January 2016

NESTED ASSOCIATION MAPPING TO IDENTIFY SEED COMPOSITION QTL IN DIVERSE SOYBEAN LINES

Mohammad Wali Salari
Purdue University

PURDUE UNIVERSITY
GRADUATE SCHOOL
Thesis/Dissertation Acceptance

This is to certify that the thesis/dissertation prepared

By Mohammad Wali Salari

Entitled

NESTED ASSOCIATION MAPPING TO IDENTIFY SEED COMPOSITION QTL IN DIVERSE SOYBEAN LINES

For the degree of Doctor of Philosophy

Is approved by the final examining committee:

Katy Martin Rainey
Chair

Mitch Tuinstra

Shaun N Casteel

Kevin T McNamara

To the best of my knowledge and as understood by the student in the Thesis/Dissertation Agreement, Publication Delay, and Certification Disclaimer (Graduate School Form 32), this thesis/dissertation adheres to the provisions of Purdue University's "Policy of Integrity in Research" and the use of copyright material.

Approved by Major Professor(s): Katy Martin Rainey

Approved by: Joe M Anderson 7/18/2016

Head of the Departmental Graduate Program Date

NESTED ASSOCIATION MAPPING TO IDENTIFY SEED COMPOSITION QTL IN DIVERSE SOYBEAN LINES

A Dissertation

Submitted to the Faculty

of

Purdue University

by

Mohammad Wali Salari

In Partial Fulfillment of the

Requirements for the Degree

of

Doctor of Philosophy

August 2016

Purdue University

West Lafayette, Indiana

I would like to dedicate this thesis to my parents who have always encouraged me to pursue my higher education.

ACKNOWLEDGEMENTS

I would like to thank my advisor, Dr. Katy Martin Rainey, for her guidance during all stages of my study. I am grateful for all her patience, advice, discussions and, giving me the opportunity to earn my Ph.D. degree with the Soybean Breeding Program. I am thankful to my advisory committee members: Dr. Mitch Tuinstra, Dr. Shaun Casteel and Dr. Kevin McNamara for their encouragement, advisory discussion, and help. I am grateful to USAID program for providing financial support for my Ph.D. program through the Strengthening Afghan Agriculture Faculties (SAAF) project. I would also like to give my deepest thanks to my close family members, my parents, my brothers, and sisters, who encouraged me to follow my Ph.D. program while away from them.

Finally, I would like to thank all members of Dr. Rainey's research program who helped me with various aspects of my research.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

GWAS	Genome Wide Association Study
LD	Linkage Disequilibrium
SNP	Single Nucleotide Polymorphism
QTL	Quantitative Trait Loci
RFO	Raffinose Family Oligosaccharides
RSO	Raffinose Series Oligosaccharides
UDP	Uridine Diphosphate
SAS	Statistical Analysis System
IA	Iowa
IN	Indiana
NE	Nebraska
IL	Illinois
Std	Standard deviation
AEA	Average Environment Axis (AEA)
AEC	Average Environment Coordinate
PCA	Principal Component Analysis
GGE	Genotype Plus Genotype by Environment Interaction

GEI	Genotype by Environment Interaction
NIRS	Near-Infrared Spectroscopy

ABSTRACT

Salari, Mohammad Wali . Ph.D., Purdue University, August 2016. Nested Association Mapping to Identify Seed Composition QTL in Diverse Soybean Lines: Major Professor: Katy Martin Rainey.

Soybeans are economically the most important legume grown worldwide. It provides quality protein and oil to food and feed markets in addition to being used for industrial products. The value of soybean could be enhanced by increasing protein, oil, and sucrose contents, while lowering anti-nutritional compounds such as oligosaccharides. Understanding the genetic and environmental factors controlling soybean seed composition is an essential prerequisite for such an endeavor. Three separate studies were initiated to understand the underlying genetics governing soybean seed compositional traits.

The first study was conducted to identify Quantitative Trait Loci (QTL) controlling seed protein and oil contents in the SoyNAM multi-parent population through a Genome-Wide Association Study (GWAS). The SoyNAM population was created by generating recombinant inbred lines from crossing the hub parent IA3023 to forty other parents representing elite public germplasm. Over 40,000 seed samples from 5486 recombinant inbred lines were evaluated in eight environments for seed protein and oil concentrations using NIR spectroscopy. Using GWAS, we identified thirteen QTL highly associated with seed protein content distributed over nine different chromosomes

and marked by 49 SNPs. Twenty-two out of 49 SNPs were located within the 39.6-40.2 Mbp region of chromosome 9, a region previously reported to be associated with seed protein content. We refined the seed protein QTL region to 0.56 Mbp compared to a previously reported 5-8 Mbp. Of the thirteen seed protein QTL, six were novel and were located on chromosomes 11, 13, 14, 15, and 18. GWAS also identified twelve QTL significantly associated with seed oil content on eight different chromosomes tagged by 109 SNPs. Six of the twelve seed oil QTL were new and were situated on chromosomes 2, 11, 15, 18, and 20. The QTL detected for protein and oil explained 15% and 23% of the phenotypic variations, respectively.

The second study was performed to identify quantitative trait loci (QTL) controlling seed sucrose, raffinose, and stachyose content in a set of 140 SoyNAM recombinant inbred lines (RILs), developed from the cross of two elite soybeans lines IA3023 and LD02-4485. Composite interval mapping (CIM) identified three QTL for sucrose content: one on chromosome 1 and two on chromosome 3. The QTL on chromosome 1 explained 10% of the phenotypic variation while the two QTL on chromosome 3 each explained 22% phenotypic variation in the sucrose content. The CIM also displayed a QTL for raffinose content on chromosome 6 and it explained 6% of phenotypic variation. This study identified novel QTL that can be validated for use in developing soybean lines with higher concentrations of sucrose and reduced levels of raffinose and stachyose.

The last study focused on Multi-Environment Trial (MET) analyses for both seed protein and oil contents. The result from the GGE-biplot analyses revealed that selection based on mean and stability was appropriate for the SoyNAM parental genotypes. The

most stable and desirable genotypes for seed protein content were LG92-1255, CL0J173-6-8, PI398881, PI561370, Prohio, PI427136, LG03-3191, PI507681B and genotypes LG03-2979, U03-100612, Prohio, LD02-4485, IA3023, LG04-4717, LG92-1255 were most desirable for seed oil content. LG94-1128 and 5M20-2-5-2 for seed protein content and NE3001 and LG05-4317 for seed oil content were unstable even though high yielding.

CHAPTER 1. GENERAL INTRODUCTION

1.1 Introduction

Soybean [*Glycine max* (L.) Merrill] is one of the oldest cultivated crops. Economically and agriculturally, it is the most important legume in the world, providing quality protein, and oil to the food and animal feed industry (Hedley 2000; Clevinger 2006). It is the second most important economic crops in the United States. It ranks third in grain production after corn (*Zea mays*) and wheat (*Triticum aestivum*), and second to corn in value (Dierking 2009). In the crop year 2015, the United States produced approximately 3.94 billion bushels of soybeans (U.S. Department of Agriculture, 2015). About 10% of the total produced soybeans were used directly for human consumption. Protein, oil, and carbohydrates of soybean seed are the most important determinant of soybeans end use. Soy protein ingredients have been gaining popularity because research showed that soy protein has health benefits. Research also found that soy oil is important for an animal meal since it produces high energy due to the presence of quality fatty acids (Dierking 2009). Due to these health benefits, soybean meal has become the most important ingredient of both humans and animal diet. The nutritional value of this quality soy meal is determined by carbohydrates components such as sucrose and Raffinose Family Oligosaccharides (RFOs). Among them, sucrose content of soybean seed is critical in soyfood industry because it adds sweetness and is easily digested by

monogastric animals. In contrast, the raffinose family oligosaccharides, which include raffinose and stachyose, are the non-desired carbohydrates because the monogastric animal cannot digest them. Considering its economic importance, soybeans has received a high priority and attention has focused on the improvement of the nutritional quality of soybeans' protein, oil and carbohydrates through genetics and as results hundreds of cultivars have been developed. These desired soybeans cultivars have been developed through research programs in which the researchers have conducted a number of genetic studies to identify and map QTL that control these traits using different mapping methods such as QTL mapping and GWAS. The research presented in this dissertation also focuses on identifying genomic regions involved in controlling protein, oil, and carbohydrates contents of soybeans seeds using both GWAS and QTL mapping methods.

1.2 Organization of the Dissertation

This dissertation is organized into five chapters. The current chapter presents a general introduction to the chapters that follow and provides an outline for the organization of the dissertation. The second chapter provides background information about soybean seed composition and the different statistical procedures that can be used to map genes/QTL controlling traits of interest. The third chapter includes genotypic and phenotypic analyses for protein and oil content of soybeans using the SoyNAM multi-parent mapping population. The RILs in this study were evaluated for protein and oil concentration using the Nested Association Mapping (NAM) technique. This technique was designed by Edward Buckler labs at Cornell University to identify and dissect the genetic architecture of complex traits in Maize. This technique combines the advantages

of high resolution from association mapping and high power to detect broad chromosome region from linkage analysis in a single unified mapping method. This method promises to identify numerous QTL that control yield and seed composition traits. NAM approach is expected to show high power to detect QTL in genome-wide association mapping approach (Holland 2007; Buckler, Holland et al. 2009; Stich 2009) and has been successfully used for genetic dissection of many complex traits in Maize (Wilson, Whitt et al. 2004; Holland 2007; Salvi, Corneti et al. 2011; Cook, McMullen et al. 2012; Meade 2012; Prado, López et al. 2014; Xiao, Tong et al. 2015; Zhang, Wu et al. 2015). The genotypic analysis of this chapter was based on Genome-Wide Association Study (GWAS). The fourth chapter contains bi-parental QTL analysis for sucrose and the Raffinose Family Oligosaccharides (RFOs) that play a key role in determining the nutritional value of soybeans in the markets. The QTL analyses for sucrose and the RFOs were conducted with QTL cartographer using the composite interval mapping method. The fifth chapter includes genotype by environment interaction analysis for protein and oil content using the parents that were used to create the Soybean Nested Association Mapping (SoyNAM) populations.

CHAPTER 2. LITERATURE REVIEW

2.1 Introduction

Soybean [*Glycine max* (L.) Merrill, 2n=40] is an important leguminous seed crop that has been grown across the world for its exceptional health and industrial benefits (Singh and Hymowitz 1999; Ghosh, Ghosh et al. 2014). It is an annual legume which typically grows 12 to 36 inches tall with dense or fewer branches depending on cultivar and growing conditions (Panthee, Pantalone et al. 2005). Soybean has a taproot system, a central root system from which other roots sprout laterally. The first root nodule appears 8-10 days after planting, depending on cultivar and growing condition (Carlson and Lersten 2004). The nodule formation, which supplies soybean plant with nitrogen, continues throughout the plant's growth stages (Panthee, Pantalone et al. 2005).

Soybean growth is divided into two stages, vegetative and reproductive. The vegetative stage begins with emergence followed by the development of four different types of leaves. The first pair of leaves is simple cotyledons which are also called seed leaves. The second pair of leaves is primary leaves. The third is called trifoliate foliage leaves and the fourth are prophylls (Carlson and Lersten 2004). The reproductive stage starts when axillary buds develop into flowers in clusters of 2 to 35 depending upon cultivar and environmental conditions such as daylength and temperature

(Carlson and Lersten 2004). Several studies reported that soybeans produce more flowers than they can develop into pods. Research reported that from 20 to 80% flowers abscise for many cultivars (Carlson and Lersten 2004). Soybean flowers are papilionaceous, white or pale purple, with a tubular calyx of five unequal sepal lobes and a five parted corolla consisting of a posterior banner petal, two lateral wing petals, and two anterior keel petals (Panthee, Pantalone et al. 2005).

Soybeans have two distinct growth habits, determinant, and indeterminate. Soybeans with a determinant growth habit stop vegetative growth on the main stem soon after flowering begins while indeterminate soybeans continue producing nodes on the main stem and branches until the start of seed filling stage (Pedersen and Elbert 2004).

It is believed that soybeans have been originated from China and its domestication began in northeastern part of China in the 11th century (Hymowitz 1990; Shurtleff and Aoyagi 2010; Dwevedi and Kayastha 2011). In the early period of domestication, soybeans were not as important part of Chinese diet as it is now (Dwevedi and Kayastha 2011). They were grown primarily for fertilizers purposes, plowing them back into the soil to make it enrich for the production of other crops such as wheat and millet. Soon it became the foundation of some Asian cuisine. Soybeans were first introduced to Europe in 1712 by Englebert Kaempfer, a botanist who lived in Japan (Hymowitz 1990). Soybeans were brought to the US from China by Samuel Bowen, who worked for East India company seaman (Hymowitz 1990). In 1896, a dramatic development happened for soybean in America when a well-known American chemist, George Washington Carver, became head of the department of agriculture at Tuskegee institute in Alabama. Mr. Carver encouraged farmers to rotate their crops with soybeans and other nitrogen-fixing

legumes that would replenish the depleted soil with nitrogen and minerals (SoyStats 2012). The first scientific study of the soybean in the west was conducted by Swedish botanist Carl von Linne and named it *Glycine max* because of the unusually large nitrogen-producing nodules on its roots. Unfortunately, soybean production in the west was limited due to adverse climatic conditions (Hymowitz 1990). These days, soybean is one of the most important legume crops in research due to providing quality protein for food, livestock feed, edible oil and addition to being used for a variety of industrial products. Because of it multipurpose end use and commercial interest, attention have been paid to the improvement in genetic, agricultural engineering, pest management, agronomic practices, which lead to a drastic increase in area under soybean production across the globe. Today, the United States is the leading world soybeans commercial producer followed by Brazil, Argentina, China, India, Paraguay, and Canada (Baize 2013) Figure 2.1.

http://www.statista.com/statistics/267270/production-of-soybeans-by-countries

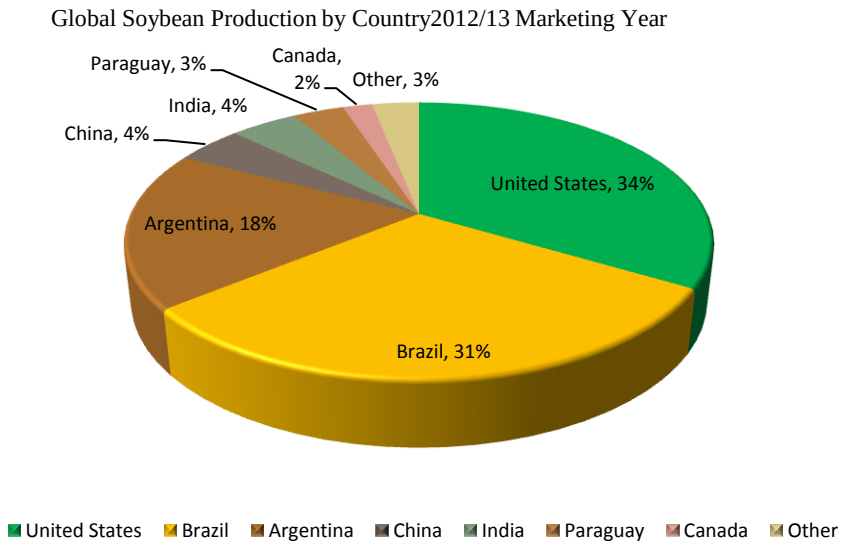


Figure 2.1. Global soybeans production by country.

2.2 Soybean Seed Composition

On average soybean seed composition comprise approximately 40% protein, 20% oil, 10% ash and other and 30% carbohydrate (Hou, Chen et al. 2008) of which about 15% is soluble carbohydrate (Figure 2.2). Soybean seed carbohydrates are divided into two main groups based on their physical and chemical properties. The first group is nonstructural carbohydrates which include oligosaccharides and polysaccharides while the second group contains structural polysaccharides that comprise dietary fiber components (Middelbos and Fahey Jr 2008; Murphy 2008). The dietary fiber consists of cell wall polysaccharide, noncellulose and structural polysaccharides such as lignin and phenolic compound (Middelbos and Fahey Jr 2008; Murphy 2008). The first group carbohydrates are also called soluble while the second group is insoluble carbohydrates.

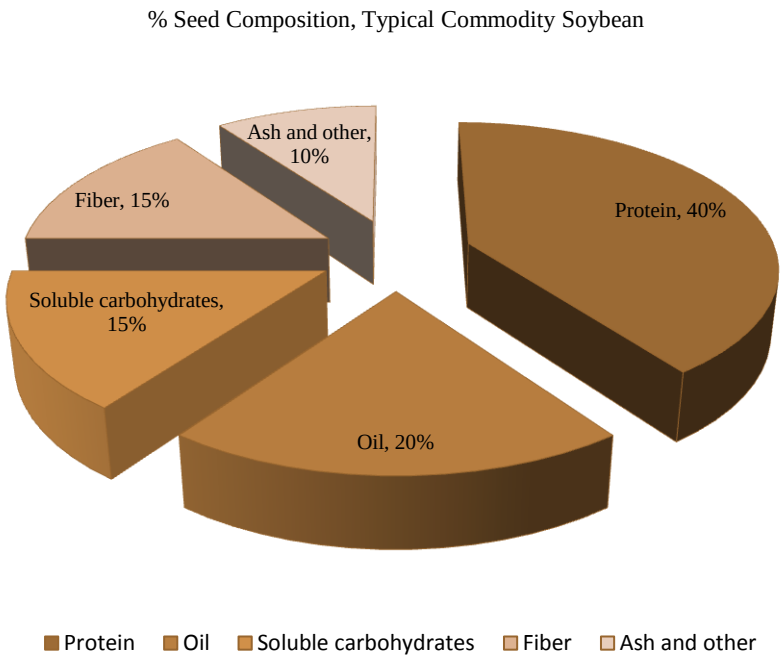


Figure 2.2. Seed composition of typical North American commodity soybean.

Soluble carbohydrates of a typical soybean seed include five major sugars such as glucose, fructose, sucrose, raffinose, and stachyose (Hou, Chen et al. 2009). Among them, the major sugars are sucrose, raffinose, and stachyose. The amount of sugars in soybean seed differs depending on soybean cultivar and growing conditions. The amount of sucrose in soybean seeds ranges from 41.3-67.5% while raffinose and stachyose make up 5.2-15.8%, 12.1-35.2%, of the total soluble sugar in soybean seed, respectively (Hou, Chen et al. 2008). Reports have confirmed trace amount of other sugars such a pinitol, myo-inositol, verbascose, galactose, in soybean seeds (Hou, Chen et al. 2008). Results from researches conducted on soybean seed carbohydrates found that sucrose, raffinose, and stachyose are important for viability and germinability of soybean seed (Middelbos and Fahey Jr 2008; Murphy 2008).

2.3 Nutritional Value of Soybeans

2.4 Soy Protein

Soybeans is commonly consumed by humans in the form of soymilk, soy protein, tofu, infant formula, miso, natto, soy flour and soy sauce (Stats 2001). They are a popular protein-rich food source in most Asian countries (Latham 1997). In the US soybean is used as feed for livestock and rarely as food for human consumption. Soybeans have been extensively used as major ingredients of non-ruminant diets throughout the world due to their high-quality protein content. Worldwide, approximately 85% of soybeans produce have been processed into soyfood. Soybean is considered an excellent source of food because it contains nine essential amino acids for humans and animal nutrition (Kwon 2009). Soybean is also known to be an excellent source of dietary fiber, and is

rich in micronutrients such as iron (Fe), zinc (Zn), and calcium (Ca) of which Ca is known to be beneficial for bone health (Messina 1999). Soybean is a good source of food for vegetarians, and a perfect protein source for children born to low-income families who often suffer from malnutrition (Kwon 2009). Soybeans seed contains a considerable amount of linolenic acid, omega-6 fatty acid, and isoflavones. Isoflavones have been implicated to play a key role in reducing diseases among humans such as a breast cancer-reducing factor (Messina 1999). It has also been known to reduce risk of developing other kinds of disease such as cervical, ovarian, lung and colon, more interestingly it is lowering a cholesterol level which reduces risk of heart related diseases (Coward, Barnes et al. 1993; Kennedy and Szuhaj 1994; Kennedy 1995; Kwon 2009).

Although soybeans have been grown mainly for protein and oil components for humans consumptions, its byproduct has been an important source of protein-rich feed for livestock, mostly for poultry and swine (Keshun 1997; Fageria, Baligar et al. 2011). Approximately 85% of the soybean produced worldwide is processed into soybean meal. Almost 98% of the soybean meal is further processed into animal feed (Hou, Chen et al. 2009; Choung 2010; Zeng, Chen et al. 2014). In addition to being used as a source of food and feed for both humans and animals, soybeans can be used for biodiesel production and are considered to be one of the most potential crops for bioenergy industry (Stats 2001).

In addition to being used as a source of food and feed for humans and animals, soy proteins with distinctive properties play important roles in plant biological function such as seed germination (Murphy 2008). Soybean seed protein exists largely in the form of storage proteins such as glycinin (primarily 11S) and β-conglycinin (primarily 7S), and

their function is to provide germinating seed with nitrogen in the form of amino nitrogen (Murphy 2008). Based on their solubility, soy protein can be classified into water soluble albumins and salt soluble globulins (Nazareth 2009). The relative proportions of these storage proteins in soybean seed depend upon genotype and the environmental conditions in which they are grown (Nazareth 2009).

Even though soybean has been considered an excellent source of food and feed for both humans and animals, it's nutritional value to monogastric animals is not optimized due to the presence of numerous naturally occurring compounds such as raffinose and stachyose which interfere with nutrient digestion and absorption (Clarke and Wiseman 2000). Raffinose and stachyose are not nutritionally available to monogastric animals because unlike in ruminants, these oligosaccharides are not hydrolyzed in the upper gut due to the absence of the α-galactosidase enzyme. In the lower intestine, the RFO's are metabolized by bacterial action leading to the production of gasses like methane, hydrogen and carbon dioxide that cause discomfort and in many cases flatulence and diarrhea. Therefore, development of soybean lines with reduced stachyose and raffinose content would improve digestibility and hence, supplying a more efficient feed source for non-ruminant.

2.5 Soy Oil

The major economic products of soybeans are protein and oil (Piper and Boote 1999; Singh and Hymowitz 1999). Soybeans have long been recognized as world's major source of edible oil for humans (Dei 2011). It represents a huge part of the vegetable oil in the market, accounting for approximately 57% of edible oil consumption globally

(SoyStats 2012). It has been used as an oilseed crop for edible oil production, which represents a large proportion of the vegetable oil in the market (SoyStats 2012). Soybean oil production has increased from 32% in mid-1960 to 56% in 2011 (SoyStats 2012). In 2010, only the United States produced over 19 billion pounds of oil and in the same year, soybean oil accounted for 68% of the U.S. edible fats and oil consumption (SoyStats 2012). Soybean oil is a useful source of feed-grade fat for animals. It has been used as high energy diets for modern breeds particularly for poultry because of its high metabolisable energy content compared with other vegetable oils (Dei 2011). Soy oil produces high energy mainly due to the high percentage of unsaturated fatty acids, which are well absorbed by the animals. The oil quality of soybean depends on its fatty acid composition that plays an important role in nutritional value, flavor and stability of the soybeans oil (Akond, Liu et al. 2014). The five fatty acids include palmitic acid, stearic acid, oleic, linolic and lino-lenic acids of which lower palmitic acid content are desirable for edible oil (Moongkanna, Nakasathien et al. 2011; Akond, Liu et al. 2014).

The correlation between seed protein and oil content is known to be negative; therefore, an increase in seed protein tends to decrease oil concentration, attributable to both environment and genotypic variation (Piper and Boote 1999). It has also been noted that temperature changes during seed filling influences seed composition and maximum seed oil content occurs when the temperature is in the range of 25-29°C and decreases when the temperature increases. Conversely, as the temperature increases seed protein content increases (Dornbos Jr and Mullen 1992; Piper and Boote 1999). Environmental stress conditions such drought during soybean seed filling can change the soybean seed chemical composition and result in increase in seed oil content (Specht, Chase et al.

2001). Considering the economic importance of the soybean seed oil and protein content, researchers have attempted to increase both constituents; however, the strong negative correlation between these two traits has made it challenging to improve both traits simultaneously (Wilcox and Cavins 1995; Cober and D Voldeng 2000; Chung, Babka et al. 2003; Panthee, Pantalone et al. 2005; Phansak 2010).

2.6 Sucrose

Sucrose also called the common table sugar, is a disaccharide made of two monosaccharide: alpha-D-glucopyranose and beta-D-fructofuranose (Dey and Dixon 1985). The two monosaccharide, alpha-D-glucopyranose and beta-D-fructofuranose, are bound through a glycosidic bond between the Carbon-1 (alpha) of glucose and the Carbon-2 (beta) of fructose (Figure 2.3) (Dey and Dixon 1985).

Sucrose has been considered a critical quality trait in soy food production (Cicek 1997), and it is the most abundant disaccharides in legumes plants (Hedley 2001). Sucrose, the primary storage form of glucose, fructose, and carbon, plays an important role in developing soybean embryos by being transported to the seed from green parts of the plant during seed development (Dey and Dixon 1985; Lowell and Kuo 1989). Studies conducted on soybean sugar contents have found a positive correlation between sucrose and oil but a negative association of each with protein.

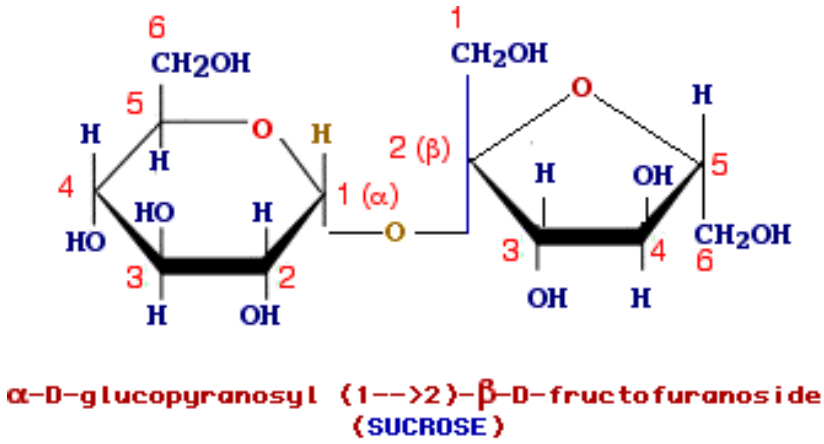


Figure 2.3. Chemical structure of sucrose.
http://wpage.unina.it/petrilli/organic/carbo.htm?html

Invertase enzyme causes sucrose to decrease rapidly after germination. It cleaves and digests sucrose to release glucose and fructose utilized in the creation of new cell in the growing embryo (Dey and Dixon 1985). Three enzymes, sucrose phosphate synthase, sucrose phosphatase and sucrose synthase are associated with sucrose synthesis in green plants (Clevinger 2006). The sucrose phosphate synthase enzyme made up of UDP-glucose and fructose 6-P, plays an important role in the regulatory control of sucrose synthesis (Dey and Dixon 1985; Clevinger 2006). Sucrose phosphatase hinders sucrose synthesis. The third enzyme, sucrose synthase, that synthesizes sucrose is abundant in higher plants and is found in nearly all plant tissues (Clevinger 2006). The most important function of sucrose synthase enzyme is the breakdown of sucrose into glucose and fructose (Dey and Dixon 1985). Sucrose synthase and alkaline invertase are the two enzymes that correspond to the accumulation of 90% total dry matter in soybean seed (Dey and Dixon 1985). Sucrose contributes to the soybeans derived food sweetness and making it a desirable food for humans and animals (Abe, Ujiie et al. 2004). High sucrose

soybean-derived food and feed is desired for monogastric animals because they can digest sucrose. This is possible because monogastric animals have an enzyme that breakdown sucrose into its bio-available components.

2.7 Raffinose Family Oligosaccharides (RFOs)

Raffinose family oligosaccharides (RFOs) are complex sugar compounds that are formed by adding D-galactose units to the D-glucose moiety of a sucrose molecule through α-(1, 6) bonds (Obendorf 1997; Tahir, Båga et al. 2012). Raffinose family oligosaccharides have been known by various names and acronyms such as RFO (Raffinose Family Oligosaccharides), RSO (Raffinose Series Oligosaccharides), and Raffinose Saccharides (Huhn 2003). RFOs include the trisaccharide raffinose, the tetrasaccharide stachyose, and the pentasaccharide verbascose (Figure 2.4.).

A typical soybean seed contains approximately 1% raffinose, and 3 to 4% stachyose (Skoneczka, Maroof et al. 2009). It is believed that the biosynthesis of raffinose family oligosaccharides in soybeans starts with initial reaction catalyzed by galactinol synthase to produce galactinol from UDP (uridine diphosphate galactose and myoinositol (Clarke and Wiseman 2000). High RFOs such as raffinose, and stachyose, are produced as a result of using galactinol to add galactosyl residues to sucrose. The RFOs biosynthesis steps are explained in the following chemical reactions and Figure 2.4.

UDP-galactose + myo-inositol

→galactinol + UDP (1)

Galactinol + sucrose

→ raffinose + myo-inositol (2)

Galactinol + raffinose

→stachyose + myo-inositol (3)

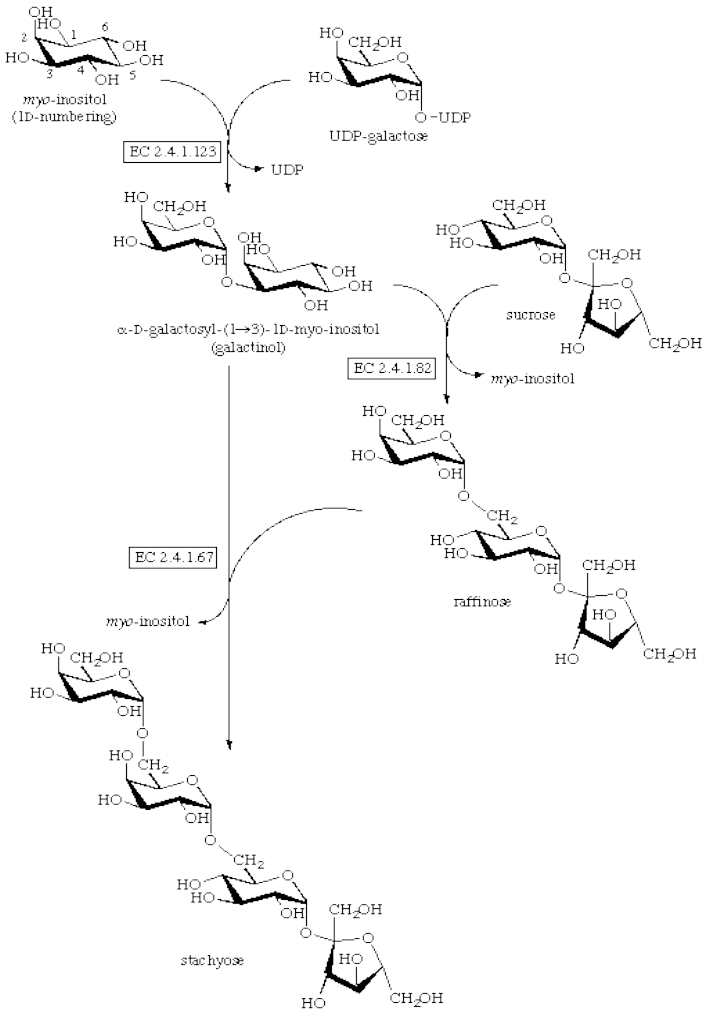


Figure 2.4. Sucrose, raffinose, and stachyose biosynthesis. http://users.bergen.org/dondew/bio/AnP/Anp1/AnP1Tri1/AnP1_Tri1_raffinose.htm.

Raffonse, [β -D-fructofuranosyl-O-α-D-galactopyranosyl- (1→6)-α-D-glucopyranoside], and stachyose ,[O-α-D-galactopyranosyl- (1→6)-O-α-Dgalactopyranosyl- (1→6)-α-D-glucopyranosyl-β-D-fructofuranoside], are the two RFOs that exist at relatively high level in grain legumes seed (Jones, DuPont et al. 1999). They have also

been detected in various parts of plants such as leaves, rhizomes, roots, seeds, stem, cotyledons, seed coats, and hypocotyls (Obendorf 1997; Bentsink, Alonso-Blanco et al. 2000). Raffinose family oligosaccharides perform a variety of function in plants (Karner, Peterbauer et al. 2004). They transport carbohydrates in the phloem, serve as storage reserves and cryoprotectants in frost-hardy plant organs (Sprenger and Keller 2000; Pennycooke, Jones et al. 2003). They are accumulated in maturing seed and play a key role in the acquisition of desiccation tolerance, storability and cold tolerance in many plant species (Horbowicz and Obendorf 1994; Pennycooke, Jones et al. 2003).

Increased demands for healthier food encouraged plant scientists to develop soybeans cultivars with higher nutritional value through research. Development of soybeans cultivars with low RFOs, high sucrose, high protein and high oil is the goal of most plant breeding programs. One possible way to develop such cultivars is to conduct genetic analysis (QTL mapping/GWAS) through which the researcher will be able to identify Quantitative Trait Loci (QTL) controlling these traits. The identified QTL then can be validated for use in developing soybean lines with higher concentration of protein, oil, sucrose and reduced levels of raffinose and stachyose.

2.8 Quantitative Trait Loci (QTL) Mapping

Mapping in terms of molecular genetics is the process in which genetic markers are arranged in order on chromosomes based on their relative genetic distance as determined by recombination frequency. The goal of genetic mapping is to identify the location of genomic regions controlling traits of interest (Bernardo 2002; Collard, Jahufer et al. 2005; Myles, Peiffer et al. 2009). The term of QTL was first used by Gelderman in 1975.

Conceptually, QTL can be a single gene or a cluster of genes that control the trait of interest. The genomic region that affects the trait of interest and the magnitude of its effect on the trait can be identified with the help of molecular markers (Bernardo 2002; Collard, Jahufer et al. 2005; Myles, Peiffer et al. 2009).

Genetic mapping is accomplished through two main approaches; (1) linkage mapping which is also called biparental mapping or family based QTL mapping , and (2) association mapping or linkage disequilibrium mapping (LD-mapping). Association mapping does not require the development of biparental mapping populations; instead it uses diverse lines from natural populations or germplasm (Abdurakhmonov and Abdukarimov 2008).

