences in population's source and size, the rate of recombination, the extend of linkage disequilibrium in the studied population, and the dependence of QTL effects on the environment (Brzostowski and Diers, 2017; Patil et al., 2017). Kim et al. (2016) studied a backcross population by using Williams 82 as the recurrent parent and PI407788A, a high protein line accession from Korea. They identified a QTL, which is from PI407788A and located on chromosome 15, associated with higher protein and lower oil content. A QTL that had major effect on seed protein and amino acid content was identified on Chr. 20 when Warrington et al. (2015) studied the population developed from the cross between Benning and Danbaekkong. Warrington et al. (2015) stated that the favorable allele was from Danbaekkong, explained approximately 55% of the seed protein's variation, and showed little negative effect on yield in the studied population. In the studies of Nichols et al. (2006), Chung et al. (2003), and Sebolt et al. (2000), the reported QTLs showed stronger negative effect on yield. Patil et al. (2017) suggested that the allele from Danbaekkong was different from the other reported QTLs or the genetic background of Danbaekkong mitigated the negative effect on yield of the QTL (Chen et al., 2008; Qi et al., 2011).

Maturity date

The flowering time and maturity of soybean are important agronomic traits. These traits allow the development and adaptation of soybean in different geographical regions. Whigham and Minor (1978) stated that these traits were affected by day length, temperature, and plant genotype. Soybean is considered short-day plant whose life cycle is shorter when days are short; although there have been some day-neutral soybean genotypes reported (Metz et al., 1985; Nissly et al., 1981).

Ten *E* loci, which are named from *E1* to *E10*, and one *J* locus have been reported to be associated with maturity in soybean (Bernard, 1971; Buzzell, 1971; Buzzell and Voldeng, 1980; McBlain and Bernard, 1987; Ray et al., 1995; Bonato and Vello, 1999; Cober and Voldeng, 2001; Cober and Morrison, 2010; Fanjiang et al., 2014; Samanfar et al., 2017). Among these, the causal genes of *E1, E2, E3*, *E4*, *E9,* and *J* have been reported at the molecular level (Li et al., 2008; Watanabe et al., 2009; Watanabe et al., 2011; Xia et al., 2012; Lu et al., 2017). The *E1* locus was reported to control maturity and flowering time in soybean by Bernard (1971). *E1* is located on chromosome 6 and encodes a transcription factor that is specific in legume (Xia et al., 2012). This transcription factor has a putative nuclear localization signal (Xia et al., 2012). *E2* is located on chromosome 10 and is a co-ortholog of *GIGANTEA*, a flowering gene in *Arabidopsis thaliana*. *E3* and *E4* are located on chromosome 19 and 20, respectively, and are *PHYA* homologs which regulate the flowering reactions to long-day conditions when there are different ratios of red-to-far-red quantum. *E1, E2, E3,* and *E4* downregulate the orthologs of *FLOWERING LOCUS T* (*GmFT2a* and *GmFT5a)* in *Arabidopsis* and delay flowering time as well as maturity under long day condition (Kong et al., 2010; Thakare et al., 2011; Watanabe et al., 2011; Xia et al., 2012; Jiang et al., 2014). Different combinations of various alleles at *E1, E3,* and *E4* determine the responses of soybean flowering, pre-flowering, and post-flowering to photoperiods, and contribute to the geographical adaptation of soybean (Tsubokura et al., 2013; Xu et al., 2013; Jiang et al., 2014). *E9* is located on chromosome 16, and the causal gene of *E9* is *FT2a*, an *Arabidopsis* FLOWERING LOCUS T's ortholog, whose recessive alleles lead to late flowering (Fanjiang et al., 2014; Zhao et al., 2016). *J* is located on chromosome 4 and is one of the orthologs of flowering time gene *ELF3* in Arabidopsis (Lu et al., 2017).

Branching traits

Chen et al. (2007) studied 154 recombinant inbred lines (RILs) from a cross between Charleston, an American semi-dwarf variety, and Dongnong 594, a high protein line, and found seven QTLs that were mapped for branch number. By using simple sequence repeat loci, Hwang et al. (2008) could differentiate Japanese and Korean cultivars which showed variation in branching phenotypes. Studying 172 RILs developed from a cross between Tokei 758 and To-8E, Sayama et al. (2010) reported five QTLs that were significantly associated with branching number. Shim et al. (2017) identified one novel QTL and confirmed three previously reported QTL associated with branching in 200 RILs developed from a cross between Jiyu69, a Chinese elite cultivar, and SS0404-T5-76, an elite high-yielding line. YANG et al. (2017) constructed a genetic linkage map of a F2 population developed from the cross between Toyomusume, a Japanese cultivar, and Suinong 10, a Chinese cultivar by using 1,306 polymorphic single nucleotide polymorphism (SNP) markers. They identified one major QTL, qBR6_1, for branching number, and located this QTL near maturity gene E1 on chromosome 6.

Genome wide association study (GWAS)

To discover QTLs for different traits of soybean, it is proven sufficient to 100 to 200 markers and a few hundred lines in mapping population which segregated from biparents (http://www.soybase.org, “SoyBase browser”, accessed 05/20/2018; Sonah et al. (2015))(www.soybase.org; Sonah et al., 2015). One drawback of conventional QTL mapping lies in its low resolution because of the limited recombination in mapping populations (Balasubramanian et al., 2009). The estimated QTL interval usually cover several cM, and this genetic distance translates into a wide genomic area with many candidate genes. By using a diverse collection of plant accessions with higher linkage disequilibrium and higher density of markers than those in family-based population mapping, a higher resolution can be achieved (Myles et al., 2009). Another drawback of population-based mapping is that it may lead to false positives at loci with allele frequencies that vary across subpopulations and have no true effects on the studied trait (Vilhjálmsson and Nordborg, 2013). This problem can be fixed by assessing and taking into account the similarity among individuals in the population (Yu et al., 2006; Segura et al., 2012).

Another solution to overcome the limitation of conventional mapping is by taking advantages of recombination events that occurred a long time ago (Huang and Han, 2014). The applications of this genome-wide association approach have successfully identified numerous candidate loci correlated with different traits in many species (Appels et al., 2013; Korte and Farlow, 2013). In human, GWAS has been used to identify loci controlling the susceptibility to a certain disease. Comparing to GWAS in crops, GWAS in human requires larger population size and higher number of markers because of the low rate of phenotypic contribution of the identified loci (Huang and Han, 2014).

GWAS has been applied successfully in many plant species including Arabidopsis, maize, rice, sorghum, foxtail millet, and soybean (Huang et al., 2010; Balint-Kurti et al., 2011; Kump et al., 2011; Tian et al., 2011; Zhao et al., 2011; Huang et al., 2012; Li et al., 2012; Jia et al., 2013; Li et al., 2013; Morris et al., 2013; Verslues et al., 2014). Among these crops, rice and maize are considered the major models due to their greatly developed and published resources (Huang and Han, 2014). Rice GWAS is slightly different from maize GWAS. The differences lie in the balance between power and resolution when mapping these selfing and outcrossing crops. Because of slow rate of LD decay, GWAS resolution in rice cannot separate a single gene (Huang and Han, 2014). For the same reason, applying sequencing and imputation of missing data in rice is very efficient and powerful. The frequencies of about 44% of the SNPs in rice genome are less than 0.05, and GWAS power to identify rare alleles is weak (Huang and Han, 2014). Using large sample size or generating many biparental populations may increase GWAS when rare alleles need to be identified. On the contrary, maize, an outcrossing crop, has rapid LD decay and high genetic diversity. Generally, GWAS resolution in maize may reach single-gene level. However, maize GWAS requires a large number of marker and is very costly because maize genome is large and has numerous repeats and paralog sequences.

To detect genotype-phenotype correlation in crop GWAS, a mixed model is usually used (Yu et al., 2006; Bradbury et al., 2007). The problem with this model is that it cannot be practically applied when a large number of markers (about 1 million SNPs) and a big population are used in GWAS (Huang and Han, 2014). To increase computation speed, Efficient Mixed-Model Association eXpedited (EMMAX) program, accelerated mixed model, compressed mixed linear model, and other advances have been used (Keurentjes et al., 2007; Zhang et al., 2009; Zhang et al., 2010; Lippert et al., 2011; Lipka et al., 2012; Seren et al., 2012; Zhou and Stephens, 2012). In addition, more statistical models have been developed to deal with false positive and false negative correlations, and provide more options be selected according to level of correlation (Korte and Farlow, 2013)

Within the loci located, there are usually several genes and only one of them is the causal gene. Therefore, more analyses and experiments should be carried out to pinpoint this gene. Additional clues about GWAS loci can be achieved by carefully examining the gene annotation, expression information, and the collection of variants (Huang and Han, 2014). To be able to conclusively point out the causal genes, more studies need to be performed.

The publication of the SoySNP50k data and other high through-put genotyping methods have improved the ability to acquire the desired marker coverage on thousands of soybean genotypes (Song et al., 2013; Sonah et al., 2015). Mamidi et al. (2014) studied a GWAS population of soybean and identified seven genomic regions that are significantly associated with iron deficiency chlorosis in this species. Performing a GWAS study for seed protein and oil in a collection of 298 soybean accessions, Hwang et al. (2014) confirmed previously reported QTL for protein and oil content and narrowed down the regions of these QTL. In 2015, Sonah et al. performed a GWAS for eight traits via genotyping-by-sequencing approach and identified SNPs significantly associated with the studied traits.

References

Acock, B., and M.C. Acock. 1987. Periodic shading and the location and timing of branches in soybean. Agron. J. 79: 949.

Agudamu, T. Yoshihira, and T. Shiraiwa. 2016. Branch development responses to planting density and yield stability in soybean cultivars. Plant Prod. Sci. 19: 331-339.

Akpertey, A., M. Belaffif, G.L. Graef, M.A.R. Mian, J.G. Shannon, P.B. Cregan, et al. 2014. Effects of selective genetic introgression from wild soybean to soybean. Crop Sci. 54: 2683-2695.

Appels, R., R. Barrero, and M. Bellgard. 2013. Advances in biotechnology and informatics to link variation in the genome to phenotypes in plants and animals. Funct. Integr. Genomics 13: 1-9.

Asanome, N., and T. Ikeda. 1998. Effect of branch direction's arrangement on soybean yield and yield components. J. Agron. Crop Sci. 181: 95-102.

Balasubramanian, S., J.O. Borevitz, J. Chory, D. Weigel, C. Schwartz, A. Singh, et al. 2009. QTL mapping in new Arabidopsis thaliana advanced intercross-recombinant inbred lines. PLoS One 4: e4318.

Balint-Kurti, P.J., K.L. Kump, S. Kresovich, J.B. Holland, P.J. Bradbury, A.R. Belcher, et al. 2011. Genome-wide association study of quantitative resistance to southern leaf blight in the maize nested association mapping population. Nat. Genet. 43: 163-168.

Bellaloui, N., H.A. Bruns, H.K. Abbas, A. Mengistu, D.K. Fisher, and K.N. Reddy. 2015. Agricultural practices altered soybean seed protein, oil, fatty acids, sugars, and minerals in the Midsouth USA. Frontiers in plant science 6: 31.

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

Board, J.E., and C.S. Kahlon. 2013. Morphological responses to low plant population differ between soybean genotypes. Crop Sci. 53: 1109-1119.

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.

Bordes, J., C. Ravel, J. Le Gouis, A. Lapierre, G. Charmet, and F. Balfourier. 2011. Use of a global wheat core collection for association analysis of flour and dough quality traits. J. Cereal Sci. 54: 137-147.

Brace, R.C., W.R. Fehr, and S.R. Schnebly. 2011. Agronomic and seed traits of soybean lines with high oleate concentration. Crop Sci. 51: 534-541.

Bradbury, P.J., Z. Zhang, D.E. Kroon, T.M. Casstevens, Y. Ramdoss, and E.S. Buckler. 2007. TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics 23: 2633-2635.

Brouwer, I.A., M.B. Katan, and P.L. Zock. 2004. Dietary alpha-linolenic acid is associated with reduced risk of fatal coronary heart disease, but increased prostate cancer risk: a meta-analysis. The Journal of Nutrition 134: 919-922.

Brumm, T.J., and C.R. Hurburgh. 1990. Estimating the processed value of soybeans. J. Am. Oil Chem. Soc. 67: 302-307.

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

Brzostowski, L.F., and B.W. Diers. 2017. Agronomic evaluation of a high protein allele from PI407788a on chromosome 15 across two soybean backgrounds. Crop Sci. 57: 2972-2978.

Buzzell, R.I. 1971. Inheritance 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 7: 13.

Cahoon, E.B. 2003. Genetic enhancement of soybean oil for industrial uses: Prospects and challenges.

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

Carrera, C.S., M.J. Martínez, J. Dardanelli, and M. Balzarini. 2011. Environmental variation and correlation of seed components in nontransgenic soybeans: Protein, oil, unsaturated fatty acids, tocopherols, and isoflavones. Crop Sci. 51: 800-809.

Carter, T.E., R. Nelson, C.H. Sneller, and Z. Cui. 2004. Genetic diversity in soybean. Soybeans: Improvement, Production and Uses, eds Boerma HR, Specht JE (Am Soc Agron, Madison, WI), pp 303–416.

Chen, P., C.H. Sneller, T. Ishibashi, and B. Cornelious. 2008. Registration of high-protein soybean germplasm line R95-1705. Journal of plant registrations 2: 58-59.

Chiari, L., N.D. Piovesan, L.K. Naoe, I.C. José, J.M.S. Viana, M.A. Moreira, et al. 2004. Genetic parameters relating isoflavone and protein content in soybean seeds. Euphytica 138: 55-60.

Cho, Y.S., and S.D. Kim. 2010. Growth parameters and seed yield compenets by seeding time and seed density of non-/few branching soybean cultivars in drained paddy field. Asian J. Plant Sci. 9: 140-145.

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

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

Cober, E.R., and H. D Voldeng. 2000. Developing high-protein, high-yield soybean populations and lines ECORC Contribution No. 991410. Crop Sci. 40: 39-42.

Cober, E.R., and M.J. Morrison. 2010. Regulation of seed yield and agronomic characters by photoperiod sensitivity and growth habit genes in soybean. Theor Appl Genet 120.

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.

Coimbra, R.R., G.V. Miranda, C.D. Cruz, D.J.H. Silva, and R.A. Vilela. 2009. Development of a Brazilian maize core collection. Genet. Mol. Biol. 32: 538-545.

Concibido, V., V.B. La, P. McLaird, N. Pineda, J. Meyer, L. Hummel, et al. 2003. Introgression of a quantitative trait locus for yield from Glycine soja into commercial soybean cultivars. Theor. Appl. Genet. 106: 575-582.

Concibido, V.C., D.A. Lange, R.L. Denny, J.H. Orf, and N.D. Young. 1997. Genome mapping of soybean cyst nematode resistance genes in 'Peking', PI 90763, and PI 88788 using DNA markers. Crop Sci. 37: 258-264.

Council, N.R. 1972. Genetic vulnerability of major cropsNational Academy of Sciences, Washington.

Cox, W.J., J.H. Cherney, and E. Shields. 2010. Soybeans compensate at low seeding rates but not at high thinning rates. Agron. J. 102: 1238-1243.

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.

Durham, D. 2003. The United Soybean Board's better bean initiative: Building United States soybean competitiveness from the inside out. AgBioForum 6: 23-26.

Eckert, H., B.J. LaVallee, B.J. Schweiger, A.J. Kinney, E.B. Cahoon, and T. Clemente. 2006. Co-expression of the borage Δ6 desaturase and the Arabidopsis Δ15 desaturase results in high accumulation of stearidonic acid in the seeds of transgenic soybean. Planta 224: 1050-1057.

Eldridge, A.C., L.T. Black, and W.J. Wolf. 1979. Carbohydrate composition of soybean flours, protein concentrates, and isolates. J. Agric. Food Chem. 27: 799-802.

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

Eskandari, M., E.R. Cober, and I. Rajcan. 2013. Genetic control of soybean seed oil: II. QTL and genes that increase oil concentration without decreasing protein or with increased seed yield. Theor. Appl. Genet. 126: 1677-1687.

Esquinas-Alcázar, J. 2005. Protecting crop genetic diversity for food security: political, ethical and technical challenges. Nat. Rev. Genet. 6: 946-953.

Fanjiang, K., N. Haiyang, C. Dong, L. Ying, W. Fangfang, W. Jialin, et al. 2014. A new dominant gene conditions early flowering and maturity in soybean. Crop Sci. 54: 2529-2535.

Fasoula, V.A., D.K. Harris, and H.R. Boerma. 2004. Validation and designation of quantitative trait loci for seed protein, seed oil, and seed weight from two soybean populations. Crop Sci. 44: 1218-1225.

Fehr, W.R. 2007. Breeding for modified fatty acid composition in soybean. Crop Sci. 47: S-72-S-87.

Foroutan-pour, K., P. Dutilleul, and D.L. Smith. 1999. Soybean canopy development as affected by population density and intercropping with corn: fractal analysis in comparison with other quantitative approaches. Crop Sci. 39: 1784-1791.

Frankel, O.H., and A.H.D. Brown. 1984. Plant genetic resources today: A critical appraisal. Crop Genetic Resources: Conservation & Evaluation (J.H.W. Holden and J.T. Williams, eds.). George Alien & Unwin Ltd., London: 249-257.

Friedman, M., and D.L. Brandon. 2001. Nutritional and health benefits of soy proteins. J. Agric. Food Chem. 49: 1069-1086.

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.

George, A.A., and B.O. De Lumen. 1991. A novel methionine-rich protein in soybean seed: identification, amino acid composition, and N-terminal sequence. Journal of Agricultural and Food Chemistry 39: 224-227.

Gizlice, Z., T.E. Carter Jnr, and J.W. Burton. 1994. Genetic base for North American public soybean cultivars released between 1947 and 1988. Crop Sci. 34: 1143-1151.

Halliburton, R. 2004. Introduction to population geneticsPearson/Prentice Hall, Upper Saddle River, NJ.

Hayes, A.J., G. Ma, G.R. Buss, and M.A.S. Maroof. 2000. Molecular marker mapping of RSV4, a gene conferring resistance to all known strains of soybean mosaic virus. Crop Sci. 40: 1434-1437.

Heatherly, L.G., and R.W. Elmore. 2004. Managing inputs for peak production. Soybeans: improvement, production, and uses. Madison (USA): Agronomy Monograph 16: 451-536.

Hildebrand, D.F., R. Li, and T. Hatanaka. 2008. Genomics of soybean oil traits. Genetics and genomics of soybean. Springer. p. 185-209.

Holbrook, C.C., P. Timper, and H.Q. Xue. 2000. Evaluation of the core collection approach for identifying resistance to Meloidogyne arenaria in peanut. Crop Sci. 40: 1172-1175.

Huang, X., X. Wei, T. Sang, Q. Zhao, Q. Feng, Y. Zhao, et al. 2010. Genome-wide association studies of 14 agronomic traits in rice landraces. Nat. Genet. 42: 961-967.

Huang, X., C. Zhu, D. Fan, Y. Lu, Q. Weng, K. Liu, et al. 2012. Genome-wide association study of flowering time and grain yield traits in a worldwide collection of rice germplasm. Nat. Genet. 44: 32-39.

Huang, X.H., and B. Han. 2014. Natural variations and genome-wide association studies in crop plants. ANNUAL REVIEWS, PALO ALTO. p. 531-551.

Hwang, E.Y., Q.J. Song, G. Jia, J.E. Specht, D.L. Hyten, J. Costa, et al. 2014. A genome-wide association study of seed protein and oil content in soybean. BMC genomics 15: 1-1.

Hymowitz, T. 1972. Relationship between the content of oil, protein, and sugar in soybean seed. Agron. J. 64: 613.

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

Hyten, D.L., Q. Song, Y. Zhu, I.Y. Choi, R.L. Nelson, J.M. Costa, et al. 2006. Impacts of genetic bottlenecks on soybean genome diversity. Proc. Natl. Acad. Sci. U.S.A. 103: 16666-16671.

Imsande, J. 2001. Selection of soybean mutants with increased concentrations of seed methionine and cysteine. Crop Sci. 41: 510-515.

Jia, G., X. Huang, H. Zhi, Y. Zhao, Q. Zhao, W. Li, et al. 2013. A haplotype map of genomic variations and genome-wide association studies of agronomic traits in foxtail millet (Setaria italica). Nat. Genet. 45: 957-961.

Jiang, B., H. Nan, Y. Gao, L. Tang, Y. Yue, S. Lu, et al. 2014. Allelic combinations of soybean maturity loci E1, E2, E3 and E4 result in diversity of maturity and adaptation to different latitudes. PLoS One 9: e106042.

Jin, Y., T. He, and B.R. Lu. 2003. Fine scale genetic structure in a wild soybean (Glycine soja) population and the implications for conservation. New Phytologist 159: 513-519.

Joseph, J.R. 1975. Oligosaccharides of food legumes: Alpha-galactosidase activity and the flatus problem. Physiological Effects of Food Carbohydrates. American Chemical Society. p. 207-222.

Kabelka, E.A., S.R. Carlson, and B.W. Diers. 2006. PI 468916 SCN resistance loci's associated effects on soybean seed yield and other agronomic traits. Crop Sci. 46: 622-629.

Keurentjes, J.J.B., R.C. Jansen, J. Fu, I.R. Terpstra, J.M. Garcia, G. van den Ackerveken, et al. 2007. Regulatory network construction in Arabidopsis by using genome-wide gene expression quantitative trait loci. Proc. Natl. Acad. Sci. U.S.A. 104: 1708-1713.

Kim, H.K., S.T. Kang, J.H. Cho, M.G. Choung, and D.Y. Suh. 2005. Quantitative trait loci associated with oligosaccharide and sucrose contents in soybean (Glycine max L.). J. Plant Biol. 48: 106-112.

Kim, K.H., K.H. Kim, K.D. Kim, D.H. Kim, D.S. Kim, T.H. Kim, et al. 2010. Whole-genome sequencing and intensive analysis of the undomesticated soybean (Glycine soja Sieb. and Zucc.) genome. Proc. Natl. Acad. Sci. U.S.A. 107: 22032-22037.

Kim, K.S., B.W. Diers, D.L. Hyten, M.A. Rouf Mian, J.G. Shannon, and R.L. Nelson. 2012. Identification of positive yield QTL alleles from exotic soybean germplasm in two backcross populations. TAG. Theoretical and applied genetics. Theoretische und angewandte Genetik 125: 1353-1369.

Kim, M.C., S. Schultz, R.L. Nelson, and B.W. Diers. 2016. Identification and fine mapping of a soybean seed protein QTL from PI407788a on chromosome 15. Crop Sci. 56: 219-225.

Klos, K.L.E., M.M. Paz, L.F. Marek, P.B. Cregan, and R.C. Shoemaker. 2000. Molecular markers useful for detecting resistance to brown stem rot in soybean. Crop Sci. 40: 1445-1452.

Kong, F., B. Liu, Z. Xia, S. Sato, B.M. Kim, and S. Watanabe. 2010. Two coordinately regulated homologs of FLOWERING LOCUS T are involved in the control of photoperiodic flowering in soybean. Plant Physiol. 154.

Korte, A., and A. Farlow. 2013. The advantages and limitations of trait analysis with GWAS: A review. Plant Methods 9: 29-29.

Kump, K.L., P.J. Bradbury, R.J. Wisser, E.S. Buckler, A.R. Belcher, M.A. Oropeza-Rosas, et al. 2011. Genome-wide association study of quantitative resistance to southern leaf blight in the maize nested association mapping population. Nat Genet 43: 163-168.

La, T.C., H.T. Nguyen, J.G. Shannon, S.M. Pathan, T. Vuong, J.D. Lee, et al. 2014. Effect of high-oleic acid soybean on seed oil, protein concentration, and yield. Crop Sci. 54: 2054-2062.

Lam, H.M., B. Wang, J. Li, M. Jian, J. Wang, G. Shao, et al. 2010. Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection. Nat. Genet. 42: 1053.

Lee, J.D., K.D. Bilyeu, and J.G. Shannon. 2007. Genetics and breeding for modified fatty acid profile in soybean seed oil. Crop Science Biotechnology 10: 201-210.

Lee, J.D., M. Woolard, D.A. Sleper, J.R. Smith, V.R. Pantalone, C.N. Nyinyi, et al. 2009. Environmental effects on oleic acid in soybean seed oil of plant introductions with elevated oleic concentration. Crop Sci. 49: 1762-1768.

Lee, J.D., J.K. Yu, Y.H. Hwang, S. Blake, Y.S. So, G.J. Lee, et al. 2008. Genetic diversity of wild soybean (Glycine soja Sieb. and Zucc.) accessions from south Korea and other countries. Crop Sci. 48: 606-616.

Lee, S.H., M.A. Bailey, M.A.R. Mian, E.R. Shipe, D.A. Ashley, W.A. Parrott, et al. 1996. 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.

LeRoy, A.R., W.R. Fehr, and S.R. Cianzio. 1991. Introgression of genes for small seed size from Glycine soja into G. max. Crop Sci. 31: 693-697.

Leske, K.L., and C.N. Coon. 1999. Nutrient content and protein and energy digestibilities of ethanol-extracted, low α-galactoside soybean meal as compared to intact soybean meal. Poult. Sci. 78: 1177-1183.

Li, G., L. Zhang, and C. Bai. 2012. Chinese Cornus officinalis : genetic resources, genetic diversity and core collection. Genet. Resour. Crop Evol. 59: 1659-1671.

Li, Y., Z.H. Yan, Y. Guan, L. Zhu, X. Ning, M.J.M. Smulders, et al. 2008. Genetic structure and diversity of cultivated soybean (Glycine max (L.) Merr.) landraces in China. Theor. Appl. Genet. 117: 857-871.

Li, Y.h., C. Zhang, M.J. Smulders, W. Li, Y.s. Ma, Q. Xu, et al. 2013. Analysis of average standardized SSR allele size supports domestication of soybean along the Yellow River. Genet. Resour. Crop Evol. 60: 763-776.

Lipka, A.E., F. Tian, Q. Wang, J. Peiffer, M. Li, P.J. Bradbury, et al. 2012. GAPIT: Genome association and prediction integrated tool. Bioinformatics 28: 2397-2399.

Lippert, C., J. Listgarten, Y. Liu, C.M. Kadie, R.I. Davidson, and D. Heckerman. 2011. FaST linear mixed models for genome-wide association studies. Nat. Meth. 8: 833-835.

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

Liu, K. 1997. Soybeans: chemistry, technology, and utilization.Chapman & Hall, New York.

Liu, W.G., M.Q. Shahid, L. Bai, Z. Lu, Y. Chen, L. Jiang, et al. 2016. Evaluation of genetic diversity and development of a core collection of wild rice (Oryza rufipogon Griff.) populations in China. PLoS One 10: e0145990.

Lu, S., X. Zhao, Y. Hu, S. Liu, H. Nan, X. Li, et al. 2017. Natural variation at the soybean J locus improves adaptation to the tropics and enhances yield. Nat. Genet. 49: 773.

Mamidi, S., R.K. Lee, J.R. Goos, and P.E. McClean. 2014. Genome-wide association studies identifies seven major regions responsible for iron deficiency chlorosis in soybean (Glycine max). PLoS One 9: e107469.

Maughan, P.J., M.A.S. 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.

Morris, G.P., P. Ramu, S.P. Deshpande, C.T. Hash, T. Shah, H.D. Upadhyaya, et al. 2013. Population genomic and genome-wide association studies of agroclimatic traits in sorghum. Proc. Natl. Acad. Sci. U.S.A. 110: 453-458.

Myles, S., J. Peiffer, P.J. Brown, E.S. Ersoz, Z. Zhang, D.E. Costich, et al. 2009. Association mapping: critical considerations shift from genotyping to experimental design. The Plant Cell 21: 2194-2202.

Nichols, D.M., K.D. Glover, S.R. Carlson, J.E. Specht, and B.W. Diers. 2006. Fine mapping of a seed protein QTL on soybean linkage group I and its correlated effects on agronomic traits. Crop Sci. 46: 834-839.

Orf, J.H., and T.C. Helms. 1994. Selection to maximize gross value per hectare within three soybean populations. Crop Sci. 34: 1163-1167.

Palomeque, L., L. Li-Jun, W. Li, B. Hedges, E.R. Cober, and I. Rajcan. 2009. QTL in mega-environments: I. Universal and specific seed yield QTL detected in a population derived from a cross of high-yielding adapted × high-yielding exotic soybean lines. Theor. Appl. Genet. 119: 417-427.

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.

Pathan, S.M., T. Vuong, K. Clark, J.D. Lee, J.G. Shannon, C.A. Roberts, et al. 2013. Genetic mapping and confirmation of quantitative trait loci for seed protein and oil contents and seed weight in soybean. Crop Sci. 53: 765-774.

Patil, G., R. Mian, T. Vuong, V.R. Pantalone, Q. Song, P. Chen, et al. 2017. Molecular mapping and genomics of soybean seed protein: a review and perspective for the future. Theor. Appl. Genet.: 1-17.

Pettersson, D., and K. Pontoppidan. 2013. Soybean meal and the potential for upgrading its feeding value by enzyme supplementation. Soybean-bio-active compounds. InTech.

Phansak, P., W. Soonsuwon, D.L. Hyten, Q. Song, P.B. Cregan, G.L. Graef, et al. 2016. Multi-population selective genotyping to identify soybean [Glycine max (L.) Merr.] seed protein and oil QTLs. G3: Genes, Genomes, Genet. 6: 1635-1648.

Qi, Z., J. Pan, X. Han, H. Qi, D. Xin, W. Li, et al. 2016. Identification of major QTLs and epistatic interactions for seed protein concentration in soybean under multiple environments based on a high-density map. Mol. Breed. 36: 55.

Qi, Z., Q. Wu, X. Han, Y. Sun, X. Du, C. Liu, et al. 2011. Soybean oil content QTL mapping and integrating with meta-analysis method for mining genes. Euphytica 179: 499-514.

Qiu, L.J., P.Y. Chen, Z.X. Liu, Y.H. Li, R.X. Guan, L.H. Wang, et al. 2011. The worldwide utilization of the Chinese soybean germplasm collection. Plant Genet. Resour. 9: 109-122.

Ramteke, R., V. Kumar, P. Murlidharan, and D.K. Agarwal. 2010. Study on genetic variability and traits interrelationship among released soybean varieties of India [Glycine max (L.) Merrill]. Electronic Journal of Plant Breeding 1: 1483-1487.

Ray, J.D., K. Hinson, J.E. Mankono, and M.F. Malo. 1995. Genetic control of a long-juvenile trait in soybean. Crop Sci. 35: 1001-1006.

Rector, B.G., J.N. All, W.A. Parrott, and H.R. Boerma. 2000. Quantitative trait loci for antibiosis resistance to corn earworm in soybean. Crop Sci. 40: 233-238.

Samanfar, B., S.J. Molnar, M. Charette, A. Schoenrock, F. Dehne, A. Golshani, et al. 2017. Mapping and identification of a potential candidate gene for a novel maturity locus, E10, in soybean. Theor. Appl. Genet. 130: 377-390.

Schon, M.K., and D.G. Blevins. 1990. Foliar boron applications increase the final number of branches and pods on branches of field-grown soybeans Plant Physiol. 92: 602-607.

Schweizer, T.F. 1978. Low molecular weight carbohydrates from leguminous seeds; a new disaccharide: galactopinitol. J. Am. Oil Chem. Soc. 29: 148-154.

Sebastian, S.A., P.S. Kerr, R.W. Pearlstein, and W.D. Hitz. 2000. Soybean germplasm with novel genes for improved digestibility. Soy in Animal Nutrition, Drackley, J.K. (Ed.), 56-73, Federation of Animal Science Societies, ISBN 18884706010, Savoy, Illinois, USA.

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

Segura, V., B.J. Vilhjálmsson, A. Platt, A. Korte, Ü. Seren, Q. Long, et al. 2012. An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. Nat. Genet. 44: 825-830.

Seren, Ü., B.J. Vilhjálmsson, M.W. Horton, D. Meng, P. Forai, Y.S. Huang, et al. 2012. GWAPP: A Web Application for Genome-Wide Association Mapping in Arabidopsis. The Plant Cell 24: 4793-4805.

Settimi, J.R., and J.E. Board. 1988. Photoperiod and planting date effects on the spatial distribution of branch development in soybean. Crop Sci. 28: 259.

Shannon, J.G., J.R. Wilcox, and A.H. Probst. 1972. Estimated gains from selection for protein and yield in the F4 generation of six soybean populations. Crop Sci. 12.

Shim, S., M.Y. Kim, J. Ha, Y.H. Lee, and S.H. Lee. 2017. Identification of QTLs for branching in soybean (Glycine max (L.) Merrill). Euphytica 213: 225.

Shivakumar, M., C. Gireesh, and A. Talukdar. 2016. Efficiency and utility of pollination without emasculation (PWE) method in intra and inter specific hybridization in soybean. Indian J. Genet. 76.

Sonah, H., L. O'Donoughue, E. Cober, I. Rajcan, and F. Belzile. 2015. Identification of loci governing eight agronomic traits using a GBS‐GWAS approach and validation by QTL mapping in soya bean. Plant Biotechnology Journal 13: 211-221.

Sonah, H., L. O'Donoughue, E.R. Cober, I. Rajcan, and F. Belzile. 2015. Identification of loci governing eight agronomic traits using a GBS‐GWAS approach and validation by QTL mapping in soya bean. Plant Biotechnol. J. 13: 211-221.

Song, Q., D.L. Hyten, G. Jia, C.V. Quigley, E.W. Fickus, R.L. Nelson, et al. 2013. Development and evaluation of SoySNP50K, a high-density genotyping array for soybean. PLoS One 8: e54985.

Song, Q., D.L. Hyten, G. Jia, C.V. Quigley, E.W. Fickus, R.L. Nelson, et al. 2015. Fingerprinting soybean germplasm and its utility in genomic research. G3: Genes, Genomes, Genet.

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

Tanksley, S.D., and S.R. McCouch. 1997. Seed banks and molecular maps: unlocking genetic potential from the wild. Science (New York, N.Y.) 277: 1063-1066.

Terry, L.I., K. Chase, T. Jarvik, J. Orf, L. Mansur, and K.G. Lark. 2000. Soybean quantitative trait loci for resistance to insects. Crop Sci. 40: 375-382.

Thakare, D., S. Kumudini, and R.D. Dinkins. 2011. The alleles at the E1 locus impact the expression pattern of two soybean FT-like genes shown to induce flowering in Arabidopsis. Planta 234: 933.

Thorne, J.C., and W.R. Fehr. 1970. Exotic germplasm for yield improvement in 2-way and 3-way soybean crosses. Crop Sci. 10: 677-678.

Tian, F., P.J. Bradbury, P.J. Brown, H. Hung, Q. Sun, S. Flint-Garcia, et al. 2011. Genome-wide association study of leaf architecture in the maize nested association mapping population. Nat Genet 43: 159-162.

Tsubokura, Y., H. Matsumura, M. Xu, B. Liu, H. Nakashima, T. Anai, et al. 2013. Genetic variation in soybean at the maturity locus E4 is involved in adaptation to long days at high latitudes. Agron. J. 3: 117-134.

Uauy, R., P. Mena, and C. Rojas. 2000. Essential fatty acids in early life: structural and functional role. Proc. Nutr. Soc. 59: 3-15.

Verslues, P.E., J.R. Lasky, T.E. Juenger, T.W. Liu, and M.N. Kumar. 2014. Genome-wide association mapping combined with reverse genetics identifies new effectors of low water potential-induced proline accumulation in Arabidopsis. Plant Physiol. 164: 144-159.

Vilhjálmsson, B.J., and M. Nordborg. 2013. The nature of confounding in genome-wide association studies. Nat. Rev. Genet. 14: 1-2.

Wang, K.J., X.H. Li, and M.F. Yan. 2014. Genetic differentiation in relation to seed weights in wild soybean species (Glycine soja Sieb. & Zucc.). Plant Syst. Evol. 300: 1729-1739.

Warner, K., and W.R. Fehr. 2008. Mid-oleic/ultra low linolenic acid soybean oil: A healthful new alternative to hydrogenated oil for frying. J. Am. Oil Chem. Soc. 85: 945-951.

Warner, K., and M. Gupta. 2005. Potato Chip Quality and Frying Oil Stability of High Oleic Acid Soybean Oil. J. Food Sci. 70: s395-s400.

Warrington, C.V., H. Abdel-Haleem, D.L. Hyten, P.B. Cregan, J.H. Orf, A.S. Killam, et al. 2015. QTL for seed protein and amino acids in the Benning × Danbaekkong soybean population. Theor. Appl. Genet. 128: 839-850.

Warrington, C.V., H. Abdel-Haleem, D.L. Hyten, P.B. Cregan, J.H. Orf, A.S. Killam, et al. 2015. QTL for seed protein and amino acids in the Benning× Danbaekkong soybean population. Theor. Appl. Genet. 128: 839-850.

Watanabe, S., K. Harada, and J. Abe. 2011. Genetic and molecular bases of photoperiod responses of flowering in soybean. Breed. Sci. 61: 531-543.

Watanabe, S., R. Hideshima, Z. Xia, Y. Tsubokura, S. Sato, and Y. Nakamoto. 2009. Map-based cloning of the gene associated with the soybean maturity locus E3. Genetics 182.

Weaver, D.B., R.L. Akridge, and C.A. Thomas. 1991. Grow habit, planting date, and row-spacing effects on late-planted soybean. Crop Sci. 31: 805-810.

Webb, D.M., B.M. Baltazar, A.P. Rao-Arelli, J. Schupp, K. Clayton, P. Keim, et al. 1995. Genetic mapping of soybean cyst nematode race-3 resistance loci in the soybean PI 437.654. Theor. Appl. Genet. 91: 574-581.

White, P.J. 2007. Fatty acid in oilseeds (Vegetable oils), p. 210-263, In C. K. Chow, ed. Fatty acids in foods and their health implications. CRC press, Marcel Dekker, Inc., New York.

2003. The use of soybean meal and full fat soybean meal by the animal feed industry. 12th Australian Soybean Conference, Toowomba, Australia.

Wilson, R.F. 2004. Seed composition. In H.R. Boerma and J.E. Specht (ed.) Soybeans: Improvement, Production, and Uses. 3rd ed. ASA, CSSA, and SSSA, Madison, WI.: 621-677.

Wilson, R.F. 2008. Soybean: market driven research needs. Genetics and genomics of soybean. Springer. p. 3-15.

Xia, Z., S. Lü, H. Wu, S. Tabata, K. Harada, S. Watanabe, et al. 2012. Positional cloning and characterization reveal the molecular basis for soybean maturity locus E1 that regulates photoperiodic flowering. Proc. Natl. Acad. Sci. U.S.A. 109: 12852-12853.

Xu, M., Z. Xu, B. Liu, F. Kong, Y. Tsubokura, and S. Watanabe. 2013. Genetic variation in four maturity genes affects photoperiod insensitivity and PHYA-regulated post-flowering responses of soybean. BMC Plant Biol. 13.