2.9 Linkage Mapping

Linkage mapping is the most common method for identifying the genetic basis of quantitative traits in plants and a useful process to study the phenotypic variation that is due to changes in DNA sequence (Myles, Peiffer et al. 2009; Soto-Cerda and Cloutier 2012). Most plant geneticists and breeders try to explain the phenotypic variation in plants in term of changes in DNA sequence (Myles, Peiffer et al. 2009; Soto-Cerda and Cloutier 2012). Family-based QTL mapping makes use of well-characterized pedigrees structure in which the mapping population is generated from the cross of individuals with known relatedness (Kloth, Thoen et al. 2012). The cross from which the mapping population is generated is called biparental cross (Kloth, Thoen et al. 2012). Abdurakhmonov et al. (2008) and Semagn, et al. (2010) provided a detailed review of the procedure and the mapping populations needed for family based QTL mapping. First of

all the researcher needs to develop experimental populations such as F_2, backcross (BC), double haploid (DH), recombinant inbred line (RILs), and near-isogenic line (NIL) that are derived from hybridization of two parental genotypes carrying trait of interest (Abdurakhmonov and Abdukarimov 2008; Semagn, Bjørnstad et al. 2010). Second, the progeny of the large experimental populations is measured for segregation of the trait of interest in different environments. Third, a number of polymorphic DNA markers, that distinguish parental genotypes from segregating genotypes in a mapping population, is identified and then the parental genotypes are screened with these markers (Abdurakhmonov and Abdukarimov 2008; Semagn, Bjørnstad et al. 2010). If the markers are identified polymorphic over the parental genotypes then all individuals of the mapping population are genotyped with these markers (Abdurakhmonov and Abdukarimov 2008; Semagn, Bjørnstad et al. 2010). Once the genotypic data collected from the mapping population is ready, marker data can be used to construct linkage map by arranging genetic markers in order on the chromosome based on their relative genetic distances between them (Abdurakhmonov and Abdukarimov 2008; Semagn, Bjørnstad et al. 2010). A linkage map is tabular or graphical depiction of marker positions on chromosomes within a linkage group. One major problem often encountered in constructing linkage map is interference. This occurs when two adjacent crossover events are not independent. This implies that the occurrence of one crossover event influences the other, making the detection of double crossover difficult. There are two commonly used map functions namely Kosambi and Haldane. Of the two map functions, the Kosambi accounts for the interference, making it the best mapping function for genetic map construction (Huehn 2011).

The construction of linkage map can be achieved through commonly used software program including Rqtl and Mapmaker (Lander, Green et al. 1987). As a result, the markers arranged along the linkage map are statistically correlated with the phenotype of individuals of a mapping population and QTL affecting the trait of interest along with the markers linked to that QTL are identified (Abdurakhmonov and Abdukarimov 2008).

Advantages of bi-parental populations mapping or family based QTL mapping is that it requires relatively fewer markers for genome coverage; no population structure; the ability to detect the effect of rare allele and high statistical power per allele (Sorrells and Yu 2009; Semagn, Bjørnstad et al. 2010). So far most of the plant QTL mapping studies have been conducted based on linkage or family based QTL mapping. This approach has some limitations. For example, occurrence of few recombination events within family, poor resolution in detecting rare QTL, only two alleles per locus can be studied simultaneously, and it requires evenly distributed markers at spacing of 10-20cm due to limited number of recombination event occurred within family (Flint-Garcia, Thuillet et al. 2005; Sorrells and Yu 2009; Semagn, Bjørnstad et al. 2010). This method is further limited by the cost associated with the longer time required to develop mapping population and evaluate a large number of genotypes (Holland 2007).

2.10 Association Mapping

The constraint posed by family based QTL mapping can be overcome with the use of population-based association study in which the gene-tagging efforts are turned from biparental crosses to natural population and from family based QTL mapping to linkage disequilibrium (Flint-Garcia, Thuillet et al. 2005; Abdurakhmonov and Abdukarimov

2008). In association mapping approach genotypic and phenotypic data are collected from a panel of mapping population in which the relatedness is not controlled by the researcher and correlation between marker and phenotype are sought within the population (Myles, Peiffer et al. 2009). Association mapping and linkage based disequilibrium association mapping are often used interchangeably in literature but they present slight differences. According to Gupta et al. (2005) association mapping refers to the significant association of a molecular marker with the phenotypic trait of interest while LD refers to a non-random association between two markers or two gene/QTL or between a QTL and a gene (Semagn, Bjørnstad et al. 2010). As a result, association mapping is one of the several applications of LD (Gupta, Rustgi et al. 2005). From statistical point of view, association refers to the covariance of the polymorphic marker and the trait of interest while LD represent the covariance of polymorphism expressed by two markers/QTL (Gupta, Rustgi et al. 2005). Association mapping is rapidly emerging as a new science being utilized as a tool to dissect complex trait in plant and offers a unique opportunity to seek complex trait variation to the sequence level by exploiting historical and recombination events at the population level (Zhu, Gore et al. 2008). This method has received special attention in the past several years because it can potentially identify a single polymorphism within a gene that causes the phenotypic differences. (Soto-Cerda and Cloutier 2012).

Abdurakhmonov and Abdukarimov (2010) have provided a general population-based mapping approach.

The overall approach for conducting association mapping in plant might be different due to different methodology chosen, but generally, it requires the following steps:

1. selection of a group of individuals or germplasm with wide coverage of genetic diversity;

1. genotyping the mapping population with available markers;

2. quantifying the extent of LD of a chosen population genome using a molecular marker data;

3. evaluating the population structure and kinship (coefficient of relatedness between pairs of each individual within a sample);

4. correlating phenotypic and genotypic/haplotypic data using an appropriate statistical approach that discloses genomic region (marker tags) positioned within close proximity of targeted trait of interest. As a result, a specific gene(s) controlling a QTL of interest can be cloned using the marker tags and annotated for an exact biological function (Abdurakhmonov and Abdukarimov 2008).

2.11 Type of Association Mapping

Association mapping broadly falls into two categories; Genome-Wide Association mapping (GWA) and Candidate Gene Association mapping (CGA) (Zhu, Gore et al. 2008). In the candidate gene association mapping approach, few genetic markers that are believed to be involved in controlling the trait of interest are genotyped and correlated with the phenotype (Zhu, Gore et al. 2008; Myles, Peiffer et al. 2009). Candidate gene mapping approach was widely used for disease–gene association in humans but has been considered inadequate approach due to failing to detect most confirmed disease genes (Risch and Merikangas 1996; Myles, Peiffer et al. 2009). This approach may work in plants but only for candidate genes whose pathways are known as well as for genes

whose role is already known in controlling the phenotype of interest (Risch and Merikangas 1996; Myles, Peiffer et al. 2009).

Due to limitations in the choice of candidate genes that are identified, the candidate gene association mapping approach always runs the risk of missing underlying nucleotides that are located in non-identified candidate genes (Hall, Tegström et al. 2010). However, candidate gene mapping approach is thought to be statistically powerful because a small genomic region is saturated with dense markers, thereby increasing the mapping resolution (Kwon and Goate 2000).

In addition, candidate genes are mainly discovered from the loss-of-function mutations in inbred lab strains; therefore, it is not clear how well such mutations describe the variation that actually underlie quantitative trait variation in natural populations (Hall, Tegström et al. 2010). Identification of SNPs between and within lines are required for candidate gene association mapping because SNPs offer the highest resolution for mapping QTL and are potentially in LD with the causative polymorphism (Semagn, Bjørnstad et al. 2010).

Due to limitations associated with candidate gene association mapping approach, one can use Genome-Wide Association mapping approach (GWA) in which the entire genome is scanned for marker-trait association with a large number of markers. In the GWA approach, the entire genome is covered with markers and a sufficient number of markers are genotyped across the genome such that functional alleles will likely be with at least one of the genotyped markers (Myles, Peiffer et al. 2009). Scanning whole genome requires high capacity DNA instruments or high oligonucleotide (oligo) arrays to

efficiently identify SNPs at a density that accurately reflects genome-wide LD structure and haplotype diversity (Semagn, Bjørnstad et al. 2010).

The association mapping approach has been used for several crops to identify QTL controlling traits of interest. The continued decrease in sequencing and genotyping costs, GWA mapping is increasingly becoming more feasible and applicable (Semagn, Bjørnstad et al. 2010). Since GWA is less dependent on prior information about the candidate genes compared to QTL mapping and candidate association mapping, it is a promising method to identify novel loci involved in complex phenotypic traits (Kloth, Thoen et al. 2012). GWA mapping is not a replacement of traditional QTL mapping; in fact, these two methods of mapping have complementary advantages and disadvantages which lead to a better understanding of causal genetic polymorphisms when they both are combined (Kloth, Thoen et al. 2012).

2.12 Population Structure Issue Associated with Association Mapping

Until recently, plant breeders were skeptical about using the association mapping approach for mapping QTL underlying quantitative traits mainly due to the false associations as a result of the confounding effects from population structure. Population structure occurs when genetically different groups in the population under study are not mating at random for at least several generations. Random mating population may not exist except in population genetic theory (Myles, Peiffer et al. 2009). Nonrandom mating generates a complex pattern of population structure and relatedness in crops and wild plants which often led to a genome-wide LD between unlinked loci (Flint-Garcia, Thuillet et al. 2005; Myles, Peiffer et al. 2009; Sneller, Mather et al. 2009).

In association mapping, mapping a trait will be problematic due to the complex structure of genetic relatedness among individuals because many genetic markers across genome will emerge associated with the phenotype when actually the markers only capture the genetic relatedness (Myles, Peiffer et al. 2009). This could be a big problem with trying to map traits subjected to local adaptation such as flowering time because variation in these phenotypes between populations is highly correlated with allele frequency differences between populations (Aranzana, Kim et al. 2005; Flint-Garcia, Thuillet et al. 2005; Buckler, Holland et al. 2009; Myles, Peiffer et al. 2009). Even for a set of common traits of agronomic interest in maize, such allele frequency differences account for an average of 9.3% of the phenotypic variation across all traits (Flint-Garcia, Thuillet et al. 2005; Myles, Peiffer et al. 2009). Populations with complex structure may show significant differences in allele frequency which might be due to genetic drift; therefore, genetic loci identified will be falsely associated with the trait of interest when there is not a real association because the markers only tag genetic relatedness. The development of a statistical model which allows accounting for population structure during association analysis has improved the application of association mapping for QTL detection in crop plants. There are two steps to account for population structure using a model-based approach; the first is to calculate the percentage of the membership of each individual to population groups using unlinked random markers. The second is to use the percentage of membership as a covariate in the model of testing associations of markers with phenotypic traits (Ersoz, Yu et al. 2009). In the unified mixed model of Yu et al. (2006), both population structure (Q) and family relatedness (K) are simultaneously

considered as covariates in the model. This model accommodates both fixed and random effects.

2.13 Nested Association Mapping

Association mapping and linkage analysis are two approaches that have been often used to dissect the genetic architecture of complex traits (Center 1995; Risch and Merikangas 1996; Holland 2007). These two methods are complementary to each other such that linkage analysis identifies broad chromosomal region of interest with low markers coverage and has high power in detecting rare QTL while association mapping uses dense markers and offers high resolution either using the candidate gene approach or the genome-wide association mapping approach (Risch and Merikangas 1996; Holland 2007). An integrated mapping approach is necessary to combine the advantages of the two mapping methods to improve mapping resolution without requiring dense marker maps (Holland 2007).

To develop such method, Yu et al. (2008) proposed Nested Association Mapping (NAM) approach which combines the advantages of the two mapping strategies in a single unified mapping population. The NAM strategy dissects complex traits at a fundamental level through creating mapping resources that enable researchers to take advantages of genetic, genomic and system biology tools (Holland 2007). This method promises to identify numerous QTL that control yield and seed composition traits. The NAM approach uses recombinant inbred lines (RILs) population derived from several crosses of parental inbreds (Holland 2007). The genome of the RILs are mosaics of chromosomal segments of their parents mainly due to diminishing chances of

recombination over short genetic distance thereby within the chromosomal segments, the linkage disequilibrium (LD) information across the parental inbreds is maintained for a given number of generation. If the parental inbreds are diverse LD will decay rapidly within the chromosomal segments of the RILs.

The Nested Association Mapping approach allows utilization of both historic and recent recombination and provides high mapping resolution (Holland 2007). In addition, using the balanced design underlying the proposed mapping strategy and systematic reshuffling of the genomes of the parental inbreds during RIL development, NAM populations are expected to show a high power to detect QTL in genome-wide association mapping approach (Holland 2007; Buckler, Holland et al. 2009; Stich 2009). According to Yu et al. (2008) and Holland et al. (2007) development of Nested Association Mapping approach requires the following steps:

1. select diverse parents and cross them to an elite line of interest;
2. develop a large set of recombinant inbred lines;
3. either sequence completely or densely genotypes the parents;
4. genotype a smaller number of tagging markers on both the parents and the RILs to define the inheritance of chromosome segments and to project high-density marker information from the parents to the RILs;
5. phenotype RILs for complex traits, and
6. conducting genome-wide association analysis relating phenotypic traits with projected high-density markers of the RILs.

Nested association mapping has been successfully used for genetic dissection of many complex traits in Maize (Wilson, Whitt et al. 2004; Holland 2007; Salvi, Corneti et al. 2011; Cook, McMullen et al. 2012; Meade 2012; Prado, López et al. 2014).

CHAPTER 3. GENOME-WIDE ASSOCIATION STUDY OF SEED PROTEIN AND OIL CONTENT IN A SOYBEAN NESTED ASSOCIATION MAPPING POPULATION

3.1 Abstract

The objectives of this study were to determine genotypic differences in soybean for protein and oil concentration, and to identify Quantitative Trait Loci (QTL) controlling these two traits in SoyNAM mapping population. A total of 5486 genotype were evaluated for seed protein and oil concentrations in 4 locations; Indiana, Iowa, Illinois, and Nebraska, for two years (2012 and 2013) in a Modified Augmented Design (MAD). Protein and oil contents in soybean seeds were estimated using NIR spectroscopy Perten DA7200 diode array instrument. The Genotype, Location and interaction sources of variation were all highly significant for both protein and oil. Locations explained the highest proportion of variation in protein (38.17%) and oil (35.33%) contents, and this was followed by genotypes, which accounted for 33.88% and 29.35% variation in the two traits respectively. Heritability estimates on a line mean basis for protein and oil concentration were 0.85, and 0.84, respectively. The phenotypic correlation between these two traits was -0.61, indicating a negative association between the two traits. GWAS identified 13 QTL highly associated with seed protein contents distributed over 9 different chromosomes and marked by 49 SNPs. Twenty two out of 49 SNPs were located in the 39.6-40.2 Mbp region of chromosome 9, a region previously

reported to be associated with seed protein content. We further refined the seed protein QTL region to 0.56 Mbp compared to a previously reported 5-8Mbp. Of the 13 seed protein QTL 6 were novel and were located on chromosomes 11 , 13, 14, 15, and 18, and the rest were previously reported QTL. GWAS also identified 12 QTL on 8 different chromosomes tagged by 109 SNPs highly significant with seed oil content. Six of the 12 seed oil QTL were novel and were situated on chromosomes 2, 11, 15, 18, and 20, and the remaining 6 were known QTL. Among the QTL detected for oil content a highly significant QTL was detected on chromosome 10 that comprised more than 90 SNPs. The QTL detected for protein and oil explained 15% and 23% of the phenotypic variations, respectively. The markers closely linked to the novel QTL could be used for marker-assisted breeding of these two traits.

3.2 Introduction

Soybean [*Glycine max* (L) Merrill] is an important leguminous seed crop that has been grown across the globe for its high protein and oil concentrations. It is one of the world's largest sources of edible oil and protein for humans and livestock, respectively (Panthee, Pantalone et al. 2005). The protein content of soybean seed has been used for both livestock and human consumption. In the US, it has been used only as livestock feed while in some Asian countries, it has been used for both human and livestock consumption (Hymowitz and Newell 1981).

Seed protein and oil content in soybean are known to be polygenic traits and is quantitatively inherited, with large effects of genotype × environment interactions (Chung, Babka et al. 2003; Phansak 2010; Hu, Liu et al. 2011; Akond, Liu et al. 2014;

Hwang, Song et al. 2014). Improving protein and oil content of soybean therefore, requires elaborate evaluation of breeding populations in several environments. This multi-environment assessment allows for quantification of the magnitude of variances that are genetic in nature compared that due to environments. Understanding genotypic variation, heritability and the interactions between genotypes and environments is critical in deciding breeding strategies for quantitative traits like protein and oil.

Past studies reported significant genotype by environment interactions (GEI) for oil and protein (Sogut 2006; Phansak 2010). Heritability estimates for protein ranging from 0.57 to 0.91 and for oil ranging from 0.51 to 0.93 have also been reported for these traits (Lee, Bailey et al. 1996). These studies asserted that effective utilization of any newly developed breeding population for trait improvement requires dissection genetic variances across environments.

Complex traits are challenging to breed for conventionally, which explains the slow progress in improving seed protein and oil in soybean. With the recent advances in genomic approaches, molecular tools are frequently being applied to elucidate such traits. For protein and oil, researchers have found several genomic regions controlling seed protein and oil concentration (Diers, Keim et al. 1992; Lee, Bailey et al. 1996; Panthee, Pantalone et al. 2005). Qiu et. al (1999) found two restriction fragment length polymorphism markers(RLFP) linked to QTL controlling protein content and one marker associated with QTL controlling oil content in a population derived from a cross of Peking and Essex on linkage group H and F (Qiu, Arelli et al. 1999). Diers et al. (1992) evaluated F_2 population derived from a cross of *G. max* and *G. soja* and found three RLFP markers linked to QTL on linkage group E, F, and I controlling seed protein of

which two makers on linkage group E and I were identified to be consistent with QTL for protein. Mansur et al. (1993) analyzed F_2 derived population from a cross between, Minsoy,(PI 27.890) and ,Noir 1, (PI 290.136) for QTL controlling seed protein and oil content and found an unlinked RFLP marker, L48, associated with seed protein content. Csanádi et. al (2001) evaluated 82 individuals of an F_2 population derived from a cross between Ma. Belle and Proto for QTL controlling protein and oil content and found four markers linked to genomic region controlling seed protein and three markers associated with oil content located on linkage group A1, B2, L, and B1. These QTL are candidates for deployment in MAS but must first be validated in different population grown in diverse locations (Panthee, Pantalone et al. 2005). To date, over 140 QTL for each seed protein and oil have been reported in a number of studies (SoyBase, the USDA, ARS Soybean Genetics and Genomics Database) (Hwang, Song et al. 2014). These QTL have been identified on many different genomic regions throughout all 20 chromosomes, however, most of them are yet to be confirmed (Hyten, Pantalone et al. 2004; Hwang, Song et al. 2014). Additionally, the population used in many studies for identifying QTL for agronomic and seed composition traits share parents that are closely related, therefore, it is important to use new and diverse parent to develop mapping population and test these populations in different environments for stable QTL that control these traits.

The main objectives of this study were to: 1), to conduct genome wide association studies to identify quantitative trait loci (QTL), particularly rare QTL associated with seed protein and seed oil concentrations in the SoyNAM mapping population; 2), to determine genetic variation for seed protein and seed oil contents; 3) and to study the magnitude of GEI and stability for the two traits across environments.

Material and Methods

3.2.1 Plant Material, SoyNAM Structure, and Experimental Design

Soybean Nested Association Mapping (SoyNAM), a technique that combines the advantages of both linkage and association mapping, were used for this experiment. The mapping populations was developed by mating IA3023, a high yielding Iowa State variety, with 40 different high yielding elite and exotic soybean lines namely, *4J105-3-4, 5M20-2-5-2, CL0J095-4-6, CL0J173-6-8, HS6-3976, LD00-3309, LD01-5907, LD02-4485, LD02-9050, LG03-2979, LG03-3191, LG00-3372, LG04-4717, LG04-6000, LG05-4292, LG05-4317, LG05-4464, LG05-4832, LG90-2550, LG92-1255, LG94-1128, LG94-1906, LG97-7012, LG98-1605, Magellan, Maverick, NE3001, Prohio, S06-13640, Skylla, TN05-3027, U03-100612, PI 398881, PI 427136, PI 437169B, PI 438164B, PI 518751, PI 561370, PI 404188A, and PI 574486*, (Figure 3.1).

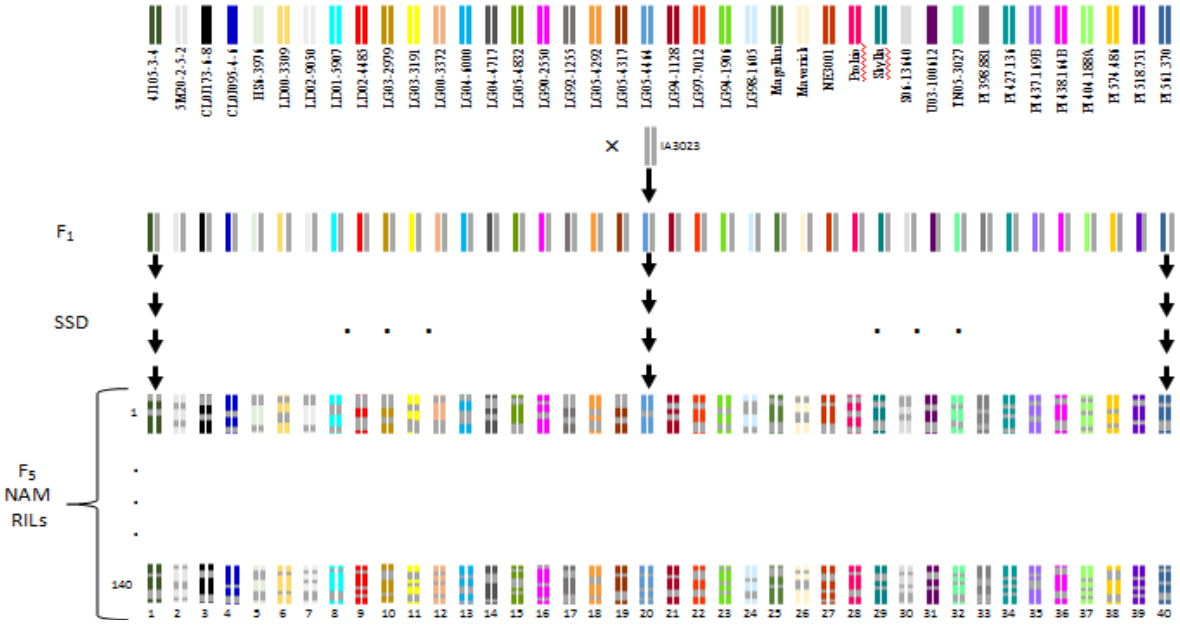


Figure 3.1. Schematic presentation of the SoyNAM structure.
(From Ben hall 2015)

The 40 elite lines included seventeen high yielding lines from eight states, fifteen lines from diverse ancestry and eight plant introductions. From each of these crosses at least 1000 F_2 lines with maturity as similar as possible to the parent IA3023 were selected and advanced up to F_5 generation, using single seed decent method. At generation F_5 each NAM family include 140 lines which were split into four sets each set having 35 recombinant inbred lines, one standard parent, two founders and three other checks. Therefore, during field planting, each set had 40 lines x 40 families which giving rise to 1600 lines per set. In each set only the checks were replicated, while the RILs were not replicated. The total number of lines planted in the field was 6400 lines (4 sets x 1600 lines). The four sets were randomly planted in two rows plots of 80 cm length. The aforementioned described field layout is a Modified Augmented Design (MAD). The experiment was replicated at four different locations; Indiana, Illinois, Nebraska, and Iowa for two years 2012 and 2013.

3.3 Data Collection and Analysis

3.3.1 Phenotypic Data Collection

In this study, we evaluated the SoyNAM population (5486 RILs of maturity group III) for variation in protein and oil contents. The phenotypic data for these two traits, which includes approximately 44,000 observations, were collected in four states, Indiana, Illinois, Nebraska, and Iowa for two years, 2012 and 2013. Approximately 300 g of seed samples were analyzed as whole grain per plot for protein and oil contents on a dry weight basis by NIR spectroscopy using a Perten DA7200 diode array instrument

equipped with collaboration equation, developed by a Perten with the assistance of the University of Minnesota (http://www.perten.com/).

Overall summary statistics for the two traits across environments were calculated using proc mixed procedure in SAS 9.4 (SAS Institute, 2014), (Table 3.1). Broad sense heritability was estimated for each trait across environments on line mean basis (Table 3.1). HSAUR2 (Hothorn and Everitt 2014) R package was used to analyze the phenotypic data for the two traits, seed protein and oil content, by environment and by populations. The phenotypic data for the two traits were also analyzed on a g/kg basis for each environment. Proc mixed model in SAS.9.4 was used to generate summary statistics for the two traits (SAS Institute, 2014), (Table 3.2; and Figure 3.7 and 3.8).

3.4 Phenotypic Data Analysis

3.4.1 Variance and Stability Analyses

The stability analysis for the two traits, seed protein and seed oil contents, were performed with the additive main effect and multiplication interaction (AMMI) in Genstat edition12th (https://kb.vsni.co.uk/Genstat/). In the AMMI model the phenotypic data were analyzed as RCBD with years considered as blocks. This is because the location data were not replicated. Generally, there were four locations; Indiana, Iowa, Illinois, and Nebraska. Each location had two years data.

Several statistical packages are available to analyze multi-environment trail data, but the most widely used one in plant breeding programs is AMMI (Agyeman, Parkes et al. 2015). AMMI model has been shown to be a powerful tool for investigating GEI analysis (Hagos and Abay 2013), since it fit in both additive (linear) and multiplicative

(bilinear) components that efficiently account for the underlying interaction (Shafii and Price 1998; Farshadfar, Poursiahbidi et al. 2012). The AMMI model combines the effect of genotype and environment from the ANOVA with principle components analysis of GEI (Ding, Tier et al. 2007).

3.4.2 AMMI Analysis of GEI

The following AMMI model was used to conduct the genotype by environment interaction (GEI). In the AMMI model, the additive portion of the variance is separated from the multiplicative variance (interaction) by analysis of variance (ANOVA) and then Principal Component Analysis (PCA), is applied to the to the interaction (residual) portion from the ANOVA to extract a new set of coordinate axes which account more effectively for the interaction patterns (Shafii and Price 1992).

$$y_{ij} = \mu + G_i + \beta_j + \Sigma\delta_n\gamma_{jn}\alpha_{jn} + e_{ij}$$

where y_{ij} is the response mean of i^{th} genotype in j^{th} environment; μ is the grand mean, G_i is the main effect of i^{th} genotype, $ß_j$ is the main effect of j^{th} environment, δ_n represents the singular value for IPCA axis n, γ_{jn} is the genotype i eigenvector value for IPCA axis n, α_{jn} is the environment j eigenvector value for IPCA axis n, and e_{ij} is the error.

3.4.3 Estimation of Heritability and Correlation by Families

The estimation of heritability and correlation is necessary for understanding the response to selection (van Kleunen and Ritland 2005). Heritability was estimated on a line mean basis for each trait across environments and for each of the 39 families using the following formula:

$$H^2 = \frac{\sigma^2 g}{\sigma^2 g + \left(\frac{\sigma^2 gy}{y}\right) + \left(\frac{\sigma^2 gl}{l}\right) + \left(\frac{\sigma e^2}{yl}\right)}$$

where H^2 represents broad sense heritability; $\sigma^2 g$ is the genetic variance for lines; $\sigma^2 gy$ is the genetic by year variance, $\sigma^2 gl$ is the genetic by location variance; σe^2 the residual variance; y and I represent year and location, respectively. The R package lme4 was used to estimate the variance components based on REML algorithm. The result of the heritability estimates is provided for each family in (Table 3.3).

Phenotypic correlation among the two traits protein and oil was calculated across environments using Pearson's correlation coefficients (r) with psych R package (Figure 3.8).

3.5 Linkage Disequilibrium (LD) Analysis

Linkage disequilibrium (LD), a non random association between various loci, is the basis of genetic association analysis for detection of gene or QTL (Hyten, Choi et al. 2007). Since GWAS measures the correlation between genotype and phenotype, LD plays key role in detecting significant association (Hyten, Choi et al. 2007).

The LD in this study was measured using correlation coefficients (r^2) between SNPs located at different physical distances. A total of 4118 SNPs spread across the 20 soybeans chromosome were first filtered in Tassel allowing minor allele frequency of 0.1 and missing data of less than 5% (Figure 3.10). Linkage disequilibrium (LD) heat map was generated for the entire genome, with heterozygous calls ignored and a default sliding window of 50 used (Figure 3.11). The filtered SNPs were used to establish LD and the LD decay rate was estimated on a genome wide and chromosome by chromosome basis (Figure 3.12 and Figure 3.13). For LD decay analysis, we generated

correlation coefficients (r^2) and pairwise distances in TASSEL and generated LD decay plots in "R version 3.0.3" (R Core Team 2014). Mean LD decay rate was calculated after every 500kb interval across all chromosomes. A line graph was used to display the mean genome-wide LD decay rates (Figure 3.12).

3.6 Population Structure

Population structure is the major factor that leads to false positive in association study. To investigate the presence of population stratification in the SoyNAM population, we conducted principal component analysis (PCA) in TASSEL using the 4118 SNPs. The number of PCs that capture the most variation in the population was determined using a scree plot, utilizing PCs and eigenvalues generated by TASSEL.

3.7 Genome Wide Association Study (GWAS)

3.7.1 Genotype and Phenotype Data

A subset of 4118 SNP markers from the Illumina SoyNAM BeadChip SNP array was used. A threshold of 0.10 for minor allele frequency (MAF) was used during SNP calling to avoid false-positives (Tabagin et al. 2009). A quality control function embedded within the NAM package was used to check for repeated markers and markers with minor allele frequency below threshold of 0.10. Imputation of missing values for markers was accomplished using a forward algorithm. This method is filling missing loci with the most likely genotype based on the previous marker, as a Markov model (Xavier 2015).

Protein and oil data for 5240 lines were used. Best linear unbiased predictors (BLUPs) were estimated for protein and oil for all entries across locations and for each location using R package Lme4 embedded in the SoyNAM package, developed by Xavier et. al (2015)(Bates 2010; Xavier 2015). The BLUP values were calculated using the following model:

$$\mathbf{y} = \mu + \mathbf{Zu} + \mathbf{Wg} + \mathbf{e}$$

where y is the vector of observed phenotype, **μ** is grand mean; Z is incident matrix for environment and **u** is a vector of random effect for environment; **W** is the incident matrix of genotype; **g** is vector of random genetic value associated with each genotype; and e is vector of the residual. The model was developed based on the assumption that $b \sim N(0, I\sigma_b^2)$ $\mathrm{u} \sim \mathrm{N}(0, \mathrm{I}\sigma_u^2)$, and $\mathbf{e} \sim \mathrm{N}(0, \mathrm{I}\sigma_e^2)$.

3.7.2 Association Analysis

The genome scan analysis for QTL associated to protein and oil was conducted in R package NAM, which is designed for association studies in *nested association mapping* (*NAM*) panels as implemented by Xavier et al. (2015). Subpopulations were used to define the stratification factor to allow different linkage phase between marker and QTL in each family. Marker effects were treated as a random to decrease the background noise (Xu and Atchley 1995). The statistical model used by this method for GWAS is:

$$\mathbf{y} = \boldsymbol{\mu} + \mathbf{Zu} + \mathbf{Wg} + \mathbf{e}$$

where **y** is the vector of observed phenotypes; μ is mean value of protein/oil across environment; Z is the incident matrix of marker effect, u is the vector of random effects

for marker; **W** incident matrix of genotype and g is the polygenetic term (estimated from the kinship matrix (K), and e is the error variance.

Prior to GWAS analysis additional quality control was accomplished for removal of repeated genotypes that could happen by genotyping error. Using this quality control function we were able to find 128 repeated lines and they were excluded from the GWAS analysis.

3.8 Result

3.8.1 Mean Differences in Soybean Protein and Oil Content

Frequency distribution of the seed protein and oil content across environments showed that the two traits were normally distributed, indicating that the seed contents of these traits are controlled by many genes (Figure 3.2) (Kang and Gauch 1996).

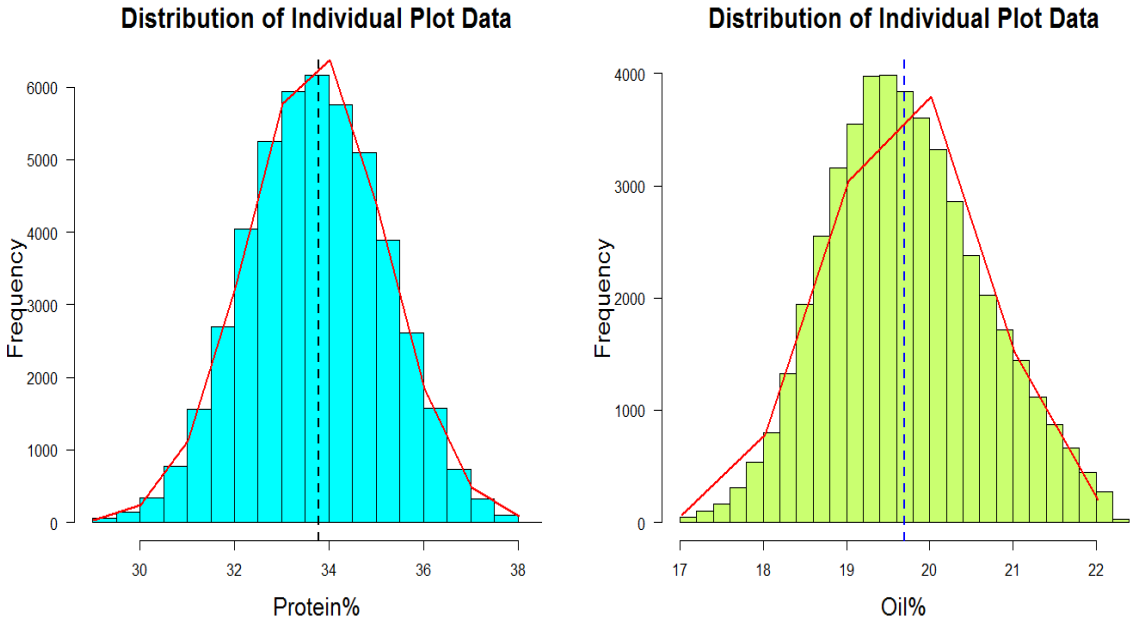


Figure 3.2. Frequency distribution of seed protein and oil content across environments; the dashed blue line represents average seed protein and seed oil content.

Protein and oil contents for the RILs varied across environments and families (Figure 3.3, 3.4, 3.5, and 3.6). The percent seed protein content in environment Nebraska 2013 was the highest followed by Indiana 2012 and Nebraska 2012, while the seed oil content in these environments was the lowest, indicating a reverse relationship among them (Figure 3.3, 3.4, 3.5, and 3.6). The negative relationship between these two traits makes it challenging to improve both traits simultaneously.

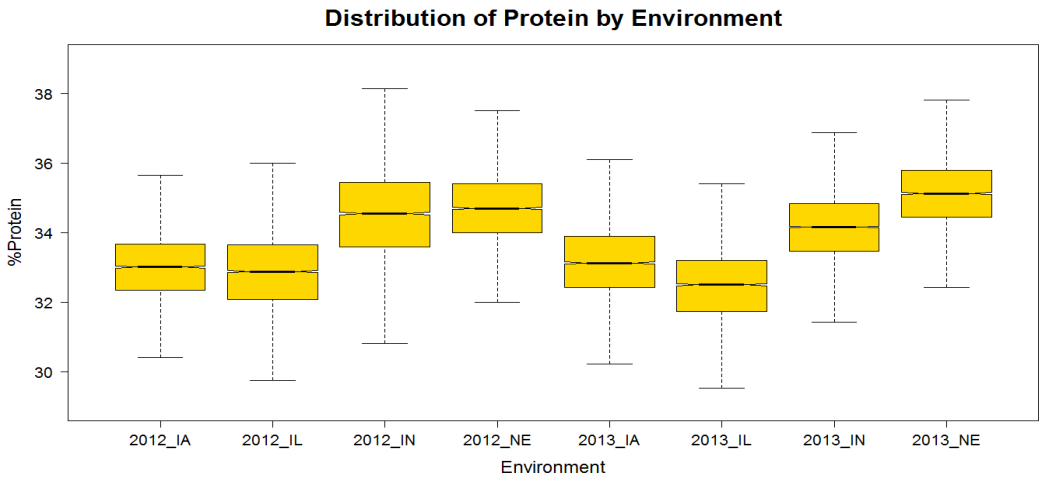


Figure 3.3. Distribution of seed protein content by environment.

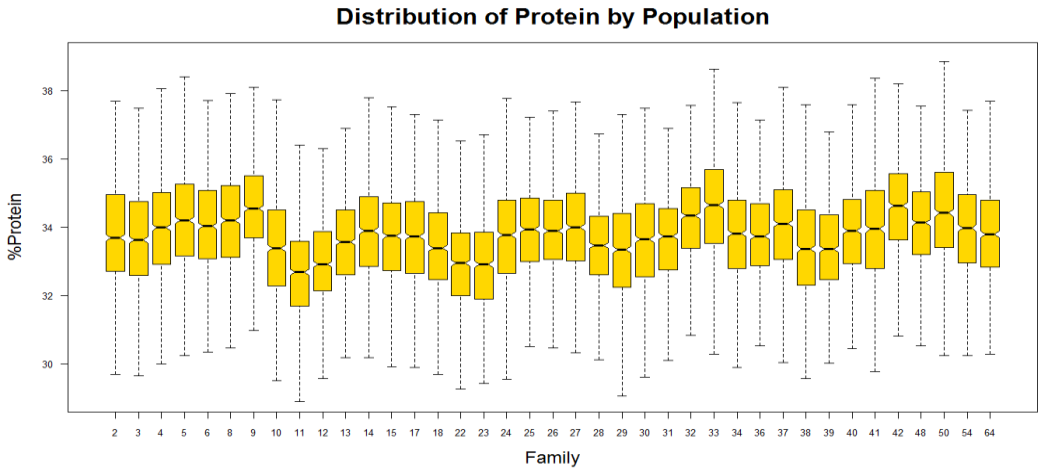


Figure 3.4. Distribution of seed protein content by populations (families).

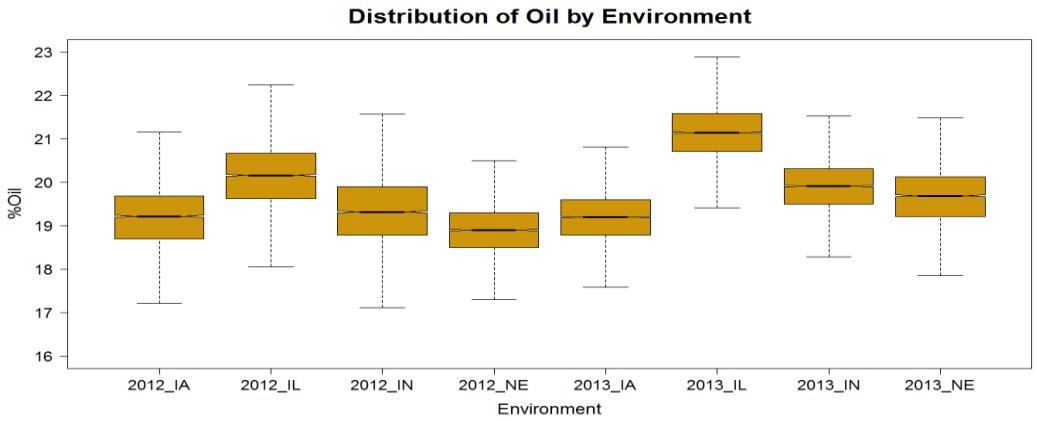


Figure 3.5. Distribution of seed oil content by environment.

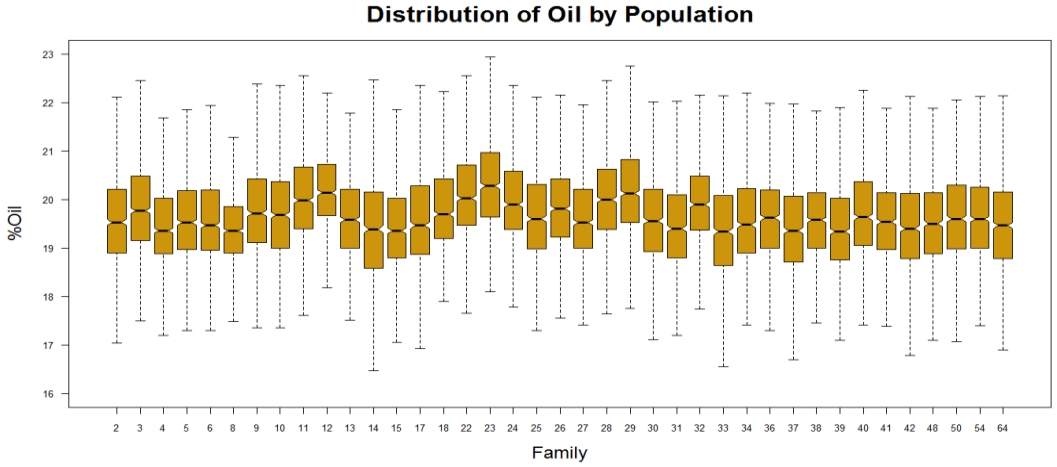


Figure 3.6. Distribution of seed oil content by populations (families).

Percent protein across environments ranged from 28.5-43.5 with a mean of 33.8 and standard deviation of 1.7 while % of oil across environments ranged from 13.0 to 23.4 with a mean of 19.7 and standard deviation of 0.97. The coefficient of variation for protein (4.3%) and oil (5%) was small, implying low experimental error (Table 3.1).

Table 3.1. Summary statistics and heritability estimates across environments for seed protein and oil.

Traits	Mean ± Std	Minimum	Maximum	CV%	Skew	Kurt	H^2
Protein (%)	33.8 ± 1.7	28.5	43.5	4.3	-0.02	-0.05	0.85
Oil (%)	19.7 ± 0.97	13.0	23.4	5.0	0.36	0.16	0.84

Std= standard deviation; H^2, implies broad sense heritability; CV, implies Coefficient of Variation.

The data were also analyzed based on a g/kg basis for both traits for each environment. The concentration of protein ranged from 286 Kg^{-1} to 435 $Kg^{-1,}$ with a mean 338 Kg^{-1} and standard deviation of 14.6 Kg^{-1} (Table 3.2 and Figure 3.7). Environment Nebraska 2013 has the highest seed protein content while the seed protein content for environment Illinois 2013 was the lowest (Figure 3.7). The concentration of oil ranged from 130 Kg^{-1} to 234 Kg^{-1} with a mean of 197 Kg^{-1} and standard deviation of 6.99 Kg^{-1} (Table 3.2). The seed oil content for environment Illinois 2013 was the highest whereas seed oil content for environment Nebraska 2012 was the smallest (Figure 3.8). The two traits indicated inverse relationship such that increase in seed protein content results in decreased seed oil content or vice versa. The analysis revealed that few lines performed well above the parents and RILs for protein concentration.

Table 3.2. Summary statistics for protein and oil for each environment.

	Protein g/kg				**Oil g/kg**			
Environment	Mean	Std	Min	Max	Mean	Std	Min	Max
2012_IA	330.3	10.3	295.4	426.9	191.7	7.5	129.9	215.6
2012_IL	328.9	11.8	285.9	413.5	201.4	7.8	150.1	227.8
2012_IN	344.9	13.8	293.7	434.2	193.5	8.3	142.1	223.5
2012_NE	347.1	10.8	309.0	419.0	188.7	6.3	149.0	212.0
2013_IA	331.8	11.0	293.2	400.7	191.9	6.5	151.5	216.0
2013_IL	324.8	11.1	285.4	408.4	211.3	6.6	156.6	234.4
2013_IN	341.3	10.6	288.5	434.6	199.1	6.2	146.6	222.8
2013_NE	351.3	10.4	309.6	424.3	196.7	6.8	153.0	221.7

IA, IL, IN, and NE, indicate, Iowa, Illinois, Indiana, and Nebraska, respectively.
Std=standard deviation, N= number of observation; Min=minimum, Max=maximum.

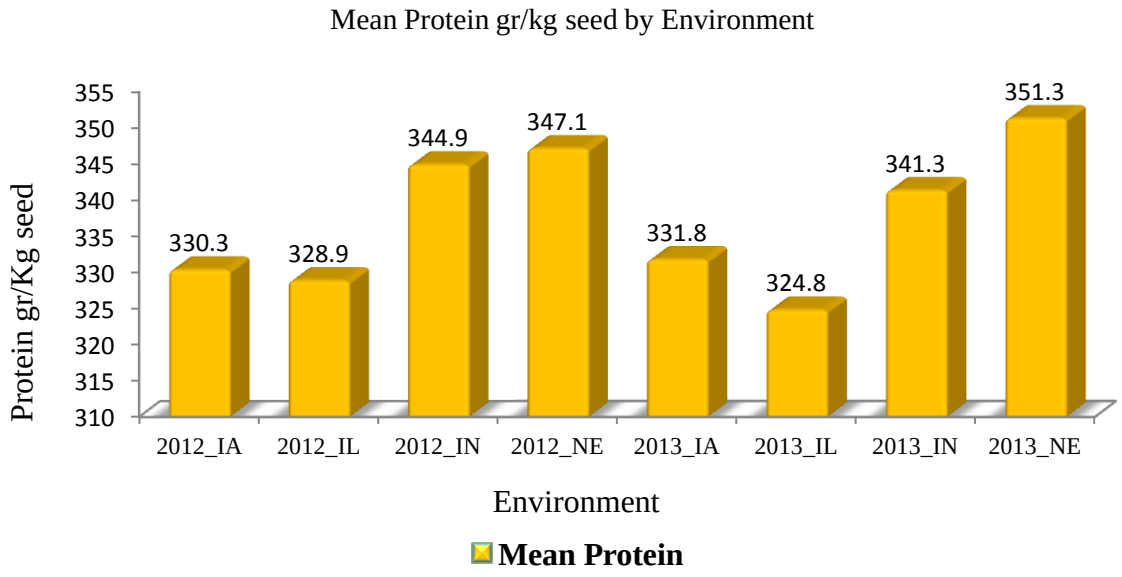


Figure 3.7. Seed protein content g/kg by environment on mean basis.
IA, IN, NE, and IL imply Iowa, Indiana, Nebraska, and Illinois, respectively.

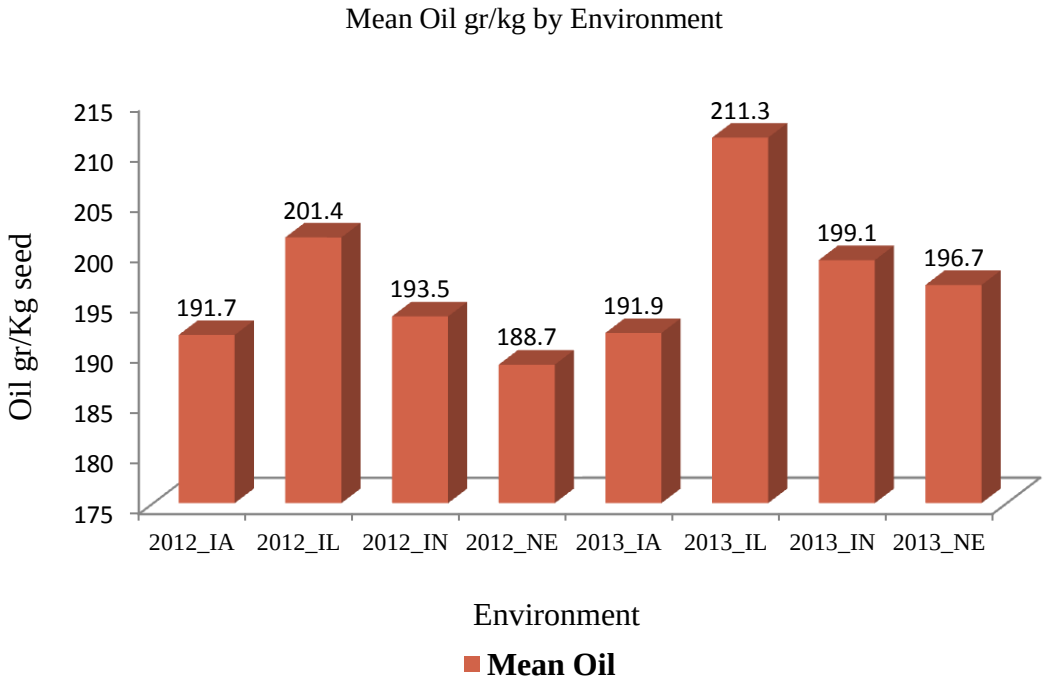


Figure 3.8. Seed oil content g/kg by environment on mean basis.
IA, IN, NE, and IL imply Iowa, Indiana, Nebraska, and Illinois, respectively.

3.8.2 Heritability Estimates and Correlations

Heritability for protein and oil, estimated based on line mean basis, was 85% and 84%, respectively (Table 3.1). The heritability seems high for both traits indicating that much of the variation in the population for these traits are due to genetic. The heritability of protein and oil were also estimated for each of the 39 SoyNAM families, which ranged from 64% to 90% and 58% to 80%, respectively (Table 3.3).

Table 3.3. Descriptive statistics for protein and oil, estimates of heritability and phenotypic correlation on family basis across environments.

	Protein					**Oil**					**Protein and Oil**
Family	Mean	Std	Min	Max	h^2	Mean	Std	Min	Max	H^2	Pheno Corr
2	337.9	9.5	310.0	359.9	0.8	196.0	5.8	181.8	209.8	0.78	-0.73
3	336.2	9.6	312.4	359.6	0.9	198.4	5.9	184.1	214.2	0.84	-0.61
4	339.8	9.9	314.0	368.1	0.8	194.8	5.6	178.7	208.2	0.85	-0.68
5	341.9	10.1	316.3	368.2	0.8	195.7	6.0	178.1	210.4	0.86	-0.69
6	340.8	9.2	315.3	364.1	0.8	195.9	5.6	181.0	210.3	0.96	-0.65
8	341.5	8.4	320.0	361.8	0.7	194.9	4.6	181.6	206.8	0.67	-0.56
9	345.6	8.8	319.4	367.4	0.8	198.1	5.7	183.2	212.4	0.74	-0.54
10	333.9	9.4	307.7	360.6	0.7	197.4	6.6	181.2	215.3	0.80	-0.61
11	326.4	10.9	297.0	356.6	0.7	200.6	6.8	184.4	218.8	0.80	-0.64
12	330.2	9.5	267.0	310.8	0.9	201.9	5.8	186.6	217.2	0.87	-0.56
13	335.1	9.2	310.3	356.8	0.8	196.2	5.8	181.5	211.0	0.87	-0.57
14	339.0	9.6	315.1	362.0	0.8	194.5	6.1	177.0	210.7	0.82	-0.59
15	337.1	9.4	312.3	359.8	0.8	194.9	6.1	178.5	209.7	0.85	-0.59
17	336.9	8.9	312.8	357.5	0.8	196.2	6.2	181.2	212.4	0.84	-0.51
18	334.4	9.1	311.0	357.2	0.8	198.4	5.8	185.0	213.3	0.78	-0.57
22	328.9	10.0	303.5	352.5	0.9	200.3	5.8	184.4	213.6	0.89	-0.59
23	329.2	10.2	264.4	312.9	0.8	203.5	5.7	188.4	218.8	0.81	-0.67
24	336.7	10.6	307.3	363.9	0.8	199.9	6.5	182.3	218.2	0.88	-0.65
25	339.4	11.4	315.0	420.2	0.9	196.2	7.9	147.4	213.0	0.92	-0.64
26	339.1	8.9	313.3	364.5	0.8	198.5	5.8	184.1	214.2	0.64	-0.63
27	339.1	9.8	313.0	366.4	0.8	196.1	6.2	181.4	210.6	0.85	-0.52
28	334.6	9.4	312.7	361.5	0.8	200.4	6.4	183.1	216.1	0.76	-0.63
29	333.5	10.3	307.0	359.3	0.8	202.4	6.0	187.3	218.2	0.71	-0.62
30	336.2	9.5	312.2	362.0	0.8	195.8	6.0	177.3	209.7	0.74	-0.73

Table 3.3 continued

31	336.1	10.3	308.2	363.5	0.8	194.6	7.1	177.2	213.0	0.78	-0.64
32	341.6	8.7	318.7	363.8	0.8	199.6	5.1	181.8	211.7	0.69	-0.59
33	345.6	12.2	313.3	375.4	0.9	192.2	7.8	172.7	211.7	0.82	-0.69
34	337.6	10.0	311.4	361.8	0.8	195.6	6.5	180.1	213.4	0.79	-0.60
36	336.9	10.1	311.9	362.6	0.6	195.2	6.6	176.5	210.4	0.77	-0.43
37	339.4	11.1	310.5	367.6	0.8	193.4	6.7	175.8	207.3	0.70	-0.59
38	334.2	8.6	313.6	357.1	0.7	196.3	5.3	182.0	209.3	0.58	-0.64
39	333.5	8.8	311.6	356.6	0.8	194.1	5.4	179.9	208.0	0.68	-0.67
40	338.9	10.9	308.5	367.6	0.8	197.2	5.7	182.5	212.7	0.65	-0.55
41	339.6	10.7	315.2	372.0	0.9	196.2	5.7	180.1	209.8	0.66	-0.63
42	345.6	10.9	316.2	374.8	0.8	194.4	7.3	175.8	211.9	0.68	-0.64
48	340.9	10.7	312.3	320.3	0.7	195.4	7.3	177.3	215.4	0.76	-0.60
50	342.7	11.9	313.4	370.9	0.8	195.7	7.2	175.8	212.2	0.78	-0.58
54	339.3	9.6	315.8	362.5	0.8	196.7	5.9	181.8	210.7	0.75	-0.69
64	337.7	10.6	308.7	364.2	0.7	194.8	7.4	173.3	213.8	0.80	-0.50

Pheno Corr implies phenotypic correlation.

Overall, negative phenotypic (r_g = -0.61**) was observed. The phenotypic correlation is depicted in Figure 3.9. The negative correlation values indicate that simultaneous improvement in both traits challenging, since improvement in one trait would result in a decrease in the other trait. On a family basis the phenotypic correlation ranged from -0.2 to -0.81 (Table 3.3). Weak correlation between the two traits implied that the families, 18, 36, 48 and 64 could be used for further genetic studies and there might be genes that act upon these traits independently.

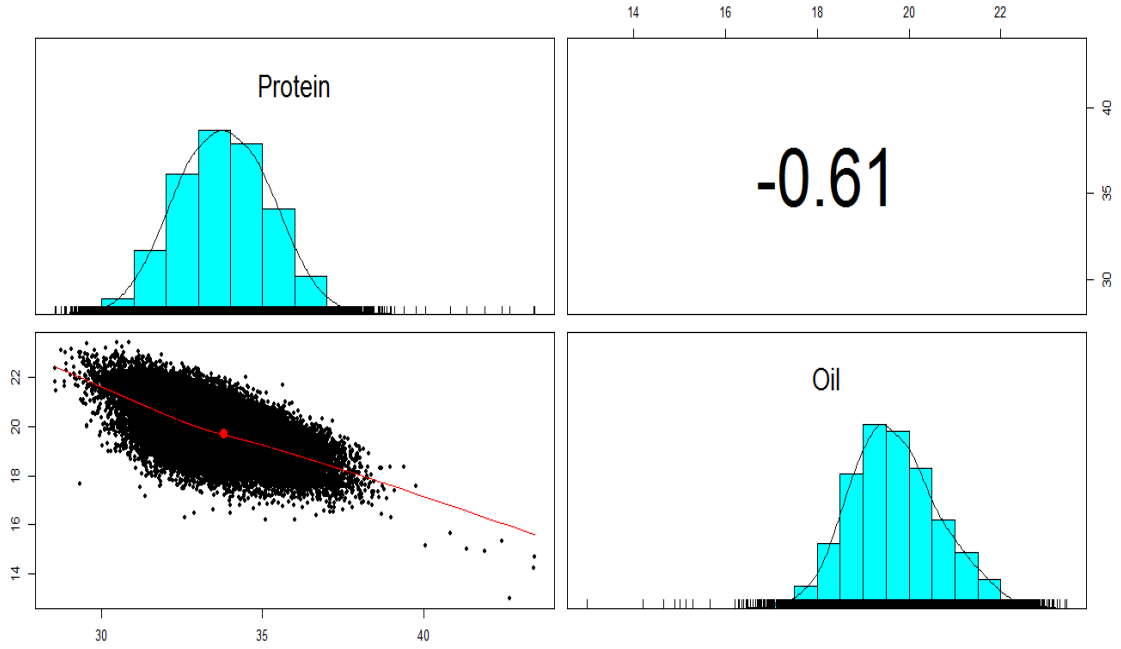


Figure 3.9. Phenotypic correlation between protein and oil across environment using Pearson's correlation coefficients (r).

3.8.3 Multi-Environment Assessment

The AMMI ANOVA revealed significant differences (P<0.001) among genotypes for both protein and oil, suggesting that the lines used were highly diverse and suitable for trait improvement (Table 3.4). Significant differences were also noticed among the four locations as well as between the two years, implying that each location and year had unique effect on genotype performance. The interaction between genotype and locations was also highly significant for both protein and oil, implying that genotype performances were specific to location (Table 3.4). Location explained the highest variation in seed protein content (38.17%) and in seed oil content (35.33%). This was followed by genotypes which accounted for 33.88% and 29.35% variation in the two traits respectively. The interaction term accounted for the least proportion of variation in

protein (13.2%) and seed oil (10.67%). The variation in seed protein content for year by location was much smaller (1.4%) compared to that of oil (12.92%), indicating that protein is more stable to seasonal variation than oil (Table 3.4).

Table 3.4. AMMI analysis of variance for protein and oil across locations.

		Protein			Oil		
Source	DF	SS	MS	R^2	SS	MS	R^2
Total	41887	92742			40973		
Genotypes	5485	31420	5.73***	33.88	12026	2.19**	29.35
Location	3	35398	11799.33**	38.17	14475	4825.00**	35.33
Year/Location	4	1333	333.25**	1.44	5295	1323.75**	12.92
Genotype*Location	16448	12245	0.74**	13.20	4370	0.27**	10.67
IPCA1	5487	5296	0.97**	43.25	1913	0.35**	43.78
IPCA2	5485	3558	0.65**	29.06	1311	0.24	30.00
Residuals	5476	3391	0.62		1146	0.21	
Error	19947	12346			4806		

*p<0.5, **p<0.01, ***p<0.001; R^2 = Variation Explained (%)

Result from the multiplicative part of the AMMI model revealed that both IPCAs for, seed protein and seed oil, were highly significant (P<0.001), indicating that they are helpful in explaining the residual multiplicative interaction (Table 3.4). Both interaction IPCAs together for protein and oil, accounted for a total of 72.31% and 73.78% of the interaction sum of square, respectively (Table 3.4). However, individually, the respective IPCA1 and IPCA2 explained 73.25% and 29.06% of the genotype by environment variation for protein, while, The IPCA1 and IPCA2 explained 73.78% and 30% of the GXE variation for oil, respectively (Table 3.4). The AMMI model selected 4 best high yielding and stable genotypes for each trait per location (Table 3.5). These genotypes are the ones that had stable and higher seed protein and seed oil contents across locations. These genotypes could be used as widely adapted genotypes with higher production of

protein and oil contents. Among the genotype for seed protein content genotype DS11-25174 from family 25 had the highest seed protein content and was the most stable genotype across locations (Table 3.5).

Table 3.5. AMMI selections of stable genotypes for protein and oil per location.

	Indiana	Nebraska	Illinois	Iowa
AMMI selections	Protein			
1	DS11-25174	DS11-25174	DS11-25174	DS11-25174
2	DS11-50332	DS11-42133	DS11-50332	DS11-33026
3	DS11-33051	DS11-41194	DS11-41194	DS11-42076
4	DS11-42076	DS11-50332	DS11-42133	DS11-33198
mean	34.32	34.92	32.67	33.14
score	4.168	1.25	0.279	-5.696
	Oil			
1	DS11-24167	DS11-11215	DS11-11215	DS11-25025
2	DS11-11230	DS11-29057	DS11-24141	DS11-29042
3	DS11-11215	DS11-24141	DS11-29057	DS11-11139
4	DS11-12038	DS11-24167	DS11-29042	DS11-25043
Mean	19.27	19.63	20.64	19.2
Score	1.992	1.844	0.928	-4.764

3.8.4 Linkage Disequilibrium (LD) Analysis and Marker Distribution

A subset of 4118 SNPs markers from the Illumina SoyNAM BeadChip SNP array with MAF >10% was used for LD analysis. The distribution of SNPs within each chromosome and across the 20 soybeans chromosomes was uneven (Figure 3.10). Chromosome 18 had the largest number of markers (290), while chromosome 9 harbored the lowest number of markers (148) (Figure 3.10).

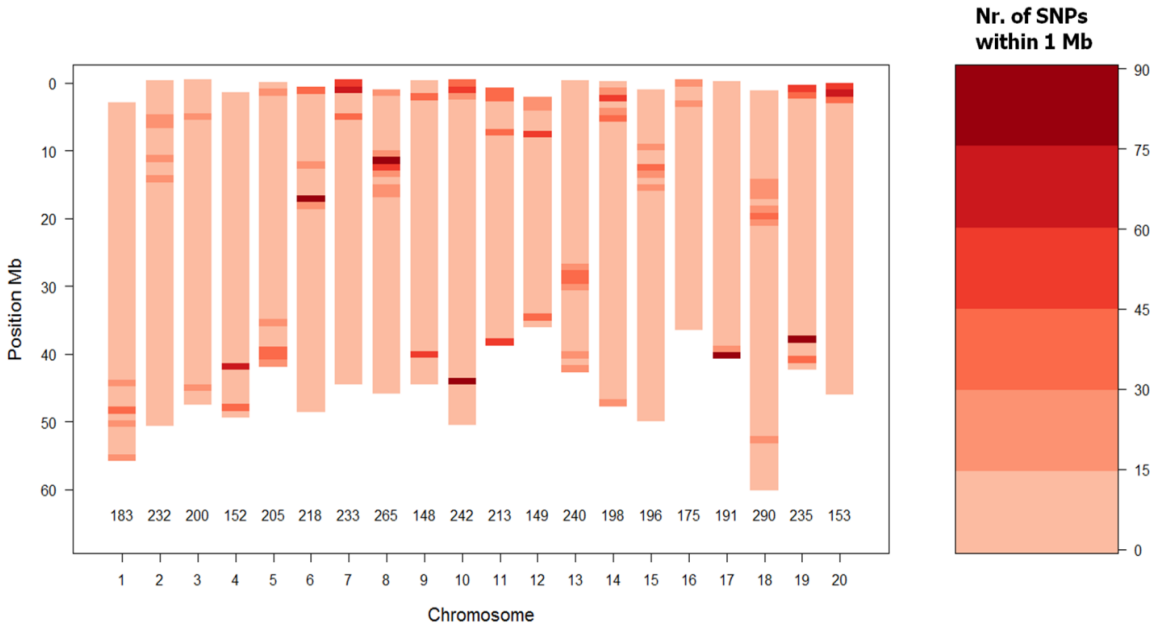


Figure 3.10. Density and distribution of (SNPs) across the 20 chromosomes of the SoyNAM mapping populations

Pattern of LD across the genome showed several haplotypes blocks anchoring SNPs that are in strong LD (Figure 3.11). The blocks in strong LD, are surrounded by genomic regions which are not in LD due to intensive recombination events. To find out the LD decay rate in SoyNAM population, r^2 was plotted against the physical distance across genome and for each chromosome (Figure 3.12 and Figure 3.13). As expected, the r^2 decreased as the distance between markers increased (Figure 3.12 and Figure 3.13). LD decay rate was different for each chromosome (Figure 3.13). The average LD decay across soybean genome was estimated between 2000-3000kb when the threshold r^2 value was set to 0.2 (Figure 3.12).

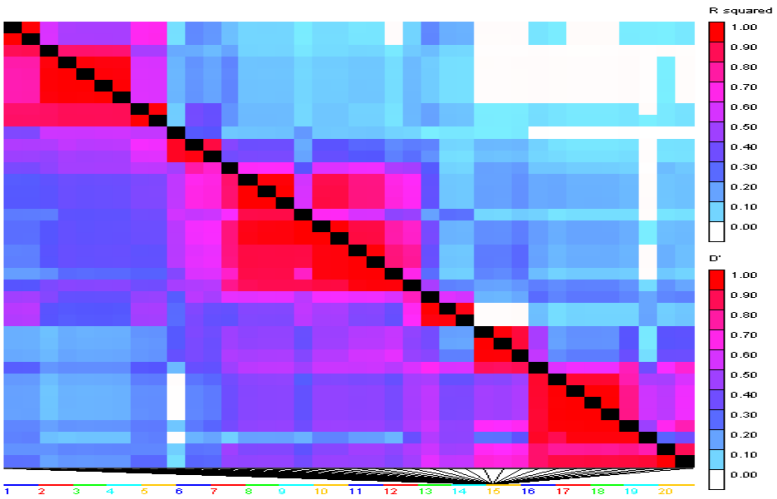


Figure 3.11. TASSEL heat map for pairwise LD between marker sites of the SoyNAM mapping populations.
LD measured using the r^2 (above diagonal) D' (below diagonal). Each cell represents the comparison of two pairs of marker sites with the color codes for the presence of significant LD. Colored bar code for the significance threshold levels in both diagonals is shown.

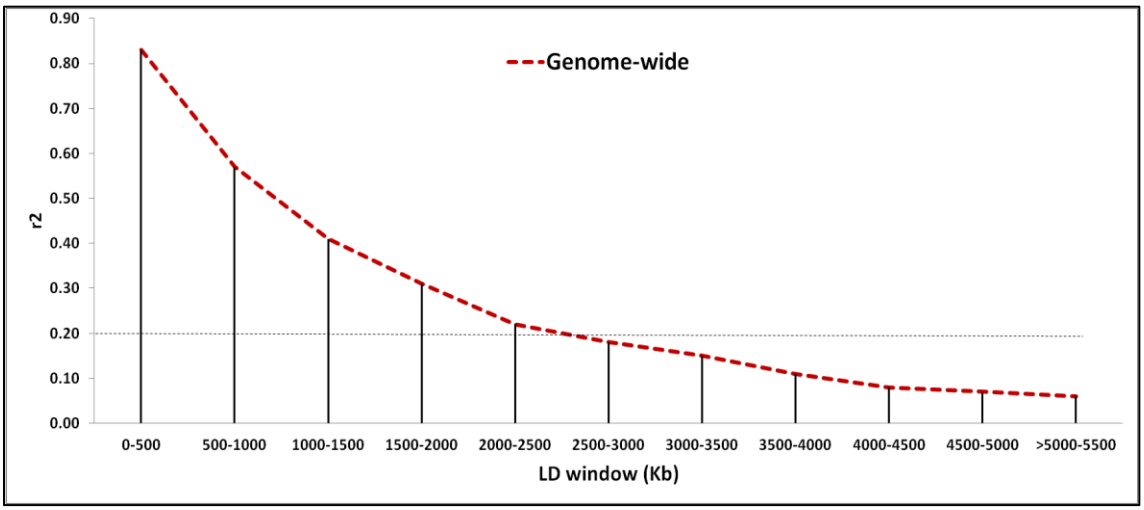


Figure 3.12. Mean LD decay rate across the soybean genome.
The LD decay rate was measured as r^2 using all pairs of SNPs located across the soybean genome. The average LD decay across soybean genome was estimated between 2000-3000kb when the threshold r^2 value was set to 0.2.