Yadav, N.S. 1996. Genetic modification of soybean oil quality. In DPS Verma, RC Shoemaker, eds, Soybean: Genetics, Molecular Biology and Biotechnology, CAB INTERNATIONAL, Wallingford, UK, pp 165-168.

Yazdi-Samadi, B. 1977. Components of developing soybean seeds: Oil, protein, sugars, starch, organic acids, and amino acids. Agron. J. 69: 481.

Yesudas, C.R., R. Bashir, M.B. Geisler, and D.A. Lightfoot. 2013. Identification of germplasm with stacked QTL underlying seed traits in an inbred soybean population from cultivars Essex and Forrest. Mol. Breed. 31: 693-703.

Yu, J., G. Pressoir, W.H. Briggs, I. Vroh Bi, M. Yamasaki, J.F. Doebley, et al. 2006. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat. Genet. 38: 203-208.

Zhang, J., and Y. Huang. 2011. Preparation and optical properties of AgGaS2 nanofilms. Cryst. Res. Technol. 46: 501-506.

Zhang, Z., E.S. Buckler, T.M. Casstevens, and P.J. Bradbury. 2009. Software engineering the mixed model for genome-wide association studies on large samples. Briefings Bioinf. 10: 664-675.

Zhang, Z., E. Ersoz, C. Lai, R.J. Todhunter, H.K. Tiwari, M.A. Gore, et al. 2010. Mixed linear model approach adapted for genome-wide association studies. Nat. Genet. 42: 355-360.

Zhao, C., R. Takeshima, J. Zhu, M. Xu, M. Sato, S. Watanabe, et al. 2016. A recessive allele for delayed flowering at the soybean maturity locus E9 is a leaky allele of FT2a, a FLOWERING LOCUS T ortholog. BMC Plant Biol. 16: 20.

Zhao, K., J. Mezey, A.M. McClung, C.D. Bustamante, S.R. McCouch, C.W. Tung, et al. 2011. Genome-wide association mapping reveals a rich genetic architecture of complex traits in Oryza sativa. Nat. Commun. 2: 467.

Zhou, X., and M. Stephens. 2012. Genome-wide efficient mixed-model analysis for association studies. Nat. Genet. 44: 821-824.

Zhou, Z., Y. Jiang, Z. Wang, Z. Gou, J. Lyu, W. Li, et al. 2015. Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean. Nat Biotech 33: 408-414.

Chapter II:

CHARACTERIZATION OF A USDA CORE COLLECTION OF WILD SOYBEAN (*Glycine soja* SIEBOLD & ZUCC.) ACCESSIONS FOR SEED COMPOSITION AND AGRONOMIC TRAITS

Abstract

The relatively low genomic variation of current U.S. soybean [*Glycine max* (L.) Merill] cultivars constrains the improvement of grain yield, seed quality, and other agronomic traits within soybean breeding programs. Recently, a substantial effort has been undertaken to introduce novel genetic diversity present in wild soybean (*Glycine soja* Siebold & Zucc.) into new elite cultivars, in both public and private applied soybean breeding programs. The objectives of this research were to evaluate the phenotypic diversity within a core collection of 80 *G. soja* plant introductions (PIs) in the United States Department of Agriculture National Genetic Resources Program that were collected in China, Japan, Russia, and South Korea, and to analyze the correlations between agronomic and seed composition traits. Field tests were conducted in Missouri and North Carolina during three years, 2013, 2014, and 2015, in a randomized complete block design (n=3). The phenotypic data collected included plant maturity date, seed weight, and the seed concentration of protein, oil, essential amino acid, fatty acid, and soluble carbohydrates. Analyzing the data from six environments, we found genotype was a significant ($p < 0.0001$) source of variation for maturity date, seed weight, seed protein and amino acids, seed oil and fatty acids, and seed carbohydrates. Significant correlations were observed between numerous traits. The core collection had lower seed weight, higher seed content of protein, linolenic acid, raffinose and stachyose but lower seed content of oil and oleic acid than those of the cultivated soybean lines that were used as checks. The amino acid profile of the core collection was significantly different from that of the checks. An association analysis revealed 19 SNP that were significantly associated with maturity, seed weight, and seed contents of aspartic acid, glutamine, palmitic acid, oleic acid, and linoleic acid. The information and data collected in this study will be invaluable in guiding soybean breeders and geneticists in selecting promising *Glycine soja* plant introductions for research and cultivar improvement.

Introduction

Soybean [*Glycine max* (L.) Merr.] seed typically has 20% oil, 40% protein, and 12% soluble carbohydrates based on dry weight (Liu, 1997). Soybean is a commodity crop and both seed oil and meal are valuable commodities in the global food supply. Soybean predominantly self-pollinates, leading to increased homozygosity and overall decreased genomic variation as compared to outcrossing species (Lam et al., 2010). The genomic variation of soybean is even less diverse due to bottleneck events during domestication and modern plant breeding (Tanksley and McCouch, 1997). Only 80 out of 45,000 unique Asian landraces account for more than 99% of 258 North American public soybean cultivars released during the period from 1947 to 1988 (Gizlice et al., 1994). In addition, 17 of the 80 Asian landraces define up to 84% of the genetic base of the 258 North American cultivars (Gizlice et al., 1994; Hyten et al., 2006). Therefore, the tendency of soybean to self-pollinate combined with the narrow genetic base of North American public soybean varieties creates a need for genetically diverse germplasm to allow for improvement of agronomic and seed quality traits.

To address the problem of narrow genetic diversity in soybean, the United States Department of Agriculture (USDA) maintains a soybean germplasm collection of 1,168 wild soybean accessions and 18,480 *G. max* accessions (Song et al., 2015]. Crop relatives and exotic germplasm are important genetic resources for improving agriculture productivity, yet wild soybean (*Glycine soja*) has been largely under-utilized in breeding efforts focused on broadening the narrow genetic background of cultivated soybeans (Jin et al., 2003; Lee et al., 2008). Wild soybean germplasm utilization is limited in breeding programs because potentially beneficial genes controlling yield and seed quality are often genetically linked and co-inherited with undesirable traits such as prostrate growth habit, hard seed coat, low seed quality, and seed shattering (Concibido et al., 2003].

Utilization of over 1100 accessions in the wild soybean germplasm collection for soybean breeding is unmanageable and impractical for public and private breeders using conventional breeding techniques and marker assisted selection (MAS). While the use of MAS has increased the utility of wild soybean to breeders, the agronomically undesirable traits of wild germplasm can be avoided during population development by backcrossing with elite varieties and by evaluating large segregating populations (Ertl and Fehr, 1985; Carpenter and Fehr, 1986; LeRoy et al., 1991; Sebolt et al., 2000; Kabelka et al., 2006; Zhang and Huang, 2011; Akpertey et al., 2014; Shivakumar et al., 2016). To address the problem, Frankel and Brown (1984) suggested the establishment of a core collection with a limited number of accessions derived from the original collection; representing about 10% of the full collection. This core collection should represent the genetic diversity of the original collection with the lowest number of redundant accessions. A core collection is easier to evaluate and more efficient to utilize. Core collections were successfully developed with multiple crops including maize (*Zea mays*), rice (*Oryza sativa*), wheat (*Triticum aestivum*), and peanut (*Arachis hypogaea*) (Holbrook et al., 2000; Coimbra et al., 2009; Bordes et al., 2011; Liu et al., 2016). Soybean core collections exist in East Asia (Qiu et al., 2013) and Brazil (Priolli et al., 2013). Domesticated soybean core collections composed of a portion of the 18,480 USDA *G. max* accessions have also been developed (Oliveira et al., 2010). Even smaller mini-core collections that represent the most diverse 1% of the accessions have been developed for multiple crops including maize (*Zea mays*), rice (*Oryza sativa*), wheat (Triticum aestivum), and peanut (*Arachis hypogaea*) (Holbrook et al., 2000; Coimbra et al., 2009; Bordes et al., 2011; Liu et al., 2016). However, to our knowledge there are no published core collections derived from the USDA *G. soja* collection.

The entire USDA *G. soja* collection has been genotyped and a core collection of genetically and phenotypically defined wild soybeans would be useful to identify and utilize accessions with favorable agronomic and seed quality traits. Even so, only one recent study examined the phenotypic variation of primarily Korean and Japanese wild soybean seed compositions from the USDA *G. soja* collection (Leamy et al., 2017). This study focuses on characterizing the phenotypic variation of maturity, seed weight, and seed compositions of accessions from a core collection representing the diversity of the entire USDA *G. soja* collection. The objectives of this study were to characterize agronomic and seed composition traits in a wild soybean core collection in replicated, multi-environment field experiments and to make these data available to other soybean breeders and researchers for variety development and genetic studies through the Genetic Resources Information Network (GRIN) maintained by the USDA. In addition, we evaluated the correlations between these traits, and identified genotypes with favorable seed composition and agronomic traits.

Materials and methods

Plant materials

The USDA soybean collection includes 1,168 *G. soja* from China, Korea, Japan and Russia (www.ars-grin.gov) and the majority was previously genotyped with a SoySNP50K beadchip (Song et al., 2013; Song et al., 2015). Analysis of the pair-wise distances among the *G. soja* accessions based on 42,509 SNPs in the beadchip showed that a total of 806 *G. soja* accessions from China, Korea, Japan, and Russia were non-redundant (Song et al., 2015). Thus, a total of 80 *G. soja* plant introductions (PIs) (Supplementary Table 1), which is approximately equal to 10% of the total number of the non-redundant *G. soja* accessions in the collection, were chosen to represent maximal diversity. The 806 accessions were clustered to a pre-defined number of clusters based on their genetic differences and one accession from each cluster was selected to form a core set. The PIs have maturity groups (MG) ranging from group 000 to group X with nearly half of the collection consisting of MG V lines (www.ars-grin.gov). The geographic range of the lines is broad consisting of lines from Eastern China (19 PIs), Japan (22 PIs), Eastern Russia (11 PIs), and South Korea (28 PIs) (www.ars-grin.gov). Seeds were obtained from the USDA Soybean Germplasm Collection via GRIN (www.ars-grin.gov). Eight *Glycine max* cultivars were planted in all Missouri locations/years as checks (Supplementary Table 2). The maturity of these checks ranges from group 0 to group VII. Because PI245331 had late maturity (MG X), this genotype was not harvested in any environment and was excluded in further analysis.

Experimental design and growth conditions

In 2013, 80 entries were planted at the Central Crops Research Station in Clayton, NC and at Bradford Farm in Columbia, MO. In 2014, a second field experiment was conducted at the Central Crops Research Station in Clayton, NC; the Sandhills Research Station in Jackson Springs, NC; and again at Bradford Farm in Columbia, MO. In 2015, the field trial was carried out at Greenley Memorial Research Center in Novelty, MO. The entire collection was planted in all Missouri locations and years and 65/80 PIs were planted in Clayton, NC and Sandhills, NC during 2013 and 2014.

In MO, all genotypes were planted in single row plots of 2.43 m in length, plot spacing was 1.22 m, and spacing was 1.52 m between rows. At the Novelty location, seeds were sown at the rate of 30 seeds m-1. At other locations, the seeds were sown at the rate of 20 seeds m-1. Plots were seeded using a four-row ALMACO cone planter with Kinze row units (ALMACO, Nevada, IA). Wild soybean seeds are hard seeds and possess extended dormancy periods (late germination) so seeds were scarified before planting by using a razor blade to make a small incision in the seed coat on the opposite side of the hilum. Lines were planted in a randomized complete block design with three replicates at all locations year-1.

In NC seeds were planted with a funnel dropper in 2.43 m long rows at 10 seeds m-1. Seeds were scarified in a coffee mill with the blades replaced with a sandpaper disk using 10 one second pulses. If the seeds were not visibly scarified 5 more one second pulses were used. Lines were planted in a randomized complete block design with three replicates at all location year^{-1}.

Measurement of agronomic traits

Plant maturity was recorded as the number of days between planting date and the date when approximately 95% of the pods' color had changed to mature pod color (R8) (Fehr et al., 1971). The maturities were determined for all Missouri plots at each location. Seed composition was determined for seed from all NC plot locations. Plants of the same plot were harvested together by hand and threshed by an Almaco single bundle thresher (ALMACO, Nevada, IA) at all locations. Seed weight was measured by randomly picking and measuring 100 seeds from each plot for all locations 3 times with replacement.

Crude protein and amino acid analysis

Approximately nine grams of soybean seeds from each plot were ground using a Thomas Wiley Mini-Mill (Thomas Scientific, Swedesboro, NJ) and filtered with a 20-mesh screen. A Labconco Freeze Dry System (Labconco, USA.) was used to lyophilize the ground powder for 48 hours. Samples containing approximately 3 g of ground seeds from each plot in all locations were sent to the University of Missouri Agricultural Experiment Station Chemical Laboratory, University of Missouri (Columbia, MO) to determine the content of crude protein. The seed protein and amino acid content (twelve amino acids) were evaluated for two out of three replicates from each location.

Crude nitrogen was determined by combustion analysis (AOAC Official method 990.03, 2006). The nitrogen content in a 200 mg subsample was measured using the Dumas method and a LECO truSpec model FP-428 nitrogen analyzer following the manufacturer's recommendations (LECO, St. Joseph, MI). The protein content of soybean seed was estimated by multiplying the total nitrogen concentration by 6.25.

The contents of twelve amino acids were measured by a single oxidation 4-hour hydrolysis method (Gehrke et al., 1987). The twelve amino acids are alanine, aspartic acid, cysteine, glutamic acid, glycine, isoleucine, leucine, lysine, methionine, proline, threonine, and valine. The hydrolyzation of the samples was carried out using 6 N HCl for 4 hours at 1450 C, and the amino acid concentration was determined by cation exchange chromatography in a Beckman 6300 Amino Acid Analyzer (Beckman Instruments, Inc., San Ramon, California).

Oil analysis

About 5 g of ground soybean seed was used to determine oil content with a XDS Rapid Content™ Analyzer (FOSS, Co., LLC, Denmark) and the ISIscan™ software. A certified 80% reflectance reference was used to create reference standard. The performance test was carried out by running four segments ten times and compiling the spectra.

Fatty acid analysis

The fatty acid profiles of total oil for each plot in Columbia and Novelty, MO were evaluated using a previously described procedure (Yoon et al. 2009). The five fatty acids that were measured are palmitic acid (C16:0), stearic acid (C18:0), oleic acid (C18:1), linoleic acid (C18:2), and linolenic acid (C18:3). The fatty acid levels were determined as a percentage of the total fatty acids in soybean seeds. The oil in 0.2 gram of ground soybean seed was extracted by placing the soybean seed powder in 2 mL of extraction buffer (chloroform:hexane:methanol [8:5:2, v/v/v]) for 12 hours. One hundred microliters of the extract was transferred to vials containing 75 µL of methylating reagent (0.25M methanolic sodium methoxide:petroleum ether:ethyl ether [1:5:2, v/v/v]). Extraction buffer was added to acquire 1 mL of sample. An Agilent (Palo Alto, CA) series 6890 capillary gas chromatograph with a flame ionization detector (275^{0}C) and an AT-silar capillary column (Alltech Associates, Deerfield, IL) was used. Standard fatty acid mixtures (Animal and Vegetable Oil Reference Mixture 1, AOCS) were used as calibration reference standards.

Sugar Analysis

The concentration of glucose, fructose, sucrose, raffinose, and stachyose were determined using an HPLC-ELSD procedure (Valliyodan et al. (2015). Approximately 90 mg of lyophilized seed powder was mixed with 900 µl HPLC-grade water (Fisher Scientific, Hampton, NH) and incubated at 55^{0} C with 250 rpm agitation for 30 minutes. After incubation, vials were vortexed, cooled down to room temperature, and blended with 900 µl HPLC grade acetonitrile (Fisher Scientific, Hampton, NH). The suspension was centrifuged for 30 minutes at 13.3 × 1000 g × min^{-1}. The supernatant was diluted five times with an acetonitrile: water mixture (65:35, v/v). The Agilent 1200 series (Agilent, USA) of HPLC- ELSD system was used with the Prevail Carbohydrate ES columns (5 µm) 250 mm × 4.6 mm, and guard columns, 7.5 mm × 4.6 mm (Grace Davison Discovery Sciences, Deerfield, IL). Sugar standards (D-fructose, D-(+) glucose, sucrose, D-(+) raffinose pentahydrate, and stachyose hydrate) were prepared in water with concentrations of 50, 100, 300, and 500 µg/mL and run to generate a standard curve for prediction.

Statistical Analysis

Each location/year was considered as a single environment (Table 1). The analysis of variance (ANOVA) was carried out by using PROC MIXED in SAS version 9.4 (SAS Institute, 2002). Genotype was used as a fixed effect to test for significant genotypic differences among accessions for all traits (Table 2 and 3; F-test P-value column). Environment was used as a fixed effect to test for significant environmental differences for all traits (Table 6). The heritability (h^2) of each trait was calculated as following (Nyquist and Baker, 1991):

$$h^2 \text{ (entry mean basis)} = \frac{\sigma_g^2}{\sigma_g^2+\sigma_{ge}^2/t+\sigma_e^2/rt},$$

$$h^2 \text{ (plot basis)} = \frac{\sigma_g^2}{\sigma_g^2+\sigma_{ge}^2+\sigma_e^2},$$

where σ_g^2 is the variance among genotypes, σ_{ge}^2 is the variance of genotype × environment interaction, σ_e^2 is experimental error, t is number of test environments, and r is number of replications.

PROC CORR of SAS (SAS Institute) was used to determine significance and correlation coefficients between studied traits based on individual genotypes' mean value across replications and locations.

Genotyping and quality control

The genotypic data including 42,509 SNP markers for all 79 genotypes was downloaded from Soybase website (http://www.soybase.org, "SoyBase browser", accessed 05/20/2018). The information of these SNP markers were retrieved from the study of Song et al. (2013). In their study, the Illumina SoySNP50k iSelect BeadChip was used for genotyping. We filtered the genotypic data by removing SNPs not located on any of the 20 chromosomes, as well as those missing rates >5% or with minor allele frequencies <5%. A total of 35,285 SNPs across 80 genotypes were used in the genome-wide association analysis with 26 studied traits.

LD estimation

TASSEL, version 5 was used to calculate squared correlation coefficients (r^2), and measure pairwise linkage disequilibrium (LD) (Bradbury et al., 2007). LD was calculated for genome wise, euchromatic and heterochromatic regions. A marker was determined to be in euchromatic or heterochromatic regions based on the physical information of the marker and these regions. The information was downloaded from Soybase. The distance between two markers, where the average r^2 reached half of its maximum value (Huang et al., 2010), was used as LD decay rate. The LD decay rate in our study was 0.22.

Genome-wide association study (GWAS)

Fixed and Random Model Circulating Probability Unification [FarmCPU; Liu et al. (2016)] method was used to perform genome-wide association analysis. We fitted the first three principal components as the covariates to correct for stratification (Price et al., 2006). We used Bonferroni test threshold as the significance cutoff of the genome-wide association study. The threshold was set as 0.05/total SNPs (-log10(P)=5.85).

Results and discussion

The collected data presented from the field experiments are categorized into two sets. Set one includes the data of 80 genotypes in three out of six studied environments from 13CLM, 14CLM, and 15NOV (Tables 2-2 and 2-4; Supplementary Tables 2-3 through 2-6). The measured traits in set one are maturity, seed weight, seed protein and amino acid content, seed oil and fatty acid composition, and seed soluble sugars composition. Set two includes seed weight, seed protein and amino acid content of 64 out of 80 PIs in all six environments (Tables 2-3, 2-5, and 2-6). The trait values in set 2 (Table 2-3) showed similar means, ranges, and variation, with the exception of seed weight, to those in set 1 (Table 2-2). The higher variation of seed weight in set 2 may be due to fewer genotypes (64 genotypes) and more environments (6 environments) were tested in set 2 comparing to 80 genotypes and three environments were tested in set 1. These results suggest that both sets of PI's showed similar phenotypic data and either of them can be used to determine significant differences among genotypes for the traits measured in the core collection. The heritability ranged from 0.52 to 0.99 (Table 2-2 and 2-3) for maturity, seed weight, seed protein and oil contents, and seed amino acids, except valine and proline. These medium to high heritability (0.52 to 0.99) indicate that genotype had a strong influence on those traits, suggesting that.

Seed weight

In both set one and set two, the 100-seed weight showed a wide phenotypic range from 0.9 g to 3.5 g, with a mean of 1.78 g (Table 2-2 and 2-3). *G. soja* has much lower seed weight ranges than cultivated soybean (11.9-16.2g 100 seed^{-1}) used as a check in this study. Substantial variation was observed for mean seed weight in both set 1 (CV = 9.08) and set 2 (CV = 12.19) (Table 2-2 and 2-3). One explanation for the wide variation of seed weight could be that a population of wild soybean was usually composed of individuals with different seed weight, and the seed weights in a population of wild soybean might show up to 2.6 fold differences (Wang et al., 2014). Genotype and GxE interaction played a significant role in seed weight variation ($P<0.001$). Our result is consistent with the study of Wang et al. (2014) in which the variation in seed weight among wild soybean genotypes could be genetically distinguished.

Protein and oil

The protein and oil contents of the core collection were in the range of 392.6 - 481.7 g kg^{-1} for protein (Table 2-2 and 2-3) and 157.6 - 175.8 g kg^{-1} for oil (Table 2-2). The means of set 1 and 2 protein concentrations were slightly different with set 1 having a mean of 438.9 g kg^{-1} and set 2 having a mean of 444.50 g kg^{-1} (Table 2-2 and 2-3). Set 1 had a mean oil content of 164.70 g kg^{-1} (Table 2-2). PI407228 had the highest seed protein content in a single plot (493 g kg^{-1}) and PI549048 had the highest seed oil content in a single plot (176 g kg^{-1}). The average contents of seed protein and oil were within the ranges of seed protein and oil of the Soybean Germplasm Collection (Wilson, 2004). Compared to the checks (Table 2-2), the seed of wild soybean lines in this collection had lower oil and higher protein content. The heritability of seed protein was 0.91 for both set 1 and set 2 (Table 2-2 and 2-3) and was 0.86 for oil (Table 2-2). These high values are similar to the entry-mean heritabilities reported by Jarquin et al. (2016) using 18,500 accessions of *G. max* in the USDA Soybean Germplasm Collection and indicate a significant potential for selecting for seed protein and oil composition in future cycles. The plot based heritabilities were also high for protein in both set one and two (0.59 and 0.53, respectively; Table 2-2 and 2-3) and for oil in set one (0.51; Table 2-2) indicating an increased possibility of selecting for these traits using a single plot per generation. These data indicate significant differences among *G. soja* PIs, which suggests more studies about the potential for novel protein and oil genetic contributions in this collection are needed.

Amino Acid Profiles

There is a substantial body of literature on soybean seed composition (Bellaloui et al., 2009; Medic et al., 2014; Yu et al., 2016; Lee et al., 2017). Even so, there have been few studies about seed content of amino acid content in soybean and wild soybean (Takahashi et al., 2003; Krishnan, 2005; Wang et al., 2015; Warrington et al., 2015). Soybean seed protein is valuable because it contains all of the essential amino acids for human and animal consumption; however, soybean seed has relatively low contents of the sulfur containing amino acids, cysteine and methionine (George and De Lumen, 1991). In this study, the contents of cysteine and methionine in wild soybean seed ranged from 14.9 - 19.2 g kg^{-1} with a mean of 16.9 g kg^{-1} and 14.0 - 16.9 g kg^{-1} with a mean of 15.1 g kg^{-1} (Table 2-2). In 2011, Banaszkiewicz (2011) reported narrower range of cysteine and methionine (15.0-17.0 g kg^{-1} and 13.6-15.8 g kg^{-1}, respectively) when the author studied soybean and soybean meals for animals. George and De Lumen (1991) reported a lower range of methionine in cultivated soybean samples that they collected. This range was 11 to 16 g kg^{-1} and most soybean samples was in the range of 12-14 g kg^{-1}. The ranges of cysteine and methionine content in this study were higher than the range reported by Warrington et al. (2015) (14.7 -16.2 g kg^{-1} for cysteine content and 13.8-14.7 g kg^{-1} for methionine content) when they studied a recombinant inbred line population developed from a cross between Benning and Danbaekkong (Warrington et al., 2015). When

Kwanyuen et al. (1997) studied a wild soybean germplasm, the seed content of cysteine and methionine(4-8 g kg^{-1} and 7-11 g kg^{-1}, respectively) showed lower ranges comparing to those in our study. These two amino acids were significantly affected by genotype ($P < 0.0001$) (Table 2-2 and 2-3). The heritability of amino acids in set one (Table 2-2) are slightly lower than those in set two (Table 2-3) because the data in set two were collected in six environments while the data in set one were collected in three environments. The heritability on an entry-mean basis of set two amino acids were higher on average, with a range from 0.55 to 0.80 (Table 2-3), compared to the set one amino acid range of 0.30 – 0.84; although both sets of data indicate the amino acid composition was strongly influenced by genotype. The high entry-mean heritability of amino acids suggest high genetic gains for amino acids can be achieved using wild soybean germplasm in breeding programs.

In terms of nutrition, the amino acid profile of soybean seed protein is a more compelling trait for selection than protein content *per se*. Due to deficiencies in methionine, lysine, and threonine of soybean protein (Pelaez and Walker, 1979; Erickson et al., 1989), these amino acids have been supplemented to improve the quality of soybean (Qi et al., 2011; Yuan et al., 2011; Wang et al., 2012). Methionine is the most limiting amino acid (Erickson et al., 1989; Wang and Li, 2012). Imsande (2001) reported roughly 100 million dollars were paid annually to supplement animal feed with methionine. In addition, the leaching of methionine supplements may lead to the formation of undesirable volatile sulfides due to bacterial degradation (George and De Lumen, 1991). Therefore, developing soybean cultivars with improved amino acid composition is desirable to both improve soybean seed value and to avoid negative environmental effects caused by supplementing amino acids.

The average percentage of nine out of twelve amino acids, based on crude protein content, showed significant differences across six studied environments (Table 2-6). The highest average content of six of the twelve amino acids were observed in Novelty, MO in 2015. Grieshop and Fahey (2001) and Karr-Lilienthal et al. (2005) suggested differences in temperature among studied environments may contribute to variation in amino acid content. Singh et al. (2016) stated that the free amino acid increased when phosphate was deficit, and the seed amino acid decreased under elevated carbon dioxide condition. Our results in this study indicate that there is significant genetic variation for amino acids in the *G. soja* core collection, and soybean researchers and breeders should be able to utilize this variation for cultivar improvement and research to understand the genetic architecture.

Fatty acids

Genotype showed significant influence on the variation of all five fatty acids ($P<0.0001$) (Table 2). The average fatty acid content of soybean oil in this study was 127.9 g kg^{-1} palmitic acid, 32.3 g kg^{-1} stearic acid, 122.1 g kg^{-1} oleic acid, 554.1 g kg^{-1} linoleic acid, and 163.8 g kg^{-1} linolenic acid. The oleic acid content (122.1 g kg^{-1}) was lower while the linolenic acid content (163.8 g kg^{-1}) was higher than values reported by Guo and Petrovic (2005) (233 g kg^{-1} and 76 g kg^{-1}, respectively) studied the oil extracted from cultivated soybean [*Glycine max* (L.) Merr.]. The higher concentrations of linolenic in wild soybean compared to that in cultivated soybean in our study was consistent with the statement of Asekova et al. (2014) and Pantalone et al. (1997). Soybean lines with high linolenic acid content are not desirable for many breeding programs because oil with a higher content of linolenic acid readily oxidizes, resulting in the formation of off-flavors when the oil is used for cooking (Yoon et al., 2009). In addition, when the oil is used as biofuel, the oxidized oil may cause viscous materials that clogs the oil filter and obstructs fuel flow (Yadav, 1996). Hydrogenation has been used to improve oil stability by reducing the number of double bonds in polyunsaturated fatty acid molecules (Yadav, 1996). This process increases the cost of oil and produces trans-fats, which is associated with increased risk of heart disease, stroke, and diabetes (Mozaffarian and Rimm, 2006). Oleic acid is more oxidatively stable, and oil with high oleic acid content is desirable for various applications such as cooking oil and biofuel. Therefore, two of the goals of soybean breeding to improve oil quality are 1) to reduce the content of linolenic acid and 2) to increase the content of oleic acid (Lee et al., 2007; Yoon et al., 2009).

Considering another aspect of linoleic acid and linolenic acid, these acids are omega-6 and omega-3 fatty acids, which are essential for human health and development (Covington, 2004). The lack of these fatty acids make the human more vulnerable to health’s problems such as heart diseases, asthma, allergies, and antimicrobial factors (Simopoulos, 2002; Simopoulos, 2008). However, the biological activities of these fatty acids in human are different. When a high amount of omega-3 fatty acids are ingested, inflammation and thrombosis may be suppressed; in contrast, high intake of omega-6 fatty acids may lead to inflammation (Asif, 2011). Due to the differences in biological activities between these two fatty acids, a heathy range of ratios of omega-6 to omega-3 (1:1 to 4:1) was reported by (Mattson and Grundy, 1985; Simopoulos, 2002). Dhakal et al. (2013) and Asekova et al. (2014) stated that this ratio in commodity soybean oil was 6:1 to 7:1. The core collection of wild soybean in this study had lower ratio (3.4:1); therefore, the wild soybean genotypes in our core collection can be used to improve the ratio of omega-6 to omega-3 in cultivated soybean for improving human health. In our study, linolenic acid was strongly correlated with seed content of protein, oil, oleic acid, and linoleic acid (0.46, -0.62, -0.54, and -0.45, respectively; Table 2-4). Because of these strong correlations, breeding for increased seed content of linolenic acid would lead to an increase in seed protein content and decreases in seed content of oil, oleic acid, and linoleic acid.

Soluble sugars

Sucrose, fructose, and glucose induce sweet taste and are easily digestible while raffinose and stachyose are indigestible by monogastric animals and cause digestion problems such as flatulence and diarrhea (Hou et al., 2009; Kumar et al., 2010). Hence, increasing sucrose, glucose, or fructose content and reducing stachyose and raffinose content in soybean seed are important for improving soybean seed quality (Yu et al., 2016). We found a higher ratio of raffinose and stachyose to sucrose compared to ratios observed in cultivated soybeans. We also found a strong correlation between maturity and stachyose (0.76; Table 2-4), a strong negative correlation between maturity and raffinose (-0.22; Table 2-4), and a strong correlation between seed weight and sucrose (0.65; Table 2-4).

The seed contents of fructose, glucose, sucrose, raffinose, and stachyose are shown on Table 2-2. Five water soluble carbohydrates, including fructose, glucose, sucrose, raffinose, and stachyose, were analyzed and significant variation was observed in sucrose, raffinose, and stachyose ($P<0.001$) (Table 2-2). Glucose, fructose, and raffinose are present at low concentrations (<15g kg^{-1}) in wild soybean seeds. The sucrose, raffinose, and stachyose carbohydrates concentrations ranged from 14.6 - 39.5 g kg^{-1} with a mean of 21.5 g kg^{-1} for sucrose, 6.6 – 9.3 g kg^{-1} with a mean of 7.8 g kg^{-1} for raffinose, and 37.2 – 58.9

g kg^{-1} with a mean of 47.8 g kg^{-1} for stachyose. Among the five studied sugars, sucrose and stachyose exhibited the highest concentrations while fructose, glucose, and raffinose were at lower concentrations. The average sucrose content was lower than those reported by Yu et al. (2016) and Hou et al. (2009) for cultivated soybeans (52.1 g kg^{-1} and 46.8 g kg^{-1}, respectively). However, the average stachyose content in this study was higher than those reported by Yu et al. (2016) and Hou et al. (2009) (39.3 g kg^{-1} and 31.7 g kg^{-1}, respectively). One possible explanation for the differences between our reported results and previous studies is that sugar profiles are strongly influenced by the environment (e.g. the CV for sugar traits is quite high, Table 2-2). The ranges of fructose, glucose, and raffinose seed content in the core collection were 5.1 – 11.6 g kg^{-1} with a mean of 7.1 g kg^{-1} for fructose, 4.3 – 6.5 g kg^{-1} with a mean of 5.1 g kg^{-1} for glucose, 6.6 – 9.3 g kg^{-1} with a mean of 7.8 g kg^{-1} for raffinose (Table 2-2). The variation and average concentrations of fructose, glucose, and raffinose in this study were similar to those in the previous reports for domesticated soybeans (Hou et al., 2009; Kumar et al., 2010; Yu et al., 2016). However, *G. soja* PIs possess unique seed profiles for sucrose and stachyose, which could be useful for understanding the underlying genetic architecture of these carbohydrates in soybean.

Correlations among Traits

Maturity showed significant but weak correlations with seed weight and crude protein (0.22 and 0.37, respectively; Table 2-4). This result is inconsistent with a recent study (Vaughn et al., 2014), which did not observe any relationship between protein levels and maturity groups using 3,258 accessions [*Glycine max* (L.) Merr.] in maturity groups I to IX from the USDA Soybean Germplasm Collection. This could be explained by the loss of genes or alleles relating to these traits during domestication and improvement selection in which cultivars were developed to achieve highest yield potential in a certain regions with specific maturity groups. Bellaloui et al. (2009) studied the relation between maturity and seed compositions of near-isogenic soybean lines derived from the cultivar Clark Johnson (1958) and Harosoy (Weiss and Stevenson, 1955). Bellaloui et al. (2009) observed a positive correlation between maturity and protein for Clark isolines and non-significant correlation between these traits for Harosoy isolines. The relationship between maturity, protein, and seed weight in wild and domesticated soybean may be refined with further studies.

The core collection seed weight shows significant associations with maturity, the content of oil, sucrose, and most fatty acids, with the exception of linoleic acid and palmitic acid (Table 2-4). Soybean seed oil and oleic acid both were positively correlated with seed weight whereas stearic acid and linolenic acid were negatively correlated with seed weight (Table 2-4). This result is consistent with the observation of Vineet et al. (2006), Guleria et al. (2008), Poeta et al. (2016), and Lee et al. (2017) in which they observed seed size had positive correlations with oil and oleic acid but a negative correlation with linolenic acid when they studied different collections of *Glycine max* accessions.

Seed protein and oil contents showed a strong negative correlation (r = – 0.66; Table 2-4). Due to the strong negative relationship between oil and protein, we also observed reciprocal relationships between protein or oil and some of their corresponding negatively and positively correlated traits; such as oleic acid, linoleic acid, linolenic acid, fructose, and sucrose (Table 2-4). The negative correlation between protein and oil has been well documented (Hymowitz, 1972; Burton, 1987; Wilcox, 1998; Chung et al., 2003; Vaughn et al., 2014; Leamy et al., 2017; Wu et al., 2017). Chung et al. (2003) reported the inverse relationship between oil and protein may be due to two traits influenced by the same genes (pleiotropy) or via traits controlled by different but strongly associated alleles. In contrast, La et al. (2014) reported a positive correlation between protein and oil contents when they studied soybean lines [*Glycine max* (L.) Merr.] with elevated oleic acid content. Other possible explanations include differences in soybean accessions or cultivation practices (Lee et al., 2017). Dornbos and McDonald (1986) and Saldivar et al. (2011) observed seed oil accumulation at early developmental stages and protein at later stages of soybean seed development. The strong negative correlation between seed protein and oil implies improving both seed protein and oil content may be difficult (Chung et al., 2003; Nichols et al., 2006). Hwang et al. (2014) recently found a biallelic SNP significantly associated with increased protein and oil content with one variant and associated with decreased content of protein and oil with the other variant.

Among the five fatty acids, oleic acid and linolenic acid showed significant correlations with seed weight, seed protein, linoleic acid, and between themselves. Brace et al. (2011) and La et al. (2014) also reported oleic acid was negatively correlated with linoleic acid and linolenic acid. These negative correlations between oleic acid content and other fatty acid contents may be due to their roles in the fatty acid biosynthesis pathway, where one fatty acid is the direct precursor of the other (Ohlrogge and Browse, 1995). Oil showed positive correlation with oleic acid and linoleic acid (0.39 and 0.40, respectively) but negative correlation with linolenic acid (-0.62) (Table 4). This result is consistent with results reported by La et al. (2014) and Wu et al. (2017) in which there was a positive correlation between oil and oleic acid but a negative correlation between oil and linolenic acid. Inconsistently, La et al. (2014) and Wu et al. (2017) reported a negative correlation between oil and oleic acid. This can be explained by the materials they used in their studies. La et al. (2014) studied soybean lines [*Glycine max* (L.) Merr.] with high and normal oleic content, and Wu et al. (2017) characterized soybean cultivars.

In this study, sucrose showed a positive correlation with raffinose but a non-significant correlation with stachyose in wild soybeans. This result is not consistent with the study of Neus et al. (2005). In their study, Neus et al. (2005) used two populations developed from crosses between two elite lines and PI200508. They observed that sucrose was negatively correlated with both raffinose and stachyose. We also found significant correlations between protein and glucose (0.29), protein and stachyose (-0.29), oil and fructose (-0.29), and oil and raffinose (-0.24). The positive correlation between protein and stachyose in the study of Neus et al. (2005) suggests that it may be challenging to develop soybean lines with low stachyose content and high protein content.

When the amino acid contents of the core collection are calculated based on seed dry weight, the amino acid contents tend to increase in parallel to an increase in protein content (Figure 1a, c; Figure 2a, c). When these amino acids were calculated based on the contents of seed protein, all amino acids, except glutamic acid, showed inverse relations with seed protein content. The exception of glutamic acid may be due to the role of this amino acid in the biosynthetic pathway. In this pathway, glutamic acid acts as basic precursor for the synthesis of arginine, aspartate, glutamine, and proline (Taiz and Zeiger, 2013). All studied amino acids showed strong and positive correlations between themselves, but negative correlation, except glutamic acid, with the content of seed crude protein (Table 5). Soybean lines with higher seed protein content typically have a lower concentration of essential amino acids (cysteine, lysine, methionine, threonine, and tryptophan) than soybean lines with lower protein content (Medic et al., 2014). These correlations may, in part, be due to the determination of soybean seed amino acid content via nitrogen assimilation during the seed filling stage (Sebastià et al., 2005)

Linkage disequilibrium

The linkage disequilibrium (LD), which was indicated as R-squared in Fig. 3 decreased to half of its highest value at 7.5kbp when LD was calculated for the whole genome. This value is lower than that in wild soybean population in the studies of Zhou et al. (2015) and Leamy et al. (2017) (27kbp and >100kbp, respectively). Because LD in wild soybean was lower than that in soybean, Leamy et al. (2017) suggested that more QTL would be found if more markers were used in wild soybean population. We also observed a difference in LD decay between euchromatic and heterochromatic regions (5.2kbp and 148.6kbp, respectively). Zhou et al. (2015) and Hyten et al. (2007) reported similar results when they studied different soybean populations.