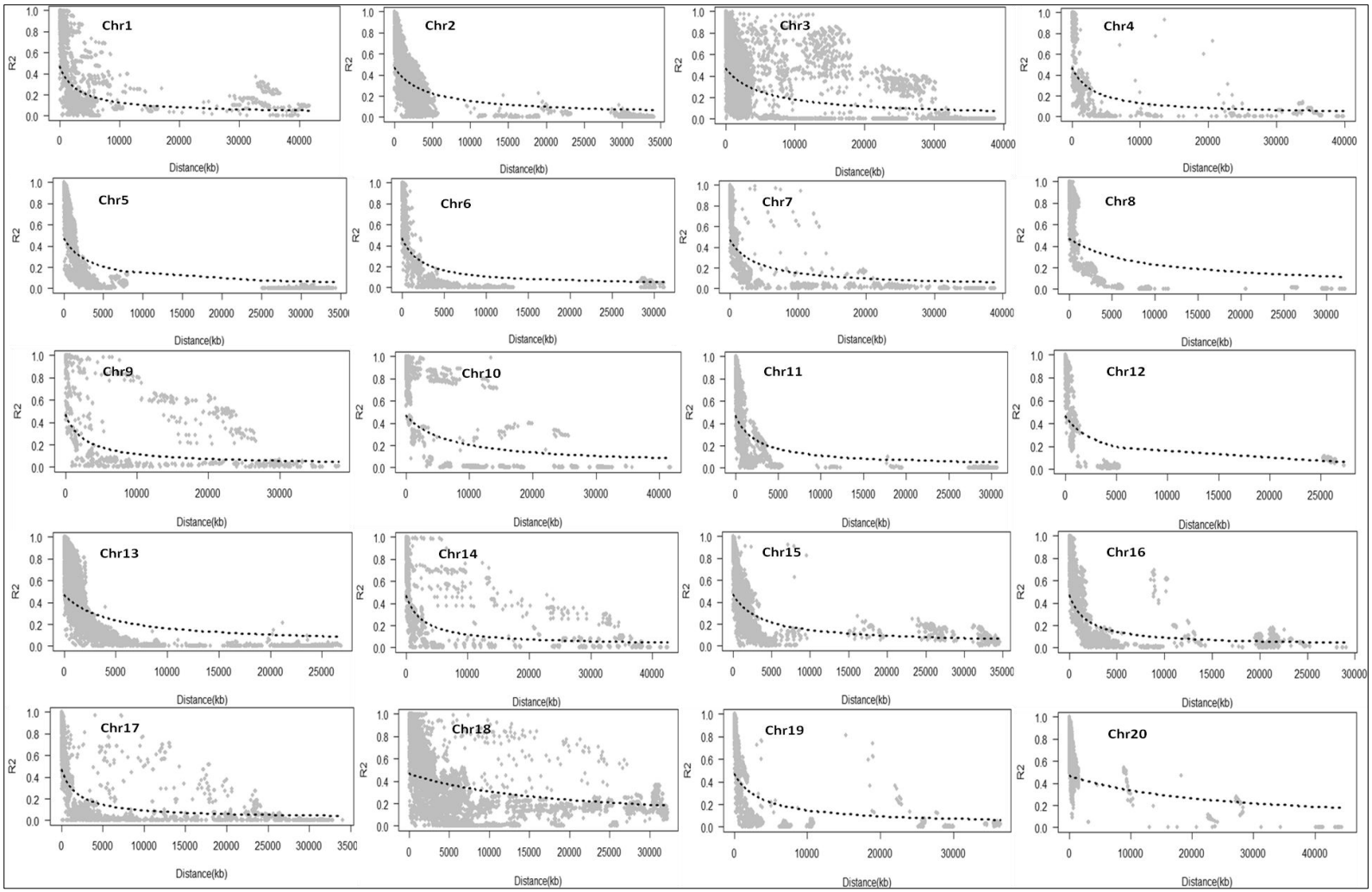


Figure 3.13. Rate of Linkage disequilibrium decay across each of the 20 chromosomes.
The LD decay rate is different for each chromosome.

3.8.5 Population Structure

Principal component analysis (PCA) based on the 4118 SNPs was conducted in TASSEL so as to analyze the structure of the SoyNAM population. We first used a scree plot to determine the number of PCs to be used in clustering of the SoyNAM population. In the scree plot, the proportion (eigenvalues) of an individual PC's contribution to total variation was plotted against the number of PCs (Figure 3.14).

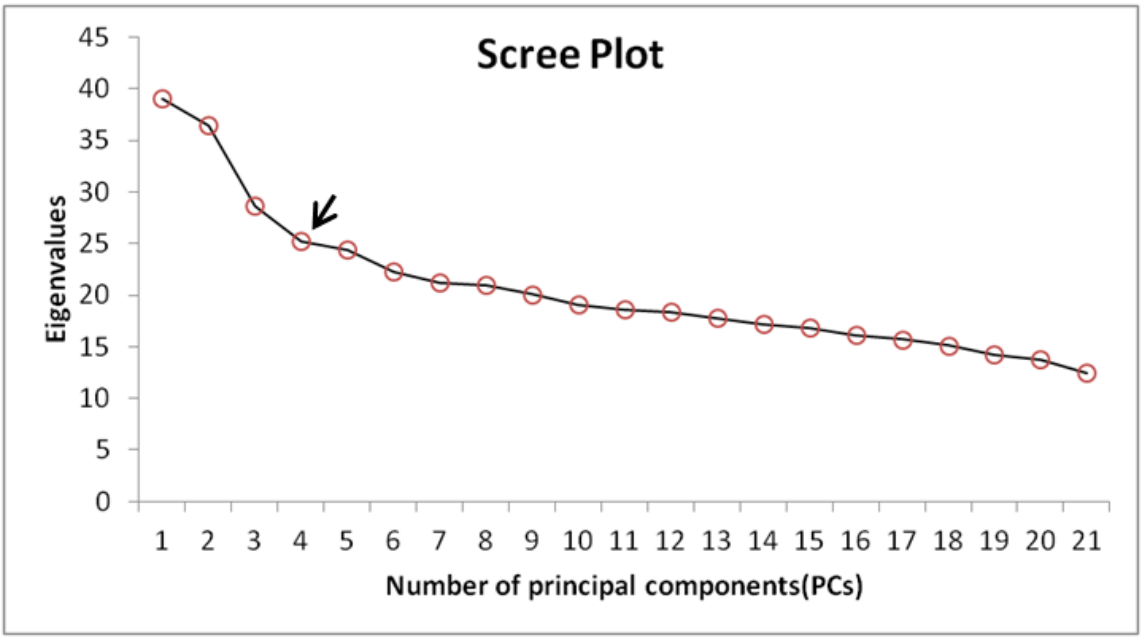


Figure 3.14. Scree plot of the PCs (X-axis) and their contribution to variance (Y-axis). Arrow indicates the "elbow" point.

The characteristic "elbow" point occurred at 4, and these first 4 PCs together accounted for 11% of total variation in the population. The 4 PCs portrayed in the scree plot indicated that they captured approximately the same amount of variation, signaling weak pattern of grouping within the population. Since each PC was approximately the same, we used the first 2 PCs which explained about 8% of total variation, defined three clusters (Figure 3.15). To account for the population stratification the NAM package,

which is based on MLM and EMMA algorithms, was used for association analysis of the two traits, protein and oil, in this study.

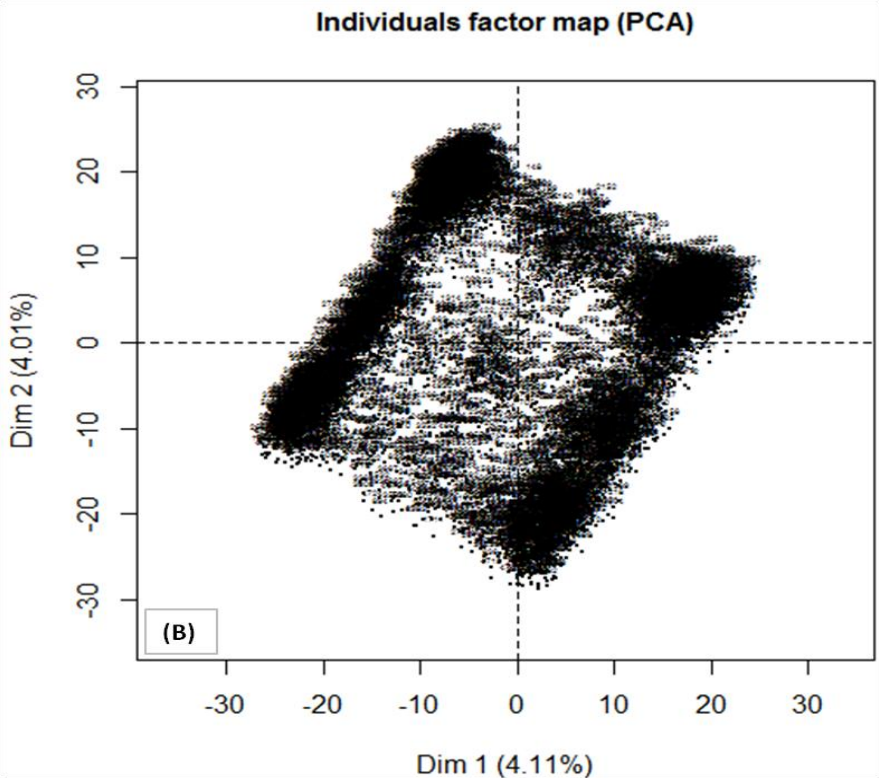


Figure 3.15. Individual factor map PCAs plot for the SoyNAM mapping population. Plot shows moderate population structure.

Stratification in the SoyNAM population might be due to growing the population in diverse environment under different growing environmental conditions, particularly due to photoperiod response. Photoperiod response is the major factor causing population stratification in soybean and it is well documented that soybean is photoperiod sensitive crop (Zhang, Singh et al. 2015).

3.8.6 Genome-Wide Association Study

The GWAS analyses were conducted for seed protein and seed oil contents, using NAM Package version (NAM 1.4.2) for each location as well as for the average data over

four locations. Using the average data over all 4 locations, a total of 49 SNPs distributed over 9 chromosomes were found to be highly associated (-logP> 4.92) with 13 QTL for seed protein (Figure 3.16).

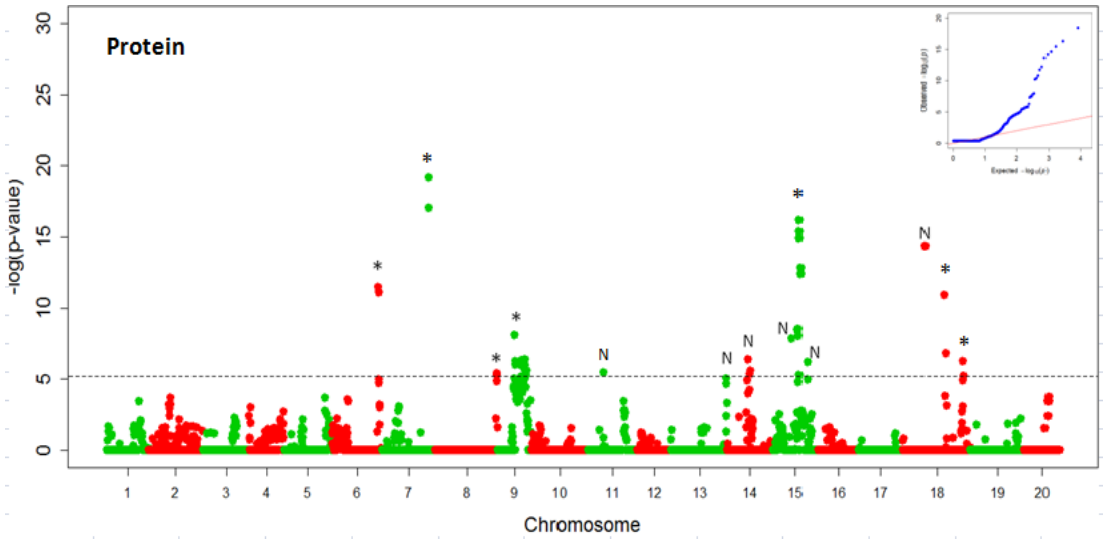


Figure 3.16. Manhattan plot for seed protein content.
The horizontal dashed line represents significant threshold and the significant threshold was set based on bonferroni correction 0.05/#marker. In the Manhattan plot N indicates novel QTL and the asterisk (*) represents previously reported QTL.

Using the Glyma.Wm82.a2 sequence browser and gene model Glyma.Wm82.a1.v1.1 at SoyBase we found that 6 out of the 13 QTL were novel and were located on chr11, chr13, chr14, chr15, and chr18. From the 49 SNPs associated with seed protein content, clusters of highly significant markers were present on chr9 and chr15 (Figure 3.16 and Figure 3.17).

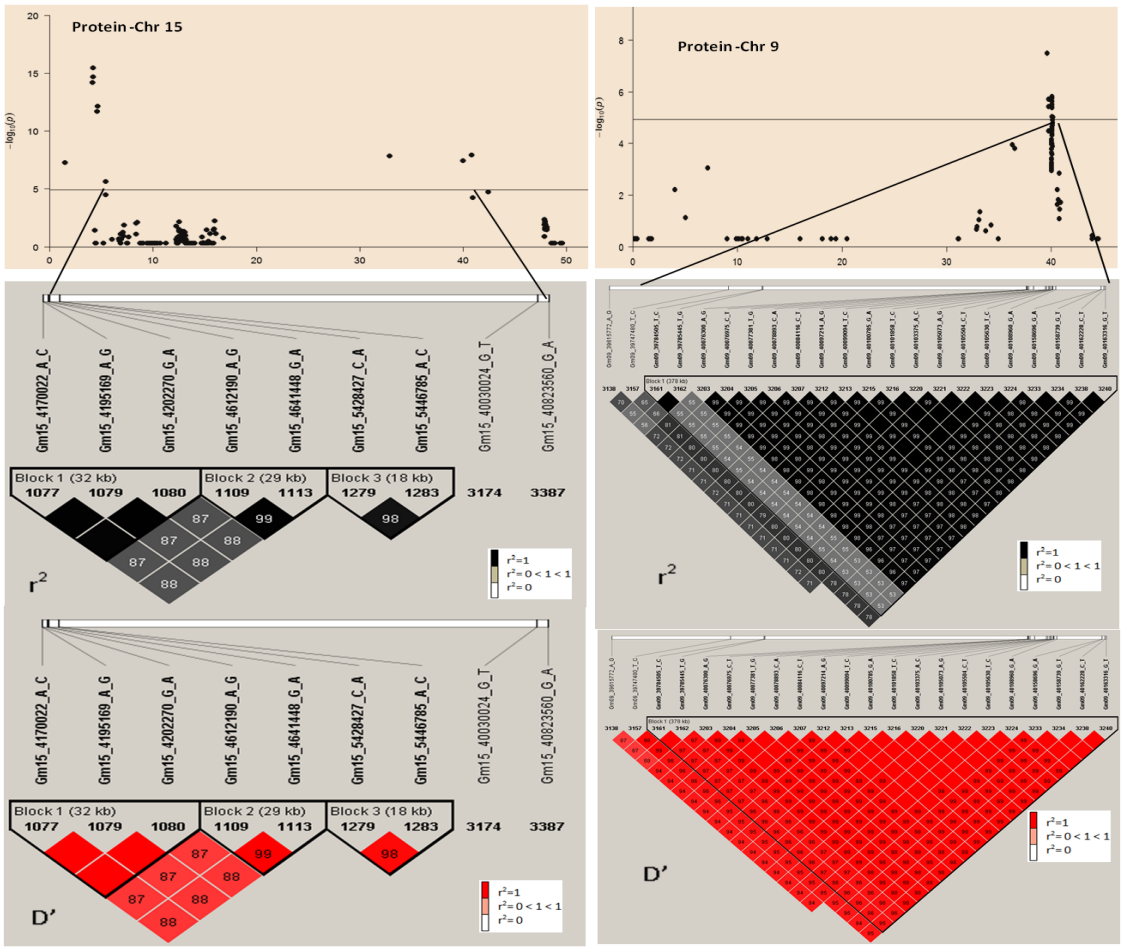


Figure 3.17. Manhattan plots show strong signal on chromosome 9 and 15.
Significant markers tagging this region are all in strong LD as indicated by D' and r^2.

Almost half (22 out of 49) of the SNPs had physical location with the 39.6-40.2 Mbp genomic region on chr9 which were in complete LD as indicated by r^2 and D' (Figure 3.17 and Figure 3.18). Lu et al. 2012 and Eskandari et al. 2013 reported seed protein QTL within 35-41Mbp and 37.5-42 Mpb respectively of the chr9, which spans the same genomic region mapped in the current study. We however, refined this genomic region from 7.705 Mbp (Lu et al. 2012) and 4.851 Mbp (Eskandari et al. 2013) to 0.56 Mbp (Figure 3.18).

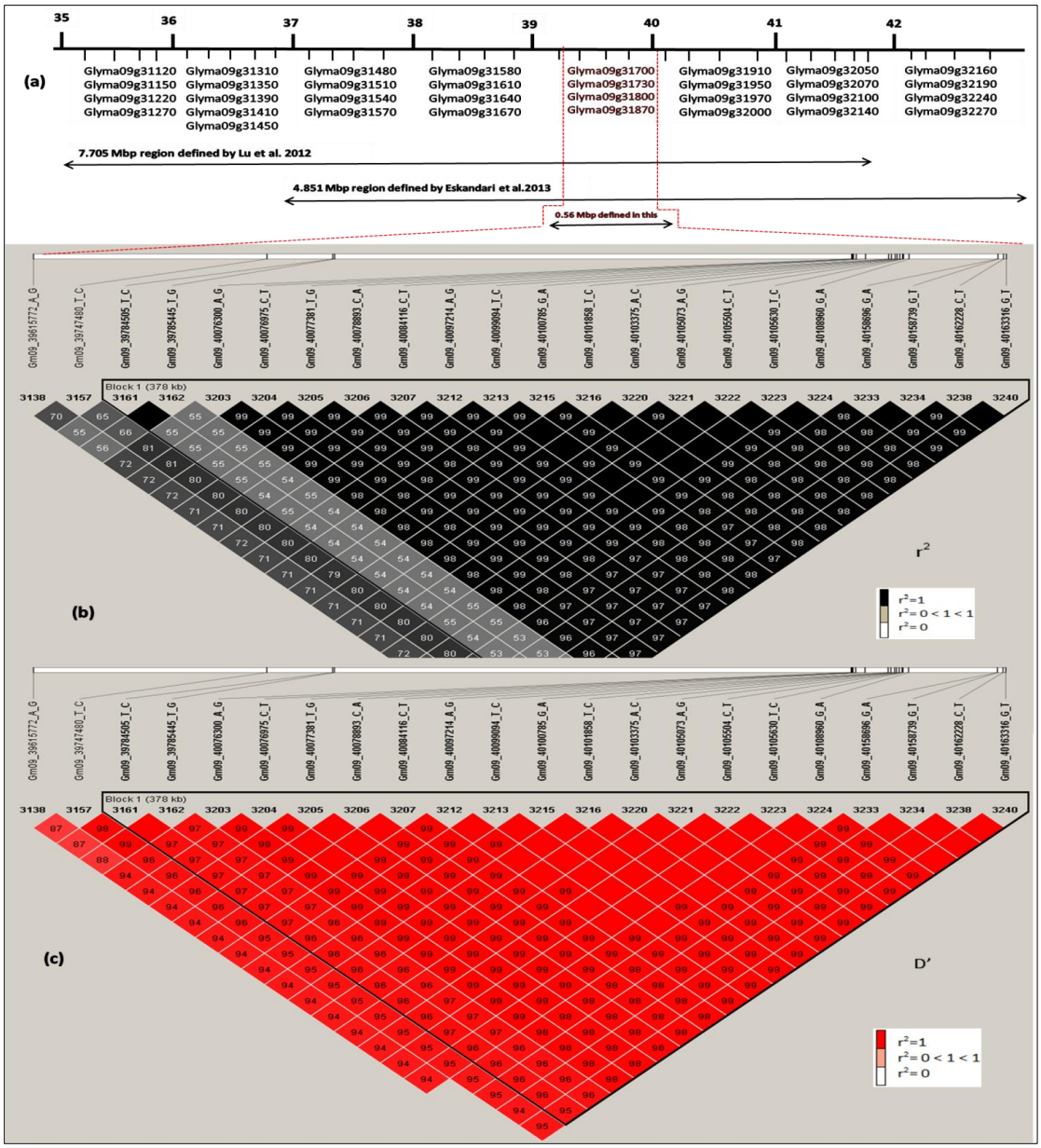


Figure 3.18. Genomic region of the seed protein QTL on chromosome 9.
Here we show the genomic region on chr9 that is believed to be associated with seed protein content. Panel a; show 7.705 and 4.851 Mbp genomic by Lu et al. 2012 and Eskandari et al. 2013 and the 0.56 Mbp genomic region identified in this study. In this study we were able to reduce the genomic region harboring QTL controlling seed protein content to a much narrow region; Panel b and c show LD based on r^2 and D'.

GWAS for oil content identified 12 QTL on 8 different chromosomes comprising 109 SNPs (Figure 3.19).

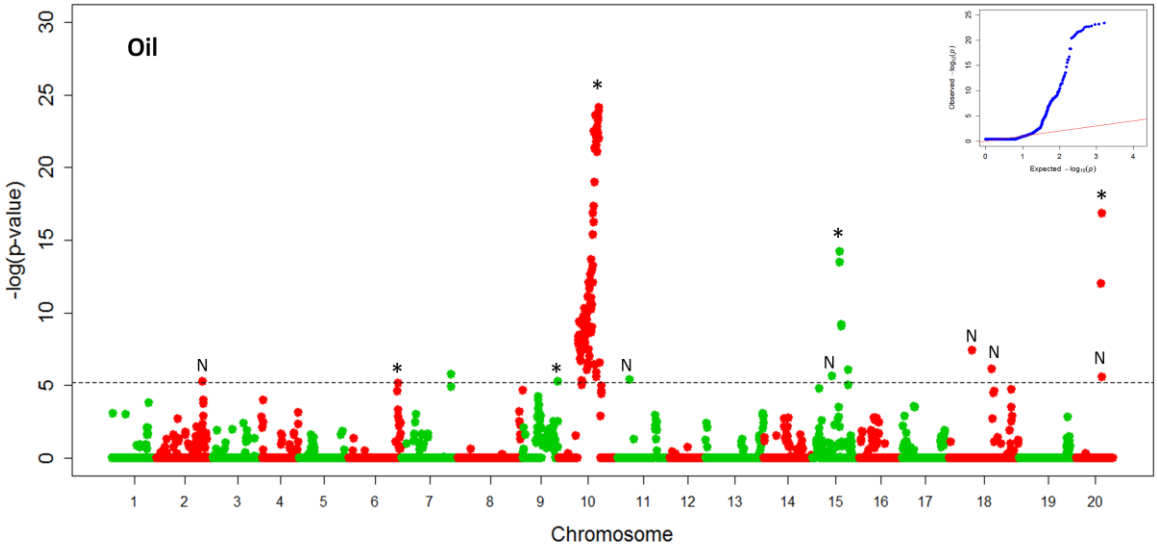


Figure 3.19. Manhattan plot for seed oil content.
The horizontal dashed line represents significant threshold and the significant threshold was set based on bonferroni correction 0.05/#marker. In the Manhattan plot N indicates novel QTL and the asterisk (*) represents previously reported QTL.

Of the total detected QTL for oil content, 6 QTL were novel and were located on chr2, chr11, chr15, chr18, and chr20. The remaining 6 QTL were known and previously reported by several different GWAS and bi-parental QTL studies. Among the QTL detected for oil content, two highly significant QTL were mapped on chr10 and 15. Of these, the QTL on chr10 comprised more than 90 SNPs (Figure 3.20).

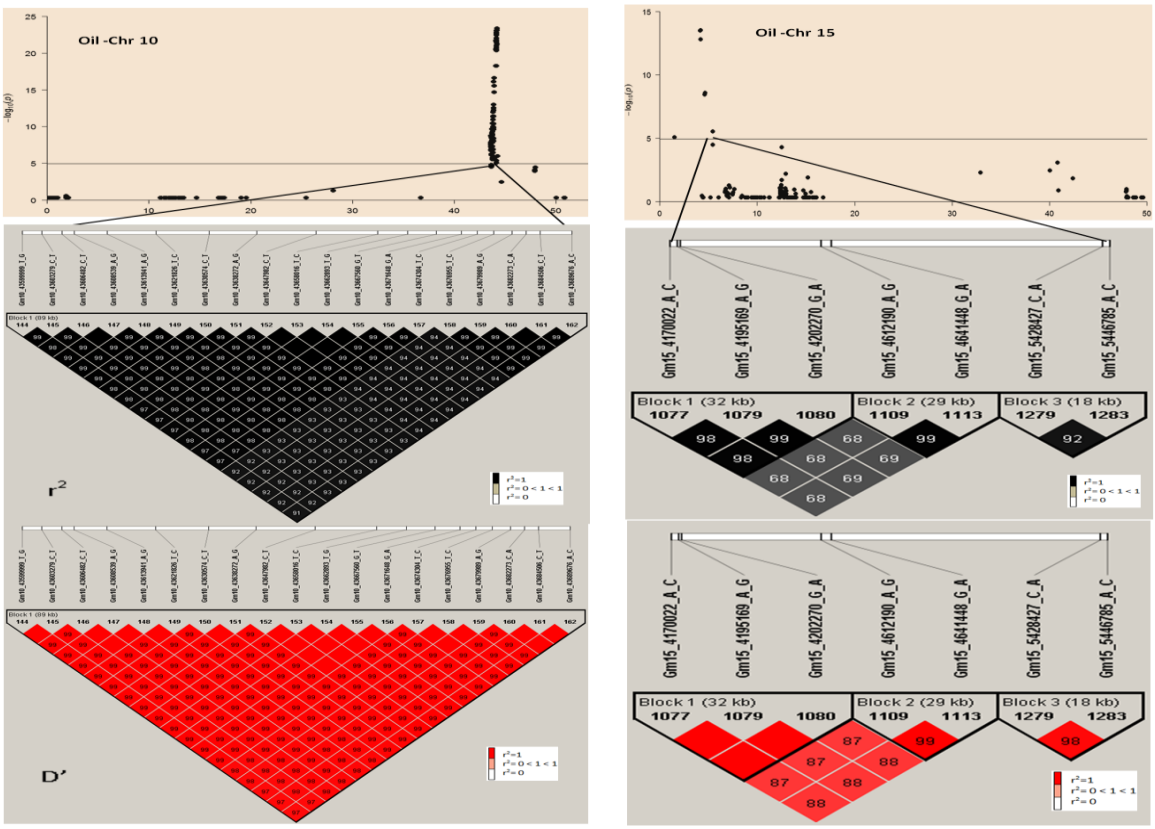


Figure 3.20. Manhattan plots show strong signal on chromosome 10 and 15. Significant markers tagging this region are all in strong LD as indicated by D' and r^2.

The SNPs identified for protein and oil explained the phenotypic variation in these two traits by 15% and 23%, respectively. Markers associated with seed protein and oil contents are presented in Table 3.6 and 3.7, respectively.

A total of 158 SNPs were detected for both seed protein and seed oil contents. Out of these, 38 SNPs were protein specific, 98 were oil specific, and 11 located on chr6, chr15, and chr18 were shared between both (Table 3.8). Allele effects estimates for markers associated with both protein and oil were negative. This suggested that the allele responsible for increased seed protein had a negative effect on seed oil production. The SNPs associated with QTL that have opposite effect for both traits could be controlled by just one pleiotropic QTL, whose two alleles have inverse effects on both traits.

Table 3.6. SNP Markers associated with seed protein content QTL.

The first column of the below table present highly associated markers (based on a -logP>3.0) and are numbered consecutively. The second column presents the status of the markers whether a marker (s) is new or has been previously reported. The third column reports whether it is also associated with seed oil content.

Protein QTL	Status	Associated trait	Chromosome	Marker Name	Physical Position(bp)	P-value	Allele effect	LOD score
1	known	Oil	6	BARC1.01_Gm06_46292681_G_T	46,292,681	3.59E-12	-0.11	9.6
	known	Oil	6	BARC1.01_Gm06_46386548_A_C	46,386,548	8.91E-12	0.11	9.2
	Known	Oil	6	BARC1.01_Gm06_46978335_G_T	46,978,335	1.17E-05	-0.07	3.4
2	known	Oil	7	BARC1.01_Gm07_7832406_T_C	7,832,406	9.49E-18	-0.07	15.0
3	known	Oil	8	BARC1.01_Gm08_45695835_C_T	45,695,835	5.42E-06	-0.06	3.7
	known	Oil	8	BARC1.01_Gm08_45765326_A_G	45,765,326	4.02E-06	0.06	3.8
4	known	Oil	9	BARC1.01_Gm09_39615772_A_G	39,615,772	8.20E-09	0.05	6.4
	known	Oil	9	BARC1.01_Gm09_39747480_T_C	39,747,480	1.03E-05	0.05	3.5
	known	Oil	9	BARC1.01_Gm09_39784505_T_C	39,784,505	5.42E-07	0.10	4.6
	known	Oil	9	BARC1.01_Gm09_39785445_T_G	39,785,445	1.07E-06	0.09	4.4
	known	Oil	9	BARC1.01_Gm09_40076300_A_G	40,076,300	7.39E-06	0.05	3.6
	known	Oil	9	BARC1.01_Gm09_40076975_C_T	40,076,975	5.03E-07	-0.06	4.7
	known	Oil	9	BARC1.01_Gm09_40077381_T_G	40,077,381	9.30E-06	0.05	3.5
	known	Oil	9	BARC1.01_Gm09_40078893_C_A	40,078,893	1.21E-06	-0.05	4.3
	known	Oil	9	BARC1.01_Gm09_40084116_C_T	40,084,116	8.80E-07	-0.06	4.4
	known	Oil	9	BARC1.01_Gm09_40097214_A_G	40,097,214	1.03E-05	0.06	3.5
	known	Oil	9	BARC1.01_Gm09_40099094_T_C	40,099,094	8.35E-06	0.06	3.5

	known	Oil	9	BARC1.01_Gm09_40100785_G_A	40,100,785	9.06E-07	-0.06	4.4
	known	Oil	9	BARC1.01_Gm09_40102780_C_T	40,102,780	6.81E-07	-0.06	4.5
	known	Oil	9	BARC1.01_Gm09_40103375_A_C	40,103,375	7.04E-06	0.06	3.6
	known	Oil	9	BARC1.01_Gm09_40105073_A_G	40,105,073	1.03E-05	0.05	3.5
	known	Oil	9	BARC1.01_Gm09_40105504_C_T	40,105,504	1.14E-06	-0.05	4.3
	known	Oil	9	BARC1.01_Gm09_40105630_T_C	40,105,630	2.76E-06	0.07	4.0
	known	Oil	9	BARC1.01_Gm09_40108960_G_A	40,108,960	4.29E-07	-0.05	4.7
	known	Oil	9	BARC1.01_Gm09_40158696_G_A	40,158,696	3.74E-06	-0.04	3.9
	known	Oil	9	BARC1.01_Gm09_40158739_G_T	40,158,739	5.08E-06	-0.04	3.7
	known	Oil	9	BARC1.01_Gm09_40162228_C_T	40,162,228	5.89E-06	-0.04	3.7
	known	Oil	9	BARC1.01_Gm09_40163316_G_T	40,163,316	2.85E-06	-0.05	4.0
5	New		11	BARC1.01_Gm11_3388809_T_C	3,388,809	3.36E-06	0.03	3.9
6	New	Oil	13	BARC1.01_Gm13_5435217_A_G	5,435,217	9.44E-06	0.07	3.5
7	New	Oil	14	BARC1.01_Gm14_35353835_C_T	35,353,835	4.22E-07	-0.09	4.7
	New	Oil	14	BARC1.01_Gm14_37947340_T_C	37,947,340	4.93E-06	0.08	3.7
8	New	Oil	15	BARC1.01_Gm15_1496570_T_G	1,496,570	1.40E-08	0.02	6.1
9	New	Oil	15	BARC1.01_Gm15_32915477_C_A	32,915,477	3.65E-09	0.09	6.7
	known	Oil	15	BARC1.01_Gm15_4170022_A_C	4,170,022	1.35E-15	0.06	12.9
	known	Oil	15	BARC1.01_Gm15_4195169_A_G	4,195,169	4.35E-16	0.07	13.4
10	known	Oil	15	BARC1.01_Gm15_4202270_G_A	4,202,270	7.00E-17	-0.07	14.1
	known	Oil	15	BARC1.01_Gm15_4612190_A_G	4,612,190	4.35E-13	0.05	10.5
	known	Oil	15	BARC1.01_Gm15_4641448_G_A	4,641,448	1.53E-13	-0.05	10.9

	known	Oil	15	BARC1.01_Gm15_5428427_C_A	5,428,427	6.56E-07	0.00	4.6
	known	Oil	15	BARC1.01_Gm15_5446785_A_C	5,446,785	1.04E-05	0.00	3.4
	known	Oil	15	BARC1.01_Gm15_40030024_G_T	40,030,024	9.13E-09	0.08	6.3
	known	Oil	15	BARC1.01_Gm15_40823560_G_A	40,823,560	2.98E-09	0.08	6.8
11	New	Oil	18	BARC1.01_Gm18_1685024_A_G	1,685,024	4.94E-15	-0.10	12.3
12	known		18	BARC1.01_Gm18_2102506_C_T	2,102,506	1.32E-11	0.08	9.0
	known	Oil	18	BARC1.01_Gm18_2396395_C_T	2,396,395	1.56E-07	-0.05	5.1
13	known	Oil	18	BARC1.01_Gm18_59588751_T_C	59,588,751	1.21E-05	-0.04	3.4
	known	Oil	18	BARC1.01_Gm18_59757800_G_T	59,757,800	5.63E-07	0.02	4.6
	known	Oil	18	BARC1.01_Gm18_60631055_A_G	60,631,055	6.47E-06	0.05	3.6

Table 3.7. SNP Markers associated with seed oil content QTL.
The first column of the below table present highly associated markers (based on a -logP>3.0) and are numbered consecutively. The second column presents the status of the markers whether a marker (s) is new or has been previously reported. The third column reports whether it is also associated with seed protein content.