Genome-wide association study

For maturity, five SNPs on chromosomes 1, 2, 6, 9, and 12 passed the threshold of the genome-wide association (Table 7; Figure 4). All of these SNPs had late maturity effect. The SNP on chromosome 6 would delay the maturity up to 10 days, and the physical location of this marker was ~4Mb away from maturity gene E1 with reference to the Williams 82 sequence (http://www.soybase.org, "SoyBase browser", accessed 05/20/2018). Another SNP on chromosome 12 that was associated with maturity and about 1 MB away from Satt317. Satt317 was reported to be significantly associated to maturity (Eskandari et al., 2013). Ten different SNPs were significantly associated with seed content of oleic and linoleic acids. These SNPs located on nine different chromosomes, with one

SNP on chromosome 2, 3 4, 7, 8, 13, 14, and 20; and two SNPs on chromosome 11. The seed content of the corresponding fatty acid could be changed by 2 to 9 g kg^{-1} due to the effect of these SNPs.

No marker was found significantly associated with seed content of protein or oil. This could be explained by the low attribution of markers to the phenotypic variance of studied traits (Gibson, 2012; Korte and Farlow, 2013). Yan et al. (2017) also stated that increasing the frequency of the genotypes containing causative alleles would increase the chance of detecting variant with major effects. The core collection in this study was chosen to represent maximum genetic diversity of 806 *G. soja* accessions; therefore, it could have low frequency of causative alleles. Due to high genetic diversity of the core collection, the extent of LD was lower than that in other GWAS studies (Hwang et al., 2014; Zhou et al., 2015; Leamy et al., 2017; Zeng et al., 2017), and a higher marker density could be required to make sure that the genome was covered adequately for a GWAS study (Nordborg and Tavaré, 2002). In 2011, Ingvarsson and Street (2011) reported that a larger sample size might be require to facilitate the discovery of associations between genetic markers and studied traits.

It's necessary to enhance the utilization of germplasm in soybean breeding for improving the genetic base of the new developed cultivars. Evaluation and characterization of the whole USDA soybean collection (1,168 *G. soja*) using multiple replications and environments would be prohibitively costly due to the very large size of the collection. In this study, the core collection consists of 80 accessions that represent the diversity of 1,168 *G. soja*. This core collection was characterized by using the data of six environments, and would provide useful information to evaluate and identify valuable parental lines. The core collections in chickpea (*Cicer arietinum* L.), groundnut (*Arachis hypogaea* L.), foxtail millet [*Setaria italic* (L.) P. Beauv.] and sorghum [*Sorghum bicolor* (L.) Moench] have provided many valuable traits. Evaluating sorghum core collection resulted in accessions resistant to grain mold, downy mildew, foliar diseases, anthracnose (Sharma et al., 2010; Sharma et al., 2012; Upadhyaya et al., 2013), and accessions tolerant to drought (Sakhi et al., 2014; Upadhyaya et al., 2017), Similarly, the core collection of foxtail millet was also the source of accessions that have drought tolerance and salinity tolerance (Krishnamurthy et al., 2016; Upadhyaya et al., 2017). Likewise, this wild soybean core collection with fewer number of accessions (80) could be evaluated more extensively for traits of agronomic and economic importance. In addition, due to the small size of the core collection, soybean breeders can simplify their management and enhance their utilization of wild soybean genetic resources. The characterization of this core collection will be an important step to further study and explore the genetic resources in wild soybean and all data from this study are publicly available via GRIN. Based on the genome-wide association analysis of maturity, seed weight, and seed compositions, we observed markers with significant association with the studied traits. The information about these markers suggests some new candidate genes and assist further studies to identify genes controlling the maturity, seed weight, as well as seed content of aspartic acid, glutamine, palmitic acid, oleic acid, and linoleic acid.

References

Akpertey, A., M. Belaffif, G.L. Graef, M.A.R. Mian, J.G. Shannon, P.B. Cregan, et al. 2014. Effects of selective genetic introgression from wild soybean to soybean. Crop Sci. 54: 2683-2695.

Asekova, S., J.H. Chae, B.K. Ha, K.H. Dhakal, G. Chung, J.G. Shannon, et al. 2014. Stability of elevated [alpha]-linolenic acid derived from wild soybean (Glycine soja Sieb. & Zucc.) across environments. Euphytica: International Journal of Plant Breeding 195: 409.

Asif, M. 2011. Health effects of omega-3,6,9 fatty acids: Perilla frutescens is a good example of plant oils. Orient. Pharm. Exp. Med. 11: 51-59.

Banaszkiewicz, T. 2011. Nutritional value of soybean meal. In: H. El-Shemy, editor Soybean and Nutrition. InTech, Rijeka. p. Ch. 01.

Bellaloui, N., J.R. Smith, J.D. Ray, and A.M. Gillen. 2009. Effect of maturity on seed composition in the early soybean production system as measured on near-isogenic soybean lines. Crop Sci. 49: 608-620.

Bordes, J., C. Ravel, J. Le Gouis, A. Lapierre, G. Charmet, and F. Balfourier. 2011. Use of a global wheat core collection for association analysis of flour and dough quality traits. J. Cereal Sci. 54: 137-147.

Brace, R.C., W.R. Fehr, and S.R. Schnebly. 2011. Agronomic and seed traits of soybean lines with high oleate concentration. Crop Sci. 51: 534-541.

Bradbury, P.J., Z. Zhang, D.E. Kroon, T.M. Casstevens, Y. Ramdoss, and E.S. Buckler. 2007. TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics 23: 2633-2635.

Burton, J.W. 1987. Quantitative genetics: Results relevant to soybean breeding. p. 211–247. In J.R.Wilcox (ed.) Soybeans: Improvement production and uses. 2nd ed. Agron. Monogr. 16. ASA, CSSA, and SSSA, Madison, WI. Agronomy (USA).

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

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

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

Coimbra, R.R., G.V. Miranda, C.D. Cruz, D.J.H. Silva, and R.A. Vilela. 2009. Development of a Brazilian maize core collection. Genet. Mol. Biol. 32: 538-545.

Covington, M.B. 2004. Omega-3 fatty acids. Atlantic 1.

Dhakal, K.H., J.D. Lee, Y.S. Jeong, H.S. Kim, J.G. Shannon, and Y.H. Hwang. 2013. Stability of linolenic acid in seed oil of soybean accessions with elevated linolenic acid concentration. J. Food, Agric. Environ. 11: 80-85.

Dornbos, D.L., and M.B. McDonald. 1986. Mass and composition of developing soybean seeds at five reproductive growth stages. Crop Sci. 26: 624-630.

Erickson, P.S., D.J. Schauff, and M.R. Murphy. 1989. Diet digestibility and growth of holstein calves fed acidified milk replacers containing soy protein concentrate. J. Dairy Sci. 72: 1528-1533.

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

Eskandari, M., E.R. Cober, and I. Rajcan. 2013. Genetic control of soybean seed oil: II. QTL and genes that increase oil concentration without decreasing protein or with increased seed yield. Theor. Appl. Genet. 126: 1677-1687.

Frankel, O.H., and A.H.D. Brown. 1984. Plant genetic resources today: A critical appraisal. Crop Genetic Resources: Conservation & Evaluation (J.H.W. Holden and J.T. Williams, eds.). George Alien & Unwin Ltd., London: 249-257.

George, A.A., and B.O. De Lumen. 1991. A novel methionine-rich protein in soybean seed: identification, amino acid composition, and N-terminal sequence. J. Agric. Food Chem. 39: 224-227.

Gibson, G. 2012. Rare and common variants: twenty arguments. Nat. Rev. Genet. 13: 135-145.

Gizlice, Z., T.E. Carter Jnr, and J.W. Burton. 1994. Genetic base for North American public soybean cultivars released between 1947 and 1988. Crop Sci. 34: 1143-1151.

Grieshop, C.M., and G.C. Fahey. 2001. Comparison of quality characteristics of soybeans from Brazil, China, and the United States. J. Agric. Food Chem. 49: 2669-2673.

Guleria, S., S.K. Sharma, B.S. Gill, and S.K. Munshi. 2008. Distribution and biochemical composition of large and small seeds of soybean (Glycine max L.). J. Sci. Food Agric. 88: 269-272.

Guo, A., and Z. Petrovic. 2005. Vegetable based polyols. p. 110–130. In S.Z. Erham (ed.) Industrial uses of vegetable oils. AOCS Press, Champaign, Illinois. .

Holbrook, C.C., P. Timper, and H.Q. Xue. 2000. Evaluation of the core collection approach for identifying resistance to Meloidogyne arenaria in peanut. Crop Sci. 40: 1172-1175.

Hou, A., P. Chen, J. Alloatti, D. Li, L. Mozzoni, B. Zhang, et al. 2009. Genetic variability of seed sugar content in worldwide soybean germplasm collections. Crop Sci. 49: 903-912.

Huang, X.H., X.H. Wei, T. Sang, Q.A. Zhao, Q. Feng, and Y. Zhao. 2010. Genome-wide association studies of 14 agronomic traits in rice landraces. Nat. Genet. 42.

Hwang, E.Y., Q.J. Song, G. Jia, J.E. Specht, D.L. Hyten, J. Costa, et al. 2014. A genome-wide association study of seed protein and oil content in soybean. BMC genomics 15: 1-1.

Hymowitz, T. 1972. Relationship between the content of oil, protein, and sugar in soybean seed. Agron. J. 64: 613.

Hyten, D.L., I.Y. Choi, Q. Song, R.C. Shoemaker, R.L. Nelson, J.M. Costa, et al. 2007. Highly variable patterns of linkage disequilibrium in multiple soybean populations. Genetics 175: 1937-1944.

Hyten, D.L., Q.J. Song, Y.L. Zhu, I.Y. Choi, R.L. Nelson, and J.M. Costa. 2006. Impacts of genetic bottlenecks on soybean genome diversity. Proc. Natl. Acad. Sci. U.S.A. 103.

Imsande, J. 2001. Selection of soybean mutants with increased concentrations of seed methionine and cysteine. Crop Sci. 41: 510-515.

Ingvarsson, P.K., and N.R. Street. 2011. Association genetics of complex traits in plants. New Phytol. 189: 909-922.

Jarquin, D., J. Specht, and A. Lorenz. 2016. Prospects of genomic prediction in the USDA soybean germplasm collection: Historical data creates robust models for enhancing selection of accessions. G3: Genes, Genomes, Genet. 6: 2329-2341.

Jin, Y., T. He, and B.R. Lu. 2003. Fine scale genetic structure in a wild soybean (Glycine soja) population and the implications for conservation. New Phytologist 159: 513-519.

Johnson, H.W. 1958. Registration of soybean varieties: VI. Agron. J. .

Kabelka, E.A., S.R. Carlson, and B.W. Diers. 2006. PI 468916 SCN resistance loci's associated effects on soybean seed yield and other agronomic traits. Crop Sci. 46: 622-629.

Karr-Lilienthal, L.K., C.M. Grieshop, J.K. Spears, and G.C. Fahey. 2005. Amino acid, carbohydrare, and fat composition of soybean meals prepared at 55 commercial U.S. soybean processing plants. J. Agric. Food Chem. 53: 2146-2150.

Korte, A., and A. Farlow. 2013. The advantages and limitations of trait analysis with GWAS: a review. Plant Methods 9: 1-9.

Krishnamurthy, L., H.D. Upadhyaya, J. Kashiwagi, R. Purushothaman, S.L. Dwivedi, and V. Vadez. 2016. Variation in drought-tolerance components and their interrelationships in the core collection of foxtail millet (Setaria italica) germplasm. Crop Pasture Sci. 67: 834-846.

Krishnan, H.B. 2005. Engineering soybean for enhanced sulfur amino acid content. Crop Sci. 45: 454-461.

Kumar, V., A. Rani, L. Goyal, A.K. Dixit, J.G. Manjaya, J. Dev, et al. 2010. Sucrose and raffinose family oligosaccharides (RFOs) in soybean seeds as influenced by genotype and growing location. J. Agric. Food Chem. 58: 5081-5085.

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

La, T.C., H.T. Nguyen, J.G. Shannon, S.M. Pathan, T. Vuong, J.D. Lee, et al. 2014. Effect of high-oleic acid soybean on seed oil, protein concentration, and yield. Crop Sci. 54: 2054-2062.

Lam, H.M., B. Wang, J. Li, M. Jian, J. Wang, G. Shao, et al. 2010. Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection. Nat. Genet. 42: 1053.

Leamy, L.J., H. Zhang, C. Li, C.Y. Chen, and B.H. Song. 2017. A genome-wide association study of seed composition traits in wild soybean (Glycine soja). BMC Genomics 18: 18.

Lee, J., Y.-S. Hwang, S.T. Kim, W.-B. Yoon, W.Y. Han, I.-K. Kang, et al. 2017. Seed coat color and seed weight contribute differential responses of targeted metabolites in soybean seeds. Food Chemistry 214: 248-258.

Lee, J., Y.S. Hwang, S.T. Kim, W.B. Yoon, W.Y. Han, I.K. Kang, et al. 2017. Seed coat color and seed weight contribute differential responses of targeted metabolites in soybean seeds. Food Chem. 214: 248-258.

Lee, J.D., K.D. Bilyeu, and J.G. Shannon. 2007. Genetics and breeding for modified fatty acid profile in soybean seed oil. Crop Science Biotechnology 10: 201-210.

Lee, J.D., J.K. Yu, Y.H. Hwang, S. Blake, Y.S. So, G.J. Lee, et al. 2008. Genetic diversity of wild soybean (Glycine soja Sieb. and Zucc.) accessions from south Korea and other countries. Crop Sci. 48: 606-616.

LeRoy, A.R., W.R. Fehr, and S.R. Cianzio. 1991. Introgression of genes for small seed size from Glycine soja into G. max. Crop Sci. 31: 693-697.

Liu, K. 1997. Soybeans: chemistry, technology, and utilization.Chapman & Hall, New York.

Liu, W.G., M.Q. Shahid, L. Bai, Z. Lu, Y. Chen, L. Jiang, et al. 2016. Evaluation of genetic diversity and development of a core collection of wild rice (Oryza rufipogon Griff.) populations in China. PLoS One 10: e0145990.

Mattson, F.H., and S.M. Grundy. 1985. Comparison of effects of dietary saturated, monounsaturated, and polyunsaturated fatty acids on plasma lipids and lipoproteins in man. J. Lipid Res. 26: 194-202.

Medic, J., C. Atkinson, and C.R. Hurburgh. 2014. Current knowledge in soybean composition. J. Am. Oil Chem. Soc. 91: 363-384.

Mozaffarian, D., and E.B. Rimm. 2006. Fish intake, contaminants, and human health: Evaluating the risks and the benefits. JAMA 296: 1885-1899.

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.

Nichols, D.M., K.D. Glover, S.R. Carlson, J.E. Specht, and B.W. Diers. 2006. Fine mapping of a seed protein QTL on soybean linkage group I and its correlated effects on agronomic traits. Crop Sci. 46: 834-839.

Nordborg, M., and S. Tavaré. 2002. Linkage disequilibrium: what history has to tell us. Trends Genet. 18: 83-90.

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

Ohlrogge, J., and J. Browse. 1995. Lipid biosynthesis. The Plant Cell 7: 957-970.

Pantalone, V.R., G.J. Rebetzke, J.W. Burton, and R.F. Wilson. 1997. Genetic regulation of linolenic acid concentration in wild soybean Glycine soja accessions. J. Am. Oil Chem. Soc. 74: 159-163.

Pelaez, R., and D.M. Walker. 1979. Milk replacers for preruminant lambs: limiting amino acids in two soybean protein isolates determined with a change-over design. Aust. J. Agric. Res. 30: 125-134.

Poeta, F., L. Borrás, and J.L. Rotundo. 2016. Variation in seed protein concentration and seed size affects soybean crop growth and development. Crop Sci. 56: 3196-3208.

Price, A.L., N.J. Patterson, R.M. Plenge, M.E. Weinblatt, N.A. Shadick, and D. Reich. 2006. Principal components analysis corrects for stratification in genome-wide association studies. Nat. Genet. 38: 904-909.

Qi, Z., Q. Wu, X. Han, Y. Sun, X. Du, C. Liu, et al. 2011. Soybean oil content QTL mapping and integrating with meta-analysis method for mining genes. Euphytica 179: 499-514.

Sakhi, S., S. Rehman, O. Kazutoshi, A. Shahzad, and M. Jamil. 2014. Evaluation of sorghum (Sorghum bicolor) core collection for drought tolerance: Pollen fertility and mean performance of yield traits and its components at reproductive stage. Int. J. Agric. Biol. 16: 251-260.

Saldivar, X., Y.J. Wang, P. Chen, and A. Hou. 2011. Changes in chemical composition during soybean seed development. Food Chem. 124: 1369-1375.

Sebastià, C.H., F. Marsolais, C. Saravitz, D. Israel, R.E. Dewey, and S.C. Huber. 2005. Free amino acid profiles suggest a possible role for asparagine in the control of storage-product accumulation in developing seeds of low- and high-protein soybean lines. J. Exp. Bot. 56: 1951-1963.

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

Sharma, R., H.D. Upadhyaya, S.V. Manjunatha, V.P. Rao, and R.P. Thakur. 2012. Resistance to foliar diseases in a mini-core collection of sorghum germplasm. Plant Dis. 96: 1629-1633.

Sharma, R.K., V.P. Rao, H.D. Upadhyaya, V.G. Reddy, and R.P. Thakur. 2010. Resistance to grain mold and downy mildew in a mini-core collection of sorghum germplasm. Plant Dis. 94: 439-444.

Shivakumar, M., C. Gireesh, and A. Talukdar. 2016. Efficiency and utility of pollination without emasculation (PWE) method in intra and inter specific hybridization in soybean. Indian J. Genet. 76.

Simopoulos, A.P. 2002. The importance of the ratio of omega-6/omega-3 essential fatty acids. Biomed. Pharmacother. 56: 365-379.

Simopoulos, A.P. 2008. The importance of the omega-6/omega-3 fatty acid ratio in cardiovascular disease and other chronic diseases. Exp. Biol. Med. 233: 674-688.

Singh, S.K., J.Y. Barnaby, V.R. Reddy, and R.C. Sicher. 2016. Varying response of the concentration and yield of soybean seed mineral elements, carbohydrates, organic acids, amino acids, protein, and oil to phosphorus starvation and CO(2) enrichment. Frontiers in Plant Science 7: 1967.

Song, Q., D.L. Hyten, G. Jia, C.V. Quigley, E.W. Fickus, R.L. Nelson, et al. 2013. Development and evaluation of SoySNP50K, a high-density genotyping array for soybean. PLoS One 8: e54985.

Song, Q., D.L. Hyten, G. Jia, C.V. Quigley, E.W. Fickus, R.L. Nelson, et al. 2015. Fingerprinting soybean germplasm and its utility in genomic research. G3: Genes, Genomes, Genet.

Taiz, L., and E. Zeiger. 2013. Plant Physiology. 5. ed, Sunderland: Sinauer Associates. 782 p.

Takahashi, M., Y. Uematsu, K. Kashiwaba, K. Yagasaki, M. Hajika, R. Matsunaga, et al. 2003. Accumulation of high levels of free amino acids in soybean seeds through integration of mutations conferring seed protein deficiency. Planta 217: 577-586.

Tanksley, S.D., and S.R. McCouch. 1997. Seed banks and molecular maps: unlocking genetic potential from the wild. Science (New York, N.Y.) 277: 1063-1066.

Upadhyaya, H.D., S.L. Dwivedi, M. Vetriventhan, L. Krishnamurthy, and S.K. Singh. 2017. Post-flowering drought tolerance using managed stress trials, adjustment to flowering, and mini core collection in sorghum. Crop Sci. 57: 310-321.

Upadhyaya, H.D., Y.H. Wang, R. Sharma, and S.K. Sharma. 2013. Identification of genetic markers linked to anthracnose resistance in sorghum using association analysis. Theor. Appl. Genet. 126: 1649-1657.

Vaughn, J.N., R.L. Nelson, Q. Song, P.B. Cregan, and Z. Li. 2014. The genetic architecture of seed composition in soybean is refined by genome-wide association scans across multiple populations. G3: Genes, Genomes, Genet. 4: 2283-2294.

Vineet, K., R. Anita, S. Shweta, and S.M. Hussain. 2006. Influence of growing environment on the biochemical composition and physical characteristics of soybean seed. J. Food Compos. Anal. 19: 188-195.

Wang, K.-J., X.-H. Li, and M.-F. Yan. 2014. Genetic differentiation in relation to seed weights in wild soybean species (Glycine soja Sieb. & Zucc.). Plant Systematics and Evolution 300: 1729-1739.

Wang, K.J., and X.H. Li. 2012. Genetic characterization and gene flow in different geographical-distance neighbouring natural populations of wild soybean (Glycine soja Sieb. & Zucc.) and implications for protection from GM soybeans. Euphytica 186: 817-830.

Wang, K.J., X.H. Li, and M.F. Yan. 2014. Genetic differentiation in relation to seed weights in wild soybean species (Glycine soja Sieb. & Zucc.). Plant Syst. Evol. 300: 1729-1739.

Wang, K.J., H.L. Xiang, T. Yamashita, and T. Yoshihito. 2012. Single nucleotide mutation leading to an amino acid substitution in the variant Tik soybean Kunitz trypsin inhibitor (SKTI) identified in Chinese wild soybean (Glycine sojaSieb. & Zucc.). Plant Syst. Evol. 298.

Wang, X., G.L. Jiang, Q. Song, P.B. Cregan, R.A. Scott, J. Zhang, et al. 2015. Quantitative trait locus analysis of seed sulfur-containing amino acids in two recombinant inbred line populations of soybean. Euphytica 201: 293-305.

Warrington, C.V., H. Abdel-Haleem, D.L. Hyten, P.B. Cregan, J.H. Orf, A.S. Killam, et al. 2015. QTL for seed protein and amino acids in the Benning × Danbaekkong soybean population. Theor. Appl. Genet.

Warrington, C.V., H. Abdel-Haleem, D.L. Hyten, P.B. Cregan, J.H. Orf, A.S. Killam, et al. 2015. QTL for seed protein and amino acids in the Benning × Danbaekkong soybean population. Theor. Appl. Genet. 128: 839-850.

Weiss, M.G., and T.M. Stevenson. 1955. Registration of soybean varieties: V. Agron. J. 47:541–543.

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

Wilson, R.F. 2004. Seed composition. In H.R. Boerma and J.E. Specht (ed.) Soybeans: Improvement, Production, and Uses. 3rd ed. ASA, CSSA, and SSSA, Madison, WI.: 621-677.

Wu, T., X.H. Yang, S. Sun, C. Wang, Y. Wang, H.C. Jia, et al. 2017. Temporal–spatial characterization of seed proteins and oil in widely grown soybean cultivars across a century of breeding in China. Crop Sci. 57: 748-759.

Yadav, N.S. 1996. Genetic modification of soybean oil quality. In DPS Verma, RC Shoemaker, eds, Soybean: Genetics, Molecular Biology and Biotechnology, CAB INTERNATIONAL, Wallingford, UK, pp 165-168.

Yan, L., N. Hofmann, S. Li, M.E. Ferreira, B. Song, G. Jiang, et al. 2017. Identification of QTL with large effect on seed weight in a selective population of soybean with genome-wide association and fixation index analyses. BMC genomics 18: 529.

Yoon, M.S., J. Lee, C.Y. Kim, J.H. Kang, E.G. Cho, and H.J. Baek. 2009. DNA profiling and genetic diversity of Korean soybean (Glycine max (L.) Merrill) landraces by SSR markers. Euphytica 165: 69-77.

Yu, X., F. Yuan, X. Fu, and D. Zhu. 2016. Profiling and relationship of water-soluble sugar and protein compositions in soybean seeds. Food Chem. 196: 776-782.

Yu, X., F. Yuan, X. Fu, and D. Zhu. 2016. Profiling and relationship of water-soluble sugar and protein compositions in soybean seeds. Food Chemistry 196: 776-782.

Yuan, Y.C., S.Y. Gong, H.J. Yang, Y.C. Lin, D.H. Yu, and Z. Luo. 2011. Effects of supplementation of crystalline or coated lysine and/or methionine on growth performance and feed utilization of the Chinese sucker, Myxocyprinus asiaticus. Aquaculture 316: 31-36.

Zeng, A., P. Chen, K. Korth, F. Hancock, A. Pereira, K. Brye, et al. 2017. Genome-wide association study (GWAS) of salt tolerance in worldwide soybean germplasm lines. Mol. Breed. 37: 30.

Zhang, J., and Y. Huang. 2011. Preparation and optical properties of AgGaS2 nanofilms. Cryst. Res. Technol. 46: 501-506.

Zhou, Z., Y. Jiang, Z. Wang, Z. Gou, J. Lyu, W. Li, et al. 2015. Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean. Nat. Biotech. 33: 408-414.

Tables and figures

Table 2-1. Summary information of environments used in the study

Code	Year	Location	Latitude	Longitude	Soil type	Number of genotypes tested
13CLA	2013	Clayton, NC	35º38’ N	78º27’ W	Johns fine sandy loam soil	64
13CLM	2013	Columbia, MO	38º59’ N	92º12’ W	Mexico silt loam	80
14CLA	2014	Clayton, NC	35˚38’ N	78˚27’ W	Johns fine sandy loam soil	64
14CLM	2014	Columbia, MO	38º59’ N	92º12’ W	Mexico silt loam	80
14SAN	2014	Jackson Springs, NC	35º11’ N	79º40’ W	deep sandy soil	64
15NOV	2015	Novelty, MO	40º01’ N	92º11’ W	Putnam silt loam	80

Table 2-2. Summarized information about maturity, seed weight and seed compositions of 80 *G. soja* plant introductions across three environments, 13CLM, 14CLM, and 15NOV.

Trait	Range	Check’s range	Mean	h^2 (entry-mean basis)	h^2 (plot basis)	CV (%)	P-value[‡]	LSD$_{0.05}$
Maturity, day	102-174	116-172	147	0.99	0.97	1.8	<0.0001	1
Seed weight, g 100seed^{-1}	0.9-3.5	11.9-16.2	1.8	0.94	0.79	9.1	<0.0001	0.2
Crude protein, g kg^{-1} [§]	392.6-481.7	360.2-402.5	438.9	0.91	0.59	3.1	<0.0001	13.8
Oil, g kg^{-1} [§]	157.6-175.8	214.0-258.9	164.7	0.86	0.51	1.6	<0.0001	2.6
C16:0, g kg^{-1} [¶]	110.5-140.1	107.3-114.0	127.9	0.65	0.17	6.5	<0.0001	8.5
C18:0, g kg^{-1} [¶]	28.1-38.5	31.3-42.4	32.3	0.54	0.13	10.7	<0.0001	3.5
C18:1, g kg^{-1} [¶]	107.4-161.9	175.1-271.9	122.1	0.63	0.22	9.5	<0.0001	11.8
C18:2, g kg^{-1} [¶]	510.7-578	522.2-588.5	554.1	0.66	0.20	2.5	<0.0001	13.7
C18:3, g kg^{-1} [¶]	120.3-184.7	11.6-94.5	163.8	0.52	0.14	9.6	<0.0001	15.9
Fructose, g kg^{-1} [§]	5.1-11.6	4.2-7.3	7.1	0.21	0.04	29.5	0.0803	2.1
Glucose, g kg^{-1} [§]	4.3-6.5	4.5-6.0	5.1	0.13	0.02	19.5	0.3368	1.2
Sucrose, g kg^{-1} [§]	14.6-39.5	43.5-69.0	21.5	0.90	0.58	15.4	<0.0001	3.3
Raffinose, g kg^{-1} [§]	6.6-9.3	6.7-8.6	7.8	0.50	0.15	11.2	0.0002	0.9

Trait	Range	Check's range	Mean	h^2 (entry-mean basis)	h^2 (plot basis)	CV (%)	P-value[‡]	$\text{LSD}_{0.05}$
Stachyose, g kg^{-1} [§]	37.2-58.9	35.9-47.1	47.8	0.84	0.38	10.7	<0.0001	5.0
Alanine, g kg^{-1} [#]	38.9-42.9	41.8-43.9	40.8	0.63	0.16	3.2	<0.0001	1.6
Aspartic acid, g kg^{-1} [#]	104.7-116.6	109.5-117.1	111.0	0.55	0.12	3.2	<0.0001	4.3
Cysteine, g kg^{-1} [#]	14.9-19.2	15.3-16.5	16.9	0.84	0.42	4.1	<0.0001	0.8
Glutamic acid, g kg^{-1} [#]	158.3-181.6	168.3-177.3	171.9	0.64	0.17	3.6	<0.0001	7.5
Glycine, g kg^{-1} [#]	39.3-44.4	42.1-43.9	42.4	0.69	0.20	3.2	<0.0001	1.6
Isoleucine, g kg^{-1} [#]	41.1-46.0	44.8-47.7	43.2	0.59	0.13	3.5	<0.0001	1.8
Leucine, g kg^{-1} [#]	68.6-75.6	74.7-78.9	72.2	0.62	0.14	3.3	<0.0001	2.9
Lysine, g kg^{-1} [#]	60.9-67.3	63.9-67.6	64.2	0.62	0.15	3.2	<0.0001	2.4
Methionine, g kg^{-1} [#]	14.0-16.9	15.1-16.5	15.1	0.74	0.23	4.7	<0.0001	0.9
Proline, g kg^{-1} [#]	46.2-52.1	48.4-51.0	49.6	0.47	0.09	3.9	<0.0001	2.3
Threonine, g kg^{-1} [#]	35.1-39.6	38.2-40.1	37.3	0.69	0.19	3.6	<0.0001	1.6
Valine, g kg^{-1} [#]	43.8-48.9	47.3-50.0	46.4	0.30	0.05	4.3	<0.0001	2.4

[‡] P-value of F-test for genotypic source of variance

[§] Seed compositions were calculated as the proportion of each seed composition for seed dry weight

[¶] Fatty acids for each genotype was calculated as the proportion of each fatty acid for the total oil fraction

[#] Amino acid for each genotype was calculated as the proportion of each amino acid for the total protein fraction

Table 2-3. Summarized information about seed weight and seed compositions of 64 *G. soja* plant introductions across six studied environments, 13CLA, 13CLM, 14CLA, 14CLM, 14SAN, 15NOV.

Trait	Range	Mean	Maximum	h^2 (entry-mean basis)	h^2 (plot basis)	CV (%)	P-value[‡]	LSD$_{0.05}$
Seed weight, g 100seed^{-1}	0.9-3.5	1.8	3.5	0.78	0.33	12.19	<0.0001	0.2
Crude protein, g kg^{-1} [§]	392.6-481.7	444.5	481.7	0.91	0.53	3.14	<0.0001	11.3
Alanine, g kg^{-1} [#]	38.6-42.5	40.5	42.5	0.71	0.17	3.05	<0.0001	1.1
Aspartic acid, g kg^{-1} [#]	106.9-116.6	111.0	116.6	0.61	0.12	3.05	<0.0001	3.1
Cysteine, g kg^{-1} [#]	14.6-19.2	16.6	19.2	0.80	0.31	4.52	<0.0001	0.7
Glutamic Acid, g kg^{-1} [#]	160.5-181.6	169.3	181.6	0.69	0.17	3.43	<0.0001	5.3
Glycine, g kg^{-1} [#]	39.5-44.1	42.1	44.1	0.79	0.24	3.00	<0.0001	1.1
Isoleucine, g kg^{-1} [#]	40.9-46.0	43.0	46.0	0.69	0.16	3.43	<0.0001	1.3
Leucine, g kg^{-1} [#]	69.4-75.5	72.4	75.5	0.66	0.14	3.14	<0.0001	2.1
Lysine, g kg^{-1} [#]	60.2-67.2	63.6	67.2	0.70	0.16	3.07	<0.0001	1.8
Methionine, g kg^{-1} [#]	13.8-16.9	15.0	16.9	0.72	0.19	4.46	<0.0001	0.6
Proline, g kg^{-1} [#]	47.2-51.5	48.6	51.5	0.55	0.10	3.57	<0.0001	1.6
Threonine, g kg^{-1} [#]	35.0-39.3	36.9	39.3	0.67	0.16	3.38	<0.0001	1.1
Valine, g kg^{-1} [#]	44.3-48.7	46.2	48.7	0.63	0.13	3.97	<0.0001	1.7

[‡] P-value of F-test for genotypic source of variance

[§] Seed compositions were calculated as the proportion of each seed composition for seed dry weight

[#] Amino acid for each genotype was calculated as the proportion of each amino acid for the total protein fraction

Table 2-4. Trait correlations based on the means of 79 genotypes for maturity, seed weight, seed oil and protein, fatty acids, and soluble carbohydrates over three studied environments, 13CLM, 14CLM, and 15NOV.

	Seed weight	Oil†	CP†	C16:0‡	C18:0‡	C18:1‡	C18:2‡	C18:3‡	Fructose †	Glucose †	Sucrose †	Raffinose †	Stachyose †
Maturity	0.22*	ns	0.37***	-0.39***	-0.41***	ns	ns	0.36**	0.56***	ns	ns	-0.22*	0.76***
Seed weight		0.42***	ns	ns	-0.22*	0.45***	ns	-0.4***	ns	ns	0.65***	ns	ns
Oil †			-0.66***	ns	ns	0.39***	0.4***	-0.62***	-0.29**	ns	0.47***	-0.24*	ns
CP †				ns	ns	-0.28*	-0.23*	0.46***	0.24*	-0.29**	-0.48***	ns	0.29*
C16:0 ‡					ns	ns	-0.29**	ns	ns	ns	ns	ns	-0.35**
C18:0 ‡						ns	ns	-0.27*	-0.3**	ns	ns	ns	ns
C18:1 ‡							-0.35**	-0.54***	ns	ns	0.38***	ns	ns
C18:2 ‡								-0.45***	ns	ns	ns	ns	ns
C18:3 ‡									0.29**	ns	-0.41***	ns	0.26*
Fructose †										-0.1ns	ns	ns	0.36**
Glucose †											0.27*	ns	ns
Sucrose †												0.32**	ns
Raffinose †													ns

* Significant at the 0.05 probability level
** Significant at the 0.01 probability level
*** Significant at the 0.001 probability level
ns: Non-significant
CP, crude protein

† Seed compositions were calculated as the proportion of each seed composition for seed dry weight
‡ Fatty acids for each genotype was calculated as the proportion of each fatty acid for the total oil fraction

Table 2-5. The correlations between seed protein and amino acid composition based on 63 genotype means over six studied environments, 13CLA, 13CLM, 14CLA, 14CLM, 14SAN, 15NOV. Amino acid contents were determined based on seed crude protein content.

	Alanine	Aspartic Acid	Cysteine	Glutamic Acid	Glycine	Isoleucine	Leucine	Lysine	Methionine	Proline	Threonine	Valine
Crude protein	-0.72***	-0.37**	-0.29*	ns	-0.69***	-0.53***	-0.48***	-0.68***	-0.58***	-0.39**	-0.75***	-0.40**
Alanine		0.70***	0.42***	0.45***	0.83***	0.8***	0.76***	0.87***	0.64***	0.65***	0.92***	0.75***
Aspartic Acid			0.48***	0.84***	0.79***	0.83***	0.85***	0.72***	0.6***	0.86***	0.63***	0.70***
Cysteine				0.25*	0.47***	0.39**	ns	0.54***	0.58***	0.45***	0.34**	0.26*
Glutamic Acid					0.65***	0.69***	0.71***	0.52***	0.45***	0.79***	0.36**	0.58***
Glycine						0.79***	0.70***	0.81***	0.67***	0.75***	0.73***	0.68***
Isoleucine							0.88***	0.73***	0.54***	0.80***	0.64***	0.87***
Leucine								0.74***	0.54***	0.78***	0.72***	0.80***
Lysine									0.82***	0.67***	0.84***	0.63***
Methionine										0.49***	0.66***	0.35**
Proline											0.58***	0.70***
Threonine												0.58***

* Significant at the 0.05 probability level

** Significant at the 0.01 probability level

*** Significant at the 0.001 probability level

Table 2-6. Summary of crude protein content (g kg-1) and amino acid contents (g kg-1) that was calculated as the proportion of each amino acid for the total protein fraction across 6 studied environments. Means on the same column with different letters are significantly different (LSD0.05).

Trait	Crude protein	Alanine	Aspartic Acid	Cysteine	Glutamic Acid	Glycine	Isoleucine	Leucine	Lysine	Methionine	Proline	Threonine	Valine
13CLA	429.8 ^{c}	40.8 ^{ab}	112.7 ^{a}	17.8 ^{a}	162.8 ^{c}	42.0 ^{ab}	42.7 ^{b}	72.9 ^{a}	63.2 ^{ab}	15.6 ^{b}	47.4 ^{b}	37.4 ^{a}	46.5 ^{a}
13CLM	431.9 ^{c}	40.6 ^{ab}	110.9 ^{ab}	17.3 ^{ab}	172.1 ^{ab}	42.4 ^{ab}	42.8 ^{b}	71.9 ^{a}	63.8 ^{a}	14.5 ^{c}	48.9 ^{b}	37.6 ^{a}	45.8 ^{a}
14CLA	460.3 ^{a}	39.7 ^{b}	110.3 ^{ab}	15.3 ^{c}	166.6 ^{bc}	41.4 ^{b}	42.6 ^{b}	72.4 ^{a}	61.9 ^{b}	14.4 ^{c}	47.3 ^{b}	35.8 ^{a}	45.7 ^{a}
14CLM	433.6 ^{c}	41.1 ^{a}	108.6 ^{b}	16.9 ^{ab}	166.3 ^{bc}	42.4 ^{a}	42.1 ^{b}	71.3 ^{a}	65.0 ^{a}	14.6 ^{c}	50.9 ^{a}	37.2 ^{a}	45.3 ^{a}
14SAN	460.0 ^{a}	39.9 ^{b}	110.7 ^{ab}	15.7 ^{c}	166.9 ^{bc}	41.7 ^{ab}	42.9 ^{b}	72.7 ^{a}	62.3 ^{b}	14.5 ^{c}	47.6 ^{b}	36.1 ^{a}	45.8 ^{a}
15NOV	451.7 ^{b}	40.8 ^{ab}	113.2 ^{a}	17.1 ^{ab}	175.8 ^{ab}	42.2 ^{ab}	44.6 ^{a}	73.3 ^{a}	64.4 ^{a}	16.5 ^{a}	48.5 ^{b}	37.4 ^{a}	48.0 ^{a}

Table 2-7. SNPs significantly associated with studied traits in core collection of wild soybeans

Trait	SNP	Chr	bp[†]	P-value	$-\log_{10}P$	Allelic effect[‡]
Maturity, day	ss715579536	1	4,554,193	5.03E-09	8.30	2.79
Maturity, day	ss715582748	2	41,350,974	6.58E-08	7.18	2.54
Maturity, day	ss715594007	6	24,465,069	6.91E-13	12.16	10.71
Maturity, day	ss715603290	9	22,054,158	1.10E-14	13.96	8.18
Maturity, day	ss715612498	12	34,900,919	3.14E-10	9.50	6.02
Maturity, day	ss715603525	9	32,425,584	4.48E-07	6.35	0.33
Seed weight, g $100seed^{-1}$	ss715586629	3	45,263,747	1.49E-12	11.83	-0.38
Aspartic acid, g kg^{-1} [#]	ss715616006	13	38,026,151	5.84E-07	6.23	-1.81
Glutamine, g kg^{-1} [#]	ss715616006	13	38,026,151	9.26E-07	6.03	-3.40
C16:0, g kg^{-1} [¶]	ss715637594	20	35,574,654	6.25E-07	6.20	-4.05
C18:1, g kg^{-1} [¶]	ss715585752	3	36,701,848	7.16E-07	6.15	2.36
C18:1, g kg^{-1} [¶]	ss715598001	7	40,940,646	4.10E-07	6.39	-2.92
C18:1, g kg^{-1} [¶]	ss715602718	8	7,864,681	2.99E-10	9.52	-5.98
C18:1, g kg^{-1} [¶]	ss715618857	14	43,083,528	9.92E-07	6.00	3.54
C18:2, g kg^{-1} [¶]	ss715584162	2	9,308,478	1.05E-06	5.98	-3.26
C18:2, g kg^{-1} [¶]	ss715588279	4	46,118,547	1.01E-09	8.99	-9.34
C18:2, g kg^{-1} [¶]	ss715609933	11	27,298,217	1.55E-06	5.81	2.08
C18:2, g kg^{-1} [¶]	ss715610306	11	32,304,429	3.82E-07	6.42	-6.37
C18:2, g kg^{-1} [¶]	ss715616705	13	17,001,366	1.29E-06	5.89	-2.05

† The physical location (in base pairs, bp)

‡ Allelic effect, difference in mean between genotypes with major allele and minor allele

§ Seed compositions were calculated as the proportion of each seed composition for seed dry weight

# Amino acid for each genotype was calculated as the proportion of each amino acid for the total protein fraction

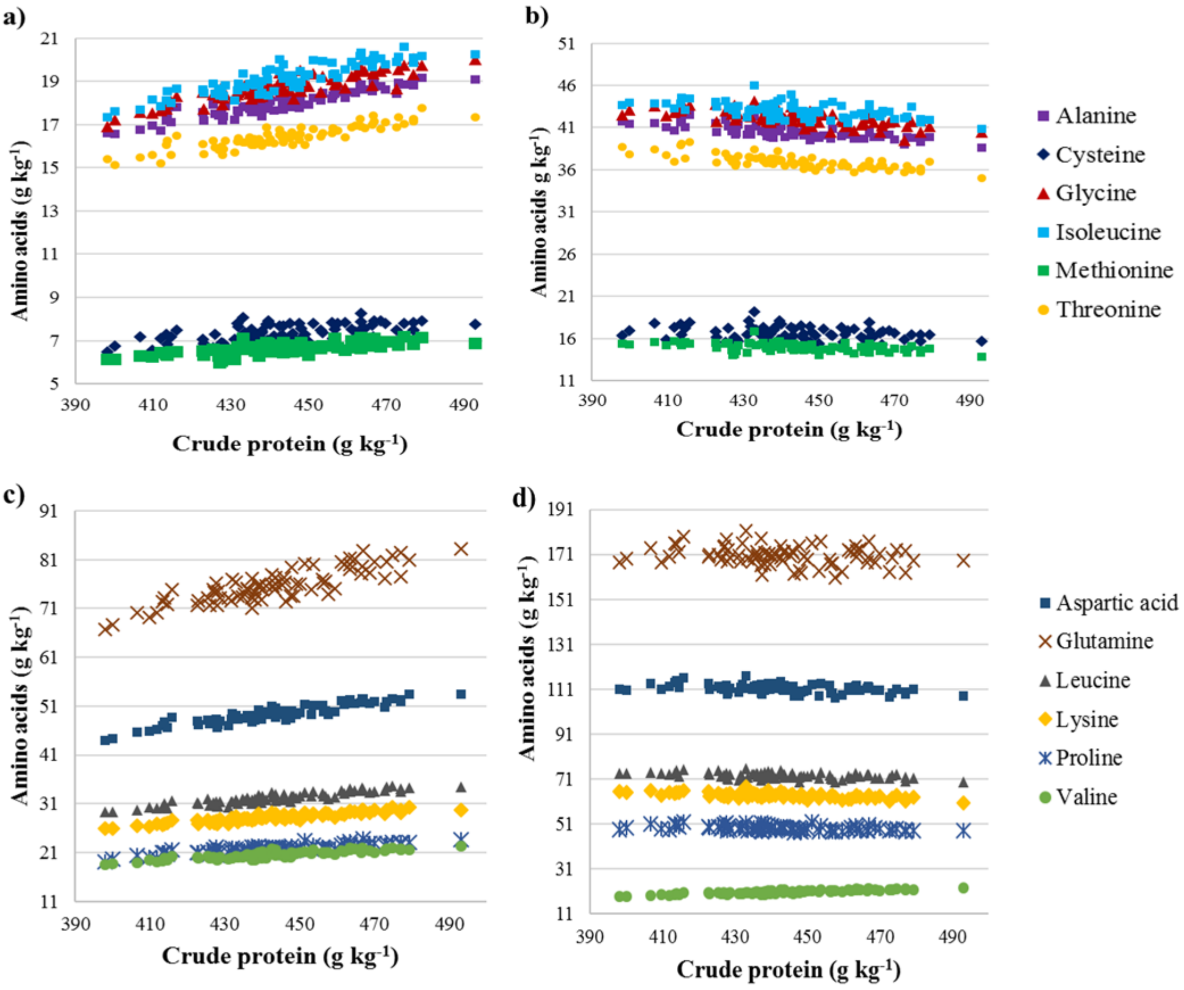


Figure 2-1. The average amino acid as the proportion of each amino acid for the whole soybeans (a,c) and the total protein fraction (b, d) of genotypes with varying protein percentage (n = 79 PIs, three environments, 13CLM, 14CLM, and 15NOV). The crude protein was calculated as the proportion of protein for seed dry weight

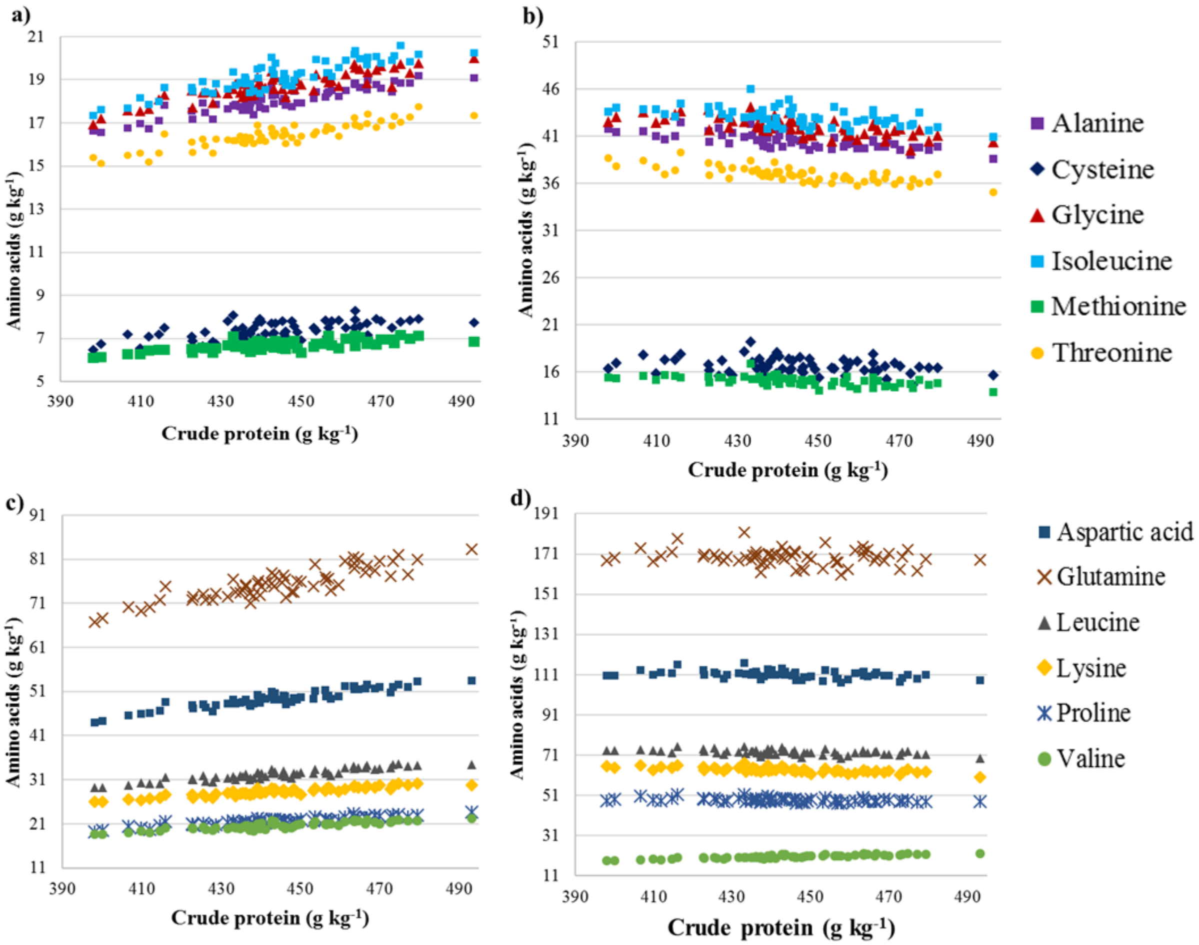
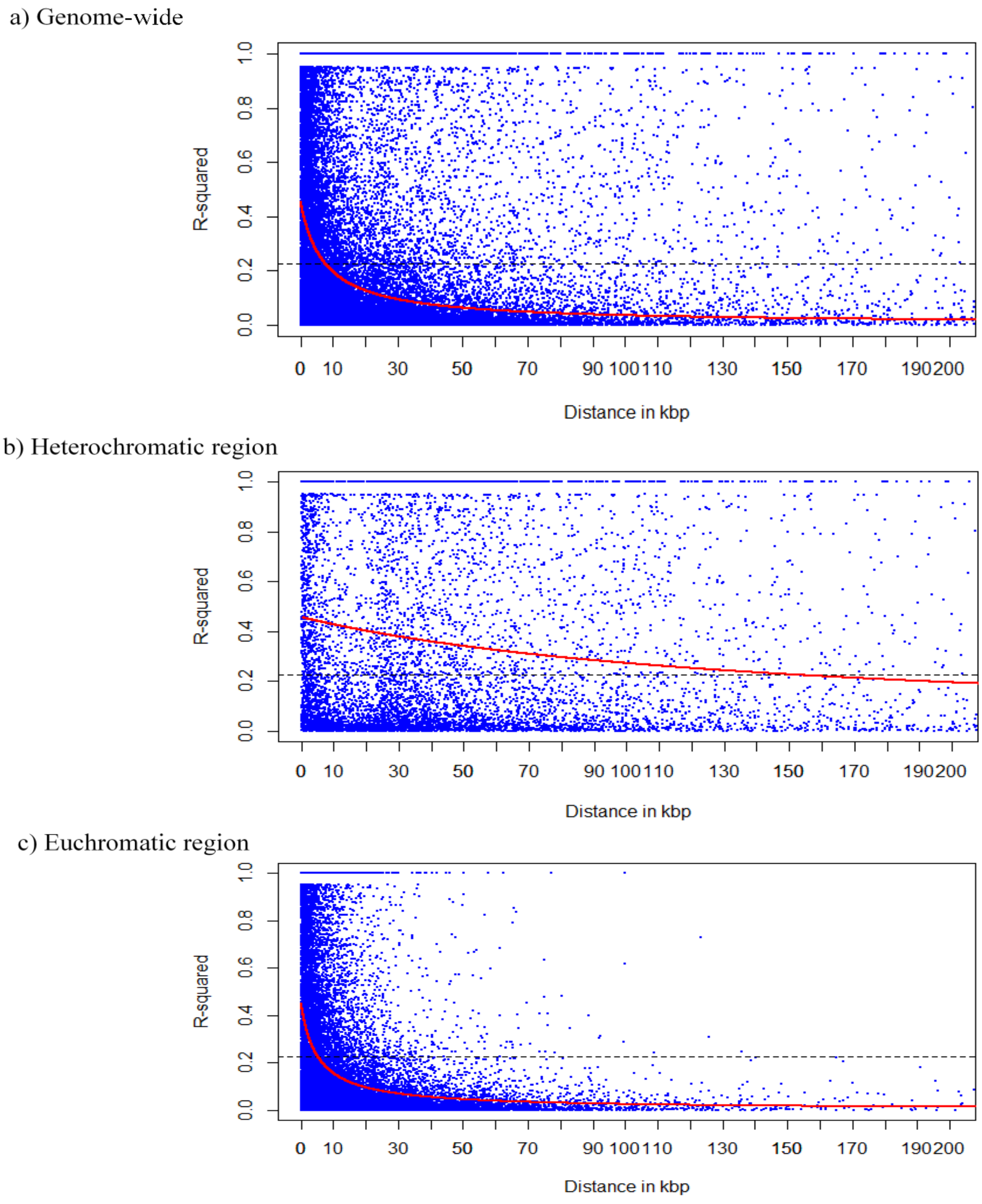


Figure 2-2. The average amino acid as the proportion of each amino acid for the whole soybeans (a,c) and the total protein fraction (b, d) of genotypes with varying protein percentage (n = 63 PIs, six environments, 13CLA, 13CLM, 14CLA, 14CLM, 14SAN, 15NOV). The crude protein was calculated as the proportion of protein for seed dry weight

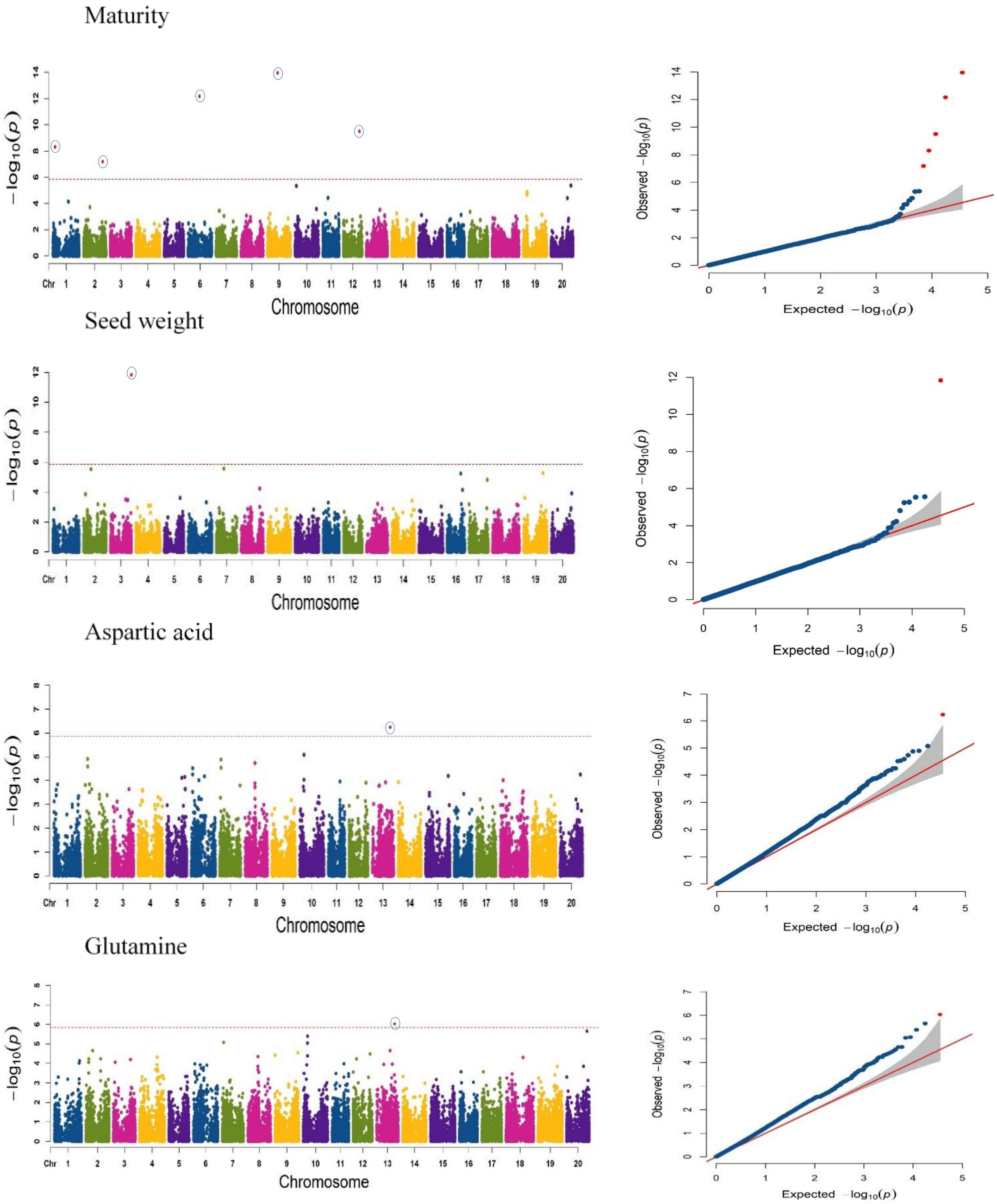


a) Genome-wide

b) Heterochromatic region

c) Euchromatic region

Figure 2-3. The average LD decay of a) the whole genome b) heterochromatic regions and c) euchromatic regions estimated for the core collection of wild soybean.

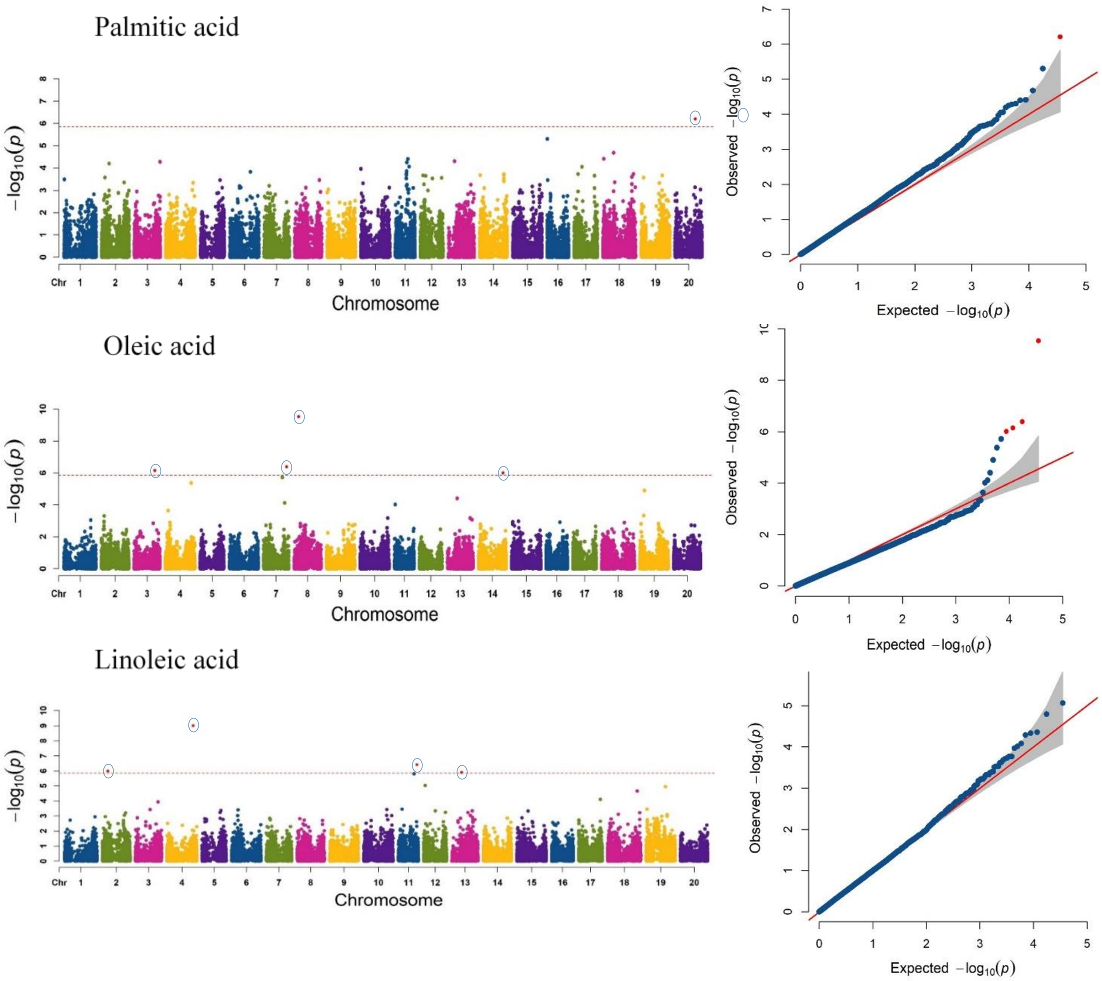

Maturity

Seed weight

Aspartic acid

Glutamine

Figure 2-4. Manhattan plots for maturity, seed weight, and seed content of aspartic acid and glutamine. The yellow horizontal line indicates the genome-wide threshold (-log10(P)=5.85).



Palmitic acid

Oleic acid

Linoleic acid

Figure 2-5. Manhattan plots and QQ plots for seed content of palmitic acid, oleic acid, and linoleic acid. The yellow horizontal line shows the genome-wide threshold (–log10(P)=5.85).

Supplementary tables

Supplemental Table 0-1. Summary information of wild soybean genotypes in mini-core collection

Name	Origin	Cultivar	MG	Descriptors
PI479746B	China	GD 50062	II	PTANBl BBlBl Flk
PI597448D	China		Z	N PTVaNBl BBlBl
PI458536	China		Z	PTENBl BBlBl
PI447003A	China		Z	PTANBl BBlBl Flk
PI101404A	China		II	PTENBl BBlBl
PI407300	China		V	PTENBl BlBl
PI407096	Japan		VII	PTANBl BlBl
PI245331	Taiwan		X	PTENBl LbBlBl Sflk
PI507761	Russia	VIR 8455	I	N PTANBl BBlBl Flk
PI639635	Russian Federation	USSURIISKAYA		N PTESspBl BBlBl Na
PI342622A	Russia		I	PTVaNBl BBlBl Flk
PI522235B	Russia	VIR 8525	I	D PTSaNBl BBlBl Flk Sna
PI522233	Russia	VIR 8522	I	D PTVaNBl BBlBl Flk Na
PI549032	China	ZYD 2632	III	N DpTVaNBl BBlBl Flk
PI479768	China	Long 79-3313-1	Z	PTANBl BBlBl Flk
PI464890B	China	Gong di No. 2019	I	PTENBl BBlBl Flk
PI479752	China	GD 50388-2	I	PTVaNBl BBlBl
PI407206	South Korea	K18	V	PTVaNBl BlBl
PI424082	South Korea	74082	V	PTVaNBl BlBl
PI424102A	South Korea	74101	V	PTVaNBl BlBl
PI424004B	South Korea	74001	II	PTENBl LbBlBl Flk
PI424045	South Korea	74060	V	PTANBl BlBl
PI424070B	South Korea	74051	V	PTANBl BlBl
PI424025B	South Korea	74023	V	PTVaNBl BlBl
PI378686B	Japan		VI	PTANBl BlBl
PI339871A	South Korea		V	PTENBl BlBl
PI522226	Russia	VIR 8511	ZZZ	D PTVaNBl BBlBl Flk
PI407195	South Korea	K12-A	IV	PTVaNBl BBlBl Flk
PI407059	Japan		VII	PTANBl BBlBl Flk
PI407020	Japan		V	PTVaNBl BlBl
PI407042	Japan		V	PTANBl BlBl
PI593983	Japan	Hidaka-6	III	N PTANBl BBlBl Flk
PI407052	Japan		V	PTVaNBl BlBl

Name	Origin	Cultivar	MG	Descriptors
PI378697A	Japan		V	PTENBl BlBl
PI407038	Japan		V	PTVaNBl BBlBl Flk
PI366122	Japan		IV	PTSaNBl DBlBl Flk
PI407157	Japan		VI	PTENBl BlBl
PI507641	Japan	NIAR 060014	V	PTVaN
PI378683	Japan		VI	PTANBl BlBl
PI378684B	Japan		VI	PTANBl BlBl
PI507656	Japan	NIAR 090011	VII	PTAN
PI507618	Japan	NIAR 040015	V	PTSaNBl
PI378690	Japan		VII	PTSaNBl BlBl
PI378696B	Japan		VI	PTANBl BlBl
PI407156	Japan		VI	PTENBl BlBl
PI507624	Japan	NIAR 050002	VII	PTVaNBl DBlBl Flk
PI407085	Japan		VI	PTVaNBl BlBl
PI424083A	South Korea	74083	V	PTANBl BlBl
PI424116	South Korea	74116	IV	PTENBl LbBlBl Flk
PI479751	China	GD 50351-1	III	PTVaNBl BBlBl Flk
PI407240	South Korea	K35-A	V	PTANBl BlBl
PI407171	South Korea	K2-F	IV	PTANBl BBlBl Flk
PI407314	South Korea		V	PTANBl BlBl
PI562547	South Korea	KC1	V	N PTVaNBl BBlBl
PI407179	South Korea	K6-A	V	PTANBl BlBl
PI407231	South Korea	K47-C	V	PTVaNBl BlBl
PI424035	South Korea	74038	V	PTANBl BlBl
PI407214	South Korea	K23-A	V	PTANBl BlBl
PI407228	South Korea	K46-D	V	PTANBl BlBl
PI562551	South Korea	KC26	V	N PTVaNBl BBlBl
PI407248	South Korea	K37-D	V	PTVaNBl BlBl
PI562553	South Korea	KD5	V	N PTVaNBl BBlBl
PI407287	Japan	Tsuru mame	V	PTSaNBl BBlBl
PI597460A	China		IV	D PTVaNBl BBlBl Flk
PI483466	China		V	PTVaNBl BlBl
PI597462B	China		IV	D PTANBl BBlBl Flk
PI407191	South Korea	K57	V	PTVaNBl BBlBl
PI522209B	Russia	VIR 8493	II	D PTVaNBl BBlBl Flk
PI424007	South Korea	74005	V	PTVaNBl BlBl

Name	Origin	Cultivar	MG	Descriptors
PI424123	South Korea	74137	V	PTVaNBl BlBl
PI562565	South Korea	KF13	IV	N PTANBl BBlBl
PI549048	China		III	N PLtENBl BBlBl Flk
PI549046	China	ZYD 3728	III	D PNgENBl SBlBl Flk
PI597461B	China		V	D WTANBl BBlBl Flk
PI639623A	Russian Federation	KBL 552		N PTANBl BBlBl Na
PI639588B	Russian Federation	KA 325		D PTANBl BBlBl Na
PI597458C	China		V	N WTANBl BBlBl Flk
PI562561	South Korea	KE29	V	N PTANBl BBlBl
PI639621	Russian Federation	KT 156		D PTANBl BBlBl
PI639586	Russian Federation	KB 14		D PTVaSspBl BBlBl Na

'The 'Descriptors' column represents phenotypic descriptors as follows: stem termination|flower color|pubesence color|pubesence form|pubesence density|pod wall color|seed coat luster|seed coat color|hilum color

Supplemental Table 0-2. Summary information of checks used in the study

Cultivar	IA3023	IA4005	NKS39-U2	MN0095	MN1410	Ellis	Dillon	NC-Raleigh
MG	III	IV	IV	0	I	V	VI	VII

Supplemental Table 0-3. Analysis of variance of maturity, seed size, and seed composition in 80 wild soybean genotypes grown at three environments, at Columbia, MO in 2013 and 2014, and at Novelty in 2015. Fatty acids were calculated based on oil content. Other seed compositions were calculated based on seed dry weight.

	Df [†]	MS[§]	F-Value	P-value[‡]
Maturity				
Genotype	78	2835.22	122.26	<0.0001
Environment	2	5481.79	211.7	<0.0001
Environment*Genotype	154	23.53	16.35	<0.0001
Replication(Environment)	6	4.74	3.3	0.0035
Residual	423	1.44		
Seed weight				
Genotype	78	2.08	17.55	<0.0001
Environment	2	0.04	0.23	0.7962
Environment*Genotype	154	0.12	4.64	<0.0001
Replication(Environment)	6	0.07	2.59	0.0178
Residual	425	0.03		
Protein				
Genotype	78	7535.95	2.26	<0.0001
Environment	2	39359.00	8.07	0.0048

	Df [†]	MS[§]	F-Value	P-value[‡]
Environment*Genotype	154	3358.44	2.09	<0.0001
Replication(Environment)	6	3211.78	2	0.0642
Residual	426	1604.22		
Oil				
Genotype	78	590.69	1.26	0.1119
Environment	2	4751.66	8.89	0.0012
Environment*Genotype	154	472.17	2.66	<0.0001
Replication(Environment)	6	250.81	1.41	0.2079
Residual	429	177.48		
Alanine				
Genotype	78	2.03	5.32	<0.0001
Environment	2	10.31	21.99	<0.0001
Environment*Genotype	151	0.38	1.38	0.0148
Replication(Environment)	5	0.71	2.56	0.0281
Residual	222	0.28		
Aspartic				
Genotype	78	21.58	4.84	<0.0001
Environment	2	136.01	39.27	<0.0001
Environment*Genotype	151	4.48	1.61	0.0006
Replication(Environment)	5	3.35	1.2	0.3081
Residual	222	2.79		
Cysteine				
Genotype	78	0.89	7.25	<0.0001
Environment	2	0.49	3.38	0.0642
Environment*Genotype	151	0.12	1.87	<0.0001
Replication(Environment)	5	0.23	3.5	0.0046
Residual	222	0.07		
Glutamic				
Genotype	78	68.51	5.2	<0.0001
Environment	2	367.83	26.72	<0.0001
Environment*Genotype	151	13.24	1.7	0.0001
Replication(Environment)	5	19.77	2.54	0.0291
Residual	222	7.77		
Glycine				
Genotype	78	2.12	5.1	<0.0001
Environment	2	5.85	15.6	<0.0001
Environment*Genotype	151	0.42	1.32	0.0293
Replication(Environment)	5	0.40	1.28	0.2756
Residual	222	0.31		
Isoleucine				

	Df [†]	MS[§]	F-Value	P-value[‡]
Genotype	78	2.78	4.55	<0.0001
Environment	2	32.75	26.89	0.0001
Environment*Genotype	151	0.61	1.4	0.0107
Replication(Environment)	5	2.39	5.49	<0.0001
Residual	222	0.44		
Leucine				
Genotype	78	8.24	4.78	<0.0001
Environment	2	54.34	19.5	0.0003
Environment*Genotype	151	1.73	1.55	0.0015
Replication(Environment)	5	5.12	4.58	0.0005
Residual	222	1.12		
Lysine				
Genotype	78	5.76	5.92	<0.0001
Environment	2	29.11	29.19	<0.0001
Environment*Genotype	151	0.98	1.39	0.0136
Replication(Environment)	5	1.29	1.82	0.1097
Residual	222	0.71		
Methionine				
Genotype	78	0.41	4.05	<0.0001
Environment	2	17.65	167.82	<0.0001
Environment*Genotype	151	0.10	0.99	0.5190
Replication(Environment)	5	0.11	1.12	0.3508
Residual	222	0.10		
Proline				
Genotype	78	4.47	4.06	<0.0001
Environment	2	17.91	8.89	0.0062
Environment*Genotype	151	1.10	1.38	0.0139
Replication(Environment)	5	3.83	4.8	0.0003
Residual	222	0.80		
Threonine				
Genotype	78	1.47	4.21	<0.0001
Environment	2	6.79	6.82	0.0193
Environment*Genotype	151	0.35	1.19	0.1231
Replication(Environment)	5	2.13	7.23	<0.0001
Residual	222	0.29		
Valine				
Genotype	78	3.86	4.28	<0.0001
Environment	2	39.60	8.67	0.0156
Environment*Genotype	151	0.90	1.29	0.0440
Replication(Environment)	5	10.88	15.5	<0.0001

	Df [†]	MS[§]	F-Value	P-value[‡]
Residual	222	0.70		
Fructose				
Genotype	78	10.30	1.31	0.0798
Environment	2	648.56	9.12	0.0123
Environment*Genotype	153	7.93	1.93	<0.0001
Replication(Environment)	6	69.11	16.82	<0.0001
Residual	418	4.11		
Glucose				
Genotype	78	1.27	1.17	0.2091
Environment	2	11.96	0.77	0.5014
Environment*Genotype	141	1.10	1.3	0.0408
Replication(Environment)	6	16.33	19.2	<0.0001
Residual	223	0.85		
Sucrose				
Genotype	78	186.97	10.46	<0.0001
Environment	2	3022.22	11.28	0.0079
Environment*Genotype	153	18.01	2.1	<0.0001
Replication(Environment)	6	264.91	30.88	<0.0001
Residual	420	8.58		
Raffinose				
Genotype	78	3.26	1.97	0.0002
Environment	2	241.58	13.39	0.0046
Environment*Genotype	153	1.67	2.51	<0.0001
Replication(Environment)	6	17.46	26.19	<0.0001
Residual	420	0.67		
Stachyose				
Genotype	78	175.77	6.35	<0.0001
Environment	2	771.88	1.04	0.4074
Environment*Genotype	153	27.79	1.47	0.0014
Replication(Environment)	6	748.46	39.63	<0.0001
Residual	420	18.89		
Palmitic				
Genotype	78	1.90	2.88	<0.0001
Environment	2	3.64	0.93	0.4464
Environment*Genotype	152	0.66	0.96	0.6162
Replication(Environment)	6	4.02	5.83	<0.0001
Residual	415	0.69		
Stearic				
Genotype	78	0.31	2.16	<0.0001
Environment	2	0.57	3.59	0.0782

	Df [†]	MS[§]	F-Value	P-value[‡]
Environment*Genotype	152	0.14	1.18	0.1040
Replication(Environment)	6	0.14	1.14	0.3374
Residual	415	0.12		
Oleic				
Genotype	78	7.18	2.68	<0.0001
Environment	2	9.53	2.58	0.1116
Environment*Genotype	152	2.70	2.01	<0.0001
Replication(Environment)	6	2.41	1.79	0.0989
Residual	415	1.34		
Linoleic				
Genotype	78	6.75	2.94	<0.0001
Environment	2	15.32	3.63	0.0788
Environment*Genotype	152	2.30	1.25	0.0421
Replication(Environment)	6	3.81	2.07	0.0553
Residual	415	1.84		
Linolenic				
Genotype	78	8.97	2.08	<0.0001
Environment	2	3.52	0.92	0.4147
Environment*Genotype	152	4.35	1.75	<0.0001
Replication(Environment)	6	2.02	0.81	0.5612
Residual	415	2.49		

[†] Degree of freedom

[§] Mean square

[‡] P-value of F-test for the source of variance

Supplemental Table 0-4. Analysis of variance of maturity, seed size, and seed composition in 80 wild soybean genotypes grown at Columbia, MO in 2013. Fatty acids were calculated based on oil content. Other seed compositions were calculated based on seed dry weight.

	Df [†]	MS[§]	F-Value	P-value[‡]
Maturity				
Genotype	78	1178.72	427.37	<0.0001
Replication	2	10.81	3.92	0.0219
Residual	150	2.76		
Seed weight				
Genotype	78	1.05	32.55	<0.0001
Replication	2	0.06	1.84	0.1622
Residual	152	0.03		
Protein				
Genotype	78	7538.91	4.86	<0.0001
Replication	2	6491.92	4.19	0.017

	Df [†]	MS[§]	F-Value	P-value[‡]
Residual	154	1550.91		
Oil				
Genotype	78	940.46	5.35	<0.0001
Replication	2	523.84	2.98	0.0537
Residual	154	175.71		
Alanine				
Genotype	76	0.80	4.31	<0.0001
Replication	1	0.83	4.42	0.0388
Residual	75	0.19		
Aspartic				
Genotype	76	9.14	3.93	<0.0001
Replication	1	3.30	1.42	0.2374
Residual	75	2.33		
Cysteine				
Genotype	76	0.39	7.03	<0.0001
Replication	1	0.37	6.6	0.0122
Residual	75	0.06		
Glutamic				
Genotype	76	0.99	4.6	<0.0001
Replication	1	0.30	2.47	0.1205
Residual	75	0.15		
Glycine				
Genotype	76	29.84	4.51	<0.0001
Replication	1	15.99	3.27	0.0744
Residual	75	6.48		
Isoleucine				
Genotype	76	0.87	4.13	<0.0001
Replication	1	0.63	0.23	0.6307
Residual	75	0.19		
Leucine				
Genotype	76	1.12	4.57	<0.0001
Replication	1	0.06	0.24	0.6256
Residual	75	0.27		
Lysine				
Genotype	76	3.39	4.66	<0.0001
Replication	1	0.18	5.62	0.0203
Residual	75	0.74		
Methionine				
Genotype	76	2.14	4.85	<0.0001
Replication	1	2.58	0.08	0.777

	Df [†]	MS[§]	F-Value	P-value[‡]
Residual	75	0.46		
Proline				
Genotype	76	0.19	2.33	0.0002
Replication	1	0.00	8.78	0.0041
Residual	75	0.04		
Threonine				
Genotype	76	1.70	3.88	<0.0001
Replication	1	6.40	11.67	0.001
Residual	75	0.73		
Valine				
Genotype	76	0.64	2.77	<0.0001
Replication	1	1.92	0.9	0.3469
Residual	75	0.16		
Fructose				
Genotype	76	1.80	1.49	0.0189
Replication	1	0.58	9.9	<0.0001
Residual	75	0.65		
Glucose				
Genotype	78	16.19	1	0.5134
Replication	2	107.50	6.54	0.0031
Residual	151	10.86		
Sucrose				
Genotype	68	2.39	9.98	<0.0001
Replication	2	15.71	36.54	<0.0001
Residual	48	2.40		
Raffinose				
Genotype	78	88.11	3.32	<0.0001
Replication	2	322.61	26.56	<0.0001
Residual	151	8.83		
Stachyose				
Genotype	78	4.27	6.12	<0.0001
Replication	2	34.19	58.69	<0.0001
Residual	151	1.29		
Palmitic				
Genotype	78	72.57	1.53	0.0142
Replication	2	695.47	12.83	<0.0001
Residual	151	11.85		
Stearic				
Genotype	77	0.93	1.42	0.0344
Replication	2	7.85	1.01	0.3668

	Df [†]	MS[§]	F-Value	P-value[‡]
Residual	150	0.61		
Oleic				
Genotype	77	0.19	4.82	<0.0001
Replication	2	0.13	2.73	0.0682
Residual	150	0.13		
Linoleic				
Genotype	77	6.66	2.58	<0.0001
Replication	2	3.78	2.55	0.0815
Residual	150	1.38		
Linolenic				
Genotype	77	4.89	2.83	<0.0001
Replication	2	4.83	1.88	0.1563
Residual	150	1.90		

[†] Degree of freedom

[§] Mean square

[‡] P-value of F-test for the source of variance

Supplemental Table 0-5. Analysis of variance of maturity, seed size, and seed composition in 80 wild soybean genotypes grown at Columbia, MO in 2014. Fatty acids were calculated based on oil content. Other seed compositions were calculated based on seed dry weight.