Oil QTL	Status	Associated trait	Chromosome	Marker Name	Physical Position (bp)	P-value	Allele effect	LOD score
1	New	Protein	2	BARC1.01_Gm02_6011261_T_C	6,011,261	5.79E-06	-0.072	3.7
2	Known	Protein	6	BARC1.01_Gm06_46386548_A_C	46,386,548	7.02E-06	-0.048	3.6
3	Known		9	BARC1.01_Gm09_7163703_A_C	7,163,703	5.76E-06	0.040	3.7
4	Known		10	BARC1.01_Gm10_43599999_T_G	43,599,999	1.37E-08	0.009	6.1
	Known		10	BARC1.01_Gm10_43603279_C_T	43,603,279	4.76E-09	-0.017	6.6

Known	10	BARC1.01_Gm10_43606482_C_T	43,606,482	8.22E-09	0.010	6.3
Known	10	BARC1.01_Gm10_43608539_A_G	43,608,539	4.29E-10	0.017	7.6
Known	10	BARC1.01_Gm10_43613941_A_G	43,613,941	4.91E-09	0.006	6.6
Known	10	BARC1.01_Gm10_43621826_T_C	43,621,826	3.80E-09	0.000	6.7
Known	10	BARC1.01_Gm10_43630574_C_T	43,630,574	6.91E-09	-0.023	6.4
Known	10	BARC1.01_Gm10_43638272_A_G	43,638,272	4.09E-08	0.005	5.7
Known	10	BARC1.01_Gm10_43647982_C_T	43,647,982	1.55E-07	-0.010	5.1
Known	10	BARC1.01_Gm10_43658016_T_C	43,658,016	2.10E-08	0.004	6.0
Known	10	BARC1.01_Gm10_43662893_T_G	43,662,893	5.74E-08	-0.004	5.6
Known	10	BARC1.01_Gm10_43667560_G_T	43,667,560	2.18E-07	-0.014	5.0
Known	10	BARC1.01_Gm10_43671648_G_A	43,671,648	2.10E-08	-0.022	6.0
Known	10	BARC1.01_Gm10_43674304_T_C	43,674,304	6.12E-10	0.023	7.4
Known	10	BARC1.01_Gm10_43676955_T_C	43,676,955	1.03E-05	-0.019	3.5
Known	10	BARC1.01_Gm10_43679989_A_G	43,679,989	9.91E-06	-0.013	3.5
Known	10	BARC1.01_Gm10_43682273_C_A	43,682,273	4.54E-06	0.008	3.8
Known	10	BARC1.01_Gm10_43684506_C_T	43,684,506	6.01E-06	0.011	3.7
Known	10	BARC1.01_Gm10_43689676_A_C	43,689,676	6.13E-06	-0.011	3.7
Known	10	BARC1.01_Gm10_43692191_G_A	43,692,191	3.10E-08	-0.016	5.8
Known	10	BARC1.01_Gm10_43694296_A_G	43,694,296	3.10E-09	0.028	6.8
Known	10	BARC1.01_Gm10_43697533_C_T	43,697,533	7.99E-10	-0.031	7.3
Known	10	BARC1.01_Gm10_43714296_C_T	43,714,296	8.80E-10	-0.024	7.3
Known	10	BARC1.01_Gm10_43716784_A_G	43,716,784	1.94E-10	0.025	7.9

Known	10	BARC1.01_Gm10_43722532_G_A	43,722,532	9.86E-10	-0.030	7.2
Known	10	BARC1.01_Gm10_43725982_T_C	43,725,982	4.56E-10	0.016	7.5
Known	10	BARC1.01_Gm10_43729702_G_A	43,729,702	3.43E-10	-0.034	7.7
Known	10	BARC1.01_Gm10_43735348_A_C	43,735,348	5.09E-11	0.023	8.5
Known	10	BARC1.01_Gm10_43743280_G_A	43,743,280	1.94E-09	-0.024	6.9
Known	10	BARC1.01_Gm10_43755306_T_G	43,755,306	3.90E-09	0.023	6.7
Known	10	BARC1.01_Gm10_43757437_G_A	43,757,437	1.28E-08	-0.025	6.2
Known	10	BARC1.01_Gm10_43762210_C_T	43,762,210	2.19E-09	-0.038	6.9
Known	10	BARC1.01_Gm10_43767529_A_G	43,767,529	3.74E-10	0.036	7.6
Known	10	BARC1.01_Gm10_43773903_T_G	43,773,903	7.64E-10	0.028	7.3
Known	10	BARC1.01_Gm10_43776707_C_T	43,776,707	2.77E-10	-0.055	7.8
Known	10	BARC1.01_Gm10_43779401_G_A	43,779,401	3.09E-10	-0.055	7.7
Known	10	BARC1.01_Gm10_43783537_T_C	43,783,537	1.23E-10	0.045	8.1
Known	10	BARC1.01_Gm10_43790810_T_C	43,790,810	8.83E-07	0.015	4.4
Known	10	BARC1.01_Gm10_43793452_A_G	43,793,452	2.75E-07	0.020	4.9
Known	10	BARC1.01_Gm10_43799226_T_C	43,799,226	5.42E-07	0.031	4.6
Known	10	BARC1.01_Gm10_43808630_G_A	43,808,630	8.27E-12	-0.097	9.2
Known	10	BARC1.01_Gm10_43815883_G_A	43,815,883	3.63E-08	-0.059	5.7
Known	10	BARC1.01_Gm10_43818041_C_T	43,818,041	8.87E-13	-0.101	10.2
Known	10	BARC1.01_Gm10_43821942_T_C	43,821,942	2.77E-09	0.058	6.8
Known	10	BARC1.01_Gm10_43825392_A_G	43,825,392	3.32E-08	0.044	5.8
Known	10	BARC1.01_Gm10_43828130_A_C	43,828,130	2.52E-12	0.089	9.7

Known	10	BARC1.01_Gm10_43833979_A_G	43,833,979	1.00E-07	0.050	5.3
Known	10	BARC1.01_Gm10_43838442_C_A	43,838,442	1.10E-09	-0.070	7.2
Known	10	BARC1.01_Gm10_43841675_A_G	43,841,675	4.42E-11	0.083	8.5
Known	10	BARC1.01_Gm10_43856969_C_T	43,856,969	2.69E-11	-0.079	8.7
Known	10	BARC1.01_Gm10_43859917_G_A	43,859,917	9.91E-13	-0.092	10.1
Known	10	BARC1.01_Gm10_43863467_A_G	43,863,467	2.34E-13	0.094	10.7
Known	10	BARC1.01_Gm10_43868983_C_T	43,868,983	2.40E-09	-0.057	6.9
Known	10	BARC1.01_Gm10_43872139_C_T	43,872,139	1.01E-11	-0.077	9.1
Known	10	BARC1.01_Gm10_43880346_C_T	43,880,346	2.28E-14	-0.103	11.7
Known	10	BARC1.01_Gm10_43882385_G_A	43,882,385	5.94E-11	-0.069	8.4
Known	10	BARC1.01_Gm10_43885571_C_T	43,885,571	1.30E-13	-0.085	11.0
Known	10	BARC1.01_Gm10_43890845_C_T	43,890,845	9.89E-10	-0.056	7.2
Known	10	BARC1.01_Gm10_43903647_C_T	43,903,647	3.06E-11	-0.059	8.7
Known	10	BARC1.01_Gm10_43908174_G_A	43,908,174	8.65E-13	-0.085	10.2
Known	10	BARC1.01_Gm10_43913576_T_C	43,913,576	6.20E-14	0.069	11.3
Known	10	BARC1.01_Gm10_43919402_G_T	43,919,402	1.48E-17	-0.097	14.8
Known	10	BARC1.01_Gm10_43921575_T_C	43,921,575	4.25E-16	0.079	13.4
Known	10	BARC1.01_Gm10_43929636_A_G	43,929,636	6.11E-17	0.075	14.2
Known	10	BARC1.01_Gm10_43934881_G_A	43,934,881	4.61E-18	-0.092	15.3
Known	10	BARC1.01_Gm10_44114545_T_C	44,114,545	3.54E-23	0.074	20.3
Known	10	BARC1.01_Gm10_44118289_G_T	44,118,289	5.39E-22	-0.080	19.2
Known	10	BARC1.01_Gm10_44120764_T_C	44,120,764	1.01E-19	0.064	16.9

Known	10	BARC1.01_Gm10_44124696_G_A	44,124,696	3.57E-07	0.088	4.8
Known	10	BARC1.01_Gm10_44146333_A_C	44,146,333	1.07E-19	0.050	16.9
Known	10	BARC1.01_Gm10_44148529_C_T	44,148,529	5.89E-23	-0.088	20.1
Known	10	BARC1.01_Gm10_44151052_A_G	44,151,052	2.85E-24	0.088	21.4
Known	10	BARC1.01_Gm10_44155722_G_A	44,155,722	5.91E-22	-0.086	19.1
Known	10	BARC1.01_Gm10_44158333_C_A	44,158,333	1.32E-06	0.084	4.3
Known	10	BARC1.01_Gm10_44161160_C_T	44,161,160	2.78E-06	0.080	4.0
Known	10	BARC1.01_Gm10_44163504_G_A	44,163,504	1.83E-22	-0.083	19.6
Known	10	BARC1.01_Gm10_44166650_T_G	44,166,650	3.07E-22	0.079	19.4
Known	10	BARC1.01_Gm10_44169310_T_C	44,169,310	8.62E-22	0.071	19.0
Known	10	BARC1.01_Gm10_44172388_A_G	44,172,388	1.63E-23	0.091	20.7
Known	10	BARC1.01_Gm10_44176284_T_C	44,176,284	3.02E-23	0.082	20.4
Known	10	BARC1.01_Gm10_44182198_C_A	44,182,198	4.45E-23	-0.093	20.2
Known	10	BARC1.01_Gm10_44185202_T_C	44,185,202	3.85E-24	0.085	21.3
Known	10	BARC1.01_Gm10_44187665_C_A	44,187,665	4.00E-24	-0.101	21.3
Known	10	BARC1.01_Gm10_44189871_C_A	44,189,871	6.01E-24	-0.098	21.1
Known	10	BARC1.01_Gm10_44196956_T_C	44,196,956	1.44E-24	0.096	21.7
Known	10	BARC1.01_Gm10_44199135_T_C	44,199,135	1.39E-24	0.096	21.7
Known	10	BARC1.01_Gm10_44213424_A_C	44,213,424	7.51E-25	0.098	22.0
Known	10	BARC1.01_Gm10_44227168_A_C	44,227,168	1.12E-22	0.088	19.8
Known	10	BARC1.01_Gm10_44287415_G_A	44,287,415	2.82E-07	0.083	4.9
Known	10	BARC1.01_Gm10_44500915_T_C	44,500,915	8.09E-34	0.066	30.8

	Known		10	BARC1.01_Gm10_44630777_C_A	44,630,777	4.58E-37	-0.065	34.0
5	Known	Protein	10	BARC1.01_Gm10_47987331_C_T	47,987,331	1.08E-05	-0.005	3.4
6	New	Protein	11	BARC1.01_Gm11_18651414_A_G	18,651,414	4.07E-06	0.027	3.8
7	New		15	BARC1.01_Gm15_1496570_T_G	1,496,570	2.48E-06	-0.009	4.0
	Known	Protein	15	BARC1.01_Gm15_4170022_A_C	4,170,022	6.77E-15	-0.041	12.2
	Known	Protein	15	BARC1.01_Gm15_4195169_A_G	4,195,169	3.30E-14	-0.041	11.5
	Known	Protein	15	BARC1.01_Gm15_4202270_G_A	4,202,270	6.16E-15	0.039	12.3
8	Known	Protein	15	BARC1.01_Gm15_4612190_A_G	4,612,190	8.65E-10	-0.024	7.3
	Known	Protein	15	BARC1.01_Gm15_4641448_G_A	4,641,448	6.79E-10	0.021	7.4
	Known	Protein	15	BARC1.01_Gm15_5428427_C_A	5,428,427	8.33E-07	0.025	4.5
	Known	Protein	15	BARC1.01_Gm15_5446785_A_C	5,446,785	1.01E-05	-0.024	3.5
9	New	Protein	18	BARC1.01_Gm18_1685024_A_G	1,685,024	3.71E-08	0.045	5.7
10	New	Protein	18	BARC1.01_Gm18_2102506_C_T	2,102,506	7.23E-07	-0.034	4.5
11	Known		20	BARC1.01_Gm20_42993516_T_C	42,993,516	1.00E-12	0.049	10.1
	Known		20	BARC1.01_Gm20_42999237_C_A	42,999,237	1.59E-17	-0.062	14.8
12	New		20	BARC1.01_Gm20_46120144_C_T	46,120,144	2.68E-06	0.018	4.0

Table 3.8. SNP Markers shared between seed protein and oil contents QTL.

Marker Name	Chromosome	Physical Position(bp)	Seed Protein Content			Seed Oil Content		
			P-value	Allele effect (%)	LOD score	P-value	Allele effect (%)	LOD score
BARC1.01_Gm06_46386548_A_C	6	46386548	7.02E-06	-0.048	3.6	8.91E-12	0.109	9.2
BARC1.01_Gm15_1496570_T_G	15	1496570	2.48E-06	-0.009	4.0	1.40E-08	0.015	6.1

BARC1.01_Gm15_4170022_A_C	15	4170022	6.77E-15	-0.041	12.2	1.35E-15	0.060	12.9
BARC1.01_Gm15_4195169_A_G	15	4195169	3.3E-14	-0.041	11.5	4.35E-16	0.069	13.4
BARC1.01_Gm15_4202270_G_A	15	4202270	6.16E-15	0.039	12.2	7.00E-17	-0.071	14.1
BARC1.01_Gm15_4612190_A_G	15	4612190	8.65E-10	-0.024	7.3	4.35E-13	0.045	10.5
BARC1.01_Gm15_4641448_G_A	15	4641448	6.79E-10	0.021	7.4	1.53E-13	-0.046	10.9
BARC1.01_Gm15_5428427_C_A	15	5428427	8.33E-07	0.025	4.5	6.56E-07	0.000	4.6
BARC1.01_Gm15_5446785_A_C	15	5446785	1.01E-05	-0.024	3.5	1.04E-05	-0.002	3.4
BARC1.01_Gm18_1685024_A_G	18	1685024	3.71E-08	0.045	5.7	4.94E-15	-0.100	12.3
BARC1.01_Gm18_2102506_C_T	18	2102506	7.23E-07	-0.034	4.5	1.32E-11	0.082	9.0

GWAS analysis was also conducted for each environment to find the stability of the QTL across locations. Genome scan using NAM GWAS identified variable number of SNPs for both traits in each of the four locations (Table 3.9). Variability, in the number of detected QTL for each location and trait suggested that most of these QTL were location specific. Most of these QTL were identified in two or three locations but not all four locations. These results verified most of the seed protein and seed oil contents QTL reported at SoyBase.

Table 3.9. SNPs identified for each location for controlling % seed protein and oil contents.

Location	Trait	Number of SNPs indentified	Chromosome	Phenotypic variance explained (%)
Iowa	Protein	17	6,7,10,13,15	6
	Oil	83	10,11,13,15,20	11
Illinois	Protein	27	6,7,8,9,15,18	9
	Oil	102	2,6,7,9,10,15,18,20	21

Indiana	Protein	26	6,7,8,9,10,14,15,18	13
	Oil	50	6,8,10,15,18,20	14
Nebraska	Protein	19	5,6,7,14,15,18	7
	Oil	100	6,10,15,16,18,20	16

The GWAS scan conducted for each environment identified six SNPs for seed protein and thirty three SNPs for seed oil that were consistently identified in all the four locations and the combined data across locations (Table 3.10). The rest SNPs associated with genomic regions controlling seed protein and seed oil were expressed in some but not in all locations.

Table 3.10. SNP Markers associated with both seed protein and oil contents QTL that were consistently identified in all the four locations and the combined data across locations.

Trait	Illinois	Iowa	Indiana	Nebraska
Protein	BARC1.01_Gm06_46292681_G_T	BARC1.01_Gm06_46292681_G_T	BARC1.01_Gm06_46292681_G_T	BARC1.01_Gm06_46292681_G_T
	BARC1.01_Gm06_46386548_A_C	BARC1.01_Gm06_46386548_A_C	BARC1.01_Gm06_46386548_A_C	BARC1.01_Gm06_46386548_A_C
	BARC1.01_Gm07_7832406_T_C	BARC1.01_Gm07_7832406_T_C	BARC1.01_Gm07_7832406_T_C	BARC1.01_Gm07_7832406_T_C
	BARC1.01_Gm15_4195169_A_G	BARC1.01_Gm15_4195169_A_G	BARC1.01_Gm15_4195169_A_G	BARC1.01_Gm15_4195169_A_G
	BARC1.01_Gm15_4202270_G_A	BARC1.01_Gm15_4202270_G_A	BARC1.01_Gm15_4202270_G_A	BARC1.01_Gm15_4202270_G_A
Oil	BARC1.01_Gm10_43818041_C_T	BARC1.01_Gm10_43818041_C_T	BARC1.01_Gm10_43818041_C_T	BARC1.01_Gm10_43818041_C_T
	BARC1.01_Gm10_43828130_A_C	BARC1.01_Gm10_43828130_A_C	BARC1.01_Gm10_43828130_A_C	BARC1.01_Gm10_43828130_A_C
	BARC1.01_Gm10_43841675_A_G	BARC1.01_Gm10_43841675_A_G	BARC1.01_Gm10_43841675_A_G	BARC1.01_Gm10_43841675_A_G
	BARC1.01_Gm10_43859917_G_A	BARC1.01_Gm10_43859917_G_A	BARC1.01_Gm10_43859917_G_A	BARC1.01_Gm10_43859917_G_A

BARC1.01_Gm10_43863467_A_G	BARC1.01_Gm10_43863467_A_G	BARC1.01_Gm10_43863467_A_G	BARC1.01_Gm10_43863467_A_G
BARC1.01_Gm10_43880346_C_T	BARC1.01_Gm10_43880346_C_T	BARC1.01_Gm10_43880346_C_T	BARC1.01_Gm10_43880346_C_T
BARC1.01_Gm10_43885571_C_T	BARC1.01_Gm10_43885571_C_T	BARC1.01_Gm10_43885571_C_T	BARC1.01_Gm10_43885571_C_T
BARC1.01_Gm10_43903647_C_T	BARC1.01_Gm10_43903647_C_T	BARC1.01_Gm10_43903647_C_T	BARC1.01_Gm10_43903647_C_T
BARC1.01_Gm10_43908174_G_A	BARC1.01_Gm10_43908174_G_A	BARC1.01_Gm10_43908174_G_A	BARC1.01_Gm10_43908174_G_A
BARC1.01_Gm10_43913576_T_C	BARC1.01_Gm10_43913576_T_C	BARC1.01_Gm10_43913576_T_C	BARC1.01_Gm10_43913576_T_C
BARC1.01_Gm10_43919402_G_T	BARC1.01_Gm10_43919402_G_T	BARC1.01_Gm10_43919402_G_T	BARC1.01_Gm10_43919402_G_T
BARC1.01_Gm10_43921575_T_C	BARC1.01_Gm10_43921575_T_C	BARC1.01_Gm10_43921575_T_C	BARC1.01_Gm10_43921575_T_C
BARC1.01_Gm10_43929636_A_G	BARC1.01_Gm10_43929636_A_G	BARC1.01_Gm10_43929636_A_G	BARC1.01_Gm10_43929636_A_G
BARC1.01_Gm10_43934881_G_A	BARC1.01_Gm10_43934881_G_A	BARC1.01_Gm10_43934881_G_A	BARC1.01_Gm10_43934881_G_A
BARC1.01_Gm10_44114545_T_C	BARC1.01_Gm10_44114545_T_C	BARC1.01_Gm10_44114545_T_C	BARC1.01_Gm10_44114545_T_C
BARC1.01_Gm10_44118289_G_T	BARC1.01_Gm10_44118289_G_T	BARC1.01_Gm10_44118289_G_T	BARC1.01_Gm10_44118289_G_T
BARC1.01_Gm10_44120764_T_C	BARC1.01_Gm10_44120764_T_C	BARC1.01_Gm10_44120764_T_C	BARC1.01_Gm10_44120764_T_C
BARC1.01_Gm10_44146333_A_C	BARC1.01_Gm10_44146333_A_C	BARC1.01_Gm10_44146333_A_C	BARC1.01_Gm10_44146333_A_C
BARC1.01_Gm10_44148529_C_T	BARC1.01_Gm10_44148529_C_T	BARC1.01_Gm10_44148529_C_T	BARC1.01_Gm10_44148529_C_T
BARC1.01_Gm10_44151052_A_G	BARC1.01_Gm10_44151052_A_G	BARC1.01_Gm10_44151052_A_G	BARC1.01_Gm10_44151052_A_G
BARC1.01_Gm10_44155722_G_A	BARC1.01_Gm10_44155722_G_A	BARC1.01_Gm10_44155722_G_A	BARC1.01_Gm10_44155722_G_A
BARC1.01_Gm10_44163504_G_A	BARC1.01_Gm10_44163504_G_A	BARC1.01_Gm10_44163504_G_A	BARC1.01_Gm10_44163504_G_A
BARC1.01_Gm10_44166650_T_G	BARC1.01_Gm10_44166650_T_G	BARC1.01_Gm10_44166650_T_G	BARC1.01_Gm10_44166650_T_G
BARC1.01_Gm10_44169310_T_C	BARC1.01_Gm10_44169310_T_C	BARC1.01_Gm10_44169310_T_C	BARC1.01_Gm10_44169310_T_C

BARC1.01_Gm10_44172388_A_G	BARC1.01_Gm10_44172388_A_G	BARC1.01_Gm10_44172388_A_G	BARC1.01_Gm10_44172388_A_G
BARC1.01_Gm10_44176284_T_C	BARC1.01_Gm10_44176284_T_C	BARC1.01_Gm10_44176284_T_C	BARC1.01_Gm10_44176284_T_C
BARC1.01_Gm10_44182198_C_A	BARC1.01_Gm10_44182198_C_A	BARC1.01_Gm10_44182198_C_A	BARC1.01_Gm10_44182198_C_A
BARC1.01_Gm10_44185202_T_C	BARC1.01_Gm10_44185202_T_C	BARC1.01_Gm10_44185202_T_C	BARC1.01_Gm10_44185202_T_C
BARC1.01_Gm10_44187665_C_A	BARC1.01_Gm10_44187665_C_A	BARC1.01_Gm10_44187665_C_A	BARC1.01_Gm10_44187665_C_A
BARC1.01_Gm10_44189871_C_A	BARC1.01_Gm10_44189871_C_A	BARC1.01_Gm10_44189871_C_A	BARC1.01_Gm10_44189871_C_A
BARC1.01_Gm10_44196956_T_C	BARC1.01_Gm10_44196956_T_C	BARC1.01_Gm10_44196956_T_C	BARC1.01_Gm10_44196956_T_C
BARC1.01_Gm10_44199135_T_C	BARC1.01_Gm10_44199135_T_C	BARC1.01_Gm10_44199135_T_C	BARC1.01_Gm10_44199135_T_C
BARC1.01_Gm10_44213424_A_C	BARC1.01_Gm10_44213424_A_C	BARC1.01_Gm10_44213424_A_C	BARC1.01_Gm10_44213424_A_C

The results of our GWAS analysis using the NAM method confirmed most of the QTL that were reported by previous studies for the two traits (Diers, Keim et al. 1992; Shoemaker and Specht 1995; Csanadi, Vollmann et al. 2001; Zeng, Chen et al. 2014).

3.8.7 Discussion

Soybean Nested Association Mapping (SoyNAM) is the best approach for dissecting complex trait since it combines high power in detecting rare QTL from linkage and high resolution from association mapping. The diverse 40 elite parents used to develop SoyNAM mapping population represent an excellent source of genetic variation for the application of GWAS.

3.8.8 Phenotypic Differences, Heritability, and Correlation

The average protein and oil concentration in this study were 338 Kg^{-1} and 197 Kg^{-1} which is a little lower than the typical average protein and oil concentration 400 Kg^{-1}, and 200 Kg^{-1}, respectively (Panthee, Pantalone et al. 2005), (Table 3.2).

3.8.9 Multi-Environment Analysis

The multi environment analysis of the seed protein and seed oil contents showed significant variation ($P<0.001$) among genotype across locations, which is consistent with most previous studies. Sudarić et al. (2006) performed AMMI analysis for seed protein and seed oil contents using combined data from 15 environments and found a significant GEI. These authors found that locations accounted for large proportion of the total variance for protein content. Lee et al. (2003) conducted GE interaction analysis for isoflvones in soybean and found that environment and GE had the highest effects on

genotype performance and accounted for most of the variation in isoflovens contents. Zhe et al. (2010) reported significant GE interaction for seed composition and other agronomic traits. Most studies attributed the GE interaction for seed composition, particularly for seed protein and seed oil, to the effects of fluctuating temperature. Research conducted by Schnebly and Fehr et al. (1993) on soybeans seed fatty acid concentration indicated that higher environmental temperature affects fatty acid composition. Gurmu et al. (2009) reported positive correlation between higher temperature and % oil. Results from a study conducted by Kumar et al. (2006) on seven Indian cultivars reported significant GE interaction for genotypes, and genotype by location interaction for seed protein and seed oil contents. These results largely agree with the findings of the present study.

Heritability for protein and oil, estimated based on line mean basis, were 85% and 84%, respectively which is in range reported by other studies. Lee et al. (1996) estimated heritability for seed protein contents ranging from 0.57 to 0.91 and for seed oil contents ranging from 0.51 to 0.93. The observed high heritability in this study suggests that selection for these traits would result in high genetic gain, and that means families with high heritability would play key role in increasing protein and oil concentration.

Phenotypic correlation between protein and oil was -0.61. Past studies also found protein and oil to be negatively correlated (Hwang, Song et al. 2014; Bandillo, Jarquin et al. 2015).

3.8.10 Linkage Disequilibrium

LD level mostly indicated by r^2 is an important factor to consider while conducting GWAS. It plays key role in association analysis because the extent of LD can help determine the density of markers required for effective GWAS. In our study, using the SoyNAM population, the extent of LD declined to 0.2 within 2000-3000kb implying moderate LD decay rate (Figure 3.12). The extent of LD varies between different soybean mapping populations due to factors such as mating system, selection, domestication, funding event, genetic diversity, and population stratification (Hyten, Choi et al. 2007). LD decay rate in our study falls in the range of LD decay rates reported by other studies. LD decay rate in the study conducted by Hwang et al. (2014) for soybean seed protein and oil contents declined to 0.2 within 6000-8000kb much slower than the LD decay rate in our study. LD decay rate in the study conducted by Voung ea al. (2015) for soybean cyst nematode declined to 0.2 within 250kb, faster than the rate reported in the present study. LD decay rate in our study was in strong agreement with the LD decay rate reported by Zhang ea al. (2015) for sudden death syndrome trait in soybean. The moderate LD decay rate in the SoyNAM population implies that the population is genetically diversity and the number of SNPs (4118) used in this study are dense enough to capture the genetic variation in the SoyNAM population.

3.8.11 Seed Protein and Oil Contents QTL

The main objective of this analysis was to identify QTL controlling seed protein and oil contents in the SoyNAM population using GWAS. A number of seed protein and oil QTL have been reported at various positions across the soybean genome. Most of the

previously reported soybean seed protein and oil content QTL were identified via linkage analysis and therefore, their precise genomic regions could not be determined. The recent advances in soybean genetic map (release of genetic map version 4.0) made it possible to narrow the genomic region of the previously reported seed protein and oil QTL.

Using information from the genetic map version 4 at SoyBase, we were able to compare the physical locations of the previously reported QTL with positions of the markers identified in this study. Consequently, we aligned the 13 genomic regions associated with seed protein content identified in this study with previously reported QTL positions. Based on the alignment, 7 of the13 genomic regions were previously known and the remaining 6 QTL were novel (Figure 3.16). Out of the 12 seed oil content QTL, 6 were previously reported and the rest were novel (Figure 3.19). We were also able to detect a well known QTL for seed protein content on chromosome 15 which were identified in almost all previously conducted GWAS and QTL mapping studies (Hwang, Song et al. 2014; Vaughn, Nelson et al. 2014). Surprisingly, we did not detect the major seed protein QTL known to be located on chromosome 20 (Diers, Keim et al. 1992; Brummer, Graef et al. 1997; Chung, Babka et al. 2003). One possible reason could be that the parents used for creating the SoyNAM population may not carry the rare allele controlling the seed protein content on chromosome 20.

3.8.12 Refining the Candidate Region for Protein on Chromosome 9

SoyNAM method which takes advantage of both QTL and association mapping, uses the power from QTL mapping and resolution from GWAS, generates more precise QTL position. Using this approach we identified 7 known and 6 novel QTL for seed

protein content and 6 known and 6 new QTL for seed oil content with high level of significance. From the known QTL, a QTL located on chromosome 9 was identified to be associated with seed protein content and co-located with previously reported QTL responsible for significant pleiotropic effects on protein and oil (Eskandari, Cober et al. 2013; Lu, Wen et al. 2013).

Previously, Lu et al. (2012) mapped this QTL within 35 Mbp to 41.7 Mbp region (Figure 3.18a). Another study conducted by Eskandari et al. (2013) reported the same QTL within 37 Mbp to 42 Mbp region (Figure 3.18a). In the present study we narrowed down this genomic region to 0.56 Mbp (39.6 Mbp to 40.2 Mbp) (Figure 3.18a). The QTL identified in this study for seed protein content comprised large clusters of markers all in one large LD block as determined by r^2 and D'(Figure 3.18b and 3.18c). We believe that the QTL identified in all three studies is the same QTL that controls seed protein content with significant effect on seed oil content. The genomic regions defined by Eskandari et al. (2013) and Lu et al. (2012), contained several putative model genes. Our refined genomic region contained only four candidate genes: Glyma09g31700, Glyma09g31730, Glyma09g31800, and Glyma09g31870 (Figure 3.18a). One of these genes may likely be the gene that control soybean seed protein content. The QTL identified in this study is believed to have one allele associated with higher seed protein and higher seed oil content and the alternative allele with lower seed protein and lower seed oil content (Hwang, Song et al. 2014). This might be an interesting QTL to those breeding for seed composition (Hwang, Song et al. 2014).

3.8.13 Conclusion

The main objective of this study was to identify QTL controlling seed protein and seed oil contents in SoyNAM population using GWAS. SoyNAM has been the biggest mapping population ever created in the history of soybean breeding program. The aim of developing such big mapping population was to increase the number of recombination events and resolution to identify rare QTL associated with seed protein and oil contents. Using 4118 markers and 5240 RILs, we were able to identify many previously reported and novel QTL for both seed protein and oil contents. We further refined the previously reported genomic region for seed protein content on chromosome 9 and narrowed it down to a genomic region where the causative gene might be located. The novel QTL identified in this study for both seed protein and oil contents could be used by plant breeders as source of genetic variation for further improvement of the soybean seed protein and oil contents.

CHAPTER 4. MAPPING QTL CONTROLLING SOYBEAN SEED SUCROSE AND OLIGOSACCHARIDES IN A SINGLE FAMILY OF SOYBEAN NESTED ASSOCIATION MAPPING (SOYNAM) POPULATION

4.1 Abstract

Soybean meal value of monogastric animals is determined, in part, by sucrose and raffinose family oligosaccharides (RFOs), which include raffinose and stachyose. Among them, only sucrose is desirable, while raffinose and stachyose are the non-digestive carbohydrates that cause flatulence and abdominal discomfort. Developing soybean lines with improved seed sucrose and reduced RFOs will enhance soybean meal value in the market. The objective of this study was to identify quantitative trait loci (QTL) controlling seed sucrose, raffinose, and stachyose content in a set of 140 SoyNAM recombinant inbred lines (RILs), developed from the cross of two elite soybeans lines IA3023 and LD02-4485. A total of 3038 SNP markers from the Illumina SoyNAM BeadChip SNP were used to map the QTL for sucrose and the RFOs, raffinose, and stachyose. ANOVA revealed significant genotypic differences ($P<0.001$) for sucrose, raffinose and stachyose contents across years. Composite interval mapping (CIM) identified three QTL for sucrose content one on chromosome 1 and two on chromosome 3. The QTL on chromosome 1 explained 10% of the phenotypic variation while the two QTL on chromosome 3 each explained 22% phenotypic variation in the sucrose content. A QTL for raffinose content was detected on chromosome 6 and it explained 6% of

phenotypic variation. CIM did not identify any significant QTL for stachyose content. This study identified novel QTL that can be validated for use in developing soybean lines with higher concentrations of sucrose and reduced levels of raffinose and stachyose.

Introduction

Soybean [*Glycine max* (L.) Merrill] belonging to legume family is a miracle and versatile crop that has been widely grown across the world for food, feed, and industrial use. Over the past century, it has been recognized as world's major source of vegetable oil and vegetable protein (Zeng, Chen et al. 2014). Saldivar et al. (2011) classified soybeans into two types: oil beans and food beans according to their end uses. Oil beans are used for vegetable oil and protein production such as defatted soy flour and soy protein concentrate while food beans are converted to various soy products.

Past studies have focused elucidating genetic control of protein and oil contents in soybeans seed but limited information exists for carbohydrates (Maughan, Maroof et al. 2000). Genetic analysis of carbohydrates is challenging because the trait is polygenic and environmental factors like temperature confounds the expression of the trait. For instance, high temperature has been shown to reduce the seed sucrose content. Developments of soybean lines with improved digestibility are crucial to the livestock and broiler chickens that are intensively fed with soybeans. Development of soybean lines with decreased soybean seed stachyose level (< 1%) would result in increase sucrose level, and therefore, would create a more efficient feed source for non-ruminant (Skoneczka, Maroof et al. 2009).