	Df [†]	MS[§]	F-Value	P-value[‡]
Maturity				
Genotype	78	945.46	959.11	<0.0001
Replication	2	2.30	2.33	0.1007
Residual	136	0.99		
Seed weight				
Genotype	78	0.65	26.3	<0.0001
Replication	2	0.05	1.84	0.1623
Residual	136	0.02		
Protein				
Genotype	78	4811.13	2.12	<0.0001
Replication	2	2132.76	0.94	0.3931
Residual	134	2268.42		
Oil				
Genotype	78	174.83	1.32	0.0772
Replication	2	70.46	0.53	0.5879
Residual	136	132.14		
Alanine				
Genotype	77	1.01	3.68	<0.0001
Replication	2	1.01	3.69	0.0299
Residual	72	0.27		
Aspartic				
Genotype	77	10.17	3.48	<0.0001
Replication	2	6.01	2.06	0.1351
Residual	72	2.92		
Cysteine				
Genotype	77	0.30	3.78	<0.0001
Replication	2	0.06	0.72	0.4894
Residual	72	0.08		
Glutamic				
Genotype	77	0.81	3.96	<0.0001
Replication	2	0.21	4.09	0.0207
Residual	72	0.23		
Glycine				
Genotype	77	30.34	2.98	<0.0001
Replication	2	31.37	0.13	0.8817
Residual	72	7.67		
Isoleucine				

	Df [†]	MS[§]	F-Value	P-value[‡]
Genotype	77	1.06	2.63	<0.0001
Replication	2	0.05	2.36	0.1017
Residual	72	0.36		
Leucine				
Genotype	77	1.31	3.15	<0.0001
Replication	2	1.17	9.11	0.0003
Residual	72	0.50		
Lysine				
Genotype	77	3.99	3.67	<0.0001
Replication	2	11.56	1.47	0.2369
Residual	72	1.27		
Methionine				
Genotype	77	2.90	3.17	<0.0001
Replication	2	1.16	1	0.3719
Residual	72	0.79		
Proline				
Genotype	77	0.17	2.29	0.0002
Replication	2	0.05	1.79	0.1737
Residual	72	0.05		
Threonine				
Genotype	77	2.51	2.65	<0.0001
Replication	2	1.97	5.47	0.0062
Residual	72	1.10		
Valine				
Genotype	77	0.75	2.53	<0.0001
Replication	2	1.54	23.24	<0.0001
Residual	72	0.28		
Fructose				
Genotype	77	1.84	3.35	<0.0001
Replication	2	16.94	169.54	<0.0001
Residual	72	0.73		
Sucrose				
Genotype	77	1.97	7.86	<0.0001
Replication	2	99.81	37.24	<0.0001
Residual	130	0.59		
Raffinose				
Genotype	77	68.25	3.97	<0.0001
Replication	2	323.31	36.96	<0.0001
Residual	132	8.68		
Stachyose				

	Df [†]	MS[§]	F-Value	P-value[‡]
Genotype	77	1.53	4.35	<0.0001
Replication	2	14.26	57.21	<0.0001
Residual	132	0.39		
Palmitic				
Genotype	77	86.40	2.01	0.0002
Replication	2	1135.00	5.32	0.0060
Residual	132	19.84		
Stearic				
Genotype	77	1.35	2.6	<0.0001
Replication	2	3.57	1.5	0.2263
Residual	130	0.67		
Oleic				
Genotype	77	0.24	3.01	<0.0001
Replication	2	0.14	1.22	0.2994
Residual	130	0.09		
Linoleic				
Genotype	77	3.94	1.98	0.0003
Replication	2	1.59	0.22	0.8019
Residual	130	1.31		
Linolenic				
Genotype	77	3.71	2.65	<0.0001
Replication	2	0.41	0.03	0.9748
Residual	130	1.87		

[†] Degree of freedom

[§] Mean square

[‡] P-value of F-test for the source of variance

Supplemental Table 0-6. Analysis of variance of maturity, seed size, and seed composition in 80 wild soybean genotypes grown at Novelty, MO in 2015. Fatty acids were calculated based on oil content. Other seed compositions were calculated based on seed dry weight.

	Df [†]	MS[§]	F-Value	P-value[‡]
Maturity				
Genotype	76	835.75	1869.28	<0.0001
Replication	2	1.12	2.51	0.0847
Residual	137	0.45		
Seed weight				
Genotype	76	0.66	33.62	<0.0001
Replication	2	0.10	4.87	0.0091
Residual	137	0.02		

	Df [†]	MS[§]	F-Value	P-value[‡]
Protein				
Genotype	76	2420.76	2.38	<0.0001
Replication	2	1010.66	0.99	0.3734
Residual	138	1018.77		
Oil				
Genotype	76	638.50	2.85	<0.0001
Replication	2	158.14	0.71	0.4951
Residual	139	223.80		
Alanine				
Genotype	76	1.03	2.75	<0.0001
Replication	2	0.36	0.96	0.3873
Residual	75	0.37		
Aspartic				
Genotype	76	11.74	3.77	<0.0001
Replication	2	0.72	0.23	0.7933
Residual	75	3.11		
Cysteine				
Genotype	76	0.47	7.50	<0.0001
Replication	2	0.34	5.32	0.0069
Residual	75	0.06		
Glutamic				
Genotype	76	1.11	3.92	<0.0001
Replication	2	0.26	1.10	0.5244
Residual	75	0.40		
Glycine				
Genotype	76	35.92	2.65	<0.0001
Replication	2	10.06	1.63	0.3386
Residual	75	9.16		
Isoleucine				
Genotype	76	1.05	2.96	<0.0001
Replication	2	0.64	8.80	0.2037
Residual	75	0.40		
Leucine				
Genotype	76	1.61	3.31	<0.0001
Replication	2	4.78	0.86	0.0004
Residual	75	0.54		
Lysine				
Genotype	76	4.47	3.21	<0.0001
Replication	2	1.16	0.88	0.4292
Residual	75	1.35		

	Df [†]	MS[§]	F-Value	P-value[‡]
Methionine				
Genotype	76	2.80	1.22	<0.0001
Replication	2	0.76	1.10	0.4208
Residual	75	0.87		
Proline				
Genotype	76	0.25	4.52	0.1953
Replication	2	0.23	7.59	0.3396
Residual	75	0.21		
Threonine				
Genotype	76	2.62	1.85	<0.0001
Replication	2	4.39	6.46	0.0010
Residual	75	0.58		
Valine				
Genotype	76	0.80	2.91	0.0043
Replication	2	2.81	13.68	0.0026
Residual	75	0.44		
Fructose				
Genotype	76	2.12	1869.28	<0.0001
Replication	2	9.96	2.51	<0.0001
Residual	75	0.73		
Glucose				
Genotype	76	8.36	1.71	<0.0001
Replication	2	0.01	0.65	0.0847
Residual	137	0.00		
Sucrose				
Genotype	75	0.50	8.45	0.0064
Replication	2	0.19	18.15	0.5223
Residual	97	0.29		
Raffinose				
Genotype	76	69.28	3.69	<0.0001
Replication	2	148.83	15.56	<0.0001
Residual	137	8.20		
Stachyose				
Genotype	76	0.94	3.00	<0.0001
Replication	2	3.95	16.13	<0.0001
Residual	137	0.25		
Palmitic				
Genotype	76	77.17	1.17	<0.0001
Replication	2	414.91	0.81	<0.0001
Residual	137	25.72		

	Df [†]	MS[§]	F-Value	P-value[‡]
Stearic				
Genotype	76	0.93	1.26	0.2124
Replication	2	0.64	1.05	0.4472
Residual	135	0.79		
Oleic				
Genotype	76	0.17	1.89	0.1206
Replication	2	0.14	1.39	0.3544
Residual	135	0.14		
Linoleic				
Genotype	76	2.51	1.69	0.0007
Replication	2	1.86	3.55	0.2519
Residual	135	1.33		
Linolenic				
Genotype	76	2.95	1.69	0.0040
Replication	2	6.19	0.78	0.0315
Residual	135	1.74		

[†] Degree of freedom

[§] Mean square

[‡] P-value of F-test for the source of variance

Chapter III:

QTL MAPPING FOR SEED PROTEIN AND OIL IN THE OSAGE × PI593983 SOYBEAN POPULATION

Abstract

Soybean seed contains a relatively high percentage of protein and oil, and is one of the main sources of protein and oil for industrial purposes, as well as food processing and livestock feed. The objectives of this study were to identify quantitative trait loci (QTLs) associated with the contents of seed protein and oil in a recombinant inbred line (RIL) population developed from one single F_2 plant from the cross between Osage and PI593983. Field tests were carried out in Missouri for two years during 2016 and 2017, in a randomized complete block design (n=2). The RIL population was phenotyped for seed protein and oil contents in multiple environments using near-infrared spectroscopy. Both protein and oil contents showed high heritabilities. Seed protein and seed oil were negatively correlated (–0.77). A total of 4,374 polymorphic markers were used to construct a genetic linkage map, and nine QTLs for protein content, explained 7.6 to 36.7% of variance, and seven QTLs for oil content, explained for 7.8 to 29.7% of variance, were detected using composite interval mapping. The identified QTLs from PI593983 (*Glycine soja* Sieb. and Zucc) for protein and oil contents in this study further verify the QTLs that were previously reported.

Introduction

Soybean [*Glycine max* (L.) Merr.] is one of the most valuable grain legumes in the world. Typically, soybean seed has 200 g kg^{-1} oil and 400 g kg^{-1} protein based on dry weight (Liu, 1997; Wilson, 2004). The soybean seed contents of protein and oil affect the price of soybean, and soybean with high content of protein and oil is preferable by soybean processors and consumers (Brumm and Hurburgh, 1990; Orf and Helms, 1994).

Soybean oil is used not only as cooking oil but also as an ingredient in salad dressings, mayonnaise, and margarine (Wilson, 2004; Wilson, 2008). Soybean oil is also used in industrial products such as biodiesel fuel and plastics (http://www.soystats.com, accessed 05/20/2018). Because of the nutritional value and health benefits of soybean food such as edamame, soymilk, tofu and natto, Lee et al. (2015) and Shi et al. (2010) stated that soybean breeders, especially in Japan and Korea, had got more interest in soy-food-driven traits including protein, oil and other ingredients. In the US, soybean meal, the high protein product containing soybean seed after the oil removal process, is the major ingredient of livestock feed since soy protein is the cheapest protein that provides a complete panel of amino acids that are required by livestock (Chiari et al., 2004; Yesudas et al., 2013). Pettersson and Pontoppidan (2013) reported that soybean meal and oil accounted for approximately 60% and 40% of soybean's value, respectively. Willis (2003) stated that soybean seed with ≥ 415 g kg^{-1} protein content and 220 g kg^{-1} based on dry weight was the standard to achieve soybean meal with ≥ 475 g kg^{-1} protein.

The contents of oil and protein in soybean seed range from 81 g kg^{-1} to 279 g kg^{-1} and from 341 g kg^{-1} to 568 g kg^{-1}, respectively; however, there are only a few accessions with high contents of both protein and oil in the USDA Soybean Germplasm Collection

(Chung et al., 2003; Wilson, 2004). These traits are quantitative traits, controlled by many genes, and strongly affected by environment (Lee et al., 2007). Chung et al. (2003) stated that the contents of soybean protein and oil varied significantly in different regions of the USA due to the different environmental conditions that affect protein and oil content in soybean. Specht et al. (2001) reported the depression of seed protein due to drought stress. Bellaloui et al. (2015) stated that cool temperature had a negative effect on seed storage protein, and Carrera et al. (2011) showed that oil content increases when the nighttime temperature during pod fill is warm.

Although QTLs associated with higher protein without decreasing oil content have been widely reported (Lee et al., 1996; Eskandari et al., 2013), a strong negative correlation between these traits has also been reported (Shannon et al., 1972; Cober and D Voldeng, 2000; Nichols et al., 2006; Ramteke et al., 2010; Phansak et al., 2016; Smallwood et al., 2017). In addition, yield has been shown to be negatively associated with protein content but positively associated with oil content (Sebolt et al., 2000; Chung et al., 2003; Nichols et al., 2006; Leamy et al., 2017; Wu et al., 2017). Because of the negative correlation between protein and oil content, it has been challenging for soybean breeders to increase the content of either seed oil or protein without decreasing the other. Chung et al. (2003) suggested that a pleiotropic QTL or tight association between high protein allele(s) and low oil allele(s) might lead to the strong negative correlation.

There have been 322 QTLs and 240 QTLs associated with the content of seed oil and protein, respectively (http://www.soybase.org, “SoyBase browser”, verified April 4th, 2018). These QTLs have been located on all chromosomes (http://www.soybase.org, “SoyBase browser”, accessed 05/20/2018). Brummer et al. (1997) studied eight different soybean populations developed in Mid-west USA in multiple environments. They reported nine QTLs significantly associated with seed protein content and seven QTLs significantly associated with seed oil content. Hyten et al. (2004) evaluated 131 F_6-derived recombinant inbred lines developed from a cross between Essex and Williams in six different environments. They identified four QTLs for protein and six QTLs for oil. Wang et al. (2014) studied two different populations (SD02-4-59 × A02-381100 and SD02-911 × SD00-1501) in multiple environments and detected 11 QTLs for protein content and 16 QTLs for oil content. Although many QTLs for protein and oil content have been reported, few of these QTL have been confirmed (Panthee et al., 2005; Pathan et al., 2013; Phansak et al., 2016). Diers et al. (1992) studied a population developed from a cross between an experimental line of Iowa State University [*Glycine max* (L.) Merr.] and a wild soybean plant introduction PI468916 (*G. soja* Sieb. and Zucc.). They reported two QTLs, one on Chr. 15 and one on Chr. 20, associated with high protein. Nichols et al. (2006) and Fasoula et al. (2004) confirmed these QTLs by following the guidelines that was suggested by Soybean Genetics Committee (http://soybase.org/). The QTLs on Chr. 15 and 20 were named cqSeed protein-001 (Fasoula et al., 2004) and cqSeed protein-003 (Nichols et al., 2006), respectively. These QTL with large effect on protein content have been located in these two locations with different significance levels (Sebolt et al., 2000; Chung et al., 2003; Wang et al., 2014; Warrington et al., 2015; Kim et al., 2016; Phansak et al., 2016; Qi et al., 2016). The difference in significance levels could be explained by the differences in the respective population's source and size, the rate of recombination, the extent of linkage disequilibrium in the studied population, and the dependence of QTLs' effects on the environment (Brzostowski and Diers, 2017; Patil et al., 2017). Kim et al. (2016) studied a backcross population by using Williams 82 as the recurrent parent and PI407788A, a high protein line accession from Korea. They identified a QTL, which is from PI407788A, associated with higher protein and lower oil content. A QTL that had major effect on seed protein and amino acid content was identified on Chr. 20 when Warrington et al. (2015) studied the population developed from the cross between Benning and Danbaekkong. Warrington et al. (2015) stated that the favorable allele was from Danbaekkong, explained approximately 55% of the seed protein's variation, and showed little negative effect on yield in the studied population. In the studies of Nichols et al. (2006), Chung et al. (2003), and Sebolt et al. (2000), the reported QTLs showed stronger negative effect on yield. Patil et al. (2017) suggested that the allele from Danbaekkong was different from the other reported QTLs or the genetic background of Danbaekkong mitigated the negative effect on yield of the QTL. (Chen et al., 2008; Qi et al., 2011).

The objectives of this study were to identify QTLs for seed oil and protein content in a RIL mapping population of Osage × PI593983 in four environments and to identify new QTLs significantly associated with seed protein and oil content.

Materials and methods

Plant materials and field experiment

The studied population started from a cross between Osage [*Glycine max* (L.) Merr.] and PI593983 (*G. soja* Sieb. and Zucc.) in North Carolina in 2011. During the winter of 2011/2012, the F_1 generation was grown at a USDA-ARS winter nursery in Isabela, Puerto Rico (coordinates: 18^{o}30'N, 67^{o}1'W; soil type: Coto clay). The F_2

generation was grown in Columbia, MO during the summer of 2012 and a single F_2 plant was selected and the F_3 seeds were harvested from this single plant. During the summer of 2013, 338 F_3 plants were grown and harvested individually. In 2014, 338 $F_{3:4}$ inbred lines were grown at Bradford Farm in Columbia, MO (coordinates: 38^{o}59'N, 92^{o}12'W; soil type: Mexico silt loam), and one plant was randomly collected from each line. The $F_{4:5}$ seeds were sent to winter nursery in Isabela, Puerto Rico for seed increase. In 2015, 164 RILs were randomly selected from 338 RILs and planted at Bay Farm in Columbia, MO. Due to extreme weather, the germination rate was low and data was not collected.

In 2016, 164 $F_{5:6}$ RILs were planted at Greenley Memorial Research Center in Novelty, MO (coordinates: 40^{o}01'N, 92^{o}11'W; soil type: Putnam silt loam) and at the Hundley-Whaley Research Center in Albany, MO (coordinates: 40^{o}15'N, 94^{o}19'W; soil type: Grundy silt loam). In 2017, the field experiment was conducted at Bradford Farm in Columbia, MO (coordinates: 38^{o}59'N, 92^{o}12'W; soil type: Mexico silt loam) and at Greenley Memorial Research Center in Novelty, MO. In all years and locations, the 164 RILs were planted in two-row plots. Plot dimensions were 2.44m by 2.29m. Seeds were sown at the rate of 41 seeds m^{-1}. The RILs were planted in a randomized complete block design with two replications at all environments. The population was planted by using a four-row ALMACO cone planter with Kinze row units (ALMACO, Nevada, IA) and four rows spaced at 0.76m. Seed was harvested at R8 by an ALMACO SPC-40 plot combine (ALMACO, Inc. Nevada, IA).

Protein and oil analysis

Approximately 5 g of ground soybean seed was used to calculate reflectance spectra by using XDS-NIRS Rapid Content™ Analyzer (FOSS Analytical, Slangerupgade, Denmark) and the ISIscan™ software. The spectra were used to calculate the contents of seed protein and oil using the equations which were previously developed (Choung et al., 2001) based on the spectra from standard samples, calibration, and validation assessments. The calibration database includes soybeans from all over the US and Canada in 2010 and were ground with a Foss Knifetec grinder (5-1-5 second burst).

A certified 80% reflectance reference was used to create reference standard. The performance test was carried out by running four segments ten times and compiling the spectra.

Statistical Analysis

The analysis of variance (ANOVA) was carried out by using PROC MIXED in SAS version 9.4 (SAS Institute, 2002). Genotype was used as a fixed effect to test for significant genotypic differences among accessions for all traits (Table 1; F-test P-value column). The heritability (h^2) of each trait was calculated as following (Nyquist and Baker, 1991):

$$h^2 \text{ (entry mean basis)} = \frac{\sigma_g^2}{\sigma_g^2+\sigma_{ge}^2/t+\sigma_e^2/rt},$$

$$h^2 \text{ (plot basis)} = \frac{\sigma_g^2}{\sigma_g^2+\sigma_{ge}^2+\sigma_e^2},$$

where σ_g^2 is the variance among genotypes, σ_{ge}^2 is the variance of genotype × environment interaction, σ_e^2 is experimental error, t is number of test environments, and r is number of replications.

PROC CORR of SAS (SAS Institute) was used to determine significance and correlation coefficients between oil and protein contents based on means of the RILs across replications and environments. PROC TTEST of SAS (SAS Institute) was used to determine the differences between RILs with homozygous alleles from Osage and PI593983 at the same loci. Box-and-whisker plots were done using Microsoft Excel^{TM}.

DNA isolation and Genotyping-By-Sequencing

DNA was isolated from ~40 mg of lyophilized leaf tissue from a pool of 5-10 plants per RIL using the DNeasy Plant Mini kit (QIAGEN, Valencia, CA) according to the manufacturer's instructions. DNA samples were then submitted to the Institute for Genomic Diversity (IGD) at Cornell University, where genotyping by sequencing (GBS) libraries were created (Elshire et al., 2011) using ApeKI, DNA ligase, and appropriate Illumina adapters. IGD carried out all library construction, Illumina sequencing, read mapping, and SNP calling using TASSEL.

A total of 548,086,161 reads were produced for 2 blanks (no sample), 8 *G. soja* lines including the two parents of the RIL population, and 164 RILs. One RIL was determined to be not derived from the cross by PCA analysis and was dropped. A total of 64.1% of the reads were found to map to single positions in the 'Williams 82' Wm82.a2.v1 reference sequence [Schmutz et al. (2010); http://phytozome.jgi.doe.gov/], using the BWA 0.7.8-r455 program (Li and Durbin, 2009). The TASSEL 5.0 pipeline was used to call SNPs, resulting in 170,463 raw SNPs and 139,012 filtered SNP positions in total, which had 6.687 and 7.019 mean site depth in the raw and filtered datasets, respectively.

SNP dataset quality control

Allele frequencies were called using TASSEL software and SNPs filtered to exclude those with >80% missing data. The LinkImpute program (Money et al., 2015) with the settings of 30 high LD sites and 10 nearest neighbors was used to impute missing data. Finally parental genotypes were assigned using the ABH genotype function in TASSEL.

The ABHGenotypes function in R (Furuta et al., 2017) was then used to correct GBS related genotyping errors using the correctUnderCalledHets and correctStretches functions (settings were maxhaplength=3). Only those SNPs for which a definitive parental origin could be assigned were used for downstream genetic map creation and QTL mapping.

Linkage Map Creation

Because this RIL population is derived from a single F_2 plant, significant gaps were present and certain chromosomes had very limited segregating loci. The linkage map was constructed using only polymorphic SNPS in the F_2-derived RIL population using the R/qtl (R Foundation for Statistical Computing) software package with 4,652 SNPs. Genetic distances were estimated via the est.map function and genotyping error rate was called. Each chromosome with excessive map distances (>200 cM) were evaluated by manual removal of single markers via the droponemarker and est.map functions. In addition, chromosomes 3 and 13 were split into 3 and 2 sub-chromosomes respectively. Each of the chromosomal marker orderings was evaluated via the ripple function, and no better marker order was identified than that present in the original Wm82.a2.v1 assembly.

QTL analysis

The R/qtl software package (http://www.rqtl.org/) was used for QTL analysis. To detect the QTL, Expectation-Maximization (EM) algorithm (implemented in R/qtl) was used (Xu et al., 2000; Sen et al., 2009). Analyses were carried out by using the composite interval mapping (CIM) procedure with a 10 cM window. The empirical logarithm of odds (LOD) thresholds were calculated at the 10% level of probability with 1000 permutations for protein and oil contents (Churchill and Doerge, 1994). The percentage of phenotypic variance explained by the significant QTL was determined by effectplot function (implemented in R/qtl). The effect of each QTL was determined in R/qtl by using effectplot function, following sim.geno function with 1000 draws and an error probability of 0.01. The confidence intervals for each significant QTL was presented as 1.5-LOD by using lodint function.

Results

There were significant differences (P<0.0001) among the recombinant inbred lines for both protein and oil contents in this population. Protein content ranged from 466.2 to 543.0 g kg^{-1}, with a mean of 508.2 g kg^{-1} (Table 3-1). Oil content ranged from 161.73 to 210.05 g kg^{-1} with a mean of 183.17 g kg^{-1} (Table 3-1). Typical oil content in soybean is 200 g kg^{-1} (Liu, 1997; Wilson, 2004); the population mean for oil content in this population was lower than the typical. The heritability based on entry mean was 0.94 for seed protein and 0.92 for seed oil (Table 3-1).

The phenotypic correlation between protein and oil concentration was strong and negative (r= –0.77, P<0.0001) in this population. This value is similar to the average heritability (–0.78) and within the range of heritability (–0.66 to –0.88) that was reported by Phansak et al. (2016).

In this study, 27,248 markers were used to analyze polymorphism between parents Osage and PI593983, and among 164 RIL population (Table 3-2). After the elimination of markers with >80% missing data, without definitive parental origin, or without following the rule of F_4-derived segregation, 4,652 markers were used to construct the genetic linkage map. After removing markers for gap closure, the genetic linkage map covered 2,051.2 cM and included 4,374 markers on 20 chromosomes (Table 3-2, Figure 3-1). The average length of each chromosome was 102.6 cM, and the average distance between the included markers was 0.47 cM. There were gaps which were greater than 40 cM on Chr. 02, Chr. 08, Chr. 18, and Chr. 19. Although more than 27,000 markers were used, more than 20,000 markers with distorted segregation were excluded, and many markers were denser in some chromosome (Table 3-2; Figure 3-1; Figure 3-5). This can be partly explained by the origin of the studied population. This population derived from a single F_2 plant; therefore, approximately half of the genome would be fixed and leaded to big gaps, reduced genome coverage in the genetic linkage map.

Seven QTLs were found associated with oil content and located on Chr. 05, Chr. 14, and Chr. 20; LOD values ranged from 7.8 to 29.7, and explained for 10.6 to 38.6% of the phenotypic variance (Table 3-3). Among the seven detected QTLs, three QTLs (Oil8.1, Oil8.2, and Oil8.3) clustered on Chr. 08, and three QTLs (Oil20.1, Oil20.2, and Oil20.3) clustered on Chr. 20. On average, RILs with homozygous alleles from PI593983 at S8_8661263 had 9.6 g kg^{-1} oil lower than those with allele from Osage at the same locus (Figure 3-2). RILs with homozygous alleles from PI593983 at S8_8661263 and S20_32687273 showed significantly lower oil than those with alleles from Osage at the same loci (Table 3-4; Figure 3-2).

Eight QTLs were found associated with protein content and located on two different chromosomes, including Chr. 14 and Chr. 20; LOD values ranged from 7.6 to 36.7, and the phenotypic variation rates were between 11.8 and 48.2% (Table 3-3). Two QTLs (Pro14.1 and Pro14.2) exhibited a clustered distribution on Chr. 14, and four QTLs (Pro20.1, Pro20.2, Pro20.3, and Pro20.4) located on Chr. 20. RILs with alleles from PI593983 at S14_34985739 and S20_31308579 had significant increases in protein when they were compared to RILs with alleles from Osage at the same loci (Table 3-3 and 3-4; Figure 3-2). The significant protein and oil QTLs on Chr. 20 showed overlapped LOD–1.5 intervals and had inversed effects on seed protein and oil contents (Table 3-2; Figure 3-2). Oil20.1 and Pro20.4, whose LOD–1.5 intervals were 69.5 to 73.6 cM and 68.8 to 71.5 cM, respectively, explained 34.5% and 48.2% of total phenotypic variation for seed oil and protein contents, respectively (Table 3-3). On average, RILs carrying the allele from PI593983 of the peak SNP (S20_32687273) at Oil20.1 locus had 11.5 g kg^{-1} lower oil, while the RILs carrying the allele from PI593983 of the peak SNP (S20_31308579) at Pro20.4 locus had 23.9 g kg^{-1} higher protein (Figure 3-2).

Discussion

In the RIL population of this study, mean protein content was 508.2 g kg^{-1}. The average protein content of these RILs was much higher than the typical protein content in soybean (400 g kg^{-1}) reported by Liu (1997) and Wilson (2004) and in a collection of 600 wild soybean accessions (480 g kg^{-1}) reported by Leamy et al. (2017). And the average oil content of the RIL population in this study was 183.17 g kg^{-1}. The population mean for oil content was lower than the typical oil content (200g kg^{-1}) reported by Liu (1997) and Wilson (2004) and higher than the average oil content (110g kg^{-1}) in wild soybean collection reported by Leamy et al. (2017). The result in this study suggested that it would be a challenge to increase oil content from crosses with wild soybean.

The heritabilities for protein and oil contents in this study (0.94 and 0.92, respectively;) were higher than the heritabilities for protein and oil reported by Panthee et al. (2005) (0.54 and 0.66, respectively) when they studied a RIL population developed from a cross of N87-984-16 × TN93-99. Diers et al. (1992) reported similar heritability for oil (0.92) but much lower for protein (0.74) compared to those reported in this study (0.92 and 0.94, respectively). The heritabilities in our study were within the range of 0.84 to 0.99 in the study of Chung et al. (2003) using a 76 F¬5-derived RILs from the cross of Asgrow A3733 × PI 437088A. Chung et al. (2003), Hyten et al. (2004), and Wang et al. (2014) reported varied heritability for oil and protein (0.07 to 0.89 and 0.56 to 0.92, respectively) and they suggested that the estimates of heritability depended on the environments and populations. Burton (1987) stated that the heritabilities of seed protein and oil were high, especially when the differences between parents were high. The high heritabilities in our study suggested a high selection response for achieving genetic gain.

In this study, the genetic linkage map covered 2,051.2 cM and consisted of 23 fragments of 20 linkage groups. The markers were not evenly distributed among linkage groups, and there were gaps which could lead to the failure of significant QTL detection. Theses gaps could be the results of the fixation of about half of the genome because this population originated from one single F_2 plant (Table 3-2, Figure 3-5). Big gaps (>20.0 cM) have been reported in numerous soybean genetic linkage maps (Cregan et al., 1999; Song et al., 2004; Zhang et al., 2004; Kassem et al., 2006; Lu et al., 2015). These gaps may suggest the low recombination frequency in some regions because these genomic regions of the two parents were inverted (Kassem et al., 2006). Shultz et al. (2006) reported different regions with low recombination rate in different populations.

The strong negative correlation between the contents of seed protein and oil was observed not only in this study (r= –0.77, P<0.0001) but also in numerous studies (Burton, 1987; Chung et al., 2003; Hwang et al., 2014; Phansak et al., 2016; Smallwood et al., 2017;

Wu et al., 2017). This strong correlation and the high heritabilities of protein and oil contents suggested that the simultaneous improvement of protein and oil contents would be challenging (Recker et al., 2014). The strong and negative relationship between protein and oil contents might be the result of the tight link between loci that separately control oil and protein contents, or the result of pleiotropic effects in which a locus controls both protein and oil contents (Chung et al., 2003). Long term mating can be used to determine if the correlation between two traits is due to pleiotropy or tightly linked loci (Recker et al., 2014). Recker et al. (2014) stated that long-term and random mating could be used to distinguish if a correlation was the result of tightly linked loci or pleiotropy. If a correlation was due to tightly linked loci, long-term and random mating would increase the possibility of breaking the link between these linked loci and breeders could keep the favorable one while excluding the unfavorable one to change the correlation. However, if a correlation was due to pleiotropy, long-term and random mating would not have any effect on the correlation.

Two previously reported QTLs associated with seed protein listed in SoyBase within the LOD-1.5 support interval of the Pro4.4 QTL (www.SoyBase.org,"SoyBase browser", accessed 04/05/2018). Of these, Asekova et al. (2016) reported qCP14_1 when they studied a population developed from a cross between PI483463 (a wild soybean accession) and Hutcheson (an American cultivar). Asekova et al. (2016) reported the QTL qCP14_1 with a LOD score of 1.5 and a confidence interval of 16.2Mb-46.3Mb that explained only 7.0% of the phenotypic variation while Pro4.4 had a LOD score of 11 and a confidence interval of 32.4Mb-35.9Mb and explained up to 11.0% of the phenotypic variation (Table 3.3). RILs with homozygous alleles from wild soybean of Pro4.4 and qCP14_1 had increased protein content. Both Pro4.4 and qCP14_1 had non-significant association with oil.

Oil8.3 was located on chromosome 8, and the confidence intervals of five QTLs for oil content listed on SoyBase overlapped with the LOD-1.5 confidence interval of Oil8.3 (www.SoyBase.org, “SoyBase browser”, accessed 04/05/2018). Although Oil8.3 was significantly associated with oil and not significantly associated with protein, Oil8.3 located within the the support interval range for cqSeed protein-013, a confirmed QTL for protein content (Pathan et al., 2013).

Due to the strong and negative correlation between protein and oil contents, the overlap of Oil20.1’s and Pro20.4’s LOD–1.5 confidence intervals and their reversed effects on protein and oil contents was expected (Table 3-3; Figure 3-3). Numerous seed protein and oil QTLs have been reported on chromosome 20 (www.SoyBase.org, “SoyBase browser”, accessed 04/05/2018). Within the LOD–1.5 support intervals for Oil20.1 and Pro20.4, 16 QTLs for oil and 20 QTLs for protein have been reported (www.SoyBase.org, “SoyBase browser”, accessed 04/05/2018). Recently, Bandillo et al. (2015) reported five SNP markers that were in the intervals for Oil20.1 and Pro20.4. These five SNP markers had reversed effects on protein and oil content (Bandillo et al., 2015). Diers et al. (1992) identified a QTL on Chr. 20 that associated with both protein and oil contents when they studied a population developed from a cross between A81-356022, a *G. max* breeding line, and PI468916, an *G. soja* with high protein content from the Liaoning area, China. Nichols et al. (2006) confirmed this QTL according to the standards of Soybean Genetics Committee. This confirmed QTL,s cqSeed protein-003 and cqSeed oil-004, located within the LOD–1.5 support interval ranges of Oil20.1 and Pro20.4 (www.SoyBase.org, “SoyBase browser”, accessed 04/05/2018). Reversed correlation to protein and oil contents of cqSeed protein-003 and cqSeed oil-004 have been reported (Diers et al., 1992; Nichols et al., 2006; Warrington et al., 2015).

In this study, a recombinant inbred line (RIL) population was developed from a single F_2 plant of the cross between Osage and PI593983. The RIL population was genotyped by genotype by sequencing approach. A genetic linkage map was created by using 4,374 polymorphic SNP markers and used to identify QTLs that were significantly associated with the content of seed protein and oil. A total of nine QTL for seed protein and seven QTLs for seed oil were identified. The detetected QTLs were coincident with previously reported QTL; however, the confidence interval of the reported QTLs in our study, Oil8.3, Pro14.4, Oil20.1, and Pro20.4, were narrower than the confidence intervals of those in previous researchs (Nichols et al.; Pathan et al.; Warrington et al.; Asekova et al.). The RILs with homozygous alleles from wild soybean of Pro14.4 had increased protein content without any effect on oil content while those with homozygous alleles from wild soybean of Oil8.3 had decreased oil content without any effect on protein content. Mapping of QTLs for seed protein and oil, markers significantly associated with protein and oil content were identified. The identification of these markers suggested new candidate genes and provide information to facilitate further studies to improve protein and oil content in soybean seed as well as identify genes controlling these traits.

References

Asekova, S., K.P. Kulkarni, M. Kim, J.-H. Kim, J.T. Song, J.G. Shannon, et al. 2016. Novel quantitative trait loci for forage quality traits in a cross between PI483463 and 'Hutcheson' in soybean. Crop Science.

Bandillo, N., D. Jarquin, Q. Song, R. Nelson, P. Cregan, J. Specht, et al. 2015. A population structure and genome-wide association analysis on the USDA soybean germplasm collection. The Plant Genome 8.

Bellaloui, N., H.A. Bruns, H.K. Abbas, A. Mengistu, D.K. Fisher, and K.N. Reddy. 2015. Agricultural practices altered soybean seed protein, oil, fatty acids, sugars, and minerals in the Midsouth USA. Frontiers in plant science 6: 31.

Brumm, T.J., and C.R. Hurburgh. 1990. Estimating the processed value of soybeans. J. Am. Oil Chem. Soc. 67: 302-307.

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

Brzostowski, L.F., and B.W. Diers. 2017. Agronomic evaluation of a high protein allele from PI407788a on chromosome 15 across two soybean backgrounds. Crop Sci. 57: 2972-2978.

Burton, J.W. 1987. Quantitative genetics: Results relevant to soybean breeding. p. 211–247. In J.R.Wilcox (ed.) Soybeans: Improvement production and uses. 2nd ed. Agron. Monogr. 16. ASA, CSSA, and SSSA, Madison, WI. Agronomy (USA).

Carrera, C.S., M.J. Martínez, J. Dardanelli, and M. Balzarini. 2011. Environmental variation and correlation of seed components in nontransgenic soybeans: Protein, oil, unsaturated fatty acids, tocopherols, and isoflavones. Crop Sci. 51: 800-809.

Chen, P., C.H. Sneller, T. Ishibashi, and B. Cornelious. 2008. Registration of high-protein soybean germplasm line R95-1705. Journal of plant registrations 2: 58-59.

Chiari, L., N.D. Piovesan, L.K. Naoe, I.C. José, J.M.S. Viana, M.A. Moreira, et al. 2004. Genetic parameters relating isoflavone and protein content in soybean seeds. Euphytica 138: 55-60.

Choung, M.G., I.Y. Baek, S.T. Kang, W.Y. Han, D.C. Shin, H.P. Moon, et al. 2001. Determination of protein and oil contents in soybean seed by near infrared reflectance spectroscopy. Korean Journal of Crop Science 46: 106-111.

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

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

Cober, E.R., and H. D Voldeng. 2000. Developing high-protein, high-yield soybean populations and lines ECORC Contribution No. 991410. Crop Sci. 40: 39-42.

Cregan, P.B., T. Jarvik, A.L. Bush, R.C. Shoemaker, K.G. Lark, A.L. Kahler, et al. 1999. An integrated genetic linkage map of the soybean genome. Crop Sci. 39: 1464-1490.

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.

Elshire, R.J., J.C. Glaubitz, Q. Sun, J.A. Poland, K. Kawamoto, E.S. Buckler, et al. 2011. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS One 6: e19379.

Eskandari, M., E.R. Cober, and I. Rajcan. 2013. Genetic control of soybean seed oil: II. QTL and genes that increase oil concentration without decreasing protein or with increased seed yield. Theor. Appl. Genet. 126: 1677-1687.

Fasoula, V.A., D.K. Harris, and H.R. Boerma. 2004. Validation and designation of quantitative trait loci for seed protein, seed oil, and seed weight from two soybean populations. Crop Sci. 44: 1218-1225.

Furuta, T., M. Ashikari, K.K. Jena, K. Doi, and S. Reuscher. 2017. Adapting genotyping-by-sequencing for rice F2 populations. G3: Genes, Genomes, Genet. 7: 881-893.

Hwang, E.Y., Q.J. Song, G. Jia, J.E. Specht, D.L. Hyten, J. Costa, et al. 2014. A genome-wide association study of seed protein and oil content in soybean. BMC genomics 15: 1-1.

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

Kassem, M.A., J. Shultz, K. Meksem, Y.G. Cho, A.J. Wood, M.J. Iqbal, et al. 2006. An updated ‘Essex’by ‘Forrest’linkage map and first composite interval map of QTL underlying six soybean traits. Theor. Appl. Genet. 113: 1015-1026.

Kim, M.C., S. Schultz, R.L. Nelson, and B.W. Diers. 2016. Identification and fine mapping of a soybean seed protein QTL from PI407788a on chromosome 15. Crop Sci. 56: 219-225.

Leamy, L.J., H. Zhang, C. Li, C.Y. Chen, and B.H. Song. 2017. A genome-wide association study of seed composition traits in wild soybean (Glycine soja). BMC Genomics 18: 18.

Lee, J.D., K.D. Bilyeu, and J.G. Shannon. 2007. Genetics and breeding for modified fatty acid profile in soybean seed oil. Crop Science Biotechnology 10: 201-210.

Lee, S., T.H. Jun, A.P. Michel, and M.A. Rouf Mian. 2015. SNP markers linked to QTL conditioning plant height, lodging, and maturity in soybean. Euphytica 203: 521-532.

Lee, S.H., M.A. Bailey, M.A.R. Mian, E.R. Shipe, D.A. Ashley, W.A. Parrott, et al. 1996. 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.

Li, H., and R. Durbin. 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 25: 1754-1760.

Liu, K. 1997. Soybeans: chemistry, technology, and utilization.Chapman & Hall, New York.

Lu, S., Y. Li, J. Wang, P. Srinives, H. Nan, D. Cao, et al. 2015. QTL mapping for flowering time in different latitude in soybean. Euphytica 206: 725-736.

Money, D., K. Gardner, Z. Migicovsky, H. Schwaninger, G.Y. Zhong, and S. Myles. 2015. LinkImpute: Fast and accurate genotype imputation for nonmodel organisms. G3: Genes, Genomes, Genet. 5: 2383-2390.

Nichols, D.M., K.D. Glover, S.R. Carlson, J.E. Specht, and B.W. Diers. 2006. Fine mapping of a seed protein QTL on soybean linkage group I and its correlated effects on agronomic traits. Crop Sci. 46: 834-839.

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

Orf, J.H., and T.C. Helms. 1994. Selection to maximize gross value per hectare within three soybean populations. Crop Sci. 34: 1163-1167.

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.

Pathan, S.M., T. Vuong, K. Clark, J.D. Lee, J.G. Shannon, C.A. Roberts, et al. 2013. Genetic mapping and confirmation of quantitative trait loci for seed protein and oil contents and seed weight in soybean. Crop Sci. 53: 765-774.

Patil, G., R. Mian, T. Vuong, V.R. Pantalone, Q. Song, P. Chen, et al. 2017. Molecular mapping and genomics of soybean seed protein: a review and perspective for the future. Theor. Appl. Genet.: 1-17.

Pettersson, D., and K. Pontoppidan. 2013. Soybean meal and the potential for upgrading its feeding value by enzyme supplementation. Soybean-bio-active compounds. InTech.

Phansak, P., W. Soonsuwon, D.L. Hyten, Q. Song, P.B. Cregan, G.L. Graef, et al. 2016. Multi-population selective genotyping to identify soybean [Glycine max (L.) Merr.] seed protein and oil QTLs. G3: Genes, Genomes, Genet. 6: 1635-1648.

Qi, Z., J. Pan, X. Han, H. Qi, D. Xin, W. Li, et al. 2016. Identification of major QTLs and epistatic interactions for seed protein concentration in soybean under multiple environments based on a high-density map. Mol. Breed. 36: 55.

Qi, Z., Q. Wu, X. Han, Y. Sun, X. Du, C. Liu, et al. 2011. Soybean oil content QTL mapping and integrating with meta-analysis method for mining genes. Euphytica 179: 499-514.

Ramteke, R., V. Kumar, P. Murlidharan, and D.K. Agarwal. 2010. Study on genetic variability and traits interrelationship among released soybean varieties of India [Glycine max (L.) Merrill]. Electronic Journal of Plant Breeding 1: 1483-1487.

Recker, J.R., J.W. Burton, A. Cardinal, and L. Miranda. 2014. Genetic and phenotypic correlations of quantitative traits in two long-term, randomly mated soybean populations. Crop Sci. 54: 939-943.

Schmutz, J., S.B. Cannon, J. Schlueter, J. Ma, T. Mitros, W. Nelson, et al. 2010. Genome sequence of the palaeopolyploid soybean. Nature 463: 178-183.

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

Sen, Ś., F. Johannes, and K.W. Broman. 2009. Selective genotyping and phenotyping strategies in a complex trait context. Genetics 181: 1613-1626.

Shannon, J.G., J.R. Wilcox, and A.H. Probst. 1972. Estimated gains from selection for protein and yield in the F4 generation of six soybean populations. Crop Sci. 12.

Shi, H., P.K. Nam, and Y. Ma. 2010. Comprehensive profiling of isoflavones, phytosterols, tocopherols, minerals, crude protein, lipid, and sugar during soybean (Glycine max) germination. J. Agric. Food Chem. 58: 4970-4976.

Shultz, J.L., D. Kurunam, K. Shopinski, M.J. Iqbal, S. Kazi, K. Zobrist, et al. 2006. The Soybean Genome Database (SoyGD): a browser for display of duplicated, polyploid, regions and sequence tagged sites on the integrated physical and genetic maps of Glycine max. Nucleic Acids Res. 34: D758-D765.

Smallwood, C.J., J.D. Gillman, A.M. Saxton, H.S. Bhandari, P.A. Wadl, B.D. Fallen, et al. 2017. Identifying and exploring significant genomic regions associated with soybean yield, seed fatty acids, protein and oil. Journal of Crop Science and Biotechnology 20: 243-253.

Song, Q.J., L.F. Marek, R.C. Shoemaker, K.G. Lark, V.C. Concibido, X. Delannay, et al. 2004. A new integrated genetic linkage map of the soybean. Theor. Appl. Genet. 109: 122-128.

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

Wang, K.J., X.H. Li, and M.F. Yan. 2014. Genetic differentiation in relation to seed weights in wild soybean species (Glycine soja Sieb. & Zucc.). Plant Syst. Evol. 300: 1729-1739.

Warrington, C.V., H. Abdel-Haleem, D.L. Hyten, P.B. Cregan, J.H. Orf, A.S. Killam, et al. 2015. QTL for seed protein and amino acids in the Benning × Danbaekkong soybean population. Theor. Appl. Genet. 128: 839-850.

Warrington, C.V., H. Abdel-Haleem, D.L. Hyten, P.B. Cregan, J.H. Orf, A.S. Killam, et al. 2015. QTL for seed protein and amino acids in the Benning× Danbaekkong soybean population. Theor. Appl. Genet. 128: 839-850.

2003. The use of soybean meal and full fat soybean meal by the animal feed industry. 12th Australian Soybean Conference, Toowomba, Australia.

Wilson, R.F. 2004. Seed composition. In H.R. Boerma and J.E. Specht (ed.) Soybeans: Improvement, Production, and Uses. 3rd ed. ASA, CSSA, and SSSA, Madison, WI.: 621-677.

Wilson, R.F. 2008. Soybean: market driven research needs. Genetics and genomics of soybean. Springer. p. 3-15.

Wu, T., X.H. Yang, S. Sun, C. Wang, Y. Wang, H.C. Jia, et al. 2017. Temporal–spatial characterization of seed proteins and oil in widely grown soybean cultivars across a century of breeding in China. Crop Sci. 57: 748-759.

Xu, K., X. Xu, P.C. Ronald, and D.J. Mackill. 2000. A high-resolution linkage map of the vicinity of the rice submergence tolerance locus Sub1. Mol. Gen. Genet. 263: 681-689.

Yesudas, C.R., R. Bashir, M.B. Geisler, and D.A. Lightfoot. 2013. Identification of germplasm with stacked QTL underlying seed traits in an inbred soybean population from cultivars Essex and Forrest. Mol. Breed. 31: 693-703.

Zhang, W.K., Y.J. Wang, G.Z. Luo, J.S. Zhang, C.Y. He, and X.L. Wu. 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.

Tables and Figures

Table 3-1. Descriptive statistics for protein and oil content of the RIL’s from the cross Osage × PI593983 across four environments in 2016 and 2017 in Missouri.

Trait	Range	Mean	h^2 (entry-mean basis)	h^2 (plot basis)	CV (%)	P-value	$LSD_{0.05}$
Oil, g kg^{-1} §	161.74-210.05	183.17	0.94	0.70	1.08	<0.0001	4.70
Crude protein, g kg^{-1} §	466.20-543.02	508.18	0.92	0.65	1.49	<0.0001	8.60

§ Seed compositions were calculated as the proportion of each seed composition for seed dry weight

Table 3-2. Description of characteristics of 20 chromosomes in the genetic map

Chr	Number of markers				Length (cM)	Average spacing (cM)	Max spacing (cM)
	After imputation	Parental homozygous	Follow the rule of segregation	After removal for gap closure			
1	1,285	421	44	39	63.1	1.7	23.4
2	1,376	558	275	270	159.3	0.6	60.5
3A	1,707	909	370	365	8.4	0.4	1.8
3B				195	53.1	0.3	4.9
3C				146	47.0	0.3	5.1
4	1,464	557	264	210	74.5	0.4	8.5
5	1,198	486	231	216	111.0	0.5	20.1
6	1,249	392	165	156	57.2	0.4	3.0
7	1,207	351	143	137	95.5	0.7	24.3
8	1,565	650	241	237	143.9	0.6	67.6
9	1,209	368	14	10	61.5	6.8	31.0
10	1,219	362	180	177	113	0.6	34.2
11	1,357	614	294	253	123.9	0.5	16.0
12	1,329	685	331	312	102.3	0.3	6.4
13A	1,386	514	220	113	32.3	0.3	1.9
13B				82	29.7	0.4	2.5
14	1,973	1,245	546	522	177.0	0.3	7.8
15	981	182	55	56	57.8	1.1	19.0
16	1,387	693	314	284	123.4	0.4	6.8
17	1,110	394	221	212	75.8	0.4	3.2
18	1,391	397	178	166	99.9	0.6	59.5
19	883	75	27	22	67.9	3.2	50.5
20	1,972	1,172	539	535	173.6	0.3	6.0
Overall	27,248	11,025	4,652	4,374	2,051.2		

Table 3-3. The detected QTLs for protein and oil in the RIL population consisting of 164 RILs using 4,372 SNPs in 2016 and 2017

QTL	Chr	Position (cM)	LOD-1.5 Interval (cM)	LOD score	R^{2} [§]	Effect[¶]	Closest SNP	SNP Position (Wm82.a2.v1)	Environment	No. Genes[#]
Oil										
Oil8.1	8	36.3	34.1-41.0	11.8	14.5	-6.5	S8_7883923	7,883,923	16ALB	246
Oil8.2	8	38.9	37.0-41.0	14.5	17.2	-8.3	S8_8565390	8,565,390	16NOV	172
Oil8.3	8	40.0	38.1-42.0	19.3	26.2	-11.9	S8_8661263	8,661,263	17CLM	140
Oil8.3	8	40.0	38.1-41.0	23.2	27.9	-11.6	S8_8661263	8,661,263	17NOV	140
Oil8.3	8	40.0	38.1-41.0	18.7	25.6	-9.6	S8_8661263	8,661,263	MEAN	140
Oil14	14	27.0	22.0-29.0	7.8	10.6	-7.2	S14_5783158	5,783,158	17NOV	120
Oil20.1	20	70.3	69.3-72.6	19.1	29.3	-9.7	S20_32687273	32,687,273	16ALB	68
Oil20.2	20	71.5	69.5-72.6	29.7	38.6	-12.6	S20_33200234	33,200,234	16NOV	59
Oil20.3	20	75.4	71.5-78.0	20.6	32.4	-13.5	S20_33622818	33,622,818	17CLM	81
Oil20.1	20	70.3	68.8-71.5	17.6	26.0	-11.5	S20_32687273	32,687,273	17NOV	71
Oil20.1	20	70.3	69.5-73.6	27.5	34.5	-11.5	S20_32687273	32,687,273	MEAN	60
Protein										
Pro14.1	14	60.3	55.8-61.8	13.9	15.0	13.4	S14_11637949	11,637,949	16ALB	220
Pro14.2	14	64.9	63.4-70.0	11.7	14.6	13.9	S14_14663282	14,663,282	16NOV	101
Pro14.3	14	123.0	119.0-128.0	7.6	11.8	12.9	S14_42568158	42,568,158	17NOV	43
Pro14.4	14	100.6	95.0-106.0	11.0	14.6	12.9	S14_34985739	34,985,739	MEAN	105
Pro20.1	20	72.0	69.5-75.4	30.9	42.4	23.2	S20_33200267	33,200,267	16ALB	94
Pro20.2	20	70.3	68.8-71.5	36.7	47.9	26.1	S20_32687273	32,687,273	16NOV	89
Pro20.2	20	70.3	68.8-75.4	22.2	41.1	25.2	S20_32687273	32,687,273	17CLM	71
Pro20.3	20	56.7	55.7-60.9	21.1	35.2	23.8	S20_25575883	25,575,883	17NOV	101
Pro20.4	20	69.3	68.8-71.5	35.3	48.2	23.9	S20_31308579	31,308,579	MEAN	37

16ALB Albany 2016, 16NOV Novelty 2016, 17CLM Columbia 2017, 17NOV Novelty 2017, MEAN mean across four environments (16ALB, 16NOV, 17CLM, and 17NOV)

[§] Estimated variance in the studied trait caused by the detected QTL

[¶] Estimated effect in g kg^{-1} with respect to the Osage allele

[#] Number of genes that have been reported within LOD–1.5 support interval

Table 3-4. Descriptive characteristics of RILs carrying different alleles of S8_8661263, S14_34985739, S20_32687273, and S20_31308579

Trait	QTL	Source of allele	N	Mean	Std Dev	Std Err	Minimum	Maximum	P-value [§]
Oil (g kg^{-1})	S8_86612631	Osage	64	187.7	7.84	0.98	175	210.1	<0.0001
		PI593983	62	178.1	7.60	0.97	161.7	198.9	
	S20_32687273	Osage	69	197.2	7.99	0.96	183.7	220.2	<0.0001
		PI593983	53	185.2	6.27	0.86	169.5	196.9	
Protein (g kg^{-1})	S14_349857391	Osage	72	502.7	14.23	1.68	474.4	530.7	<0.0001
		PI593983	57	515.5	13.78	1.83	466.2	543	
	S20_326872731	Osage	69	497.4	11.69	1.41	466.2	518.6	<0.0001
		PI593983	53	521.8	8.61	1.18	497.3	543	

[§] Significance level for the RILs carrying different alleles of the studied loci, based on a two-tailed Student’s T test

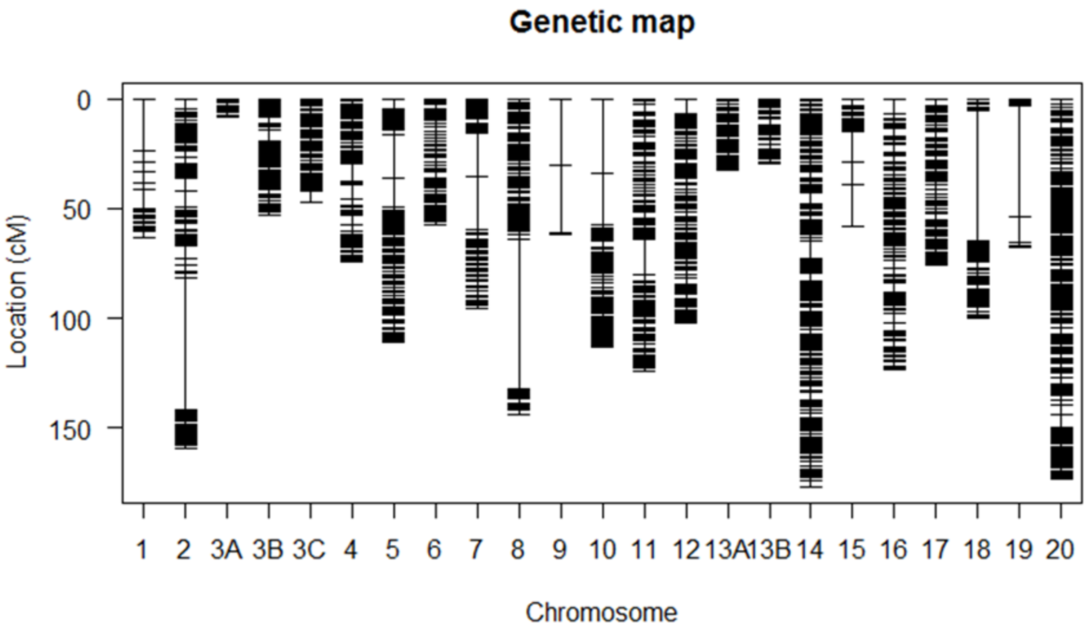


Figure 3-1. The distribution of 4,374 markers in 20 linkage groups in the RILs population developed from Osage × PI593983. The horizontal bars in each linkage group represent the mapped markers. The linkage group numbers are shown under the horizontal axis, and genetic distances between mapped markers are shown on the left of the vertical axis.

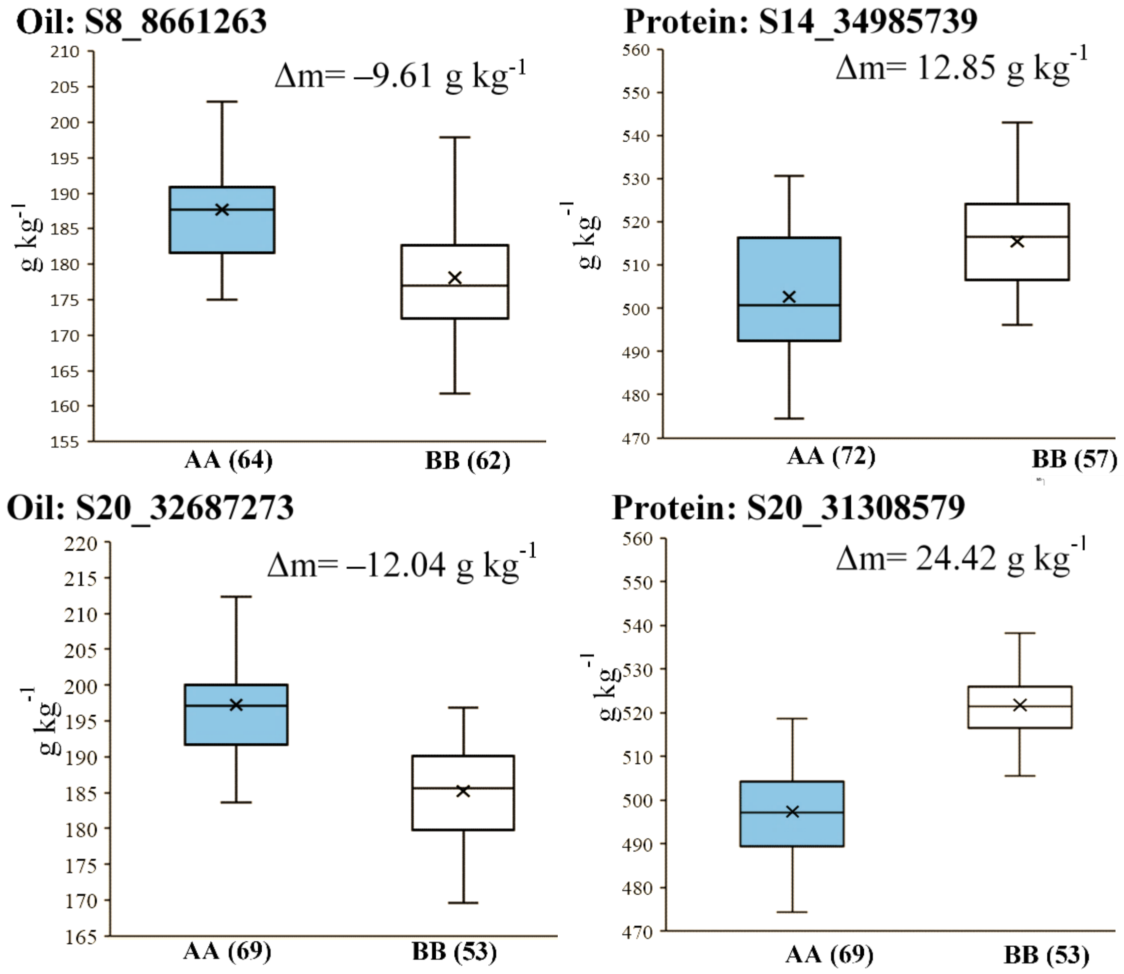


Figure 3-2. Phenotypic differences between lines carrying different homozygous alleles of (a) S8_8661263, (b) S14_34985739, (c) S20_32687273, and (d) S20_31308579 associated with oil and protein. The boxplots show the differences of (a), (c) oil content and (b), (d) protein contents across all studied environments between RILs with different homozygous alleles at the detected SNP locus (AA=Osage's homozygous alleles, BB= PI593983's homozygous alleles). The boxes show the mean (presented as ×), first and third quartiles, and Median. The numbers in the parenthesis are the numbers of RILs for each allele. The given Δm, r, and P are the difference in mean tested by the student's t-test, the Pearson correlation coefficient between genotypic and phenotypic data, and the P value of correlation, respectively.

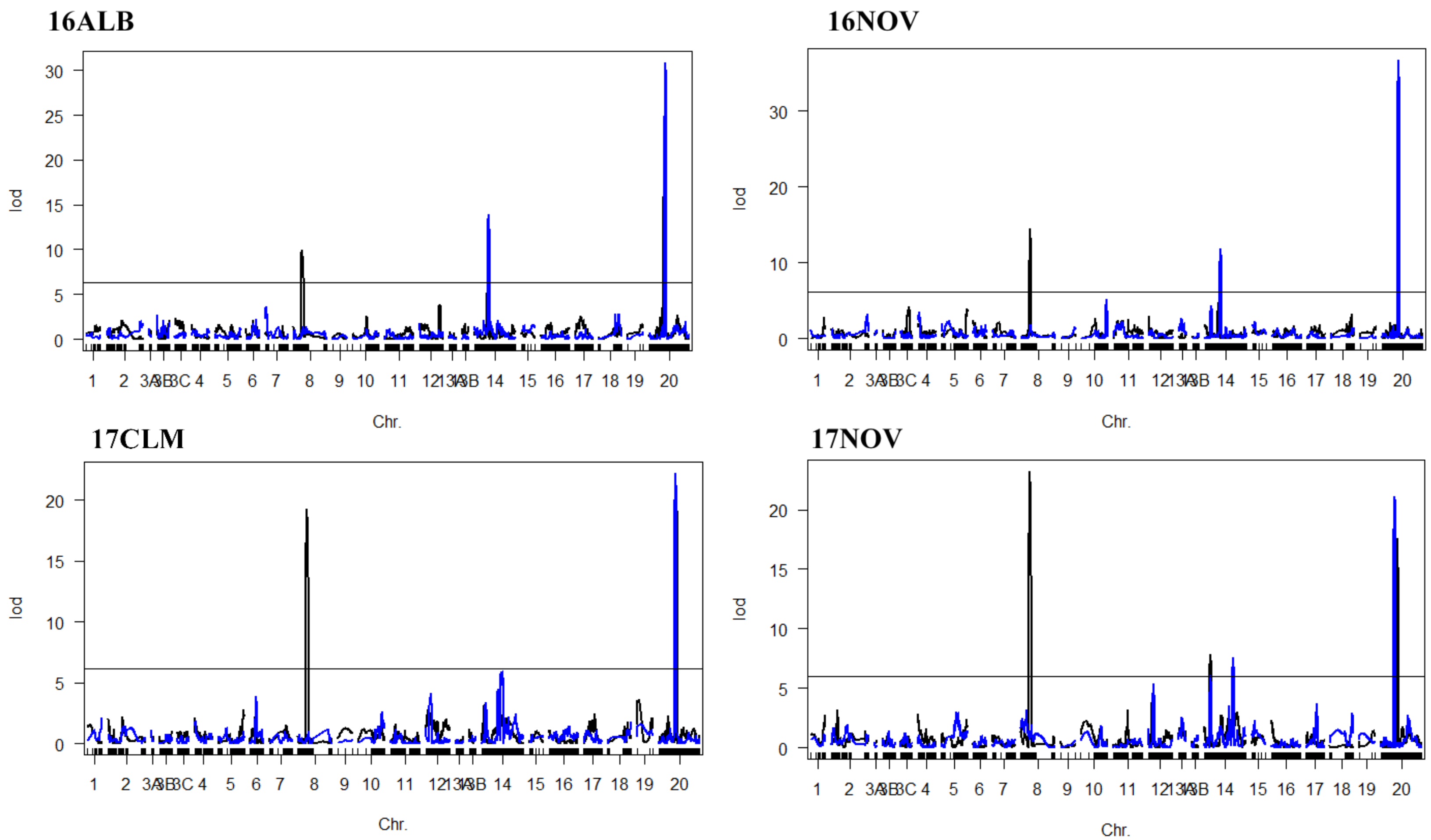


Figure 3-3. Genetic mapping of genes controlling seed protein and seed oil of the population developed from Osage × PI593983. Blue indicates seed protein and black indicates seed oil. The threshold for significant QTL is indicated by the horizontal line.

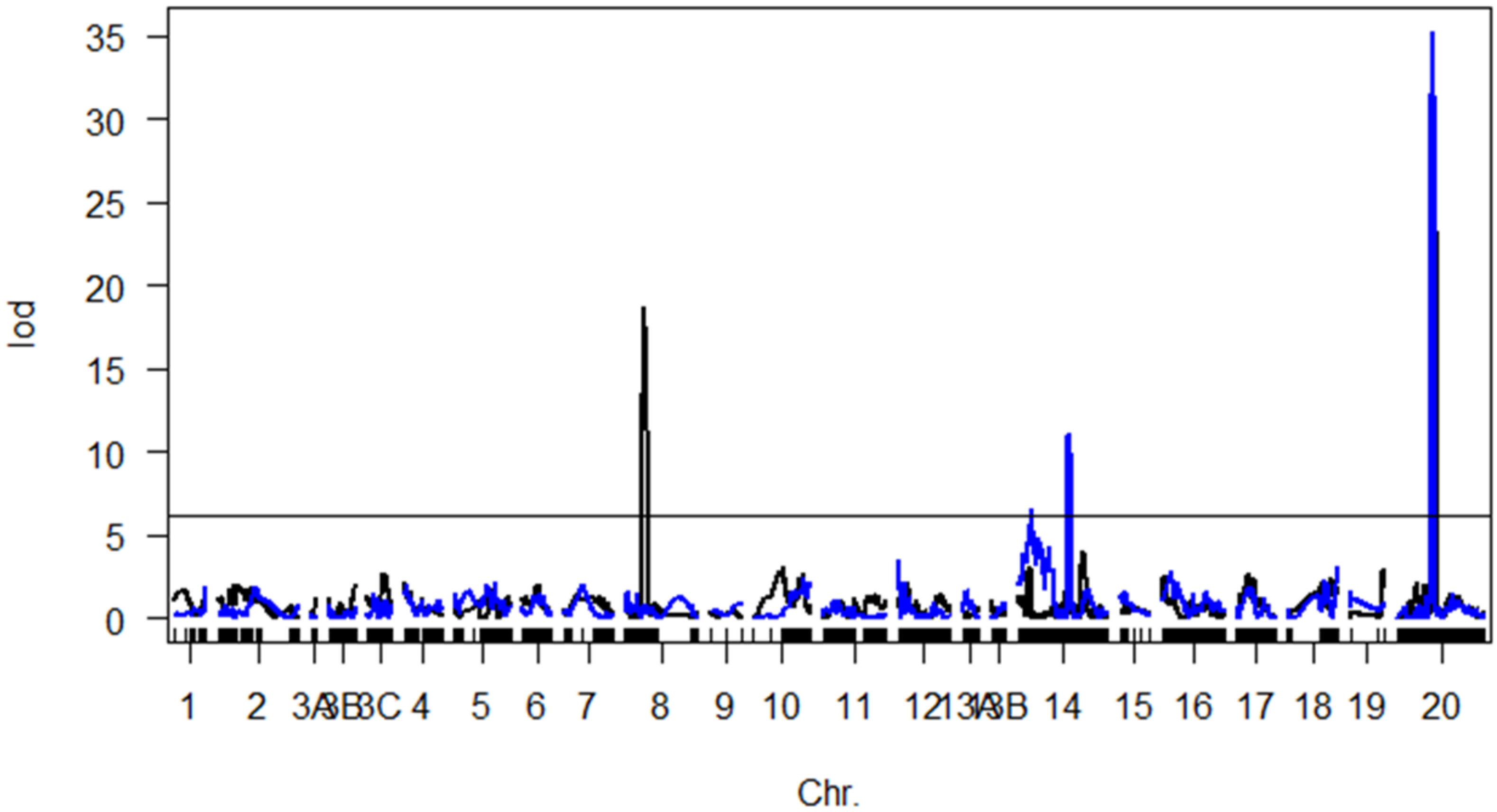


Figure 3-4. Genetic mapping of genes controlling seed protein and seed oil of the population developed from Osage × PI593983 across four studied environments, 16ALB, 16NOV, 17CLM, and 17NOV. Blue indicates seed protein and black indicates seed oil. The threshold for significant QTL is indicated by the horizontal line.

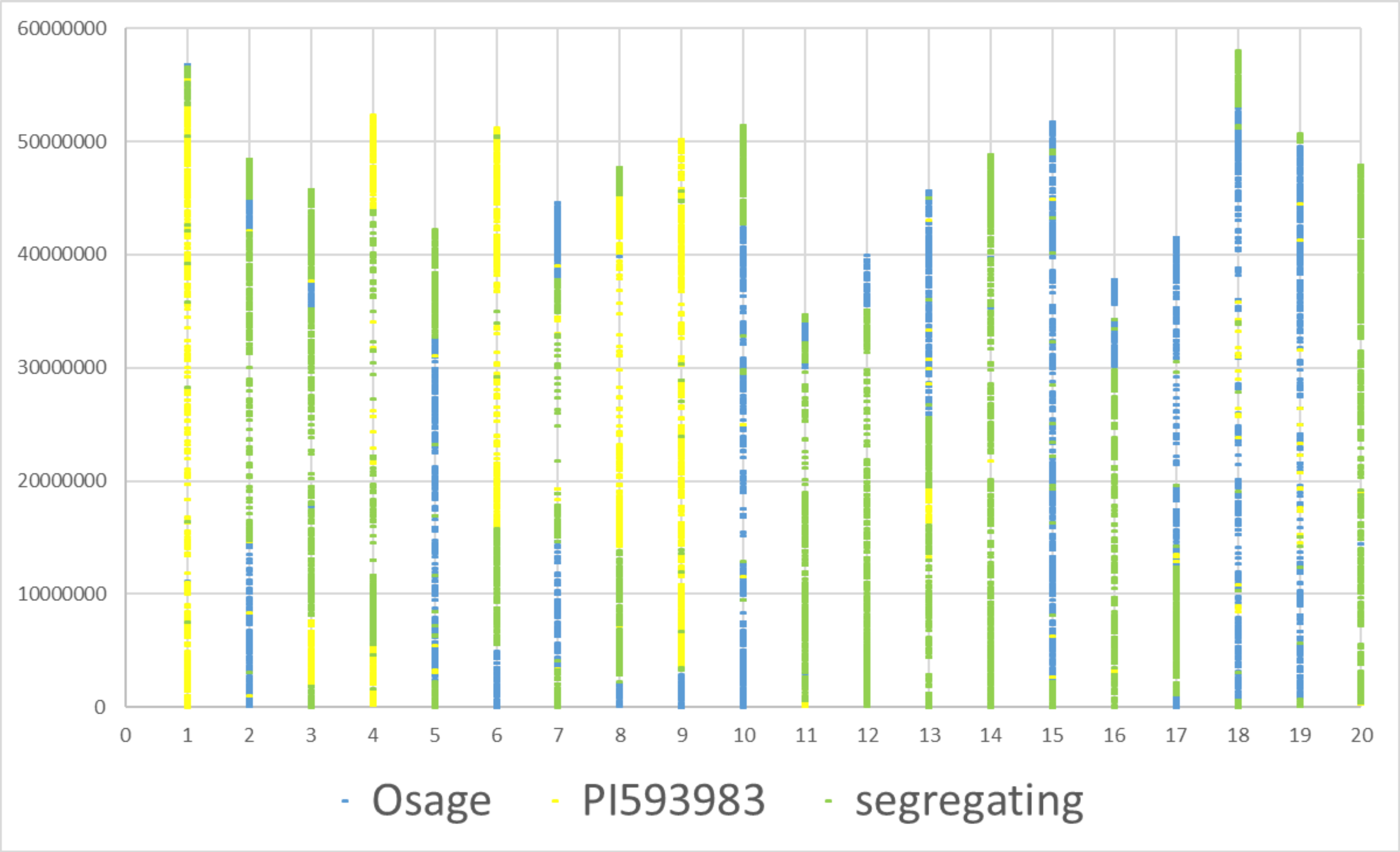


Figure 3-5. Distribution of genotyping-by-sequencing derived Osage (blue), PI593983 (yellow), and segregating (green) regions on 20 chromosomes of the physical map. The marker was determined to be from Osage or PI593983 if the frequency of major allele was >0.70. And the marker was determined to be segregating if the frequency of major allele was >0.25 and <0.70. Physical locations of markers were shown on the vertical axis in bp.

Chapter IV:

QTL MAPPING FOR MATURITY AND BRANCHING TRAITS IN THE OSAGE × PI593983 POPULATION

Abstract

In this study, we aimed to identify QTLs associated with maturity, branching traits including number of branches, total branch length, average branch length, the ratio of total branch length to height, height, lodging, and yield. The mapping population in this study included 164 F4:6 recombinant inbred lines (RILs) derived from a cross between Osage, a cultivated soybean variety, and PI593983, a wild soybean accession. The RIL population was planted in two different locations in Missouri in 2016 and 2017, and phenotyped for maturity, yield, lodging, and branching traits. Branching traits including branch number, total branch length, average branch length, and the ratio of total branch length to plant height were recorded for three randomly collected plants from the middle of each plot. Utilizing a genetic linkage map constructed using genotyping by sequencing libraries, we identified eight novel QTLs and confirmed sixteen QTLs associated with maturity (R^2 = 6.4 to 26.3%), plant height (R^2 = 7.4 to 15.5%), and total branch length (R^2 = 9.3% and 14.5%) in individual and across environments, and the ratio of total branch length to plant height (R^2 = 11.8%), yield (R^2 =12.8 and 15.7), and lodging (R^2 = 12.1 and 13.4) in individual studied environments. Sixteen QTLs for maturity, yield, and plant height confirmed previously reported QTLs, and eight QTLs have not been reported before. The results of this study will facilitate the identification of the causative genes for maturity,

height, lodging, and branching traits, and will help soybean breeder improve soybean performance by developing markers for marker-assisted selection.

Introduction

Soybean [*Glycine max* (L.) Merr.] is an important crop that serves as a major source of protein and oil for human consumption, animal feed, and use as biodiesel fuel. Soybean is classified as short day plant whose flowering is suppressed when the conditions are long-daylength and induced when the conditions are short-daylength (Watanabe et al., 2012). Maturity genes *E1*, *E2*, *E4*, and *E7* have been well reported for their responses to daylength conditions artificially induced with different red to far-red quantum ratios (Buzzell, 1971; Saindon et al., 1989; Cober et al., 1996; Cober and Voldeng, 2001). As short day plant, soybean is sensitive to photoperiod and each soybean cultivar or genotype should be grown in a certain narrow zone of latitude (Liu et al., 2018). However, Cao et al. (2017) stated that soybean, in general, was grown over a wide zone of latitude, from about 50^o N to 35^o S. Understanding the genetic and molecular mechanism underlying the wide adaptation of soybean is important for soybean breeding.

Ten *E* loci, which are named from *E1* to *E10*, and one *J* locus have been reported to be associated with maturity in soybean (Bernard, 1971; Buzzell, 1971; Buzzell and Voldeng, 1980; McBlain and Bernard, 1987; Ray et al., 1995; Bonato and Vello, 1999; Cober and Voldeng, 2001; Cober and Morrison, 2010; Fanjiang et al., 2014; Samanfar et al., 2017). Among these, the causal genes of *E1*, *E2*, *E3*, *E4*, *E9*, and *J* have been reported at the molecular level (Li et al., 2008; Watanabe et al., 2009; Watanabe et al., 2011; Xia et al., 2012; Lu et al., 2017). The *E1* locus was reported to control maturity and flowering time in soybean (Bernard, 1971). *E1* is located on chromosome 6 and encodes a transcription factor that is specific in legume (Xia et al., 2012). This transcription factor has a putative nuclear localization signal (Xia et al., 2012). *E2* is located on chromosome 10 and is a co-ortholog of *GIGANTEA*, a flowering gene in *Arabidopsis thaliana*. *E3* and *E4* are located on chromosome 19 and 20, respectively, and are *PHYA* homologs which regulate the flowering reactions to long-day conditions when there are different ratios of red-to-far-red quantum. *E1, E2, E3,* and *E4* downregulate the orthologs of *FLOWERING LOCUS T* (*GmFT2a* and *GmFT5a)* in *Arabidopsis* and delay flowering time as well as maturity under long day condition (Kong et al., 2010; Thakare et al., 2011; Watanabe et al., 2011; Xia et al., 2012; Jiang et al., 2014). Different combinations of various alleles at *E1*, *E3*, and *E4* determine the responses of soybean flowering, preflowering, and postflowering to photoperiods, and contribute to the geographical adaptation of soybean (Tsubokura et al., 2013; Xu et al., 2013; Jiang et al., 2014). *E9* is located on chromosome 16, and the causal gene of *E9* is *FT2a*, an *Arabidopsis* FLOWERING LOCUS T's ortholog, whose recessive alleles lead to late flowering (Fanjiang et al., 2014; Zhao et al., 2016). *J* is located on chromosome 4 and is one of the orthologs of flowering time gene *ELF3* in Arabidopsis (Lu et al., 2017).

Branching traits, determine the growth and development of branches, have important effects on the canopy architecture, lodging resistance, light reception, and density of plants (Board and Kahlon, 2013; Yang et al., 2017). Narrow-row/high-density planting (>25 plants m^{-2} and 40-50cm interrow spacing) is widely applied in the USA (Heatherly and Elmore, 2004); however, lower plant densities (<20 plants m^{2}) are used in Korea to avoid lodging and disease, and lower the costs of seed and labor (Cho and Kim, 2010). Agudamu et al. (2016) and Cox et al. (2010) showed that the decrease in plant density was associated with the increase in seed yield from branches, and this increase in branch seed yield compensated for the reduction of main stem number per planting area due to lower plant density.

Branching in soybean can be affected by environmental conditions including planting date, row spacing, plant density, and nutrient availability (Acock and Acock, 1987; Weaver et al., 1991; Asanome and Ikeda, 1998; Foroutan-pour et al., 1999). Acock and Acock (1987) showed that an increase in number of weeks that plants were in shading condition would lead to decreases in the number and length of branches. In their study, Acock and Acock (1987) used 66% shade cloth which transmitted 34% of light to cover the plants and create shading condition. Settimi and Board (1988) found that planting date with optimal photoperiod resulted in an improvement of the branch distribution in soybean. Schon and Blevins (1990) reported that the number of branches in soybean increased when boron was used. Board and Kahlon (2013) stated that the variation in branch development under subnormal density resulted in differences in yield. Shim et al. (2017) observed differences in branch development among soybean varieties in which American/Chinese soybean varieties had fewer branches than Korean/Japanese soybean varieties. Because of the influences of environment on soybean branching phenotype, the genomic studies about this trait are limited. Nelson (1996) found that branching trait in soybean was controlled by different alleles at two different loci (*Br1*, *Br2*). Chen et al. (2007) studied 154 recombinant inbred lines (RILs) from a cross between Charleston, an American semi-dwaft variety, and Dongnong 594, a high protein line, and found seven QTLs that were mapped for branch number. By using simple sequence repeat loci, Japanese and Korean cultivars, which showed variation in branching phenotypes, could be differentiated (Hwang et al, 2008). Studying 172 RILs developed from a cross between Tokei 758 and To-8E, Sayama et al. (2010) reported five QTLs that were significantly associated with branching number. Shim et al. (2017) identified one novel QTL and confirmed three previously reported QTL associated with branching in 200 RILs developed from a cross between Jiyu69, a Chinese elite cultivar, and SS0404-T5-76, an elite high-yielding line. Yang et al. (2017) constructed a genetic linkage map of a F_2 population developed from the cross between Toyomusume, a Japanese cultivar, and Suinong 10, a Chinese cultivar by using 1,306 polymorphic single nucleotide polymorphism (SNP) markers. They identified one major QTL, *qBR6_1*, for branching number, and located this QTL near maturity gene *E1* on chromosome 6.