So far a number of quantitative trait loci (QTL) associated with soybean seed soluble sugar have been identified. The first QTL controlling sucrose in soybeans seed

was identified by Maughan et al. (2000) on chromosomes 5, 7, 8, 13, 15, 19, and 20. Kim et al (2006) reported four QTL located on chromosomes 2, 11, and 19, highly associated with seed sucrose content in RILs population developed from the cross of *'Keunolkong' 9 'Shinpaldalkong'*(Kim, Klein et al. 2005; Wang, Chen et al. 2014). Kim et al. (2006) also mapped two QTL associated with seed sucrose content on chromosomes 12 and 16. Stachyose and high sucrose QTL were also mapped on chromosome 6 in two separate QTL mapping studies conducted by Skoneczka et al (2009) in population derived from the cross of *PI 87013 X PI 200508* and *PI243545 × PI200508.* Recently (Zeng, Chen et al. 2015) mapped two QTL for stachyose content on chromosome 10 and 11 in RILs population of the Osage cultivar derived from 'Hartz 5545' x 'KS4895'. These literature surveys revealed multiple QTL, reaffirming the polygenic nature of these traits, and the need to for further dissection to give more insights into the underlying mechanisms. The present study utilized one family of SoyNAM population to identify QTL controlling sucrose and oligosaccharides.

4.2 Materials and Methods

4.2.1 Plant Material

A total of 140 recombinant inbreed lines (RILs) from the cross between two elite soybean lines IA3023 and LD02-4485 were used in this study. This RILs population is subset of the 40 families of SoyNAM population. The SoyNAM mapping population was developed by mating IA3023, a high yielding Iowa State variety, with 40 different high yielding elite and exotic soybean lines, followed line derivations through single seed descent (SSD) method to generate F_5 lines. The SoyNAM project was developed under a

collaborative umbrella of several universities with overall objectives of mapping genes/QTL and other genetic factors that controls yield potential, agronomic traits, and seed composition traits in soybeans. For Further information about SoyNAM project please refer to Soybase http://soybase.org/SoyNAM/. We selected family 12 of the SoyNAM for use in the present study based on results of a preliminary screen of the 40 SoyNAM founders for sucrose, raffinose, and stachyose contents from two locations, Indiana, and Illinois using High Performance Liquid Chromatography (HPLC). The HPLC data showed that the two parents (D02-4485 and IA3023) of family 12 had the best contrast for high sucrose (Table 4.1), therefore, the 140 RILs developed from the cross of these parents were selected for QTL mapping.

Table 4.1. SoyNAM parent screened for percent high sucrose content across two locations using HPLC.

Genotype	Indiana 2012	Indiana 2013	Illinois 2012	Illinois 2013	$\bar{x}$
[#1]IA3023	**5.1**	**4.4**	**6.2**	**5.6**	**5.3**
4J105-3-4	5.2	5.3	5.0	4.9	5.1
5M20-2-5-2	6.4	5.3	5.2	5.2	5.5
CL0J095-4-6	6.0	5.8	6.0	5.0	5.7
CL0J173-6-8	7.9	6.1	7.8	6.8	7.2
HS6-3976	6.6	4.3	7.7	6.2	6.2
Prohio	5.8	5.1	5.3	4.8	5.3
LD00-3309	6.6	6.0	6.7	5.7	6.3
LD01-5907	8.5	7.7	6.8	7.3	7.6
[#2]LD02-4485	**8.7**	**7.7**	**7.5**	**7.4**	**7.8**
LD02-9050	8.7	5.6	7.1	6.9	7.1
Magellan	8.1	6.8	7.1	6.2	7.1
Maverick	7.4	6.9	7.0	6.3	6.9
S06-13640	8.6	7.3	7.3	6.0	7.3
NE3001	7.4	7.7	7.3	7.5	7.5
Skylla	7.6	8.2	6.3	6.9	7.3
U03-100612	5.9	5.9	5.2	6.1	5.8
LG03-2979	5.0	4.8	6.6	5.2	5.4
LG03-3191	7.4	6.0	6.8	5.1	6.3
LG04-4717	5.5	6.2	5.0	4.7	5.4

Table 4.1 continued					
LG05-4292	7.1	5.9	7.3	6.9	6.8
LG05-4317	5.4	5.4	5.9	4.3	5.3
LG05-4464	7.5	6.3	6.9	5.8	6.6
LG05-4832	7.8	7.4	5.8	5.7	6.7
LG90-2550	6.4	5.7	5.9	4.4	5.6
LG92-1255	6.3	5.3	5.8	5.1	5.6
LG94-1128	7.0	5.6	5.3	5.2	5.8
LG94-1906	7.5	6.2	6.6	6.6	6.7
LG97-7012	7.6	5.3	6.1	5.1	6.0
LG98-1605	5.9	5.8	6.3	5.5	5.9
LG00-3372	7.4	7.0	6.5	6.4	6.8
LG04-6000	4.5	5.9	7.1	6.6	6.0
PI398881	7.8	6.9	7.2	7.7	7.4
PI427136	7.4	5.9	6.7	6.1	6.5
PI437169B	5.7	4.4	5.5	4.7	5.1
PI507681B	6.7	5.1	6.6	7.4	6.5
PI518751	7.1	6.0	6.6	6.2	6.5
PI561370	5.5	3.8	5.4	4.4	4.8
PI404188A	7.4	6.0	7.4	6.6	6.9
PI574486	7.5	5.7	6.9	7.5	6.9

[#1] and [#2] were the two contrasting parents chosen for this study; $\bar{x}$ represents mean sucrose content

4.2.2 Experimental Design

The experiment followed a modified augmented design used in the larger SoyNAM population. The two years (2012 and 2013) trials included the 140 RILs planted in two rows plot of 80cm length at Purdue University Agronomy Center for Research and Education (ACRE).

4.2.3 Phenotype Data

4.2.3.1 Soluble Sugars Determination

We sampled 10 healthy seeds per genotype from the two years' trials and sent it to Molecular Genetics and Soybean Genomics Laboratory (Nguyen Laboratory) at the

University of Missouri for quantification of sucrose, raffinose, and stachyose. The sugar contents of sucrose, raffinose, and stachyose were determined for each sample using HPLC protocol described by Valliyodan and Shi et al. 2015. The HPLC method used in this study has been equipped with an evaporative light scattering detector (ELSD) that can separate, identify, and quantify several sugars, including sucrose, raffinose, and stachyose. This method has been successfully used to quantify sucrose and oligosaccharides in soybeans (Valliyodan, Shi et al. 2015).

4.2.4 Genotype Data

A subset of 4118 SNP markers from the Illumina SoyNAM BeadChip SNP array were selected for the QTL mapping. The markers were initially tested for segregation distortion in "R/qtl package" (Broman, Wu et al. 2003) , with an adjustment for multiple testing using a Bonferroni correction at alpha=0.05. A total of 1080 SNPs were found to be distorted and were dropped from the analysis. Finally, 3038 SNP markers were used for QTL mapping.

4.3 Statistical Analysis

4.3.1 Phenotypic Assessment

Best linear unbiased predictors (BLUPs) were estimated for sucrose, raffinose and stachyose using 'lme4' package in 'R'(Bates, Maechler et al. 2014). The BLUP values were calculated using the following model:

$$\mathrm{BLUP} = \mu + \mathrm{Xb} + \mathrm{Bu} + \varepsilon$$

where μ is the grand mean; b is random effect of lines; u is fixed effect of the year; ε is the residual.

Analysis of variance was conducted for each trait based on the following linear mixed model using 'lme4' package in 'R'.

$$A_{ij} = \mu + G_i + Y_j + e_{ij},$$

where A_{ij} is the observed value of the i^{th} genotype in the j^{th} year, μ is the general mean, G_i and Y_j are the effects of the genotype, and year, and e_{ij} is the residual effect. There were not enough degrees of freedom to estimate G x Y interaction because individual year trial was not replicated.

4.3.2 Repeatability Estimation and Correlation Determination

The estimation of repeatability is necessary for understanding the response to selection (van Kleunen and Ritland 2005). Repeatability of each trait was estimated on a line mean basis across years using the following equation:

$$\text{Repeatability} = R = \sigma^2 g / [(\sigma^2 g + (\sigma^2 e)/r]$$

where R represents repeatability; $\sigma^2 g$ is the genetic variance for lines; $\sigma^2 e$ is the residual variance; r is the year variance. The R package lme4 was used to estimate the variance components based on REML algorithm.

Correlations among the three carbohydrates sucrose, Raffinose, and Stachyose was calculated across years using Pearson's correlation coefficients ® with following equation using psych R package.

$$r_{(x,y)} = \text{(Pearson's) coefficient of correlation} = COV(x,y) / \sqrt{[\ \sigma^2 (x) \times \sigma^2 (y)]}$$

where $r_{(x,y)}$ is the Pearson's correlation coefficients; COV(x, y) is the covariance between the two traits x and y; σ^2 is the variance for traits x and y.

4.3.3 QTL Analyses

QTL analysis was conducted by composite interval mapping (CIM) for all three traits in *QTL Cartographer* v. 2.5 (Zeng 1994; Wang, Basten et al. 2006). The CIM was performed following a standard model 6. The five most significant background markers for inclusion in the CIM model were selected by backward stepwise regression. Walking speed was set at 2cM and a window size of 10 cM. A total of 1000 permutations were performed for each trait using average data across years to establish genome-wide LOD significance threshold at a 0.05 probability (Churchill and Doerge 1994). QTL were considered to exist only at positions where a LOD score exceeded the corresponding significance threshold (Churchill and Doerge 1994). The percentage of variation explained (R^2) for all significant QTL were determined at their peak LOD values.

4.4 Result

4.4.1 ANOVA, Heritability Estimates, and Correlation

Descriptive statistics for all three carbohydrates are presented in Table 4.2. Mean seed sucrose and raffinose content differed between the two years, while stachyose content remained nearly the same (Figure 4.1, Table 4.2). This observation suggested that seed sucrose and raffinose were more influenced by environmental variation than stachyose.

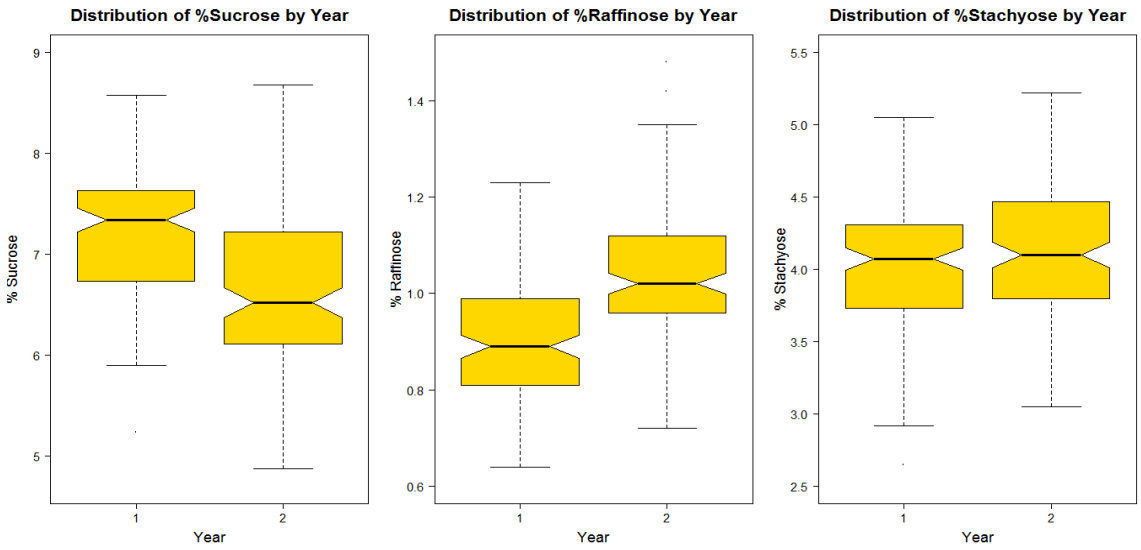


Figure 4.1. Distribution of sucrose, raffinose, and stachyose content by year in 140 RILs.

Frequency distribution showed that all three traits are normally distributed across years, indicating that the seed contents of the three traits are controlled by many genes (Figure 4.2). The range of sucrose, raffinose and stachyose contents in the mapping population exceeded the mean values of the two parents, suggesting the presence of transgressive segregation for these traits. The mean seed sucrose content for parent LDO2-4485 was higher than the population mean, while that of parent IA3023 was below the population mean (Figure 4.2). For raffinose, the mean values were close together for the two parents, indicating less parental contrast for this trait. The distribution for stachyose on the other hand showed parent IA3023 to have a mean value nearer the population mean compared to parent LDO2-4485 (Figure 4.2).

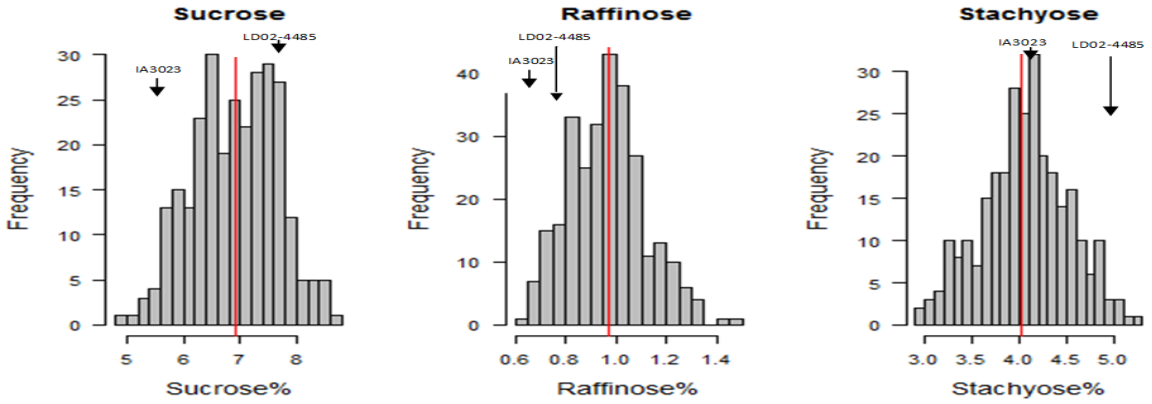


Figure 4.2. Frequency distribution of seed sucrose, raffinose, and stachyose content in a population of 140 RILs derived from the cross of IA3023 and LD02-4485.
The vertical red lines represent overall mean value for each trait.

Percent of sucrose content in 2012 ranged from 5.2-8.6 with a mean of 7.2 and standard deviation of 0.7, while in 2013 it ranged from 4.9-8.7 with mean of 6.6 and standard deviation of 0.7. The coefficient of variation for sucrose was 11% in 2013, and 9% in 2012. These results revealed a higher variation for sucrose content in 2013 than in 2012. Percent of raffinose content in 2012 ranged from 0.6-1.2 with mean of 0.9 and standard deviation of 0.1, whereas in 2013 it ranged from 0.7-1.6 with mean of 1.0 and standard deviation of 0.1. Coefficient of variation for raffinose content in 2013 was higher than that in 2012. Percent of stachyose content in 2012 ranged from 1.7-5.0 with mean of 4.0, standard deviation of 0.4, and a CV of 10, while in 2013 it ranged from 3.0-5.2 with mean of 4.1, standard deviation of 0.5 and a CV of 12 consequently, variation in stachyose content in 2012 was higher compared to that in 2013 (Table 4.2).

Table 4.2. Summary statistics for sucrose, raffinose and stachyose measured over two years at ACRE Indiana.

Trait	Year	N	Mean	Std	Range	Difference	CV%
Sucrose	(2012)	142	7.2	0.7	5.2-8.6	3.4	9
	(2013)	142	6.6	0.7	4.9-8.7	3.8	11
Raffinose	(2012)	142	0.9	0.1	0.6-1.2	0.6	11
	(2013)	142	1.0	0.1	0.7-1.6	0.9	10

Table 4.2 continued

Stachyose	(2012)	142	4.0	0.4	1.7-5.0	3.3	10
	(2013)	142	4.1	0.5	3.0-5.2	2.2	12

CV= Coefficient of variation; N=number of observation, Std= standard deviation.

Analysis of variance revealed significant differences for sucrose, raffinose and stachyose contents among the genotypes and the years (Table 4.3). The genetic component accounted for 50.3%, 49.8%, and 57.9% of total variation in sucrose, raffinose, and stachyose, respectively. This suggested that variability for these traits are largely under genetic control and are therefore amendable to selection. Variation in seed stachyose content for years was much smaller (1.3%) compared to that of sucrose (14.5%), and raffinose (19.6%) (Table 4.3), corroborating the earlier observation in Figure 4.1 which suggested that stachyose is more stable to environmental fluctuations. Repeatability, for sucrose, raffinose, and stachyose content were 30%, 38% and 30%, respectively (Table 4.3). Just as revealed by ANOVA, the observed moderate to high repeatability values suggested that genetic variation for the three traits are repeatable over time and can be exploited for improvement.

Table 4.3. Analysis of variance for soybean seed sucrose, raffinose, and stachyose content of 142 genotypes grown in two Indiana environments for two years (2012 and 2013).

		Sucrose			Raffinose			Stachyose		
Source	DF	SS	MS	R^2	SS	MS	R^2	SS	MS	R^2
Genotypes	141	77.2	0.55**	50.3	3.27	0.024**	49.8	35.3	0.25*	57.9
Year	1	22.3	22.3**	14.5	1.29	1.29***	19.6	0.8	0.8*	1.3
Residuals	141	53.7	0.38		2.0	0.014		24.8	0.17	
Total	283	153.2			6.56			60.9		
LSD		1.24			0.23			0.82		
SED		0.62			0.11			0.41		
R		0.30			0.38			0.30		
σ^2G			0.082			0.005			0.037	

*, **, *** represents the significant level of 0.05, 0.01 and 0.001, respectively; R = repeatability; R^2= phenotypic variation explained (%).

Correlation analysis showed that stachyose was significantly and positively correlated with sucrose ($r = 0.33$, $P \leq 0.001$) and raffinose ($r = 0.28$, $P \leq 0.001$), yet sucrose and raffinose were positively but weakly correlated ($r = 0.09$, $P \leq 0.027$) (Table 4.4). The observed positive correlations among these traits suggested that these traits improving one of them may simultaneously enhance the others.

Table 4.4. Correlation between the three carbohydrate contents in soybean seeds of the 142 genotypes.

Correlation	r	r^2	P-value
Sucrose vs. Raffinose	0.09ns	0.008	$P \leq 0.027$
Sucrose vs. Stachyose	0.33***	0.108	$P \leq 0.001$
Raffinose vs. Stachyose	0.28***	0.078	$P \leq 0.001$

*, **, *** represents the significant level of 0.05, 0.01 and 0.001, respectively; ns, not significant.

4.4.2 QTL Mapping

Composite interval mapping (CIM) identified four QTL affecting seed sucrose and raffinose content (Figures 4.3 and 4.4). A summary of significant QTL is presented in Table 4.5. Of the QTL identified for sucrose, one was located on chromosome 1 at genomic position 22.8 cM and two were located on chromosome 3 at 0.63 cM and 8.15 cM positions, respectively (Table 4.5 and Figure 4.3). The QTL located on chromosome 1 accounted for 10% of the phenotypic variance and the two QTL on chromosome 3 each explained 22% of phenotypic variation in the sucrose content (Table 4.5 and Figure 4.3). These QTL also had large additive genetic effects (Table 4.5). We did not identify any significant QTL for stachyose content.

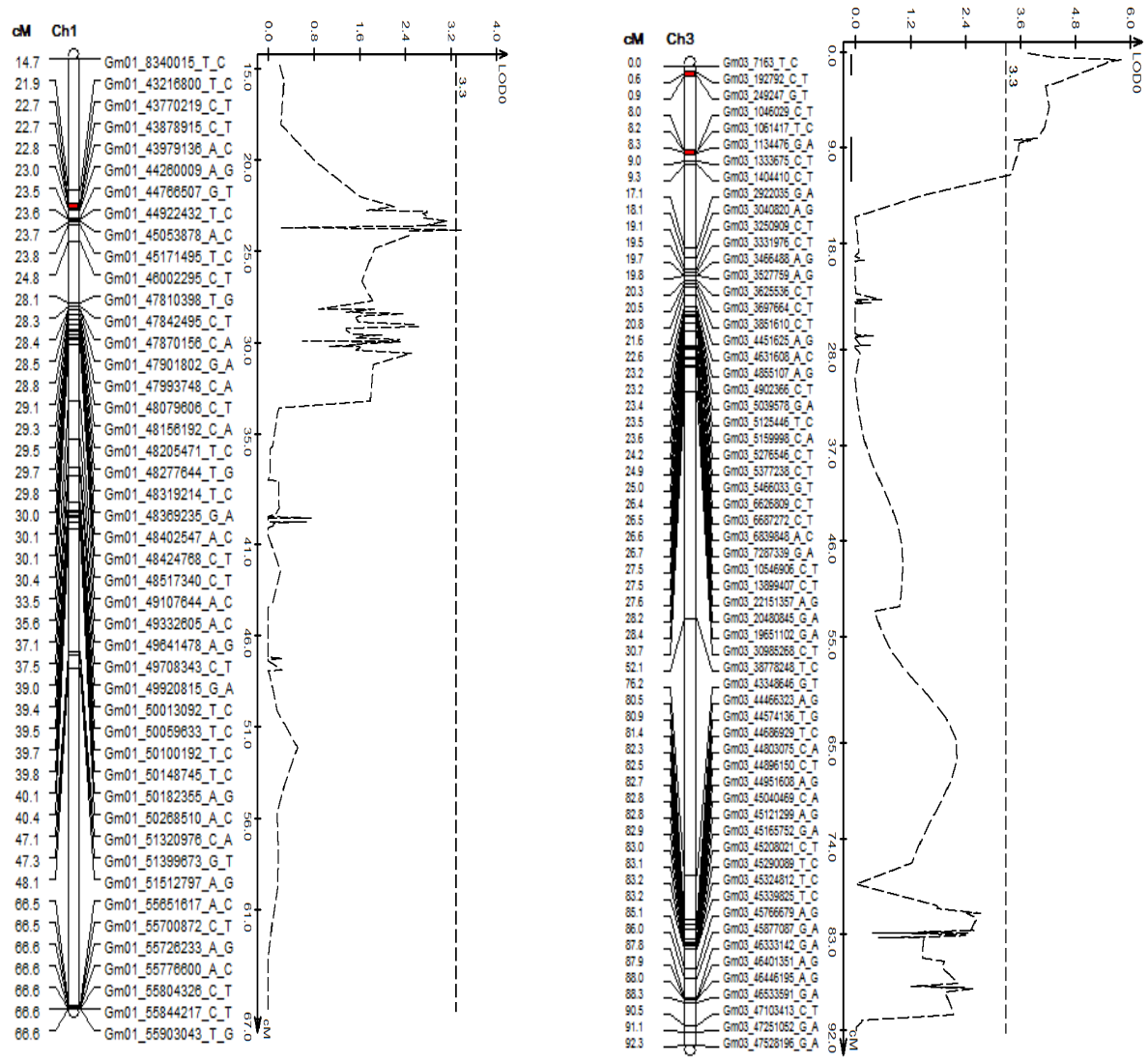


Figure 4.3. Linkage map and plots showing location of the putative sucrose QTL on chromosome 1 shown on left and chromosome 3 on the right. Highlighted in red are the locations of putative QTL controlling seed sucrose content.

Dashed vertical lines show threshold value based on 1000 permutation at probability of 0.05.

For seed raffinose content, we identified a significant (P=0.05, threshold=3.2) QTL on chromosome 6 at 69.15 genomic position and explained 6% phenotypic variation in the raffinose content (Figure 4.4 and table 4.5).

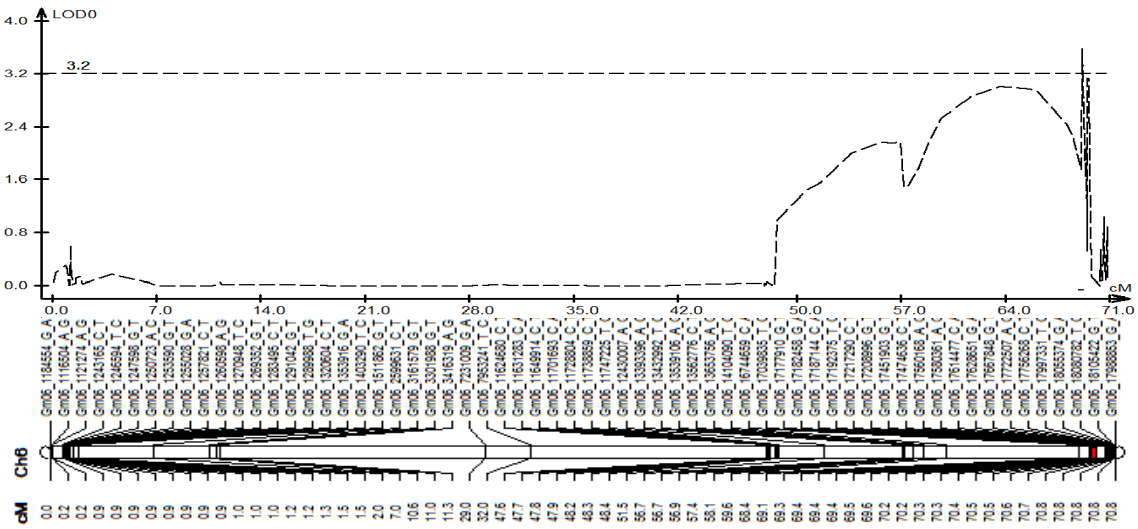


Figure 4.4. Linkage map and plots showing location of the putative raffinose QTL on chromosome 6.

Highlighted area in red in the linkage map is the locations of putative QTL controlling seed raffinose content. Dashed horizontal line shows threshold value based on 1000 permutation at probability of 0.05**.**

Table 4.5. Quantitative trait loci (QTL) associated with seed sucrose and raffinose contents.

Trait	QTL	Chromosome	Position (cM)	Additive effect	LOD	R^2
Sucrose	Gm01_43979136_A_C	1	22.8	14	3.5	10
	Gm03_192792_C_T	3	0.63	23	5.6	22
	Gm03_1061417_T_C	3	8.15	17	3.9	22
Raffinose	Gm06_17204660_T_G	6	69.51	-0.6	3.4	6

R^2, represents phenotypic variation explained (%).

4.5 Discussion

4.5.1 Phenotype Data

ANOVA result revealed that genotypic difference for sucrose and raffinose were significant at P<0.001, while that for stachyose content was significant at P <0.05, indicating that variation within genotype and years for sucrose and raffinose contents were higher than that of stachyose (Table 4.3 and Figure 4.1). Most research conducted

on soybeans seed composition reported that variables such as temperature, drought stress, planting date, genetics, environment and the interaction of genetic and environment affect soybeans seed composition (Dornbos Jr and Mullen 1992; Piper and Boote 1999; Specht, Chase et al. 2001). Environmental variation often confounds true genotypic value, making it challenging to make selection based on mean values.

4.5.2 Repeatability and Correlation

Repeatability for raffinose and stachyose contents reported in this study fell in the range reported by other studies (Jaureguy, Chen et al. 2011). Jaureguy et al. (2011) reported heritability (79%), (46%), and (73%) for sucrose, raffinose and stachyose, respectively. Cicek et al. (2011) also reported heritability (72%), (42%), and (66%) for sucrose raffinose and stachyose. The low sucrose repeatability registered might be due to environmental factors since sucrose is quantitative traits and quantitative traits are easily affected by environment factors due the involvement of many genes/QTL in controlling these traits. Indiana, experienced severe drought in year 2012 and this extreme environmental anomaly could have affected the estimates of repeatability. Overall, the repeatability values were moderate and thus predictable, suggesting that genetic improvement for these traits are possible, but environmental variance must be taken in to consideration when making selection. That is, evaluations have to be conducted across multiple environments so as to obtain an accurate phenotypic assessment for effective selection.

Correlation coefficient observed for sucrose and raffinose in this study is partially in agreement with results from previous studies. Cicek et al. (2011) reported strong

positive correlation between sucrose and raffinose while the correlation between sucrose and stachyose was strong (r=-35) and negative. Another study conducted by Huhn et al. (2003) reported a positive correlation between raffinose and stachyose but significant inverse correlation of these two with sucrose content.

It has been observed that there are variations in direction and magnitude of correlation coefficient among these three carbohydrate traits. These observations suggest that the relationships among these traits are population specific and environment dependent (Jaureguy, Chen et al. 2011).

4.5.3 QTL Analyses of Three Carbohydrates Traits

Previous studies have reported QTL for carbohydrate traits in soybean on multiple chromosomes (Maughan, Maroof et al. 2000; Kim, Kang et al. 2006). The present study, utilized the power of traditional CIM mapping methods to precisely map novel regions associated with sucrose, raffinose and stachyose. We identified a total of four QTL for sucrose and raffinose (Figure 4.3, 4.4 and Table 4.5). Among the four QTL detected, three QTL were for seed sucrose content and were located on chromosome 1 and chromosome 3 (Figure 4.3 and Table 4.5). These two QTL on chromosomes 1 and 3 explained 10% and 22% of phenotypic variation, respectively. A QTL controlling seed sucrose content on chromosome 3 was reported by (Akond, Liu et al. 2015) but at different genomic position. Therefore, we have probably detected a new allele for sucrose content on chromosome 3. The other QTL located on chromosome 1 associated with seed sucrose content was also novel. Additionally, we mapped a new QTL for seed raffinose

content on chromosome 6, and this QTL accounted for less (7%) variation in the observed phenotype.

4.6 Conclusion

The present study revealed significant variation among soybean genotypes for the three carbohydrate traits studied. Moderate repeatability was observed for these traits which indicated that the traits were predictable and that genetic improvement are possible. We uncovered four novel regions that were significantly associated with seed sucrose and raffinose content. Given the importance of these carbohydrate traits in soybean nutritional quality, our study provides more insight into the underlying genetics of these traits, and the opportunity for accelerated improvement through marker-assisted breeding.

CHAPTER 5. GENOTYPE BY ENVIRONMENT INTERACTION AND STABILITY ANALYSIS FOR PROTEIN AND OIL IN SOYNAM PARENTS

5.1 Abstract

Soybean [*Glycine max* (L.) Merrill], which has the highest protein content of all food crops, is the world's leading source of protein and oil. The objectives of this study were to determine the stability, adaptability and the magnitude of GEI for seed protein and oil contents, in 40 genetically diverse SoyNAM parental genotypes grown across different environments. Multi-environment analysis for both seed protein and oil contents revealed significant ($P<0.001$) genotype (G), environment (E) and genotype by environment interactions (GEI). The genetic component of variation for seed protein and oil content was (40.1%) and (29.1 %) while the environments explained (28.21%) and (30. %), variation in the two traits, respectively. Phenotypic and genotypic correlation between protein and oil were -0.59, and -0.66, respectively. GGE-biplot analysis revealed that selection of the SoyNAM parents for seed protein and oil contents based on mean and stability across environments was appropriate. Genotypes LG92-1255, CL0J173-6-8, PI398881, PI561370, Prohio, PI427136, LG03-3191, PI507681B for seed protein content and genotypes LG03-2979, U03-100612, Prohio, LD02-4485, IA3023, LG04-4717, LG92-1255 for seed oil content were identified as the most stable and desirable while LG94-1128 and 5M20-2-5-2 for seed protein content and NE3001 and LG05-4317 for seed oil content were unstable even though high yielding.

5.2 Introduction

Soybean [*Glycine max* (L.) Merrill, 2n=40] is an important field crop that has been grown worldwide not only for feed and food but also for industrial purposes. Protein and oil are the most important economic constituents of soybean seed composition (Piper and Boote 1999). The two traits play a key role in the economy of most soybean-growing countries.

One of the goals of most plant breeding programs is to develop plant varieties that are adapted to a broad population of environments. To develop such plant varieties, plant breeders and agronomists usually conduct multi-environment trials (METs) to select lines with the greatest yield potential, widest adaptation, and stability over a wide array of environments. The interpretation of MET results is often confounded by significant genotype by environment interaction (GEI), which challenges effective selection of genotypes. GEI occurs since gene expression is subject to modification by environmental conditions; therefore, phenotypic response of genotypes differs with environments (Kang and Gauch 1996). The difference in phenotypic expression is mainly due to environmental factors such as temperature, planting date, soil type and precipitation, which may vary from location to location and year to year. The inconsistent phenotypic response of genotypes due to these factors is called genotype by environment interaction (Baker 1988). Baker et al. (1988) defined GEI as the failure of genotypes to achieve the same relative performance in different environments due to different environmental factors.

Development of new soybeans cultivar with improved adaptation for the trait of interest requires the knowledge of GEI. GEI analysis can be used to understand genotypic

stability across environments; a genotype is considered stable if its variation among environment is small (Farshadfar, Poursiahbidi et al. 2012).

Improving protein and oil content of soybean lines is facilitated by a detailed evaluation of breeding populations in several environments. This multi-environment assessment allows for quantification of the magnitude of variances that are genetic in nature compared to those due to environments. The presence of significant GEI for quantitative traits such as protein and oil can seriously limit the feasibility of selecting superior genotypes for wider environments (Gurmu, Mohammed et al. 2009). Several studies have reported a significant GEI in soybeans for seed protein and seed oil contents.

Miladinovic et al. (2006) reported that soybeans seed oil and protein contents grown at similar environments and latitudes had a significant difference. Sogut et al. (2006) reported a significant GEI for soybean seed protein and seed oil content between years and environments. It has been noted that the same soybean cultivar grown in different years and different locations could vary significantly in seed composition (Phansak 2010). Helms et al. (1998) found that the delay in planting date would increase soybeans seed protein concentration.

Due to its economic importance, soybean breeders and agronomists have endeavored to improve seed protein and oil concentration simultaneously, however, their concurrent improvement is difficult because of their negative correlations (Wilcox and Cavins 1995; Cober and D Voldeng 2000; Chung, Babka et al. 2003; Panthee, Pantalone et al. 2005; Phansak 2010).

Having the knowledge of the magnitude of GEI and stability analysis is important for understating the response of different genotypes to different environments (Gurmu,

Mohammed et al. 2009). This kind of knowledge help in identification of stable and widely adapted and unstable but specifically adapted genotypes (Gurmu, Mohammed et al. 2009).

The objective of this study was to determine the stability, adaptability and magnitude of Genotype by Environment Interaction (GEI) for seed protein and oil contents in the genetically diverse 40 NAM parents across environments.