More than two hundred QTLs associated with flowering time and maturity and seventeen QTLs associated with branching have been detected (http://www.soybase.org, “SoyBase browser”, accessed 05/20/2018). The reported regions of these QTLs consist of many genes and the causal genes for these traits have not been fully characterized. The objective of this study was to identify QTLs associated with branching traits and maturity.

Materials and methods

Plant materials and field experiment

The US soybean cultivar Osage [*Glycine max* (L.) Merr.] and the wild soybean accession PI593983 (*G. soja* Sieb. and Zucc.) from the Hokkaido area of Japan were crossed in North Carolina in 2011. The F_1 seeds were grown at a USDA-ARS winter nursery in Isabela, Puerto Rico (coordinates: 18°30'N, 67°1'W; soil type: Coto clay) to get F_2 seed. During the summer of 2012, the F_2 population was grown in Columbia, MO. A single F_2 plant with unique growth habit was selected. This plant had productive lateral branches upright and equal to the length of the main stem. The F_3 seeds from this plant were grown in Columbia, MO during the summer of 2013. During the growing season of 2013, 338 plants were grown and harvested individually during the fall of 2013, and $F_{3:4}$ plant rows were grown at Bradford Farm in Columbia, MO (coordinates: 38°59'N, 92°12'W; soil type: Mexico silt loam). One plant was randomly selected and individually harvested from each line. The $F_{4:5}$ seeds were sent to winter nursery in Isabela, Puerto Rico for seed increase. During summer 2015, 164 $F_{4:6}$ RILs were randomly selected and planted in a replicated experiment to evaluate seed yield, branching traits, and agronomic traits (maturity, height, and lodging). Due to extreme weather, the germination rate was low and data was excluded from the analysis.

During summer 2016, 164 $F_{4:7}$ RILs were planted in replicated field experiments at Greenley Memorial Research Center in Novelty, MO (coordinates: $40^{o}01$'N, $92^{o}11$'W; soil type: Putnam silt loam) and at the Hundley-Whaley Research Center in Albany, MO (coordinates: $40^{o}15$'N, $94^{o}19$'W; soil type: Grundy silt loam). During summer 2017, the field experiment was carried out at Bradford Farm in Columbia, MO (coordinates:

38°59’N, 92°12’W; soil type: Mexico silt loam) and at Greenley Memorial Research Center in Novelty, MO. The field experimental design was a randomized complete block design with two replications. Each plot consisted of two rows with a length of 3.66m, spacing of 0.76m, and a plant interval of 1.2m. Seed were planted at the rate of 27 seeds m^{-1}. A four-row ALMACO cone planter with Kinze row units (ALMACO, Nevada, IA) was used to plant the population.

Phenotyping for maturity, yield, lodging, and branching traits

Maturity at R8 [Fehr et al., 1971) was recorded when approximately 95% of all pods became mature within each plot and scored as the number of days after September 1st. Lodging was simultaneously recorded with maturity. Lodging was scored on a 1-5 scale (1 = all plants fully erect; 5 = all plants on the ground). Seed yield and seed moisture were simultaneously measured during harvest by using an ALMACO SPC-40 plot combine (ALMACO, Inc. Nevada, IA). Seed yield was recorded and adjusted to a standard value of 13% moisture.

Three plants were randomly collected in the middle of the plot for phenotyping height and branching traits. Plant height was measured as the length from the cotyledonary node to the terminal node of the main stem at maturity. The number of primary branches, total branch length were recorded for phenotypic evaluation. Primary branches are branches with two or more nodes with at least one mature seed pod at harvest [Chen 2007; Sayama 2010].

The length of each primary branch and total length of all primary branches were manually recorded. The lowest node where the first primary branch appeared was recorded as the starting node.

Statistical Analysis

Maturity, yield, branching traits, and lodging were analyzed with a randomized complete block mixed model analysis of variance (Proc MIXED, SAS 9.4; SAS Institute, Cary, NC, USA). Genotype was fixed effect in the model, and environment and block were random effects. Correlation analysis was carried out by using PROC CORR in SAS version 9.4 (SAS Institute, Cary, NC, USA). Broad sense heritability based on entry-mean was calculated as $h^2_{\text{(entry-mean basis)}} = \sigma_G^2 / \left[\sigma_G^2 + \left(\frac{\sigma_{GE}^2}{k}\right) + \left(\frac{\sigma_e^2}{rk}\right)\right]$, and broad sense heritability based on plot was calculated as $h^2_{\text{(plot basis)}} = \sigma_G^2 / [\sigma_G^2 + \sigma_{GE}^2 + \sigma_e^2]$, where σ_G^2 is the genotypic variance, σ_{GE}^2 is the variance of genotype by environment interaction, σ_e^2 is experimental error, t is number of test environments, and r is number of replications (Nyquist and Baker, 1991).

DNA isolation and Genotyping-By-Sequencing

About 40 mg of lyophilized leaf tissue from a pool of 5-10 plants per $F_{4:5}$ RIL was used to isolate DNA by using DNeasy Plant Mini kit (QIAGEN, Valencia, CA) following the instructions of the manufacturer. DNA samples were sent to Institute for Genomic Diversity (IGD) at Cornell University to create genotyping by sequencing (GBS) libraries (Elshire et al., 2011) by using DNA ligase, ApeKI, and suitable Illumina adapters. Illumina sequencing, library construction, read mapping, and SNP calling were carried out by IGD using TASSEL.

A total of 548,086,161 reads for Osage, seven *G. soja* lines including PI593983, and the RIL population were produced. The PCA analysis showed that one RIL was not from the cross and this RIL was excluded from further analysis. The BWA 0.7.8-r455 program was used and mapped 64.1% of the reads to a single position in the reference sequence of ‘Williams 82’ Wm82.a2.v1 [(Li and Durbin; Schmutz et al.); http://phytozome.jgi.doe.gov/]. SNPs were called by using the TASSEL 5.0 pipeline to get 139,012 filtered SNP positions and 170,463 raw SNPs in total that showed 7.019 and 6.687 mean site depth in the filtered and raw datasets, respectively.

SNP dataset quality control

TASSEL software and SNPs filtered were used to calculate allele frequencies and exclude alleles with >80% missing data. To impute missing data the LinkImpute program (Money et al., 2015) with the settings of 10 nearest neighbors and 30 high LD sites was used. The ABH genotype function in TASSEL was used to assign parental genotypes. GBS related genotyping errors were corrected by using the ABHGenotypes function in R, the correctUnderCalledHets and correctStretches functions (settings were maxhaplength=3). Only SNPs with definitive parental origin were used for further genetic map creation and QTL mapping.

Genetic Linkage Map Creation

R/qtl (F Foundation for Statistical Computing) software package was used to construct the genetic linkage map by using 4,652 polymorphic SNPs. The est.map function as used to estimate genetic distances and genotyping error rate was reported. The droponemarker and est.map functions were used to manually remove single markers to evaluate chromosomes with excessive map distances (>200 cM). Chromosome 3 and chromosome 13 were divided into three and two sub-chromosomes, respectively. The ripple function was applied to evaluate all chromosomal marker orderings, and none more appropriate marker order was found than that present in the original Wm82.a2.v1 assembly.

QTL analysis

QTL analysis was carried out by using the R/qtl software package (http://www.rqtl.org/). The Expectation-Maximization (EM) algorithm, implemented in R/qtl, was used to detect the QTLs (Xu and Vogl, 2000; Sen et al., 2009). The composite interval mapping (CIM) procedure was carried out to analyses with a 10cM window. For maturity, branching traits, height, and lodging, 1,000 permutations at the 10% level of probability were carried out to calculate the empirical logarithm of odds (LOD) thresholds (Churchill and Doerge, 1994). The effectplot function, implemented in R/qtl, was used to determine the effect of each significant QTL, following sim.geno function with 1,000 draws and an error probability of 0.01. The lodint function was applied to calculate the confidence intervals, presented as 1.5-LOD, for each significant QTL.

Results

There were significant differences (P<0.0001) among the RILs for maturity, lodging, yield, plant height, and branching traits (Table 4-1). Branching traits included starting node (the node with the lowest primary branch), number of branches, average branch length, total branch length, and the ratio of total branch length to plant height. The entry-mean based heritabilities for lodging, maturity and yield ranged were 0.78, 0.93, and 0.84, respectively. These high heritabilities suggested that genetic variation accounted for a major part of the phenotypic variance in the population. The heritabilities based on entry mean were low for height, and branching traits, they ranged from 0.14 to 0.31 (Table 4-1). Maturity, lodging, yield, plant height, and branching traits showed significant differences across four studied environments (Table 4-3). Therefore, we performed QTL mapping separately based on each environment as well as across all environments.

The ratio of total branch length to height showed significant correlations with maturity, lodging, height, average branch length, and total branch length (Table 4-2). The branching traits' correlations are expected because they are partly related in the branching development. Our study showed that yield was significantly associated with lodging and maturity (-0.34 and 0.36, respectively; Table 4-2). Maturity and lodging significantly correlated with all other studied traits, except starting node (Table 4-2).

The detail information about the genetic linkage map was presented in Table 4-4 and Figure 4-1. More than 27,000 markers were used to genotype the RIL population. After filtration and quality assessment, 4,374 markers were used for genetic linkage map construction. About 23,000 markers were excluded from further analyses due to distorted segregation and fixation (Table 4-4; Figure 4-1). The exclusion of most markers could be explained by the origin of the RIL population from one single F_2 plant which had approximately half of the genome fixed. The exclusion of more than 23,000 markers would lead to big gaps in the genetic linkage map (>20 cM gap in chromosomes 1, 2, 5, 7, 8, 10, 18, and 19). In general, after filtration and quality assessment, 4,374 markers were used for genetic linkage map construction. The genetic linkage map consisted of 20 linkage groups corresponding to the twenty chromosomes of soybean (Table 4-4; Figure 4-1) and the total genetic length of this map was 2,051 cM. The average distance between adjacent markers was 0.47cM. The chromosome with the highest number of markers was chromosome 20 with 535 markers and a length of 173.6 cM. The chromosome with the lowest number of markers was chromosome 9 with 10 markers and a length of 61.5 cM.

Eleven QTLs were found associated with maturity and located on four different chromosomes, chromosomes 4, 11, 12, and 20, and explained 6.4 to 28.3% of the phenotypic variance (Table 4-5; Figure 4-4 and 4-6). Mat04, located on chromosome 4, was consistently identified in all studied environments, except in Columbia in 2016. RILs with homozygous alleles from PI593983 at this locus matured three to five days earlier than those with homozygous alleles from Osage at the same locus (Table 4-6; Figure 4-8). Three QTSs (Mat12.1, Mat12.2, and Mat12.3) were found significantly associated with maturity, located on chromosome 12 and had overlapped confidence interval (Table 4-5; Figure 4-4 and 4-6). These QTLs explained 21.5 to 28.3% of the phenotypic variance and their LOD scored ranged from 14.5 to 23.2 (Table 4-5). Seven QTLs on chromosome 20 were identified for maturity, explaining 6.4 to 10.3% of phenotypic variance. Mat20.2, Mat20.4, and Mat20.7 shared common confidence intervals. Mat20.1, Mat20.3, Mat20.5, and Mat20.6 had overlapped confidence intervals; however, these loci were from Mat20.2, Mat20.4, and Mat20.7 by a distance of more than 40cM. RILs with homozygous PI593983 alleles at each detected QTLs for maturity on chromosome 4 and chromosome 20 could advance the maturity up to five days. However, RILs with homozygous PI593983 alleles of detected QTLs on chromosome 12 delayed maturity up to 7 days. In general, RILs with homozygous alleles from PI593983 at S4_42641620, S20_29407063, and S20_39211186 had significantly earlier maturity than those with alleles from Osage at the same loci (Table 4-5; Figure 4-9).

Five QTLs, Hgt05, Hgt12, Hgt16.1, Hgt16.2, and Hgt05, were detected for height (Table 4-5; Figures 4-2 and 4-6). These QTLs explained 7.4 to 15.5% of the height variation reported in this study. Hgt16.1 and Hgt16.2 were located on the chromosome 16 but their supported intervals were separated by a distance of 8.1 cM (Table 4-5).

For branching traits, one QTL for average branch length, two QTLs for total branch length, and one QTL for the ratio of total branch length to height were detected (Table 4-5; Figure 4-3 and 4-6). These QTLs explained 9.3 to 15.5% of the variation for the corresponding traits in this study Table 4-5). Two QTLs for yield, Yld06 and Yld16, and two QTLs for lodging, Lod08 and Lod11, were reported in this study (Table 4-5). These QTLs explained 8.1 to 13.4% of the phenotypic variation for yield and lodging.

Discussion

Heritability of lodging (0.78), maturity (0.93), and yield (0.84) in this study were moderate to high. The heritability of maturity in this study was similar to those reported by Lee et al. (2015) (0.94) when they studied a RIL population from Wyandot (an American soybean variety) × PI567301B (a Chinese germplasm). The heritability of lodging in this study was higher than in the studies of Chen et al. (2017) (0.53) when they studied a RIL population developed from the cross between Zhongdou No.29, a lodging resistant cultivar, and Zhongdou No. 32, a lodging susceptible cultivar. The heritability of plant height and branch number (0.15 and 0.20, respectively) in this study was lower in the report of Chen et al. (2017) (0.68 and 0.42, respectively). The differences in these values could be explained by different contribution level of the interaction between genotype and environment to the estimation of heritability (Lee et al., 2015). The low heritability of plant height and branch number could be due to the high plot-to-plot error variance associated with estimating plant height and branch number as well as and the strong effect of environment on these traits.

Nine previously reported markers associated with QTLs for maturity listed in SoyBase located within or had support interval overlapped with the LOD-1.5 interval of Mat12.1 (www.SoyBase.org, "SoyBase browser", accessed 05/20/2018). A major flowering time QTL (qFT12.1) (R^2 = 36.4 to 38.3) was located on chromosome 12 and overlapped with the LOD-1.5 interval of Mat12.1. The QTL qFT12.1 was reported by Liu et al. (2018) when they studied 120 chromosome segment substitution lines developed from Jackson (PI548657, a cultivated soybean variety in the US) and JWS156-1 (a wild soybean accession from the Kinki area of Japan). They reported that the plants with homozygous recessive alleles (derived from JWS156-1) at qFT12.1 would have flowering time 2-4 days longer than the plants with homozygous alleles from Jackson at the same locus. The maturity QTLs in the same region of Mat12.1 indicated the consistency across different environmental conditions and different populations.

Eight QTLs for flowering time, pod maturity, and seed development on chromosome 4 shared common interval with Mat04.1 identified for maturity in this study (www.SoyBase.org, “SoyBase browser”, accessed 05/20/2018). Among these, Wang et al. (2015) reported qT-2 (R^2 = 7.3) associated with total growth duration when they studied four segregated populations developed from Xiaoheidou (maturity group III) and GR8836 (maturity group III). The support interval of qT-2 was overlapped with the LOD-1.5 support interval of Mat4.1. However, the effects of qT-2 and Mat4.1 were relatively small, and Mat4.1 was not detected in 2017 at Columbia location. The effect of Mat4.1 on maturity need to be confirmed by further research.

Mat20.7 was identified for maturity across four studied environments, 16ALB, 16NOV, 17CLM, and 17NOV (Table 4-5). This QTL had overlapped LOD-1.5 interval with Mat20.2 (Novelty, MO in 2016) and Mat20.4 (Columbia, MO in 2017) (Table 4-5). Chung et al. (2003) identified a QTL on chromosome 20 associated with maturity when they studied a population from a cross between PI437088A, a maturity group I *G. max* accession from the eastern region of the former USSR, and Asgrow A3733, a maturity group III cultivar from the north central USA. Nichols et al. (2006) followed the rules proposed by the Soybean Genetics Committee, confirmed this QTL, and designed it as cqPod mat-001. Mat20.7 and cqPod mat-001 might be the same QTL because they had physically overlapping LOD-1.5 support interval (Table 4-6; www.SoyBase.org, “SoyBase browser”, accessed 04/05/2018). Mat20.7 could be detected in multiple environments with large LOD values (7.3-9.6) and explained 8.9-9.6% of the phenotypic variation (Table 4-6) and could be considered a stable QTL.

Mat20.6 was another QTL for maturity also located on chromosome 20. Mat20.6 and Mat20.7 was separated by a distance of more than 40CM. Gm20_3880320 located within the LOD-1.5 support interval of Mat20.6 and was found associated with flowering time when Mao et al. (2017) performed association mapping on 91 soybean cultivars with maturity groups ranged from group MG000 to MGVIII. The gene *Glyma20g03988* is a homolog of *PFT1*, which encode phytochrome and flowering time regulatory protein 1 in *Arabidopsis*. *Glyma20g03988* locates 61.6 kb downstream of Gm20_3880320 and shared common genomic region with Mat20.6 (Klose et al., 2012); Mao et al. (2017). Klose et al. (2012) reported that the protein PFT1 regulated the accumulation of light-regulated genes' transcript and the floral transition.

There have been no markers listed in SoyBase that were associated with height and branch length and located within the LOD-1.5 support interval of Hgt08 and Tbr12 (www.SoyBase.org, "SoyBase browser", accessed 04/05/2018). Zhang et al. (2015) performed genome-wide association study for plant height, maturity dates, and flowering time on 309 soybean accessions from the USDA soybean Germplasm Collection (GRIN, http://www.ars-grin.gov/). They reported Gm08_42349221, a SNP associated with height and located on chromosome 8. However, Hgt08 might not be Gm08_42349221 due to their physical position being more than 3 Mbp away from each other, and Hgt08 might be a novel QTL associated with height.

For maturity, yield, lodging, plant height, and branching traits, some QTLs were detected in some environments or across environments but not the other environments in this study. This could be explained by the fact that these traits are quantitative traits and are strongly affected by the environmental conditions (Lee et al., 1996; Orf et al., 1999; Panthee et al., 2007; Liu et al., 2013). Lee et al. (2015) reported that the significant interaction between genotype and environment would affect the QTL detection, and fewer QTLs were identified across environments. They also emphasized the need to validate and confirm QTLs for quantitative traits by carrying out experiments across different environments or by mapping different populations developed from different pairs of parents. The validation of a QTL is necessary to determine if the expression of a QTL is stable across different environments (Lee et al., 2015).

Xie et al. (2010) stated that a high-density genetic map would lead to the identification of more recombination in a population and increased the accuracy of QTL mapping. In this study, 4,374 markers were integrated into twenty linkage groups corresponding to the twenty chromosomes of soybean with a total length of 2,051.2 cM and average marker density of 0.47 cM. There were uneven distribution of markers among linkage groups and big gaps (>20.0 cM) between adjacent markers that would result in the low coverage of the linkage map and fail to identify QTLs within these gaps. Kassem et al. (2006) recommended that using more markers would fixed the genetic gaps. The extensive score replication would be required for the added markers (Kassem et al., 2006). However, the population in this study was developed from a single F_2 plant that had about half of the genome fixed, and the fixed regions of the genomes would contribute to the generation of these gaps (Table 4-4; Figure 4-7). Big gaps (>20.0 cM) has been a common problem in numerous genetic linkage maps in soybean (Cregan et al., 1999; Song et al., 2004; Kassem et al., 2006; Lu et al., 2015; Cao et al., 2017). Kassem et al. (2006) suggested that the low recombination frequencies within a certain genomic regions because of the inversion within these area would cause these gaps. The low recombination rate in different genomic regions have been reported in different populations (Shultz et al., 2003; Shultz et al., 2006).

In conclusion, twelve QTLs for maturity, five QTLs for plant height, two QTLs for yield, two QTLs for lodging, two QTLs for total branch length, and one QTL for the ratio of total branch length to plant height were identified in different individual environments and/or the average data over all environments in the population used for mapping. Of the QTLs found in the mapping population, sixteen QTLs for maturity, yield, and plant height confirmed previously reported QTLs, and eight QTLs have not been reported before. The QTLs identified in our study were either confirmed or validated in different studies, and would be useful for further studies to understand the genetic mechanism of studied traits and contribute to the development of soybean production. Mat04.1, Mat12.1, and Mat20.6 were detected in multiple environments in this study and could be considered stable QTLs. These QTLs could be used for further fine mapping and cloning to reveal the mechanisms of soybean maturity.

References

Acock, B., and M.C. Acock. 1987. Periodic shading and the location and timing of branches in soybean. Agron. J. 79: 949.

Agudamu, T. Yoshihira, and T. Shiraiwa. 2016. Branch development responses to planting density and yield stability in soybean cultivars. Plant Prod. Sci. 19: 331-339.

Asanome, N., and T. Ikeda. 1998. Effect of branch direction's arrangement on soybean yield and yield components. J. Agron. Crop Sci. 181: 95-102.

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

Board, J.E., and C.S. Kahlon. 2013. Morphological responses to low plant population differ between soybean genotypes. Crop Sci. 53: 1109-1119.

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.

Buzzell, R.I. 1971. Inheritance 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 7: 13.

Cao, D., R. Takeshima, C. Zhao, B. Liu, A. Jun, and F. Kong. 2017. Molecular mechanisms of flowering under long days and stem growth habit in soybean. J. Exp. Bot. 68: 1873-1884.

Cao, Y., S. Li, X. He, F. Chang, J. Kong, J. Gai, et al. 2017. Mapping QTLs for plant height and flowering time in a Chinese summer planting soybean RIL population. Euphytica 213: 39.

Chen, H., Z. Yang, L. Chen, C. Zhang, S. Yuan, X. Zhang, et al. 2017. Combining QTL and candidate gene analysis with phenotypic model to unravel the relationship between lodging and related traits in soybean. Mol. Breed. 37: 43.

Chen, Q.S., Z.C. Zhang, C.Y. Liu, D.W. Xin, H.M. Qiu, D.P. Shan, et al. 2007. QTL analysis of major agronomic traits in soybean. Agric. Sci. China 6: 399-405.

Cho, Y.S., and S.D. Kim. 2010. Growth parameters and seed yield compenets by seeding time and seed density of non-/few branching soybean cultivars in drained paddy field. Asian J. Plant Sci. 9: 140-145.

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

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

Cober, E.R., and M.J. Morrison. 2010. Regulation of seed yield and agronomic characters by photoperiod sensitivity and growth habit genes in soybean. Theor Appl Genet 120.

Cober, E.R., J.W. Tanner, and H.D. Voldeng. 1996. Genetic control of photoperiod response in early-maturing, near-isogenic soybean lines. Crop Sci. 36: 601-605.

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.

Cox, W.J., J.H. Cherney, and E. Shields. 2010. Soybeans compensate at low seeding rates but not at high thinning rates. Agron. J. 102: 1238-1243.

Cregan, P.B., T. Jarvik, A.L. Bush, R.C. Shoemaker, K.G. Lark, A.L. Kahler, et al. 1999. An integrated genetic linkage map of the soybean genome. Crop Sci. 39: 1464-1490.

Elshire, R.J., J.C. Glaubitz, Q. Sun, J.A. Poland, K. Kawamoto, E.S. Buckler, et al. 2011. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS One 6: e19379.

Fanjiang, K., N. Haiyang, C. Dong, L. Ying, W. Fangfang, W. Jialin, et al. 2014. A new dominant gene conditions early flowering and maturity in soybean. Crop Sci. 54: 2529-2535.

Foroutan-pour, K., P. Dutilleul, and D.L. Smith. 1999. Soybean canopy development as affected by population density and intercropping with corn: fractal analysis in comparison with other quantitative approaches. Crop Sci. 39: 1784-1791.

Heatherly, L.G., and R.W. Elmore. 2004. Managing inputs for peak production. Soybeans: improvement, production, and uses. Madison (USA): Agronomy Monograph 16: 451-536.

Jiang, B., H. Nan, Y. Gao, L. Tang, Y. Yue, S. Lu, et al. 2014. Allelic combinations of soybean maturity loci E1, E2, E3 and E4 result in diversity of maturity and adaptation to different latitudes. PLoS One 9: e106042.

Kassem, M.A., J. Shultz, K. Meksem, Y.G. Cho, A.J. Wood, M.J. Iqbal, et al. 2006. An updated ‘Essex’by ‘Forrest’linkage map and first composite interval map of QTL underlying six soybean traits. Theor. Appl. Genet. 113: 1015-1026.

Klose, C., C. Büche, A.P. Fernandez, E. Schäfer, E. Zwick, and T. Kretsch. 2012. The mediator complex subunit PFT1 interferes with COP1 and HY5 in the regulation of Arabidopsis light signaling. Plant Physiol. 160: 289-307.

Kong, F., B. Liu, Z. Xia, S. Sato, B.M. Kim, and S. Watanabe. 2010. Two coordinately regulated homologs of FLOWERING LOCUS T are involved in the control of photoperiodic flowering in soybean. Plant Physiol. 154.

Lee, S., T.H. Jun, A.P. Michel, and M.A. Mian. 2015. SNP markers linked to QTL conditioning plant height, lodging, and maturity in soybean. Euphytica 203: 521-532.

Lee, S.H., M.A. Bailey, M.A.R. Mian, E.R. Shipe, D.A. Ashley, W.A. Parrott, et al. 1996. 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.

Li, H., and R. Durbin. 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 25: 1754-1760.

Li, Y., Z.H. Yan, Y. Guan, L. Zhu, X. Ning, M.J.M. Smulders, et al. 2008. Genetic structure and diversity of cultivated soybean (Glycine max (L.) Merr.) landraces in China. Theor. Appl. Genet. 117: 857-871.

Liu, D., Y. Yan, Y. Fujita, and D. Xu. 2018. A major QTL (qFT12.1) allele from wild soybean delays flowering time. Mol. Breed. 38: 45.

Liu, Y.l., Y.h. Li, J.C. Reif, M.F. Mette, Z.x. Liu, B. Liu, et al. 2013. Identification of quantitative trait loci underlying plant height and seed weight in soybean. The Plant Genome 6.

Lu, S., Y. Li, J. Wang, P. Srinives, H. Nan, D. Cao, et al. 2015. QTL mapping for flowering time in different latitude in soybean. Euphytica 206: 725-736.

Lu, S., X. Zhao, Y. Hu, S. Liu, H. Nan, X. Li, et al. 2017. Natural variation at the soybean J locus improves adaptation to the tropics and enhances yield. Nat. Genet. 49: 773.

Mao, T., J. Li, Z. Wen, T. Wu, C. Wu, S. Sun, et al. 2017. Association mapping of loci controlling genetic and environmental interaction of soybean flowering time under various photo-thermal conditions. BMC genomics 18: 415.

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.

Money, D., K. Gardner, Z. Migicovsky, H. Schwaninger, G.Y. Zhong, and S. Myles. 2015. LinkImpute: Fast and accurate genotype imputation for nonmodel organisms. G3: Genes, Genomes, Genet. 5: 2383-2390.

Nichols, D.M., K.D. Glover, S.R. Carlson, J.E. Specht, and B.W. Diers. 2006. Fine mapping of a seed protein QTL on soybean linkage group I and its correlated effects on agronomic traits. Crop Sci. 46: 834-839.

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

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

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.

Ray, J.D., K. Hinson, J.E. Mankono, and M.F. Malo. 1995. Genetic control of a long-juvenile trait in soybean. Crop Sci. 35: 1001-1006.

Saindon, G., W.D. Beversdorf, and H.D. Voldeng. 1989. Adjustment of the soybean phenology using the E4 locus. Crop Sci. 29: 1361-1365.

Samanfar, B., S.J. Molnar, M. Charette, A. Schoenrock, F. Dehne, A. Golshani, et al. 2017. Mapping and identification of a potential candidate gene for a novel maturity locus, E10, in soybean. Theor. Appl. Genet. 130: 377-390.

Sayama, T., T.Y. Hwang, H. Yamazaki, N. Yamaguchi, K. Komatsu, M. Takahashi, et al. 2010. Mapping and comparison of quantitative trait loci for soybean branching phenotype in two locations. Breed. Sci. 60: 380-389.

Schmutz, J., S.B. Cannon, J. Schlueter, J. Ma, T. Mitros, W. Nelson, et al. 2010. Genome sequence of the palaeopolyploid soybean. Nature 463: 178-183.

Schon, M.K., and D.G. Blevins. 1990. Foliar boron applications increase the final number of branches and pods on branches of field-grown soybeans Plant Physiol. 92: 602-607.

Sen, Ś., F. Johannes, and K.W. Broman. 2009. Selective genotyping and phenotyping strategies in a complex trait context. Genetics 181: 1613-1626.

Settimi, J.R., and J.E. Board. 1988. Photoperiod and planting date effects on the spatial distribution of branch development in soybean. Crop Sci. 28: 259.

Shim, S., M.Y. Kim, J. Ha, Y.H. Lee, and S.H. Lee. 2017. Identification of QTLs for branching in soybean (Glycine max (L.) Merrill). Euphytica 213: 225.

Shultz, J., K. Meksem, and D.A. Lightfoot. 2003. Evaluating physical maps by clone location comparisons. Genome Lett. 2: 98-105.

Shultz, J.L., D. Kurunam, K. Shopinski, M.J. Iqbal, S. Kazi, K. Zobrist, et al. 2006. The Soybean Genome Database (SoyGD): a browser for display of duplicated, polyploid, regions and sequence tagged sites on the integrated physical and genetic maps of Glycine max. Nucleic Acids Res. 34: D758-D765.

Song, Q.J., L.F. Marek, R.C. Shoemaker, K.G. Lark, V.C. Concibido, X. Delannay, et al. 2004. A new integrated genetic linkage map of the soybean. Theor. Appl. Genet. 109: 122-128.

Thakare, D., S. Kumudini, and R.D. Dinkins. 2011. The alleles at the E1 locus impact the expression pattern of two soybean FT-like genes shown to induce flowering in Arabidopsis. Planta 234: 933.

Tsubokura, Y., H. Matsumura, M. Xu, B. Liu, H. Nakashima, T. Anai, et al. 2013. Genetic variation in soybean at the maturity locus E4 is involved in adaptation to long days at high latitudes. Agron. J. 3: 117-134.

Wang, X., G.L. Jiang, Q. Song, P.B. Cregan, R.A. Scott, J. Zhang, et al. 2015. Quantitative trait locus analysis of seed sulfur-containing amino acids in two recombinant inbred line populations of soybean. Euphytica 201: 293-305.

Watanabe, S., K. Harada, and J. Abe. 2011. Genetic and molecular bases of photoperiod responses of flowering in soybean. Breed. Sci. 61: 531-543.

Watanabe, S., K. Harada, and J. Abe. 2012. Genetic and molecular bases of photoperiod responses of flowering in soybean. Breed. Sci. 61.

Watanabe, S., R. Hideshima, Z. Xia, Y. Tsubokura, S. Sato, and Y. Nakamoto. 2009. Map-based cloning of the gene associated with the soybean maturity locus E3. Genetics 182.

Weaver, D.B., R.L. Akridge, and C.A. Thomas. 1991. Grow habit, planting date, and row-spacing effects on late-planted soybean. Crop Sci. 31: 805-810.

Xia, Z., S. Lü, H. Wu, S. Tabata, K. Harada, S. Watanabe, et al. 2012. Positional cloning and characterization reveal the molecular basis for soybean maturity locus E1 that regulates photoperiodic flowering. Proc. Natl. Acad. Sci. U.S.A. 109: 12852-12853.

Xie, W., Q. Feng, H. Yu, X. Huang, Q. Zhao, Y. Xing, et al. 2010. Parent-independent genotyping for constructing an ultrahigh-density linkage map based on population sequencing. Proc. Natl. Acad. Sci. U.S.A. 107: 10578-10583.

Xu, M., Z. Xu, B. Liu, F. Kong, Y. Tsubokura, and S. Watanabe. 2013. Genetic variation in four maturity genes affects photoperiod insensitivity and PHYA-regulated post-flowering responses of soybean. BMC Plant Biol. 13.

Xu, S., and C. Vogl. 2000. Maximum likelihood analysis of quantitative trait loci under selective genotyping. Heredity 84: 525-537.

Yang, G., H. Zhai, W. H.Y., X.Z. Zang, S.X. Lü, Y.Y. Wang, et al. 2017. QTL effects and epistatic interaction for flowering time and branch number in a soybean mapping population of Japanese× Chinese cultivars. J. Integr. Agric. 16: 1900-1912.

Zhang, J., Q.J. Song, P.B. Cregan, R.L. Nelson, X.D. Wang, J. Wu, et al. 2015. Genome-wide association study for flowering time, maturity dates and plant height in early maturing soybean (Glycine max) germplasm. BMC Genomics 16.

Zhao, C., R. Takeshima, J. Zhu, M. Xu, M. Sato, S. Watanabe, et al. 2016. A recessive allele for delayed flowering at the soybean maturity locus E9 is a leaky allele of FT2a, a FLOWERING LOCUS T ortholog. BMC Plant Biol. 16: 20.

Tables and figures

Table 4-1. Summarized information about lodging score, maturity, yield, and branching traits of the soybean population developed from Osage × PI593983 across four environments, 16ALB, 16NOV, 17CLM, and 17NOV

Trait	Mean	Range	Check's range [*]	h^2 (entry-mean basis)	h^2 (plot basis)	CV (%)	P-value	LSD
Lod [†]	3.12	2.13-4.50	1.50-1.69	0.78	0.36	16.34	<0.0001	0.51
Mat, day [‡]	47.98	33.17-59.67	31.75-46.88	0.93	0.70	5.20	<0.0001	2.50
Yield, kg ha^{-1}	1,712.88	379.30-4,038.43	4,246.23-5,259.03	0.84	0.48	20.24	<0.0001	345.95
Ht, cm [§]	65.69	45.33-92.06	68.45-69.60	0.15	0.03	20.76	<0.0001	14.51
Starting node	4.92	3.06-10.25	3.38-5.00	0.14	0.02	65.28	<0.0001	3.42
NBr [¶]	5.62	3.62-8.52	4.02-5.71	0.20	0.03	34.85	<0.0001	2.06
TBr, cm [#]	230.15	145.86-375.52	157.90-245.71	0.25	0.04	40.92	<0.0001	99.05
Abr, cm [††]	40.63	25.12-52.65	37.75-40.87	0.25	0.05	17.69	<0.0001	7.90
BrH [‡‡]	3.65	2.30-5.42	2.66-3.49	0.31	0.06	33.93	<0.0001	1.32

[*] The checks that were used in this study were IA4005, Ellis, and Osage; [†] Lod = lodging score (1 = all plants erect, 5 = all plants down); ‡ Mat = maturity (days after September 1); § Ht = height (cm); ¶ NBr = Number of branches; # TBr = Total branch length; †† Abr = Average branches' length; ‡‡ BrH = the ratio of total branches' length to height

Table 4-2. Trait correlations based on the means of 164 genotypes for lodging score, maturity, yield, and branching traits of the soybean population developed from Osage × PI593983 across four environments, 16ALB, 16NOV, 17CLM, and 17NOV

	NBr ¶	Ht	LOD	MAT	Starting node	TBr #	BrH ‡‡	YLD
Abr ††	NS	0.69***	0.53***	0.44***	NS	0.56***	0.26***	NS
NBr ¶		NS	0.17*	0.20**	-0.17*	0.84***	0.87***	NS
Ht			0.48***	0.28***	0.20**	0.26***	-0.18*	NS
LOD				0.23**	NS	0.42***	0.22**	-0.34***
MAT					NS	0.39***	0.25***	0.36***
Starting node						NS	-0.24**	NS
TBr #							0.84***	NS
BrH ‡‡								NS

† Lod = lodging score (1 = all plants erect, 5 = all plants down); ‡ Mat = maturity (days after September 1); § Ht = height (cm); ¶ NBr = Number of branches; # TBr = Total branches' length; †† Abr = Average branches' length; ‡‡ BrH = the ratio of total branches' length to height

* Significant at the 0.05 probability level

** Significant at the 0.01 probability level

*** Significant at the 0.001 probability level

ns: Non-significant

Table 4-3. Summary of lodging score, maturity, yield, and branching traits of the soybean population developed from Osage × PI593983 across four environments, 16ALB, 16NOV, 17CLM, and 17NOV. Means on the same row with different letters are significantly different (LSD0.05).