5.3 Material and Methods

5.3.1 Plant Material

Plant materials used in this study include 40 parental genotypes that were used to create the SoyNAM population. Information about agronomic features and pedigrees of the genotypes are present in Table 5.1. Detailed information about the SoyNAM project can be accessed through the link http://soybase.org/SoyNAM/.

Table 5.1. Agronomic feature of the 40 genotypes (SoyNAM parents) used in this study. http://soybase.org/SoyNAM/.

Cultivar	Pedigree	Agronomic feature	Origin
IA3023	Dairyland DSR365 X Pioneer P9381	High yielding	Iowa
4J105-3-4	CLOJ173-6-2 X WW115926	High yielding	Purdue University
5M20-2-5-2	CLOJ173-6-8 X (OD032-3118 x LG00-6293)	High yielding	Purdue University
CL0J095-4-6	CX1705R-108 X Dwight	High yielding	Purdue University
CL0J173-6-8	Kottman X Dwight	High yielding	Purdue University
HS6-3976	HS98-7826 (2) X PI 399073	High yielding	Ohio State
Prohio	HC94-81PR X Asgrow A2506	High yielding	Ohio State University
LD00-3309	Maverick X Dwight	High yielding	University of Illinois
LD01-5907	Ina X IA3010	High yielding	University of Illinois
LD02-4485	M90-184111 X IA3010	High yielding	University of Illinois
LD02-9050	Macon X LS93-0375	High yielding	University of Illinois
Magellan	Sherman X Harper	High yielding	University of Missouri
Maverick	LN86-4668 X Resnik	High yielding	University of Missouri
S06-13640	LG99-11986 X S38-T8	High yielding	University of Missouri
NE3001	Colfax X A91-701035	High yielding	University of Nebraska

Table 5. 1 continued

Skylla	Dairyland DSR217 X Northrup King S19-90	High yielding	Michigan State University
U03-100612	MSPB6S4 X Pioneer P93B82	High yielding	University of Nebraska
LG03-2979	Rend X LG95-258	Diverse ancestry	USDA-ARS
LG03-3191	LG96-1854 X LG96-3159	Diverse ancestry	USDA-ARS
LG04-4717	LG98-5579 X A98-980047	Diverse ancestry	USDA-ARS
LG05-4292	LG94-4667 X LG97-9226	Diverse ancestry	USDA-ARS
LG05-4317	LG94-4667 X LG98-1445	Diverse ancestry	USDA-ARS
LG05-4464	LG97-8984 X A98-884037	Diverse ancestry	USDA-ARS
LG05-4832	LG98-5579 X A98-980047	Diverse ancestry	USDA-ARS
LG90-2550	LG82-8224 X LG82-8195	Diverse ancestry	USDA-ARS
LG92-1255	LG84-1291 X Asgrow A3127	Diverse ancestry	USDA-ARS
LG94-1128	LG85-3343 X LG87-1991	Diverse ancestry	USDA-ARS
LG94-1906	PI 468377 X Asgrow A3205	Diverse ancestry	USDA-ARS
LG97-7012	LG89-1525 X Asgrow A3322	Diverse ancestry	USDA-ARS
LG98-1605	LG88-8958 X LG89-771	Diverse ancestry	USDA-ARS
LG00-3372	PI 561319 X PI 574477	Diverse ancestry	USDA-ARS
LG04-6000	HS93-4118 X LG97-9912	Diverse ancestry	USDA-ARS
PI 398881	Introduction	Drought tolerance	South Korea
PI 427136	Introduction	Drought tolerance	South Korea
PI 437169B	Introduction	Drought tolerance	Russia
PI 507681 B	Introduction	Drought tolerance	China
PI 518751	Ns-Kasna X Beeson	Drought tolerance	Serbia
PI 561370	Introduction	Drought tolerance	China
PI 404188A	Introduction	Drought tolerance	China
PI 574486	Introduction	Drought tolerance	China

5.3.2 Experimental Design

The SoyNAM experiment was grown at four locations (Iowa, Nebraska, Indiana, and Illinois) in 2012 and 2013. The forty parental lines were used as replicated checks within the blocks of a Modified Augmented Design. Because the lines were replicated four times in each environment, this study extracted the seed protein and oil content data of the 40 parent lines from the overall SoyNAM experiment and analyzed the data as randomized complete block design.

5.3.3 Phenotypic Data

Seed content of protein and oil were determined from 300 g samples of whole grain per plot over the eight environments using near infrared spectroscopy (NIR) on a Perten DA7200 diode array instrument (http://www.perten.com/), conducted by Jim Specht at the University of Nebraska.

5.3.4 Multi-Environment Trial Assessment of Seed Protein and Oil Content

The purpose of the MET analysis was to determine the response of the 40 parental lines to varying environments and to determine the magnitude of genotype by environment interaction (GEI) for seed protein and oil. Year in each location was considered as a different environment with a total of eight environments (two years x four locations). Detailed environmental information for the testing environment is provided in Table 5.2.

Table 5.2. Characteristic features of study environments.
www.usclimatedata.com/climate/

Environments	AT (^{o}C)	AAR (mm)	Soil type	EL(meter)	Coordinate
IA-Ames-2012(E1)	17.3	617	Loam	334	42° 1' 50.8" N, 93° 37' 54.8" W
IA-Ames-2013(E2)	16.2	681	Loam	334	42° 1' 50.8" N, 93° 37' 54.8" W
IL-Flannagan-2012(E3)	17.7	813	Silt loam	219	40° 6' 38.1" N, 88° 12' 26.1" W
IL-Blackberry-2013(E4)	17.0	867	Silt loam	219	40° 6' 38.1" N, 88° 12' 26.1" W
IN-ACRE-2012(E5)	17.3	775	Silt clay-loam	217	40° 25' 33.1" N, 86° 54' 29.0" W
IN-ACRE-2013(E6)	16.5	908	Silt clay-loam	217	40° 25' 33.1" N, 86° 54' 29.0" W
NE-Clay center-2012(E7)	17.7	632	Silt loam	533	40° 31' 18.0" N, 98° 3' 19.1" W
NE-Clay center-2013(E8)	16.9	766	Silt loam	533	40° 31' 18.0" N, 98° 3' 19.1" W

IA, IL, IN, and NE, indicate, Iowa, Illinois, Indiana, and Nebraska, respectively; AAR=Average Annual Rainfall; AT=Average Temperature; AEL=Altitude Elevation; mm=millimeter, ^{o}C= Celsius.

Prior to stability and adaptation assessment, it is essential to perform a combined analysis of variance to determine the presence of GEI from the replicated genotypes at

various environments. A signifiant F-test implies that GEI exists and the mean performance of lines across environments varies from location to location.

5.4 Statistical Analyses

5.4.1 Analysis of Variance

Analyses of variance (ANOVA) were conducted for seed protein and oil composition using combined data from all locations and years. The ANOVA was performed with linear mixed model using GLM procedure in SAS 9.4 (SAS Institute, 2014). In the mixed model genotype and location were considered fixed effects while replication and year were random. The mixed model used to estimate variance components was as follows:

$$A_{iljk} = \mu + G_i + Y_l + E_j + R_k + (G_i \times Y_l) + (G_i \times E_j) + GYE_{ilj} + e_{ijlk}$$

where A_{iljk} is the observed value of the i genotype in Y^{th} year in the in k^{th} block in j^{th} location; μ is the grand mean; G is the genotypic effect; Y is the year effect, E is the location effect, R_k is the block effect; $(G_i \times Y_l)$, $(G_i \times E_j)$, GYE_{ilj}, represent interaction between genotype and year, interaction between genotypes and location, and interaction between genotype year and location, respectively. e_{iljk} is the residual effect.

5.4.2 Correlation Determination

The estimation of correlation is fundamental to the success of any plant breeding program since it provides information regarding response to selection (van Kleunen and Ritland 2005). Phenotypic correlation between protein and oil was calculated across

location using Pearson's correlation coefficients (r) with the following equation using the psych R package.

$r_{(x,y)}$ = (Pearson's) coefficient of correlation = COV(x,y) /√[σ^2 (x) × σ^2(y)]

Where $r_{(x,y)}$ is the Pearson's phenotypic correlation coefficients; COV(x,y) is the phenotypic covariance between the two traits x and y; σ^2 is the phenotypic variance for traits x and y.

The genetic correlation represents the additive genetic effect that is shared between a pair of traits. We determined genotypic correlations for each trait across environment. The genotypic correlation was determined based on line mean basis using REML in R software using the following formula:

$$rG = \frac{Covx, y}{\sqrt{(\sigma^2 x)(\sigma^2 y)}}$$

where rG represents genetic correlation; Cov represents covariance of the two traits; x represents the first trait and y represent the second trait; σ^2 represents variance. The dispersion of phenotypic variation (coefficient of variation) in the two traits was estimated with the following formula proposed by (Johnson, Robinson et al. 1955) as;

$$CV = [\sqrt{(\sigma^2_P / \bar{x})}] \times 100.$$

where $\bar{x}$ is the grand mean and σ^2_P is phenotypic variance.

5.4.3 MET Analyses

Prior to GEI assessment, combined analysis of variance (ANOVA) was conducted to determine the existence and magnitude of the GEIs for seed protein and oil contents. Significant F-test indicated the presence of GEIs and thus additional statistics was

calculated to determine the stability of each of the 40 genotypes over the eight environments (two years x four locations).

Stability analysis is important, particularly when the objective of a breeding program is to select genotypes for a wider array of environments.

The SoyNAM project developed from the parental genotypes evaluated in this study was conducted in four states, Iowa, Nebraska, Indiana, and Illinois, for two years (2012 and 2013) for seed protein and oil contents under various environmental conditions. Since seed protein and oil contents of soybeans are quantitatively inherited traits, therefore their evaluation under diverse environmental conditions maximize the chance of the interaction of genotypes with environment (Balestre, Santos et al. 2010). To evaluate the genotype by environment interaction, breeders must use a tool that can efficiently and accurately measure the response of these genotypes under different environments (WeiKai, Hunt et al. 2001).

Best linear unbiased predictor (BLUPs) for protein and oil contents was computed using the lme4 'R' package based on the mixed model:

$$y = Xb + Zu + e$$

where y is the phenotypic value (protein/oil), b is the fixed effects (block), u is the random effects (genotype and environment), e is the residual, while $X\ and\ Z$ are the incidence matrices. The BLUPs were estimated for each trait and were used to rank the 40 SoyNAM parental genotypes. The ranked genotyped based on BLUPs were then subjected to GGE analysis.

Several statistical packages are available to analyze multi-environment trial data, however, the stability analyses for seed protein and oil contents in this study were

performed with genotype plus genotype by environment interaction (GGE) biplot using 'R' statistical package, 'GGEBiplotGUI'(Frutos, Galindo et al. 2014; Lian and de los Campos 2015).

To visually examine genotype by environment interaction (GEI), the GGE biplot was constructed from the first two principal components (PC1 and PC2) derived from subjecting environment centered seed protein and oil contents data using the equation

$$y_{ij} = \boldsymbol{\mu} + \beta + \sum_{k=1}^{k} \lambda_k \gamma_{ik}\, \sigma_{jk}$$

embedded in the (GGE) biplot 'R' statistical package, 'GGEBiplotGUI'(Rakshit, Ganapathy et al. 2012). Where y_{ij} is the response mean of i-th genotype in j-th environment, μ is the grand mean, ß is the main effect of j-th environment, k is the number of principal components (PC) required for appropriate depiction of GGE, λ_k is the singular value of the kth PC (PC_k); and γ_{ik} and σ_{jk} are the scores of ith genotype and jth environment, respectively for PC_k (Rakshit, Ganapathy et al. 2012).

The MET data for the two traits were analyzed with the biplots tools option of GGEbiplotGUI R package with scaling set to one standard deviation, the tester centered model set to (G+GE), and singular value decomposition of position matrix (SVP) was different depending on the type of analysis. The SVP for the different types of analyses was as follows: for viewing genotype patterns or locations it was set to JK- (Row Metric Preserving), for examining relations among environments it was set to HJ-(Dual Metric Preserving), for which–won-where pattern it was set to HJ-(Dual Metric Preserving), and for the genotype mean vs stability the SVP was set to JK- (Row Metric Preserving), and ranking with reference to ideal genotype it was set to JK-(Row Metric Preserving) (Greenacre 2010).

5.5 Results

5.5.1 Variability in Seed Protein and Oil Contents

Summary statistics for the two traits were calculated using proc mixed procedure in SAS 9.4 (Table 5.3). Percent of protein and oil contents varied from location to location and years to years (Table 5.3). Detail information about the extent of the variation in seed protein and oil contents is provided in (Table 5.3). Overall, percent of mean protein content across locations and years ranged from 30.11-38.99 with the overall mean of 34.74 and standard deviation of 1.55, while the percent of mean oil content across locations and years ranged from 16.80-22.93 with the overall mean of 19.36 and standard deviation of 0.97 (Table 5.3). High standard deviation in the protein content indicates that the variation in protein content across locations and years is larger than the variation in oil content.

Table 5.3. Summary statistics for protein and oil by environment and across environments.

		Protein				Oil			
Environment	N	Mean	Std	Min	Max	Mean	Std	Min	Max
Overall	1222	34.74	1.55	30.11	38.99	19.36	0.97	16.80	22.93
2012_IA	106	34.13	1.20	31.68	37.41	18.68	0.80	16.80	20.32
2012_IL	160	34.03	1.30	30.93	37.19	19.72	0.82	17.61	21.73
2012_IN	157	35.59	1.49	31.56	38.99	18.99	0.90	16.94	21.45
2012_NE	160	35.42	1.30	31.60	37.90	18.66	0.68	17.10	20.30
2013_IA	160	33.89	1.21	30.80	36.47	18.90	0.60	17.44	20.87
2013_IL	160	33.58	1.22	30.11	36.85	20.72	0.68	18.16	22.93
2013_IN	159	35.07	1.22	31.28	37.65	19.63	0.58	17.75	21.74
2013_NE	160	36.04	1.30	32.31	38.96	19.34	0.72	17.62	20.91

IA, IL, IN, and NE, indicate, Iowa, Illinois, Indiana, and Nebraska, respectively.
Std=standard deviation, N= number of observation; Min=minimum, Max=maximum.

Frequency distribution showed that the two traits were normally distributed across location and years, indicating that the seed protein and oil contents were controlled by many genes (Figure 5.1).

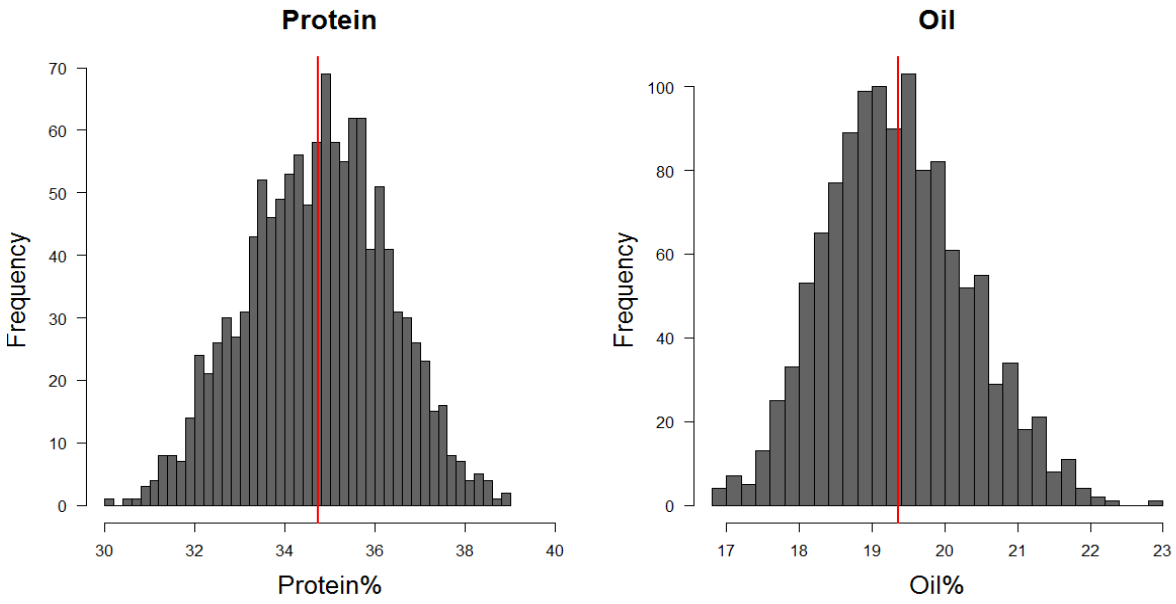


Figure 5.1. Frequency distribution of seed protein and seed oil content in SoyNAM parental genotypes vertical red lines represent overall mean value for each trait.

Combined analysis of variance (ANOVA) revealed significant differences ($P<0.001$) for seed protein content among genotypes, location, year and their respective interactions as well as the threefold interactions among genotypes, locations, and years (Table 5.4). The sum of squares due to genotypes and the environments were high, indicating that the mean seed protein and oil contents of the genotypes were different across environments and the selected environments were diverse. In term of total variation explained, genotypes accounted for 40.1% variation in protein content and 29.1% in the oil content while location explained 28.21% and 30% variation in the two traits, respectively (Table 5.4). Variation due to year, genotypes x year, genotypes x location and genotypes x location x year was significant and they explained small amount of phenotypic variation in the seed protein and oil contents compared to genotypes and locations.

Table 5.4. Analysis of variance for protein and oil across environment using the SoyNAM parents.

		Protein			Oil		
Source	DF	SS	MS	R^2	SS	MS	R^2
Genotype (G)	39	1175.2	30.1***	40.10	335	8.60***	29.10
Location (L)	3	825.9	275.3***	28.21	345.8	115.27***	30.00
Year (Y)	1	6.5	6.5***	0.22	112.8	112.8***	9.82
Replication (R)	3	1.5	0.5^{ns}	0.05	0.2	0.07^{ns}	0.02
G x L	117	161.2	1.38***	5.50	61.1	0.52***	5.31
G x Y	39	84.3	2.16***	2.87	39.1	1.00***	3.40
G x L x Y	117	166.5	1.42***	5.68	62.7	0.54***	5.45
Residuals	902	414.9	0.46		145.4	0.16	
Total	1221	2836.0			1102.1		
Mean		34.74			19.36		
Max		38.99			22.93		
Min		30.11			16.80		
LSD		0.33			0.19		
SED		0.68			0.40		
CV		2.00			2.10		

*, **, *** represents the significant level of 0.05, 0.01 and 0.001, respectively; Min=minimum, Max=maximum, LSD=least significant differences, SED= standard error of difference; CV= coefficient of variation. R^2 = phenotypic variation explained (%).

5.5.2 Correlation between Oil and Protein

The phenotypic and genotypic correlations between protein and oil were -0.59, and -0.66, respectively, reflecting that simultaneous improvement in both traits would be challenging since improvement in one trait would result in decrease in the other trait. The negative correlation between protein and oil in soybean seed could be due to a pair of tightly linked protein and oil QTL whose individual allele might increase one trait but result in decrease in the other. Or the two traits could be controlled by just one pleiotropic QTL, whose two alleles have inverse effects.

5.5.3 Stability Analysis for Protein and Oil Content Across Multiple Environments

Growing same genotypes in different environments under highly variable weather conditions often results in a mix of crossover and non-crossover types of genotype by environment interaction (GEI) (Nzuve, Githiri et al. 2013). The crossover GEI complicates breeding and selection for important traits and is a main concern to plant breeders (Lynch and Walsh 1998; WeiKai, Hunt et al. 2001; Lyimo, Pratt et al. 2012; MITROVIÃ, TRESKI et al. 2012).

To evaluate the extent of GEI, we grouped the 40 SoyNAM parental genotypes into three categories (15 top yielding, 15 moderate yielding, and 10 poor yielding), based on BLUP data (Table 5.5 and 5.6), and subjected them to GGE analysis. Genotypes in all the three categories for both, seed protein and oil contents, showed variable performance across environments, indicating presence of crossover GEI (Kaya, Akçura et al. 2006). The line graphs embedded in Tables 5.5 and 5.6 provide a clear view of GEI, showing genotypic seed protein and oil contents fluctuations due to environmental variation.

Table 5.5. Mean seed protein content of 40 SoyNAM parental lines, selected based on BLUP; seed protein content fluctuations of the parental genotypes across eight environments are displayed in the line graph.

Genotype	Code	IA-12	IA-13	IL-12	IL-13	IN-12	IN-13	NE-12	NE-13	$\bar{x}$	BLUP	Rank
PI507681B	G1	37.2	36.0	36.0	35.5	36.5	36.6	37.0	38.1	36.6	1.6	1
Prohio	G2	36.2	35.1	35.0	35.5	36.9	37.0	36.9	37.2	36.2	1.4	2
PI561370	G3	35.1	35.3	34.3	35.5	36.8	36.5	37.4	38.3	36.1	1.3	3
HS6-3976	G4	35.0	34.1	36.1	34.8	37.6	36.3	36.9	37.3	36.0	1.2	4
LG94-1128	G5	37.1	35.4	34.7	33.7	35.5	35.9	37.1	38.2	36.0	1.0	5
PI437169B	G6	35.2	35.9	35.0	34.6	36.1	35.6	36.6	38.2	35.9	1.0	6
PI427136	G7	35.1	34.8	35.8	33.1	37.0	36.5	36.8	37.4	35.8	1.0	7
5M20-2-5-2	G8	34.4	33.6	35.2	34.2	37.9	36.8	36.2	36.7	35.6	0.8	8
LG03-3191	G9	35.2	34.4	35.3	35.0	36.4	35.7	36.2	36.0	35.5	0.7	9
CL0J173-6-8	G10	34.8	35.1	34.1	34.5	37.0	35.6	35.9	37.0	35.5	0.7	10
CL0J095-4-6	G11	34.9	35.0	33.8	34.9	35.8	35.8	36.9	36.8	35.5	0.7	11
PI398881	G12	35.2	34.5	34.3	34.2	36.3	35.7	36.2	36.6	35.4	0.6	12
LG98-1605	G13	34.6	35.1	34.3	34.2	36.0	36.0	35.6	37.0	35.3	0.6	13
LG92-1255	G14	34.3	35.2	34.6	33.9	36.2	35.1	35.7	36.6	35.2	0.4	14
S06-13640	G15	34.1	33.9	36.1	34.5	36.0	35.5	35.9	35.6	35.2	0.4	15
LG03-2979	G16	35.2	34.7	33.6	33.6	35.6	35.6	35.5	36.4	35.0	0.3	16
PI404188A	G17	34.8	35.0	34.0	33.8	34.6	35.4	35.4	37.3	35.0	0.3	17
Magellan	G18	34.9	33.8	35.0	33.5	36.9	34.5	36.4	35.2	35.0	0.3	18
LG04-4717	G19	34.0	34.5	35.3	34.1	35.3	35.5	35.7	35.6	35.0	0.3	19
LD02-9050	G20	33.3	34.1	34.3	33.5	36.2	35.2	35.8	35.6	34.8	0.1	20
LG94-1906	G21	34.3	34.2	33.5	33.5	36.2	35.6	33.6	36.7	34.7	0.0	21
PI518751	G22	35.3	34.3	33.1	33.8	35.5	35.1	34.8	35.6	34.7	0.0	22
LG05-4292	G23	33.0	33.2	34.5	33.6	35.6	35.1	35.7	36.1	34.6	-0.1	23
4J105-3-4	G24	33.2	33.1	34.4	33.5	35.4	35.3	35.4	36.1	34.6	-0.1	24
LG05-4317	G25	34.4	34.4	34.0	33.6	35.5	34.4	34.8	34.9	34.5	-0.2	25
LG97-7012	G26	34.3	34.0	33.7	33.5	34.5	34.7	35.1	36.0	34.5	-0.3	26
LD00-3309	G27	33.2	32.8	33.9	34.1	36.1	34.9	35.5	34.9	34.4	-0.3	27
Maverick	G28	33.7	33.3	34.1	33.0	35.9	34.5	34.8	36.0	34.4	-0.3	28
LG90-2550	G29	34.2	33.8	32.9	33.1	36.2	34.6	35.0	35.2	34.4	-0.3	29
PI574486	G30	33.1	33.0	34.2	32.6	36.0	33.8	36.1	35.3	34.3	-0.4	30
LG04-6000	G31	33.7	32.9	33.8	32.8	36.0	34.2	35.7	34.5	34.2	-0.5	31
LG05-4464	G32	33.6	32.7	33.5	32.4	35.7	34.6	35.1	35.7	34.2	-0.5	32
LG05-4832	G33	33.7	33.7	32.5	32.8	34.6	34.4	35.4	35.9	34.1	-0.6	33
LG00-3372	G34	32.9	32.9	34.0	32.7	35.8	34.2	34.5	35.2	34.0	-0.6	34
NE3001	G35	33.1	32.4	32.3	32.4	33.7	34.5	34.0	36.0	33.6	-1.1	35
U03-100612	G36	33.7	33.1	32.3	32.3	32.9	33.8	33.8	34.6	33.3	-1.5	36
Skylla	G37	32.7	32.2	31.4	32.5	33.3	33.7	32.5	35.0	32.9	-1.7	37
IA3023	G38	32.3	32.4	32.1	31.2	33.3	33.4	33.7	34.4	32.9	-1.7	38
LD01-5907	G39	32.0	31.8	32.6	32.1	33.5	32.6	33.7	33.6	32.8	-1.8	39
LD02-4485	G40	34.2	31.6	31.6	31.1	32.2	32.8	32.1	33.4	32.4	-2.5	40

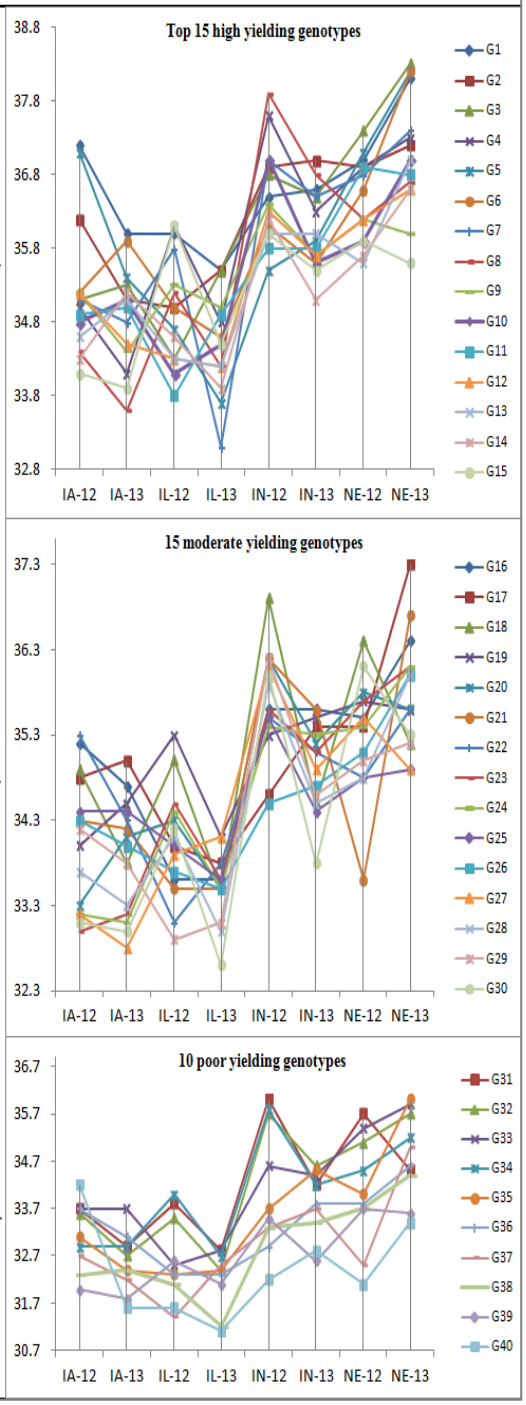
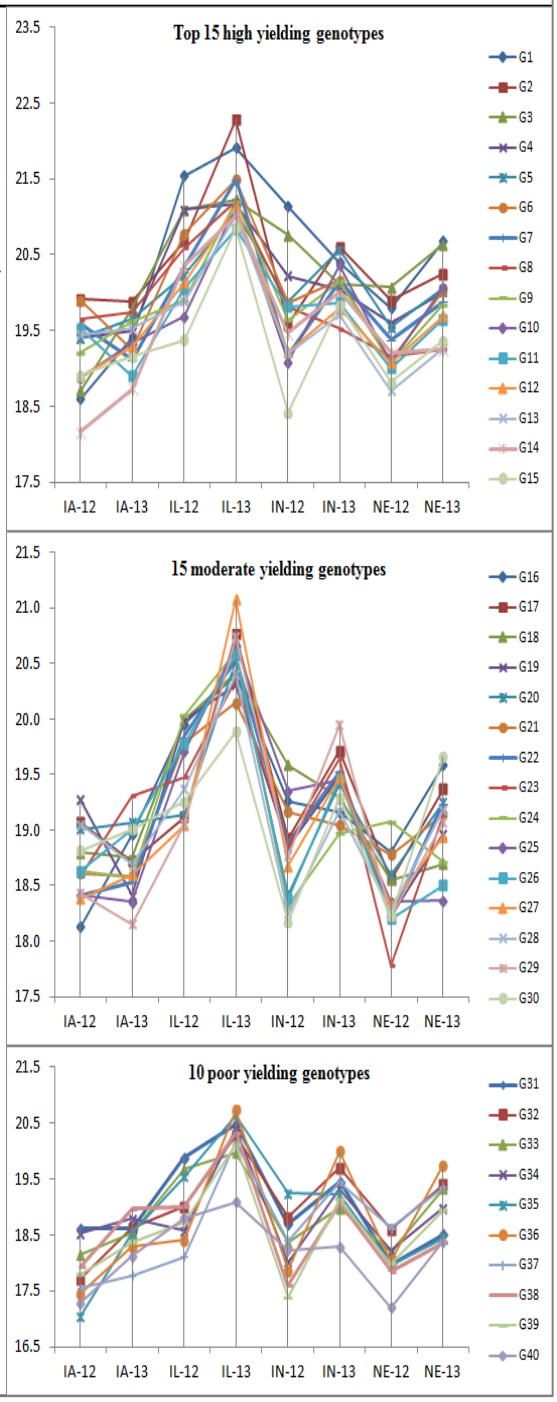


Table 5.6. Mean seed oil content of 40 SoyNAM parental lines, selected based on BLUP; seed oil content fluctuations of the parental genotypes across eight environments are shown in the line graph.

Genotype	Code	IA-12	IA-13	IL-12	IL-13	IN-12	IN-13	NE-12	NE-13	$\bar{X}$	BLUP	Rank
U03-100612	G1	18.6	19.4	21.5	21.9	21.1	20.4	19.8	20.7	20.4	1.10	1
LG05-4464	G2	19.9	19.9	20.7	22.3	19.6	20.6	19.9	20.3	20.4	0.96	2
LD02-4485	G3	18.7	19.8	21.1	21.2	20.8	20.1	20.1	20.6	20.3	0.96	3
Skylla	G4	19.4	19.5	21.1	21.2	20.2	20.0	19.6	20.0	20.1	0.70	4
LD01-5907	G5	19.4	19.7	20.2	21.1	19.9	20.6	19.6	20.0	20.1	0.70	5
IA3023	G6	19.9	19.2	20.8	21.5	19.9	20.2	19.1	20.0	20.1	0.65	6
LG92-1255	G7	19.6	19.1	20.3	21.5	19.2	20.2	19.4	19.9	19.9	0.47	7
NE3001	G8	19.6	19.7	20.6	21.2	19.8	19.5	19.2	19.2	19.9	0.47	8
LG04-4717	G9	19.2	19.6	19.9	21.2	19.6	20.2	19.1	19.8	19.8	0.44	9
LG05-4317	G10	18.9	19.3	19.7	21.1	19.1	20.4	19.1	20.1	19.7	0.33	10
LG97-7012	G11	19.5	18.9	20.1	20.8	19.8	19.9	19.0	19.6	19.7	0.30	11
LG03-2979	G12	18.9	19.3	20.1	21.1	19.2	19.8	19.1	19.7	19.7	0.28	12
4J105-3-4	G13	19.4	19.5	19.9	21.0	19.2	19.7	18.7	19.3	19.6	0.22	13
Prohio	G14	18.2	18.7	20.3	21.0	19.5	20.0	19.2	19.3	19.5	0.17	14
Maverick	G15	18.9	19.2	19.4	20.9	18.4	19.9	18.8	19.4	19.3	0.01	15
PI518751	G16	18.1	19.0	20.0	20.4	19.3	19.2	18.8	19.6	19.3	-0.01	16
LG05-4292	G17	19.1	18.7	19.1	20.8	18.9	19.7	18.6	19.4	19.3	-0.07	17
PI437169B	G18	18.8	18.7	20.0	20.4	19.6	19.3	18.6	18.7	19.3	-0.07	18
PI561370	G19	19.3	18.4	20.0	20.3	18.8	19.5	18.2	19.0	19.2	-0.14	20
LG00-3372	G20	19.0	19.1	19.1	20.5	18.4	19.4	18.6	19.2	19.2	-0.14	19
LG98-1605	G21	18.6	18.6	19.8	20.1	19.2	19.0	18.8	19.1	19.2	-0.16	22
PI398881	G22	18.4	18.5	19.8	20.5	18.9	19.5	18.2	19.2	19.1	-0.16	21
CL0J095-4-6	G23	18.6	19.3	19.5	20.4	18.9	19.6	17.8	19.1	19.1	-0.18	23
LG94-1906	G24	18.6	18.6	20.0	20.6	18.3	19.0	19.1	18.7	19.1	-0.20	24
PI404188A	G25	18.4	18.4	19.7	20.7	19.4	19.4	18.4	18.4	19.1	-0.23	25
CL0J173-6-8	G26	18.6	19.0	19.8	20.6	18.4	19.5	18.2	18.5	19.1	-0.23	26
PI427136	G27	18.4	18.6	19.0	21.1	18.7	19.5	18.3	18.9	19.1	-0.24	27
LD02-9050	G28	19.1	18.7	19.4	20.4	18.3	19.2	18.3	19.2	19.0	-0.26	28
PI574486	G29	18.4	18.2	19.1	20.7	18.8	19.9	18.2	19.1	19.0	-0.27	32
LD00-3309	G30	18.8	19.0	19.3	19.9	18.2	19.3	18.2	19.7	19.0	-0.27	29
LG05-4832	G31	18.6	18.6	19.9	20.5	18.7	19.4	18.0	18.5	19.0	-0.27	31
LG03-3191	G32	17.7	18.6	19.0	20.3	18.8	19.7	18.6	19.4	19.0	-0.27	30
LG90-2550	G33	18.1	18.5	19.7	20.0	18.4	19.0	18.2	19.3	18.9	-0.39	33
HS6-3976	G34	18.5	18.8	18.6	20.4	18.0	19.3	18.2	19.0	18.9	-0.39	35
LG94-1128	G35	17.1	18.5	19.5	20.6	19.2	19.2	18.0	18.4	18.8	-0.39	34
Magellan	G36	17.5	18.3	18.4	20.7	17.9	20.0	18.0	19.7	18.8	-0.46	36
S06-13640	G37	17.5	17.8	18.1	20.2	18.4	19.5	18.6	19.3	18.7	-0.58	37
5M20-2-5-2	G38	17.9	19.0	19.0	20.3	17.6	19.1	17.9	18.4	18.6	-0.61	38
LG04-6000	G39	17.8	18.4	18.7	20.1	17.4	19.1	18.0	18.9	18.6	-0.69	39
PI507681B	G40	17.3	18.1	18.8	19.1	18.2	18.3	17.2	18.4	18.2	-1.02	40



5.5.4 Stability Analysis for Seed Protein Content

The GGE-biplot, based on genotype focused scaling, revealed that the first two principal components PC1 (Axis 1) and PC2 (Axis2) accounted for a total of 85.79% variation in the protein mean content (Figure 5.2). All of the 15 top genotypes with high seed protein content and some of the intermediate genotypes had PC1 scores > 0 and were grouped by the biplot as adaptable or genotypes with high seed protein content (Figure 5.2).