Trait	16ALB	16NOV	17CLM	17NOV
ABr [††]	40.38 ^{ab}	41.00 ^{b}	43.13 ^{c}	38.43 ^{a}
NBr [¶]	5.29 ^{a}	5.46 ^{ab}	5.58 ^{ab}	6.07 ^{b}
Ht [§]	68.59 ^{b}	68.12 ^{b}	67.06 ^{b}	60.25 ^{a}
Lod [†]	2.95 ^{ab}	3.32 ^{c}	2.93 ^{ab}	3.17 ^{b}
Mat [‡]	53.83 ^{c}	44.96 ^{a}	44.87 ^{a}	48.42 ^{b}
Starting node	5.35 ^{b}	5.11 ^{b}	5.05 ^{b}	4.33 ^{a}
TBr [#]	217.32 ^{a}	228.21 ^{a}	241.51 ^{a}	234.19 ^{a}
BrH [‡‡]	3.36 ^{a}	3.58 ^{a}	3.62 ^{a}	3.95 ^{b}
Yield, kg ha^{-1}	32.42 ^{c}	26.66 ^{b}	25.41 ^{b}	21.68 ^{a}

16ALB, Albany 2016; 16NOV, Novelty 2016; 17CLM, Columbia 2017; 17NOV, Novelty 2017

Table 4-4. Description of characteristics of 20 chromosomes in the genetic map

Chr	**Number of markers**				**Length (cM)**	**Average spacing (cM)**	**Max spacing (cM)**
	After imputation	Parental homozygous	Follow the rule of segregation	After removal for gap closure			
1	1,285	421	44	39	63.1	1.7	23.4
2	1,376	558	275	270	159.3	0.6	60.5
3A	1,707	909	370	365	8.4	0.4	1.8
3B				195	53.1	0.3	4.9
3C				146	47.0	0.3	5.1
4	1,464	557	264	210	74.5	0.4	8.5
5	1,198	486	231	216	111.0	0.5	20.1
6	1,249	392	165	156	57.2	0.4	3.0
7	1,207	351	143	137	95.5	0.7	24.3
8	1,565	650	241	237	143.9	0.6	67.6
9	1,209	368	14	10	61.5	6.8	31
10	1,219	362	180	177	113.0	0.6	34.2
11	1,357	614	294	253	123.9	0.5	16.0
12	1,329	685	331	312	102.3	0.3	6.4
13A	1,386	514	220	113	32.3	0.3	1.9
13B				82	29.7	0.4	2.5
14	1,973	1,245	546	522	177.0	0.3	7.8
15	981	182	55	56	57.8	1.1	19.0
16	1,387	693	314	284	123.4	0.4	6.8
17	1,110	394	221	212	75.8	0.4	3.2
18	1,391	397	178	166	99.9	0.6	59.5
19	883	75	27	22	67.9	3.2	50.5
20	1,972	1,172	539	535	173.6	0.3	6.0
Overall	27,248	11,025	4,652	4,374	2,051.2		

Table 4-5. The detected QTLs maturity, yield, height, lodging score, and branching traits in the RIL population consisting of 164 RILs using 4,372 SNPs in 2016 and 2017

QTL	Chr	Position (cM)	LOD-1.5 Interval (cM)	LOD score	R^{2}§	Effect¶	Closest SNP	SNP Position (Wm82.a2.v1)	Environment	No. Genes#
Average branches length										
Abr11	11	65.0	59.4-70.0	9.4	10.1	6.1	S11_10834164	10,834,164	17NOV	152
Height										
Hgt05	5	7.9	2.0-13.0	6.5	10.6	-8.2	S5_796002	796,002	16ALB	218
Hgt12	12	18.0	16.5-23.9	8.7	13.0	8.1	S12_2882359	2,882,359	16NOV	100
Hgt16.1	16	9.3	7.0-12.2	7.3	15.5	-9.0	S16_884133	884,133	16ALB	107
Hgt16.2	16	25.9	24.3-28.7	6.6	9.9	-9.5	S16_4226137	4,226,137	17CLM	81
Hgt08	8	136.0	132.3-141.0	6.9	7.4	-1.2	S8_46144956	46,144,956	MEAN	220
Lodging										
Lod08	8	24.0	21.1-24.4	6.7	13.4	0.4	S8_5729235	5,729,235	16ALB	63
Lod11	11	76.0	75.0-82.9	7.2	12.1	0.4	S11_10834164	10,834,164	16ALB	94
Maturity										
Mat04.1	4	72.5	49.2-74.5	8.1	9.7	-3.7	S4_42641620	42,641,620	16NOV	114
Mat04.1	4	72.5	66.5.0-74.5	9.2	8.5	-3.9	S4_42641620	42,641,620	17CLM	107
Mat04.1	4	72.5	67.3-74.5	9.3	12.1	-4.7	S4_42641620	42,641,620	17NOV	114
Mat04.1	4	72.5	49.2-74.5	8.6	9.7	-3.5	S4_42641620	42,641,620	MEAN	114
Mat11.1	11	70.0	64.1-70.0	12.7	13.4	4.6	S11_10834164	10,834,164	16NOV	51
Mat12.1	12	36.1	35.4-41.5	15.9	21.5	3.9	S12_5520945	5,520,945	16ALB	65
Mat12.1	12	36.1	32.9-41.5	20.1	23.2	5.9	S12_5520945	5,520,945	16NOV	115
Mat12.2	12	33.7	32.9-38.0	14.5	28.3	7.0	S12_5008803	5,008,803	17CLM	57
Mat12.3	12	36.0	35.2-41.5	21.6	25.3	6.7	S12_5413540	5,413,540	17NOV	74
Mat12.1	12	36.1	35.2-41.5	23.1	23.2	5.3	S12_5520945	6,048,084	MEAN	74
Mat20.1	20	113.9	110.0-117.0	8.3	7.7	-2.4	S20_38907365	38,907,365	16ALB	79
Mat20.2	20	15.8	10.7-23.5	8.3	8.9	-3.0	S20_2070675	2,070,675	16NOV	79

QTL	Chr	Position (cM)	LOD-1.5 Interval (cM)	LOD score	R^{2}[§]	Effect[¶]	Closest SNP	SNP Position (Wm82.a2.v1)	Environment	No. Genes[#]
Mat20.3	20	117.0	112.9-119.6	7.8	6.4	-2.9	S20_39222363	39,222,363	16NOV	79
Mat20.4	20	55.7	17.5-69.5	10.3	9.6	-3.9	S20_25386853	25,386,853	17CLM	888
Mat20.5	20	120.3	111.2-123.5	6.18	7.6	-3.5	S20_40080375	40,080,375	17CLM	83
Mat20.6	20	116.4	50.1-119.0	6.3	6.4	-3.5	S20_39222363	39,222,363	17NOV	977
Mat20.7	20	66.8	17.5-69.5	7.9	7.3	-2.8	S20_29407063	29,407,063	MEAN	888
Mat20.6	20	116.4	112.0-120.1	8.7	12.0	-3.6	S20_39211186	39,211,186	MEAN	81
Total branch length										
Tbr01	12	36.10	31.0-41.5	7.3	14.5	65.9	S12_5520945	5,520,945	17NOV	174
Tbr02	12	31.90	31.0-38.0	6.5	9.3	32.6	S12_4631152	4,631,152	MEAN	149
Total branch length/Height										
Brh14	14	83.10	80.0-86.7	6.4	11.8	0.8	S14_19879389	19,879,389	17NOV	125
Yield										
Yld01	6	23.10	20.0-23.8	8.1	15.7	-5.9	S6_10851269	10,851,269	16NOV	88
Yld02	16	123.00	117.0-123.4	8.4	12.8	-8.8	S16_29449188	29,449,188	16ALB	105

16ALB, Albany, MO 2016; 16NOV, Novelty, MO 2016; 17CLM, Columbia, MO 2017; 17NOV, Novelty, MO 2017; MEAN mean across four environments (16ALB, 16NOV, 17CLM, and 17NOV)

[§] Estimated variance in the studied trait caused by the detected QTL

[¶] Estimated effect in g kg^{-1} with respect to the Osage allele

[#] Number of genes/QTLs that have been reported within LOD–1.5 support interval

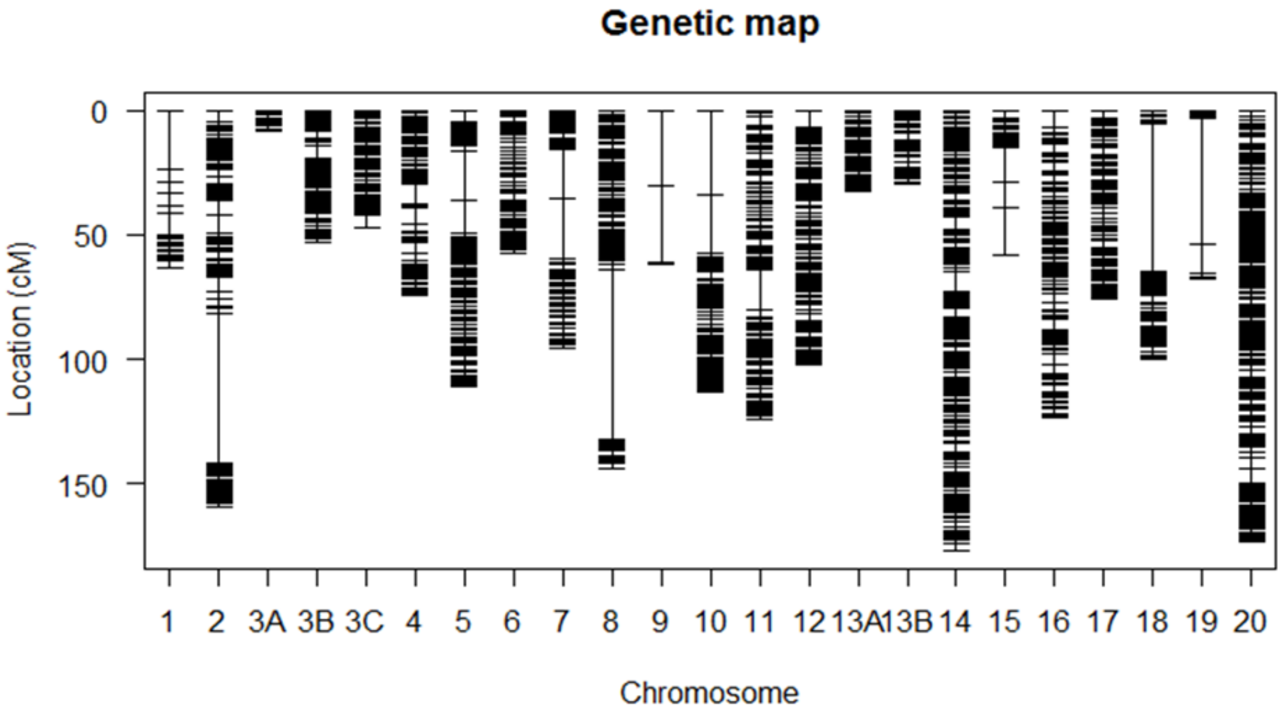


Figure 4-1. The genetic linkage map of the population developed from Osage × PI593983. The black bars in each linkage group represent the mapped markers. The numbers on the horizontal axis show the linkage group number, and the vertical axis shows the genetic distance between markers.

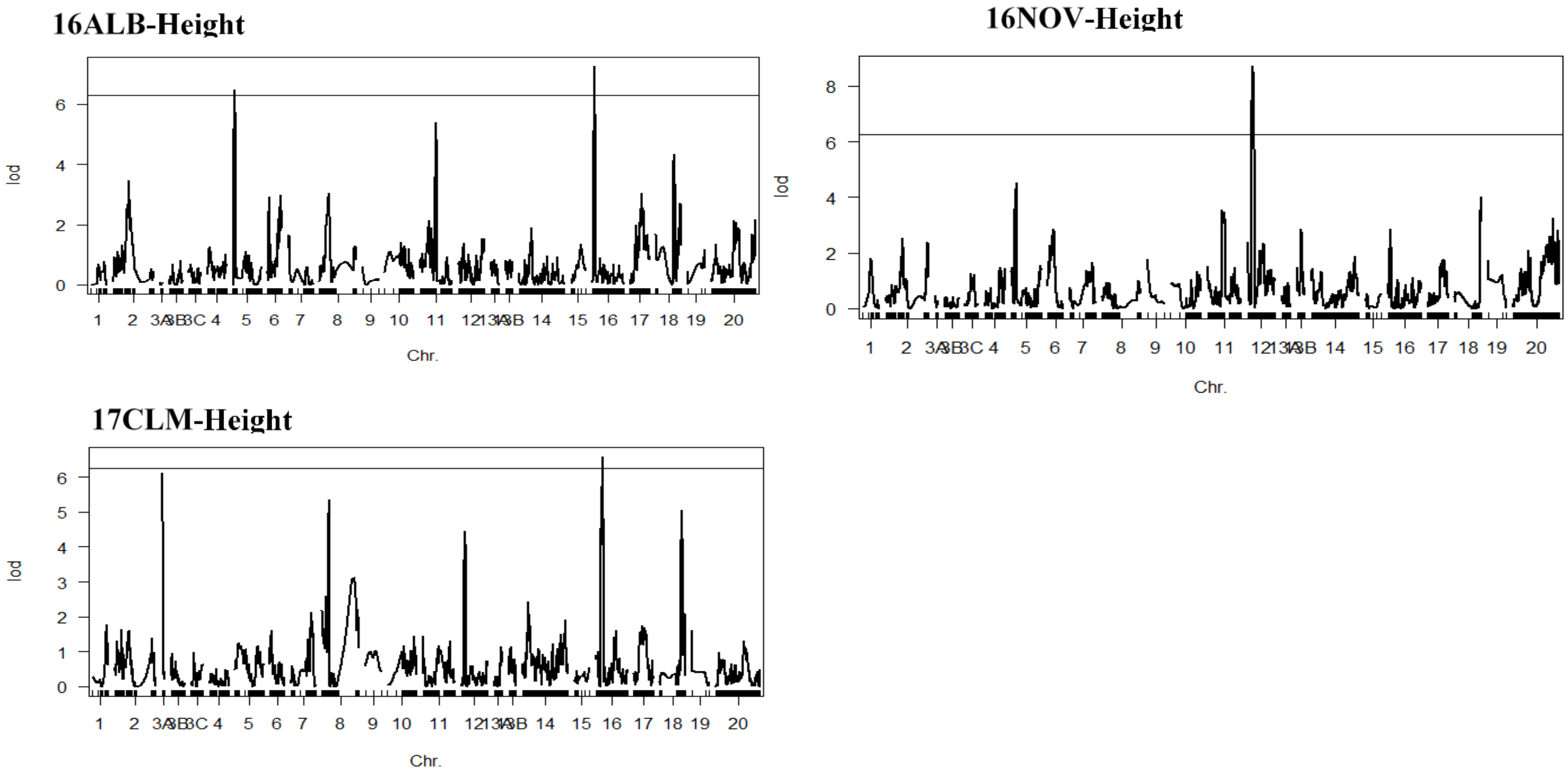


Figure 4-2. Genetic mapping of genes controlling height in 2016 and 2017 of the RILs from the cross Osage × PI593983 (16ALB, Albany, MO 2016; 16NOV, Novelty, MO 2016; 17CLM, Columbia, MO 2017). The threshold for significant QTLs is indicated by the horizontal line.

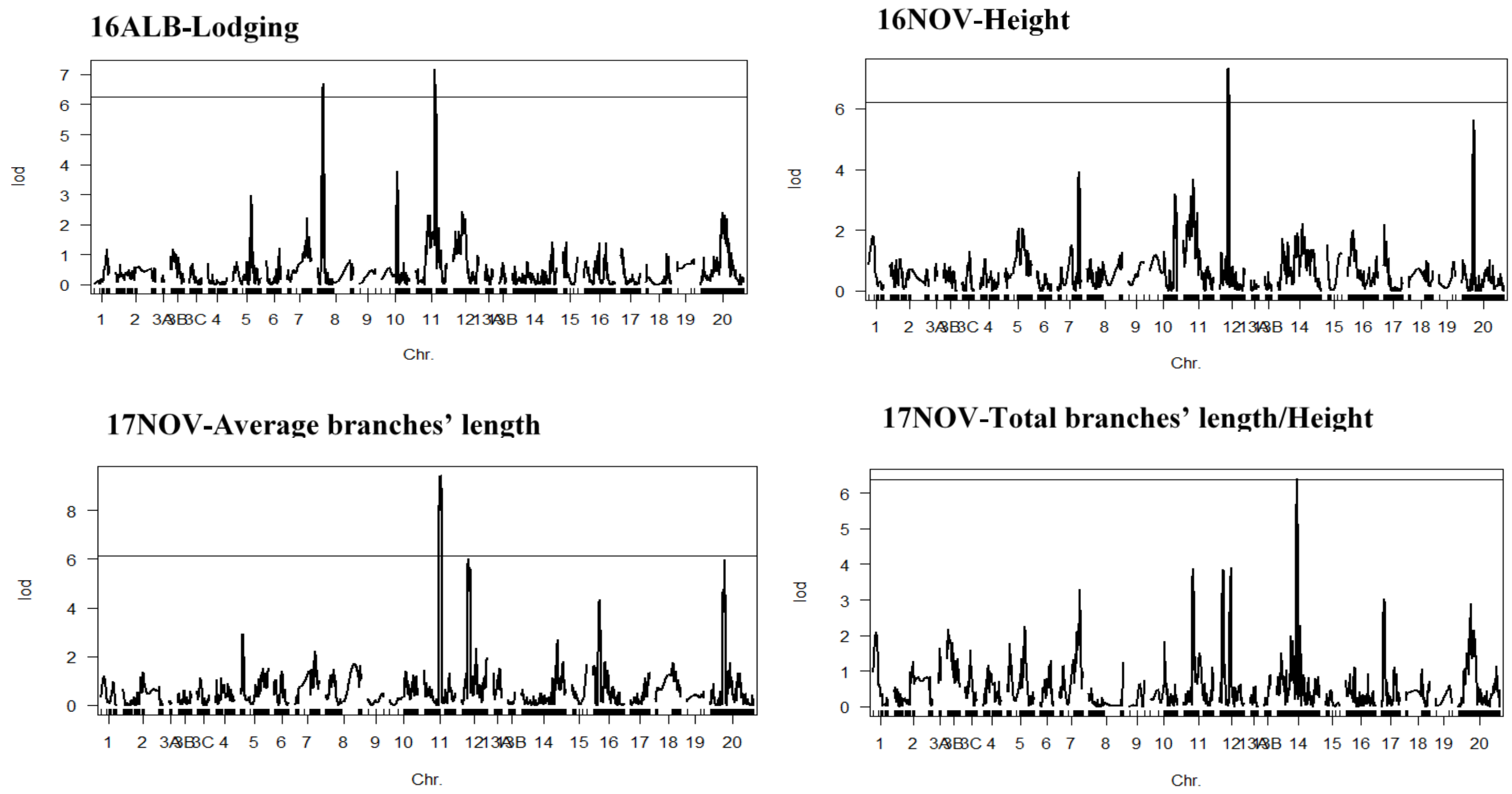


Figure 4-3. Genetic mapping of genes controlling yield, height, average branch's length, and the ratio of total branches' length to height in 2016 and 2017 in Novelty, MO of the RILs from the cross Osage × PI593983 (16NOV, Novelty, MO 2016; 17NOV, Novelty, MO 2017). The threshold for significant QTLs is indicated by the horizontal line.

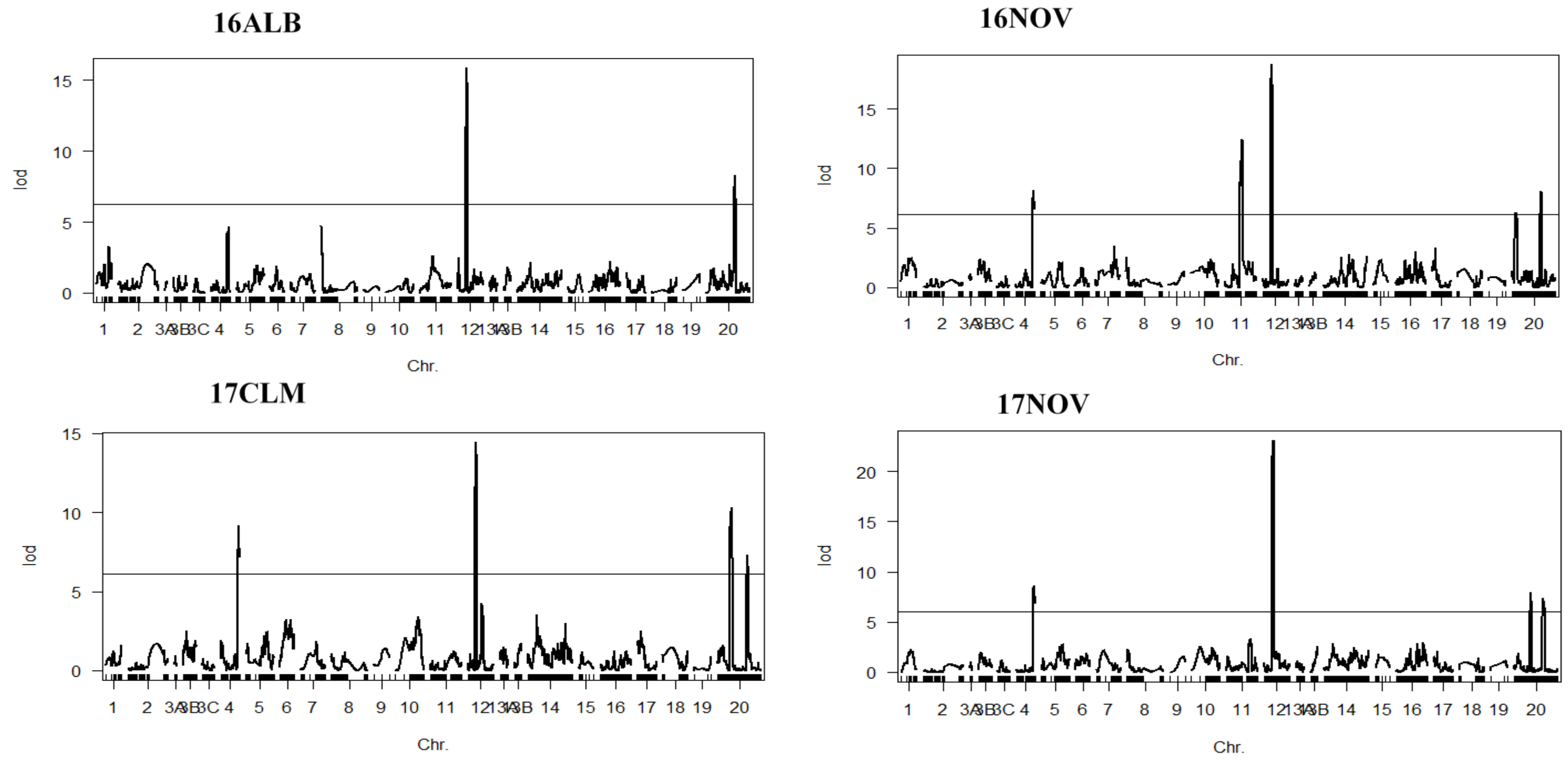


Figure 4-4. Genetic mapping of genes controlling maturity in 2016 and 2017 in Missouri of the RILs from the cross Osage × PI593983 (16ALB, Albany, MO 2016; 16NOV, Novelty, MO 2016; 17CLM, Columbia, MO 2017; 17NOV, Novelty, MO 2017). The threshold for significant QTLs is indicated by the horizontal line.

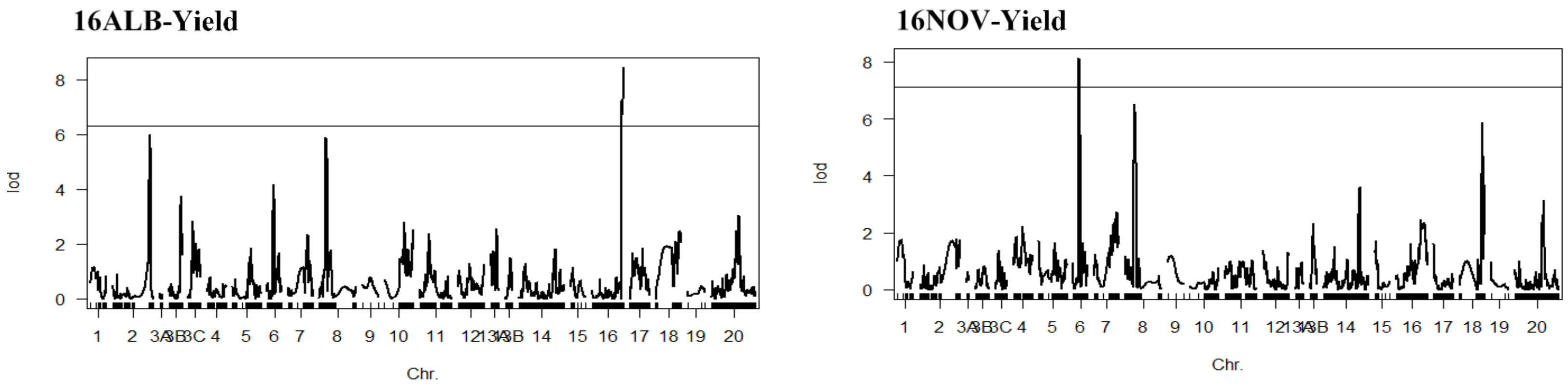


Figure 4-5. Genetic mapping of genes controlling yield in Albany, MO and Novelty, MO in 2016 of the RILs from the cross Osage × PI593983. The threshold for significant QTLs is indicated by the horizontal line.

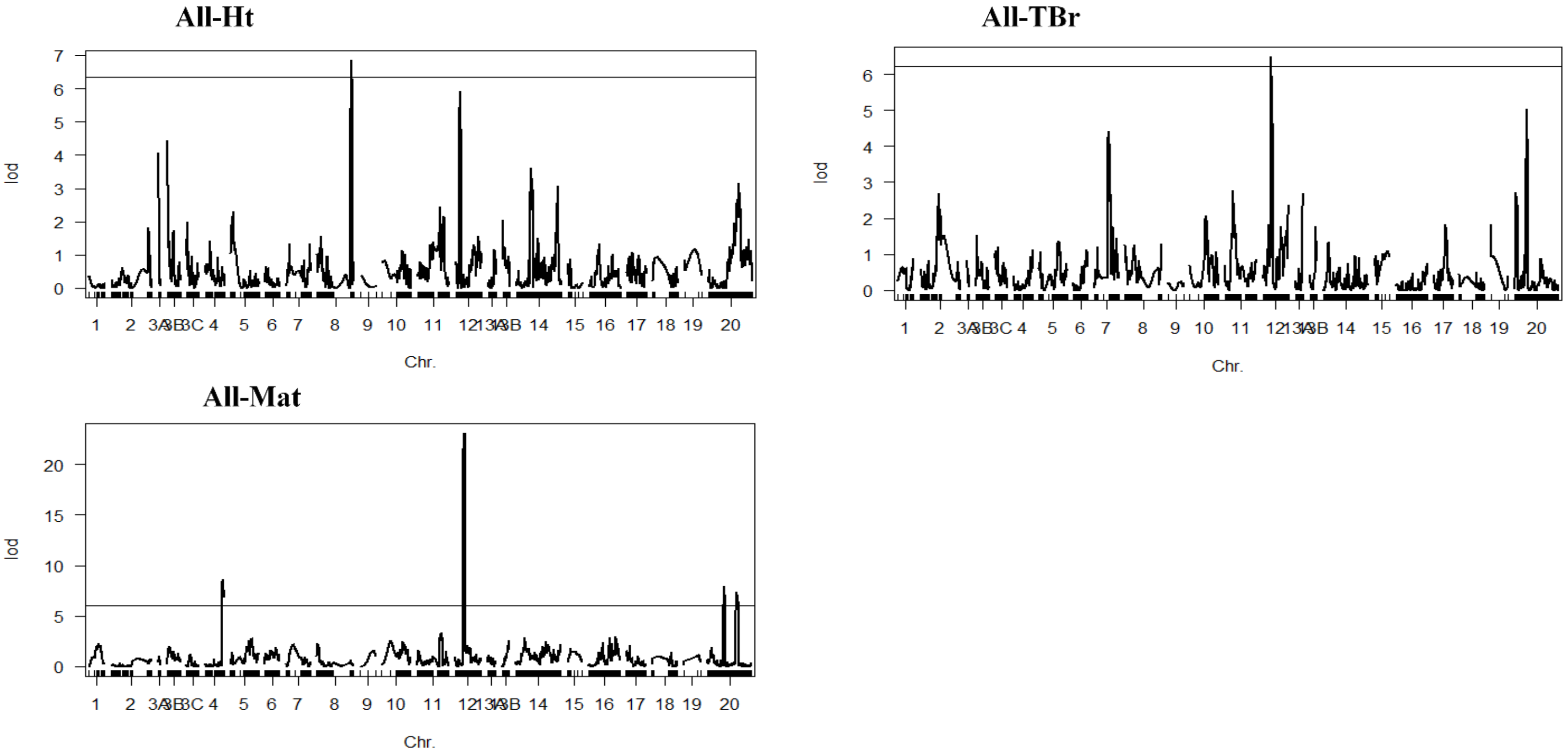


Figure 4-6. Genetic mapping of genes controlling height, total branch length, and maturity of the RILs from the cross Osage × PI593983 across four studied environments in 2016 and 2017 in Missouri. The threshold for significant QTLs is indicated by the horizontal line.

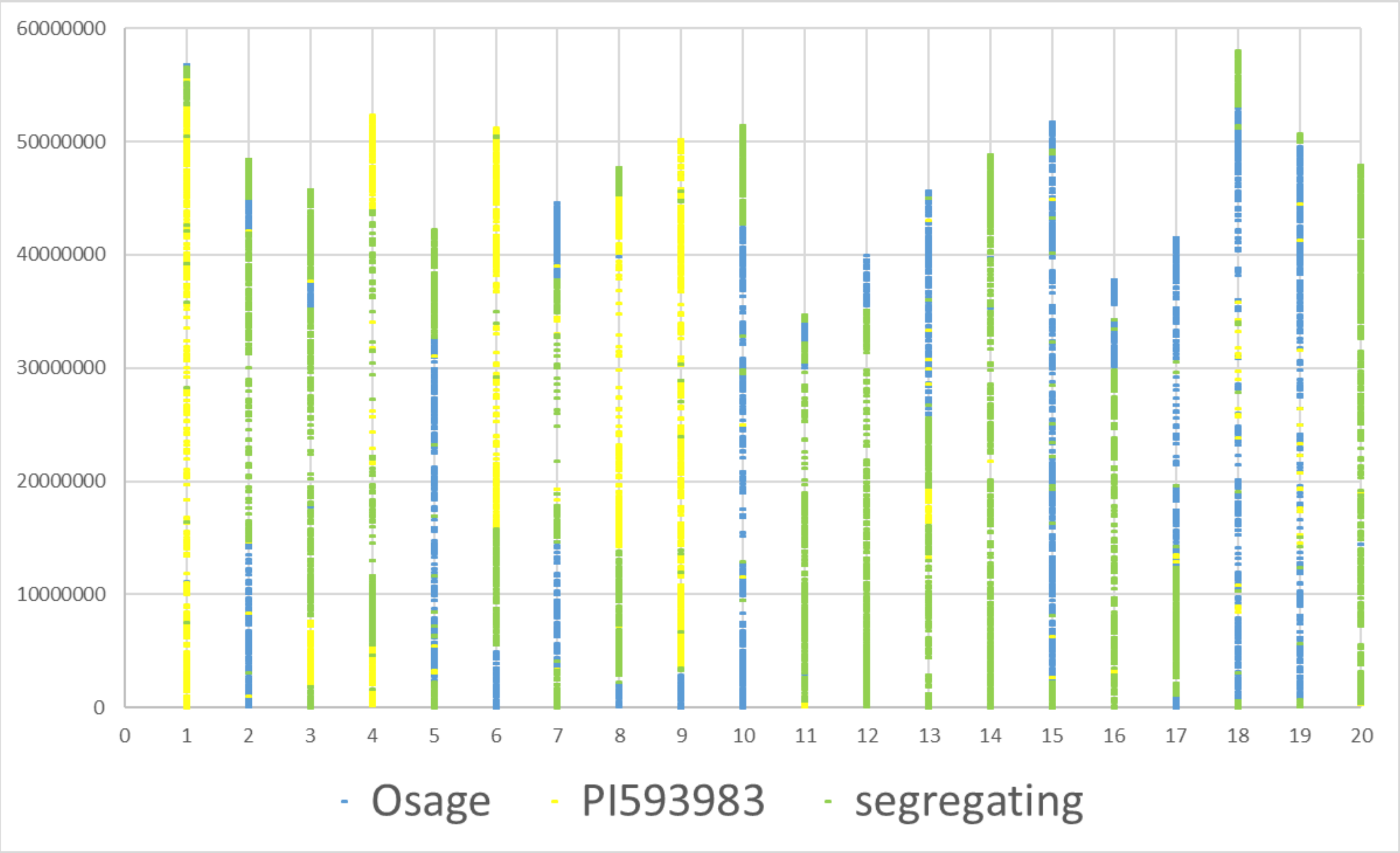


Figure 4-7. Distribution of genotyping-by-sequencing derived Osage (blue), PI593983 (yellow), and segregating (green) regions on 20 chromosomes of the physical map. The marker was determined to be from Osage or PI593983 if the frequency of major allele was >0.70. And the marker was determined to be segregating if the frequency of major allele was >0.25 and <0.70. Physical locations of markers were shown on the vertical axis in bp.

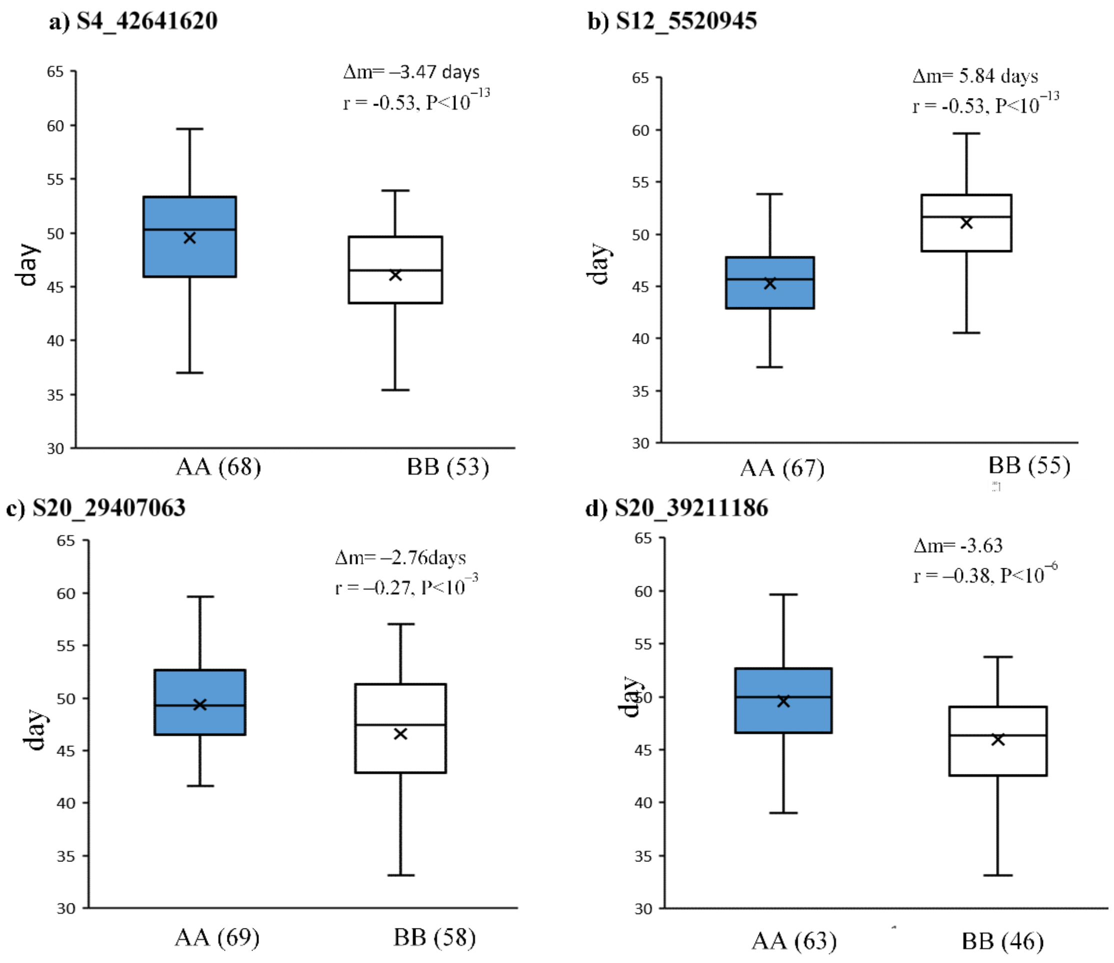


Figure 4-8. Phenotypic differences between lines carrying different homozygous alleles of (a) S4_42641620, (b) S12_6048084, (c) S20_29407063, and (d) S20_39211186 associated with maturity. The boxplots show the differences of maturity contents across all studied environments between RILs with different homozygous alleles at the detected SNP locus (AA=Osage's homozygous alleles, BB= PI593983's homozygous alleles). The boxes show the mean (presented as ×), first and third quartiles, and Median. The numbers in the parenthesis are the numbers of RILs for each allele. The given Δm, r, and P are the difference in mean tested by the student's t-test, the Pearson correlation coefficient between genotypic and phenotypic data, and the P value of correlation, respectively.

Chapter V: OVERALL SUMMARY AND FUTURE PROSPECTS

The evaluation and characterization of the core collection of wild soybean which genetically represents the entire USDA *G. soja* collection in this study will be valuable for soybean breeders to address the problem of low genetic diversity in American soybean cultivars. The results provide valuable resources for breeding materials, especially the wild soybean accessions with high contain of sulfur-containing amino acids which are deficient in soybean seed protein. The association mapping of maturity, seed weight, and seed compositions identified significant associations between markers and the studied traits. Further studies to identify the candidate genes associated with studied traits are necessary to understand the genetic mechanisms of these traits in soybean. In addition, Kim et al. (2010) suggested the re-sequencing followed by a comparative genomic analysis of wild soybean with reference to the genome of cultivated soybean would lead to the identification of a wide range of variation in nucleotide and structure between soybean cultivars and wild soybean. (Qiu et al., 2013) stated that the identified variation could be used to develop markers which would be used for introgression in soybean cultivars. Our study also emphasize the fact that breeding efforts to increase protein content alone could lead to the decrease in content of essential amino acids.

By carrying out QTL mapping for seed protein and oil in the RIL population from a single F_2 plant developed from the cross between Osage and PI593983, Oil20.1 and Pro20.4, which were located on chromosome 20 and had reversed effects on protein and oil content, emphasized the importance of cqSeed oil-004 and cqSeed protein-003 in controlling the content of seed oil and protein (Diers et al., 1992; Nichols et al., 2006; Warrington et al., 2015). While the content of protein and oil are difficult to be simultaneously increased, the identification of the high protein allele from PI593983 could be a promising candidate gene for use in breeding programs to develop soybean cultivars with high protein and without negative effect on oil content. However, the effect of the high protein allele (Pro14.4) from PI593983 on yield needs to be tested by the introgression of this gene into the genetic backgrounds of soybean cultivars and field experiments in different environments before it can be applied in breeding programs. In addition, the fine mapping of the chromosome 14 QTL (Pro14.4) would narrow the confidence interval of the QTL, assist gene cloning efforts, and provide marker information about markers that could be used to carry out marker-assisted selection.

References

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.

Kim, K.H., K.H. Kim, K.D. Kim, D.H. Kim, D.S. Kim, T.H. Kim, et al. 2010. Whole-genome sequencing and intensive analysis of the undomesticated soybean (*Glycine soja* Sieb. and Zucc.) genome. Proc. Natl. Acad. Sci. U.S.A. 107: 22032-22037.

Nichols, D.M., K.D. Glover, S.R. Carlson, J.E. Specht, and B.W. Diers. 2006. Fine mapping of a seed protein QTL on soybean linkage group I and its correlated effects on agronomic traits. Crop Sci. 46: 834-839.

Qiu, L.-J., L.-L. Xing, Y. Guo, J. Wang, S.A. Jackson, and R.-Z. Chang. 2013. A platform for soybean molecular breeding: the utilization of core collections for food security. Plant Mol Biol 83: 41-50.

Warrington, C.V., H. Abdel-Haleem, D.L. Hyten, P.B. Cregan, J.H. Orf, A.S. Killam, et al. 2015. QTL for seed protein and amino acids in the Benning × Danbaekkong soybean population. Theor. Appl. Genet. 128: 839-850.

VITA

Cao Thang La was born April 04, 1985 in Cantho, Vietnam. He graduated from Ly Tu Trong High School in 2003 and study biotechnology at Cantho University. He earned his bachelors of science in biotechnology in 2007. In 2010, he was awarded a Vietnamese Government Scholarship to pursue his M.S. degree in plant science at the University of Missouri, Missouri. After receiving the M.S. degree, he started his PhD at the University of Missouri, Missouri in the Division of Plant Science under the advisement of Dr. Andrew Scaboo.