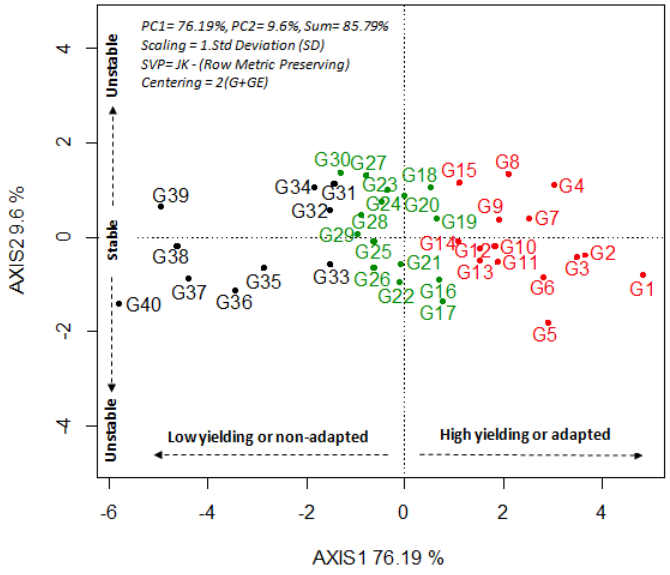


Figure 5.2. GGE-biplot for seed protein content based on genotype-focused scaling for genotypes.
G stand for genotypes. Codes of genotypes are given in Table 5.5. Genotypes G1-G15 are the top highest yielders, G16-30 intermediate, G31-G40 bottom 10 lowest yielders.

The result from the biplot analysis is in strong agreement with our BLUP selection. Although the two methods provided same results, the GGE-biplot method is preferred over BLUPs since it supplies further information on the stability of the selected genotypes. In the biplot, any genotype that has PC2 scores near or equal to zero are considered stable and those located further away from zero are considered unstable.

Accordingly, genotypes G14, G10, G12, G3, G2, G7, G9, and G1 were the most stable genotypes among the top 15 high yielding seed protein content, whereas G5 and G8 were unstable (Figure 5.2). Genotypes G25, G29, G28, and G38 were the most stable among the intermediary and poor seed protein content categories, respectively (Figure 5.2). Similar stability pattern was provided by the average environment coordination (AEC) view of the GGE-biplot (WeiKai, Hunt et al. 2001; Yan 2002) (Figure 5.3). In this method, the average PC1 and PC2 scores of all environments, shown by a small circle, defines an average environment (Figure 5.3).

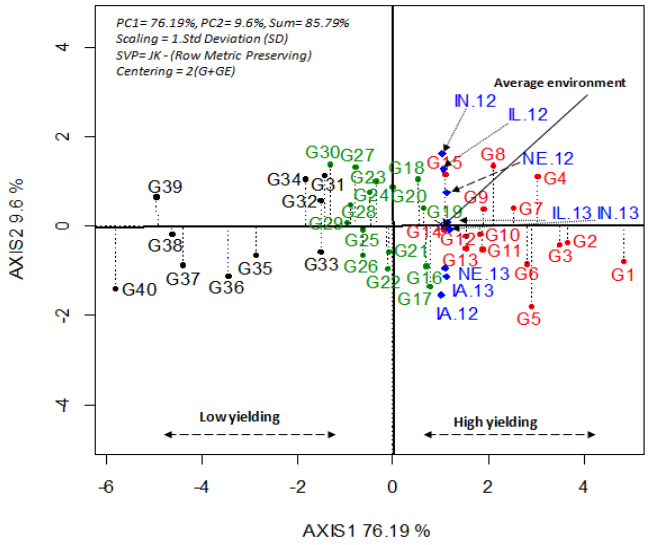


Figure 5.3. Average environment coordination (AEC) views of the GGE-biplot for seed protein content based on environment-focused scaling for the means performance and stability of genotypes.
Details of genotypes and environments are provided in Tables 5.1 and 5.2. IA, IL, IN, and NE, indicate, Iowa, Illinois, Indiana, and Nebraska, respectively; 12 and 13 represent the year 2012 and 2013.

A line passing through the average environment and the biplot origin is called the average environment axis (AEA) and serves as the value of the AEC on the horizontal axis (Kaya, Akçura et al. 2006). Genotypes closer to the AEA are considered stable while genotypes located away from AEA, are regarded unstable. In other words, as the distance

of the genotypes from the AEA increase, the chance of GEI increase and stability is reduced. The distance between the average environment marker (circle in Figure 5.3) and the biplot origin provides an estimate of the relative importance of genotype main effect vs. GEI. In this study, average environment marker (circle in Figure 5.3) is sufficient away from the biplot origin, indicating that genotypes could be selected based on seed protein content mean performance. Considering this, genotypes with mean seed protein content above average means, which include all the top 15 and some intermediate category, would be selected and the rest discarded. However, stability is important and with the GGE-biplot analysis we were able to identify and select desired genotypes based on both mean and stability. For instance, genotypes G14, G10, G12, G3, G2, G7, G9, and G1 were both stable and had higher seed protein content, while genotype G5 and G8 had higher seed protein content but unstable (Figure 5.3). We conducted environment-focused GGE-biplot analysis aiming to examine the patterns of environments (Figure 5.4).

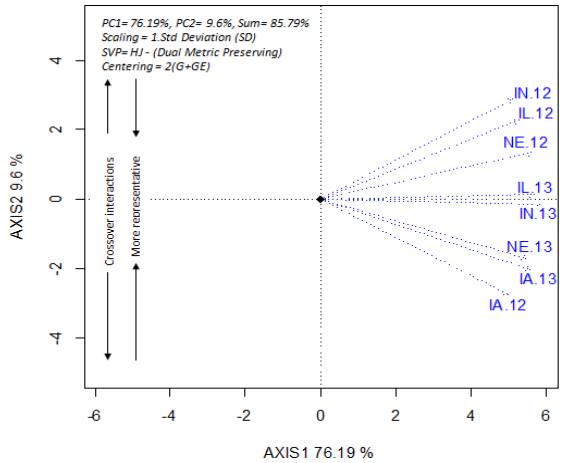


Figure 5.4. Seed protein content GGE-biplot based on environment-focused scaling for environments.
PC and E stand for principal component and environments, respectively.
Details of environments are given in Table 5.2. IA, IL, IN, and NE, indicate, Iowa, Illinois, Indiana, and Nebraska, respectively; 12 and 13 represent the year 2012 and 2013.

An environment with only positive PC1 scores, suggests that PC1signify comparative genotype-trait differences across environments, and crossover GEI is not going to happen. In contrast, PC2 with positive and negative scores represents a typical feature of crossover GEI (Yan, Hunt et al. 2000). A genotype may have large positive interactions with some environments, but at the same time, it may have large negative interactions with some other environments. In this study the environments IN.13, and IL.13 were similar; since they had similar genotype means seed protein performance (Figure 5.4). On the other hand, PC2 scores of the other six environments IN.12, IL.12, NE.12, NE.13, IA.12, and IA.13 were absolutely greater than zero, indicating large crossover interaction effects, and therefore they were considered non-representative (Figure 5.4). The PC2 scores of the two environments IN.13, and IL.13 were near zero, indicating that there will be less crossover interaction effect, and for that reason, they were considered more discriminative and representative (Figure 5.4).

The which-won-where pattern of MET data, which display polygon view of the GGE-biplot, provides visual summary of the GEI pattern of a MET data set (Yan, Hunt et al. 2000; WeiKai, Hunt et al. 2001; Gauch 2006). The polygon is formed by connecting the markers of the genotypes that are located far away from the biplot origin such that all other genotypes are kept inside the polygon. In the polygon, lines perpendicular to the sides of the polygon divide the biplot into sectors. In this study, the polygon was divided into 7 sectors and the entire environments fell in the first sector (Figure 5.5), with G1 being the vertex genotype, indicating that G1 was the highest seed protein content genotype in all environments. Genotype G1 would be preferred since it out yielded all others in all environments.

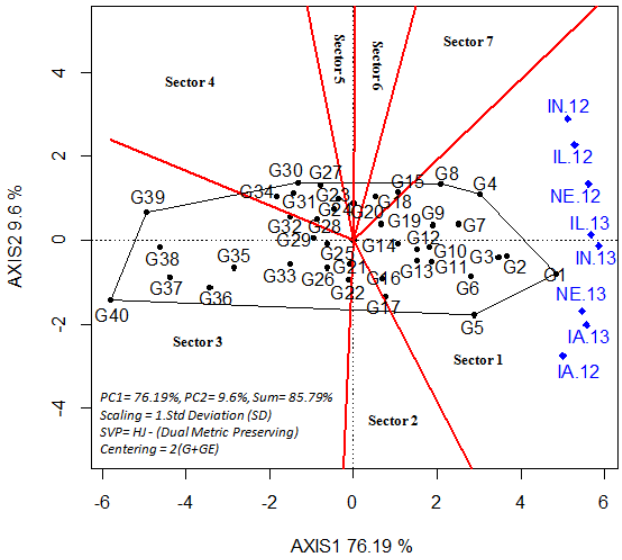


Figure 5.5. Polygon view of seed protein content GGE-biplot of the which-won where pattern for genotypes and environments.
G stand for genotype; IA, IL, IN, and NE, indicate, Iowa, Illinois, Indiana, and Nebraska, respectively; 12 and 13 represent the year 2012 and 2013. Details of genotype and the environments are provided in Tables 5.1 and 5.2.

Genotype-focused scaling GGE-biplot which compares genotypes with ideal genotype was used to identify the most desirable lines (Figure 5.6). The desired genotypes should out yield all other genotypes and should have the highest mean performance across environments with higher stability and no GEI and it is shown in the plot by an arrow in the central concentric center (Figure 5.6). The genotype focused scaling biplot, which is based on both stability and mean performance, revealed that G1 and G2 fell into the center of concentric circle. On the other hand, G3, G4, G6, and G7 were located on the next concentric circle (Figure 5.6). The genotypes in these two circles are considered the most desirable genotypes compared to the rest of the genotypes studied.

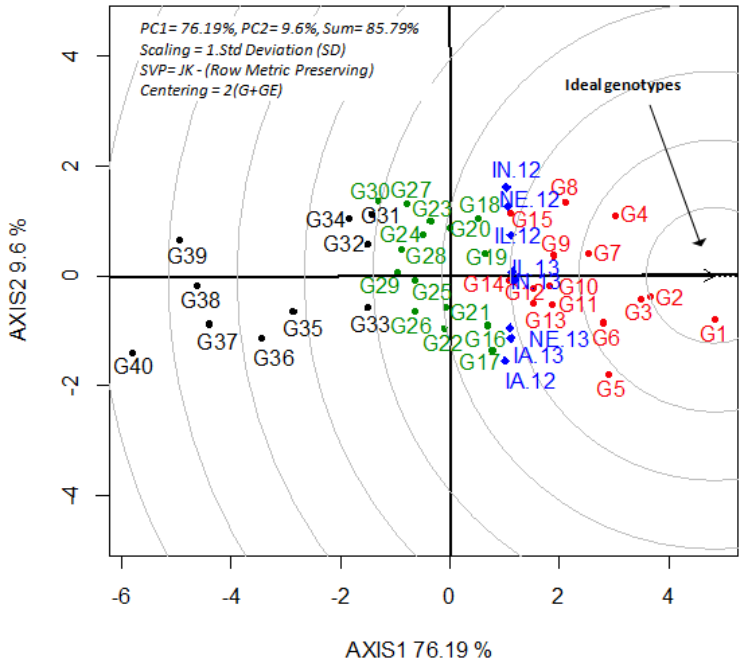


Figure 5.6. GGE-biplot for seed protein based on genotype-focused scaling for comparison of the genotypes with ideal genotype.
G stand for genotype; IA, IL, IN, and NE, indicate, Iowa, Illinois, Indiana, and Nebraska, respectively; 12 and 13 represents year 2012 and 2013. Details of genotype and the environments are provided in Tables 5.1 and 5.2.

5.5.5 Stability Analysis for Seed oil Content

The GGE-biplot, based on genotype focused scaling, revealed that the first two principal components PC1 (Axis 1) and PC2 (Axis2) explained a total of 79.51% phenotypic variation in the seed oil mean content (Figure 5.7). The PC1 scores for all top 15 high seed oil content genotype were greater than zero (PC1 > 0) therefore they were grouped by the biplot as adaptable or genotypes with the highest seed protein content (Figure5.7).

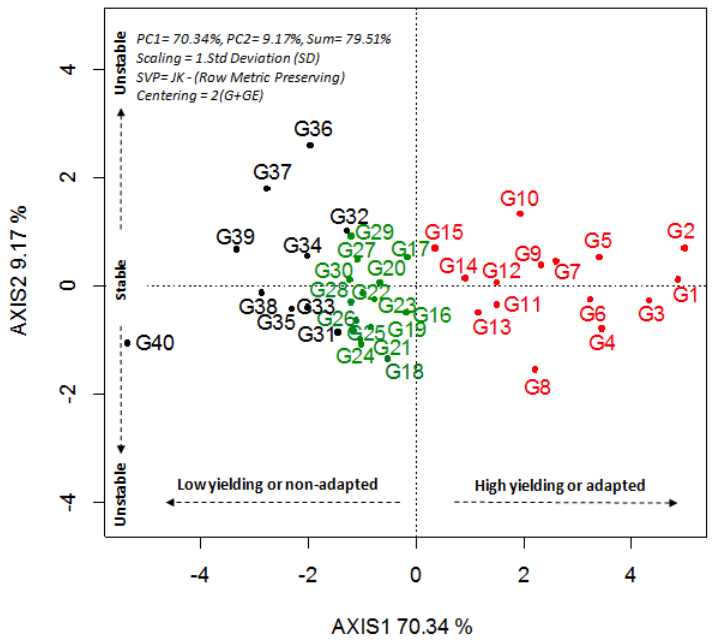


Figure 5.7. GGE-biplot for seed oil content based on genotype-focused scaling for genotypes.
G stand for genotypes. Codes of genotypes are given in Table 5.6. Genotypes G1-G15 are the top highest yielders, G16-30 intermediate, G31-G40 bottom 10 lowest yielders.

The biplot analysis result for seed oil content is consistent with our BLUP selection. In the biplot, genotypes with PC2 scores close or equal to zero are considered stable and those situated further away from zero are considered unstable, therefore, genotypes G12, G1, G14, G3, G6, G9, and G7 were the most stable genotypes among the top 15 high yielding seed oil content, while G8 and G10 were the most unstable genotypes (Figure 5.7). Genotypes G20, G30, G22, G28, G23, and G38 were stable among the intermediary and poor seed oil content categories, respectively (Figure 5.7). The average environment coordination (AEC) view of the GGE-biplot (WeiKai, Hunt et al. 2001; Yan 2002) presented similar stability patterns (Figure 5.8). In this method, the average environment, shown by a small circle in the plot, is defined by the average PC1and PC2 scores of all environments (Figure 5.8).

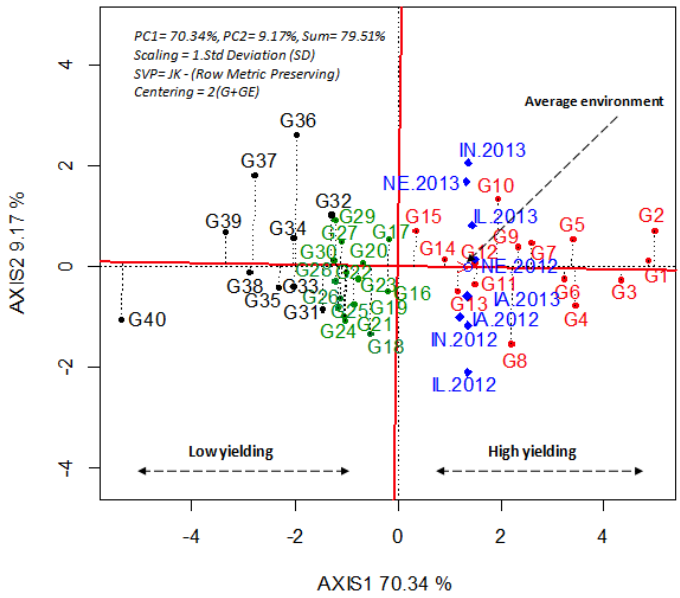


Figure 5.8. Average environment coordination (AEC) views of the GGE-biplot for seed oil content based on environment-focused scaling for the means performance and stability of genotypes.
Details of genotypes and environments are provided in Tables 5.1 and 5.2. IA, IL, IN, and NE, indicate, Iowa, Illinois, Indiana, and Nebraska, respectively; 2012 and 2013 are the years in which the experiment was conducted.

In the average environment coordination (AEC) view of the GGE-biplot genotypes closer to the average environment axis (AEA) are considered stable but as you move away from AEA, the chance of GEI increases and stability is reduced. The average environment marker is sufficiently far from the biplot origin (Figure 5.8), indicating that genotypes selection can be done based on seed oil content mean performance. In this case, all the top 15 high seed oil content genotypes, which had above-average means, would be selected and the rest would be discarded. Nevertheless, high yielding genotypes without stability are not desired. Conducting the GGE-biplot analysis, we were able to select desired genotypes based on both mean and stability. Using the GGE-biplot, we were able to identify genotypes G12, G1, G14, G3, G6, G9, and G7 both stable and high yielding, while genotypes G8, and G10 were high yielding but unstable (Figure 5.8).

The environment-focused GGE-biplot analysis was conducted to examine the environment pattern for the seed oil content (Figure 5.9). According to the plot, environments NE.2012, and NE.2013 were similar due to their genotype means performance similarity (Figure 5.9). However, the PC2 scores of the six environments IN.2012, IL.2013, IN.2013, IA.2012 and IA.2013 were absolutely greater than zero, indicating that the effect of crossover interaction would be high, and therefore they were considered unrepresentative (Figure 5.9). The PC2 scores of the two environments NE.2012, and NE.2013 were near zero, suggesting less or no crossover interaction effect, and thereby, they were considered representative (Figure 5.9).

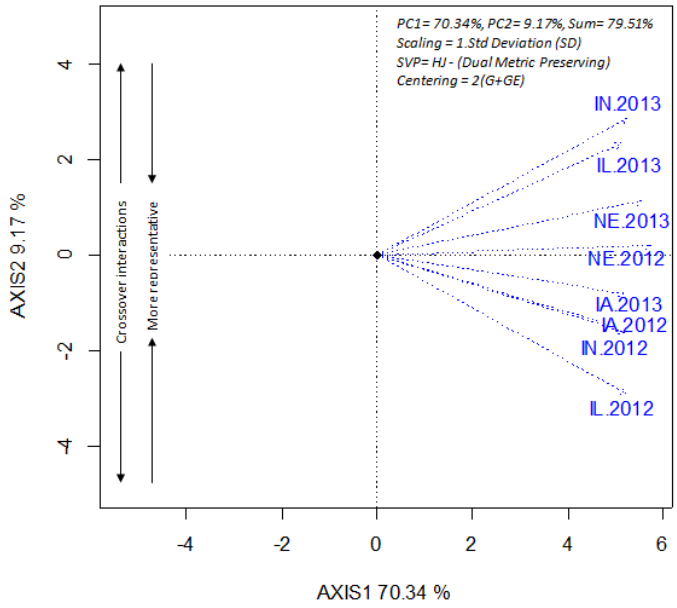


Figure 5.9. Seed oil content GGE-biplot based on environment-focused scaling for environments.
Details of environments are given in Table 5.2. IA, IL, IN, and NE, indicate, Iowa, Illinois, Indiana, and Nebraska, respectively; 2012 and 2013 are the years in which the experiment was conducted.

The which-won-where pattern revealed that the polygon was divided into 7 sectors and the entire environments fell in the first sector (Figure 5.10), with G1 being the

vertex genotype. This indicates that G1 was the highest and foremost seed oil content genotype in all environments. Genotype G1 and G2 would be desired since they performed well in all environments.

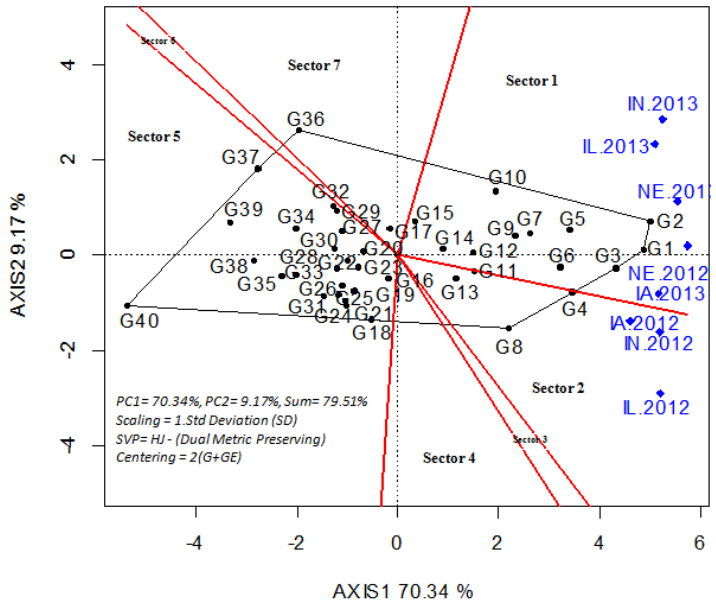


Figure 5.10. Polygon view of seed oil content GGE-biplot of the which-won where pattern for genotypes and environments.
G stand for genotype; IA, IL, IN, and NE, indicate, Iowa, Illinois, Indiana, and Nebraska, respectively; 2012 and 2013 are the years in which the experiment was implement. Details of genotype and the environments are provided in Tables 5.1 and 5.2.

The genotype-focused scaling GGE-biplot, which compares genotypes with ideal genotype and is based on both stability and mean performance, revealed that G1 and G2 fell into the center of concentric circles (Figure 5.11). These genotypes are considered the most preferred genotypes compared to the rest since they are not only high yielder but stable too.

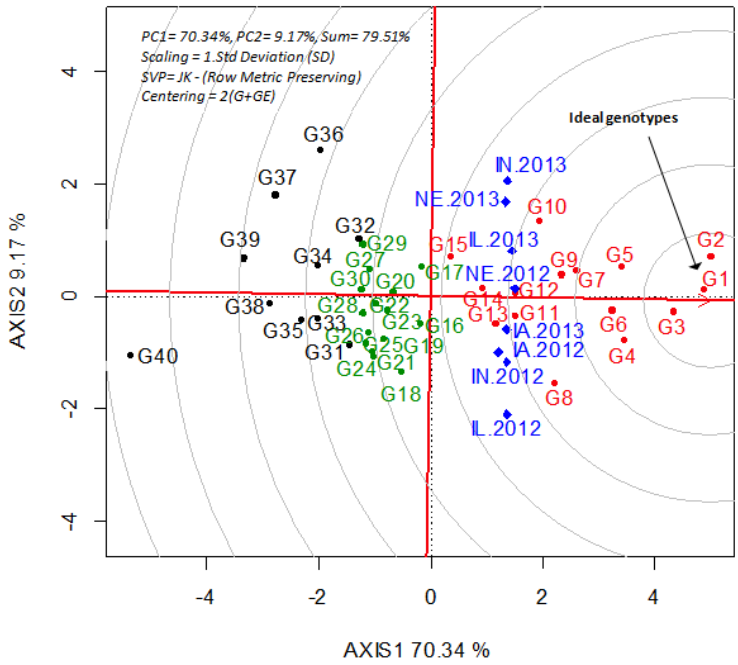


Figure 5.11. GGE-biplot for seed oil content based on genotype-focused scaling for comparison of the genotypes with ideal genotype.
G stand for genotype; IA, IL, IN, and NE, indicate, Iowa, Illinois, Indiana, and Nebraska, respectively; 2012 and 2013 are the years in which the experiment was conducted.
Details of genotype and the environments are provided in Tables 5.1 and 5.2.

5.6 Discussion

Multi-environment trials (MET) analysis is important for evaluating adaptability of genotypes in a wider array of environments. In this analysis the phenotypic variations is partitioned into components that are genetic, environmental and interactions. An understanding the relative importance of the G and E help breeders make more informed breeding decisions. In this study, the means squares for G, E and GEI were all significant, with much of the contribution to the phenotypic variations was due to genotypes. Significant G and E indicated that the genotypes were genetically diverse and environmental conditions were distinctive for both traits. This suggests that genetic improvement can be obtained, but the effect of the environment on genotype performance

should be considered as well. Significant GEI happens when ranking of genotypes change from an environment to environment (Mohammadi and Amri 2013). GEI for seed protein and seed oil contents of soybean have been reported in earlier studies. Sudarić et al. (2006) reported a significant GEI for both protein and oil contents using combined data from 15 environments. He found that locations explained the largest proportion of the total variance for protein content. Zhe et al. (2010) conducted GEI analysis for seed composition and other agronomic traits and reported a significant GE interaction. Kumar et al. (2006) reported a significant GE interaction for genotypes, environment and their interaction for seed protein and oil contents using seven Indian cultivars. These results strongly agree with findings of the present study.

In the presence of GEI, selection based on the genotype mean performance alone may not be useful because in such scenario genotype response are specific to the environment. In such situation, genotypes may be selected only for a specific environment, if interest is to select for wider array of environments then more detailed assessment of stability should be conducted (Lynch and Walsh 1998). Large GEI may decrease the heritability of quantitative traits , for that reason, it may have negative impact on genetic advance from selection. To obatin reliable results in the presence of GEI, a cultivar of interest should be tested in several environments (Lynch and Walsh 1998; WeiKai, Hunt et al. 2001). The GGE bioplot analysis is effective in dissecting all portions of MET data, providing a powerful visual image of stability, mean performance, and ideal environments for specific genotypes.

Evaluation of phenotypic and genotypic correlations revealed strong and significant correlation coefficients among the two traits, seed protein, and oil contents indicating that

simultaneous improvement of these two traits is not possible. This kind of relationship can be noticed in the GGE-biplot analysis and BLUPs selection as well. Genotype ranked number one in BLUPs selection for seed protein content is ranked the last for seed oil content and vice versa. Also, genotypes identified the highest yielding and the most stable for seed protein content were the poor yielding and most unstable genotypes in the seed oil content or vice versa. This type of relationship between the two traits, seed protein and oil contents, is due to the negative correlation between them. The opposite effect for both traits could be controlled by just one pleiotropic QTL, whose two alleles have inverse effects on both traits.

5.7 Conclusion

The present study showed that SoyNAM parental genotypes are genetically diverse for both traits. The multi-environment trial analysis (MET) revealed that variation in seed protein and oil contents performance of the SoyNAM parental genotypes was largely genetic but still influenced by the environment. Dissection of the major component sources of aviation (G+GE) for both traits using GGE-biplot depicted the possibility of identifying SoyNAM parental genotypes with broad adatapation and those that are suited for specific environments. Genotypes G14, G10, G12, G3, G2, G7, G9 and G1for seed protein content and genotypes G12, G1, G14, G3, G6, G9 and G7 for seed oil content were identified as the most stable and desired compared to the rest.

LIST OF REFERENCES

LIST OF REFERENCES

Abdurakhmonov, I. Y. and A. Abdukarimov (2008). "Application of association mapping to understanding the genetic diversity of plant germplasm resources." International journal of plant genomics 2008.

Agrawal, P. K. (1989). Carbon-13 NMR of flavonoids, Elsevier Science Ltd.

Aranzana, M. J., S. Kim, et al. (2005). "Genome-wide association mapping in Arabidopsis identifies previously known flowering time and pathogen resistance genes." PLoS Genetics 1(5): e60.

Baer, R., J. Frank, et al. (1983). "Compositional analysis of whey powders using near infrared diffuse reflectance spectroscopy." Journal of Food Science 48(3): 959-961.

Baianu, I. and J. Guo (2011). "NIR Calibrations for Soybean Seeds and Soy Food Composition Analysis: Total Carbohydrates, Oil, Proteins and Water Contents."

Birth, G., G. Dull, et al. (1985). "Nondestructive spectrophotometric determination of dry matter in onions." Journal of the American Society for Horticultural Science 110.

Boerma, H. R. and M. R. Mian (1999). Soybean quantitative trait loci and marker-assisted breeding. Proc. World Soybean Res. Conf. VI, Chicago, IL.

Bouis, H. E. (2002). "Plant breeding: a new tool for fighting micronutrient malnutrition." The Journal of nutrition 132(3): 491S-494S.

Buckler, E. S., J. B. Holland, et al. (2009). "The genetic architecture of maize flowering time." Science 325(5941): 714-718.

Burleson, S. A. (2011). Development of New and Alternative Resources for Breeding Low Phytate Soybeans, Virginia Polytechnic Institute and State University.

Campbell, K., G. Czarnecki-Maulden, et al. (1995). "Effects of animal and soy fats and proteins in the diet on fatty acid concentrations in the serum and skin of dogs." American journal of veterinary research 56(11): 1465.

Center, C. (1995). "Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results." Nature genetics 11: 241.

Cheryan, M. and J. J. Rackis (1980). "Phytic acid interactions in food systems." Critical Reviews in Food Science & Nutrition 13(4): 297-335.

Choung, M.-G. (2010). "Determination of sucrose content in soybean using near-infrared reflectance spectroscopy." Journal of the Korean Society for Applied Biological Chemistry 53(4): 478-484.

Clarke, E. and J. Wiseman (2000). "Developments in plant breeding for improved nutritional quality of soya beans II. Anti-nutritional factors." The Journal of Agricultural Science 134(2): 125-136.

Cole, J. T., G. Fahey, et al. (1999). "Soybean hulls as a dietary fiber source for dogs." Journal of animal science 77(4): 917-924.

Cromwell, G. and R. Coffey (1991). Phosphorus-a key essential nutrient, yet a possible major pollutant-its central role in animal nutrition. ALLTECH'S ANNUAL SYMPOSIUM OF BIOTECHNOOGY IN THE FEED INDUSTRY.

Davies, A. (2000). "William Herschel and the discovery of near infrared." Spectroscopy Europe 12(2): 10-17.

Dull, G. G., G. S. Birth, et al. (1989). "Near infrared analysis of soluble solids in intact cantaloupe." Journal of Food Science 54(2): 393-395.

Flint- Garcia, S. A., A. C. Thuillet, et al. (2005). "Maize association population: a high- resolution platform for quantitative trait locus dissection." The Plant Journal 44(6): 1054-1064.

Geater, C. W., W. R. Fehr, et al. (2000). "Association of soybean seed traits with physical properties of natto." Crop Science 40(6): 1529-1534.

Gillman, J. D., V. R. Pantalone, et al. (2009). "The low phytic acid phenotype in soybean line CX1834 is due to mutations in two homologs of the maize low phytic acid gene." The Plant Genome 2(2): 179-190.

Greenwood, J. S. and G. Batten (1995). "Mechanisms and regulation of mineral nutrient storage during seed development." Seed Development and Germination. Marcel Dekker, New York: 215-235.

Gupta, P. K., S. Rustgi, et al. (2005). "Linkage disequilibrium and association studies in higher plants: present status and future prospects." Plant Molecular Biology 57(4): 461-485.

Holland, J. B. (2007). "Genetic architecture of complex traits in plants." Current opinion in plant biology 10(2): 156-161.

Lee, J.-D., J. G. Shannon, et al. (2011). "Application of nondestructive measurement to improve soybean quality by near infrared reflectance spectroscopy." Soybean–Applications and Technology: 287-304.

Lemtiri-Chlieh, F., E. A. MacRobbie, et al. (2000). "Inositol hexakisphosphate is a physiological signal regulating the K+-inward rectifying conductance in guard cells." Proceedings of the National Academy of Sciences 97(15): 8687-8692.

Liener, I. E. (1994). "Implications of antinutritional components in soybean foods." Critical Reviews in Food Science & Nutrition 34(1): 31-67.

Lönnerdal, B., L. Jayawickrama, et al. (1999). "Effect of reducing the phytate content and of partially hydrolyzing the protein in soy formula on zinc and copper absorption and status in infant rhesus monkeys and rat pups." The American journal of clinical nutrition 69(3): 490-496.

Maupin, L. M. (2010). Characterization of soybean germplasm with modified phosphorus and sugar composition.

Meis, S. J., W. R. Fehr, et al. (2003). "Seed source effect on field emergence of soybean lines with reduced phytate and raffinose saccharides." Crop Science 43(4): 1336-1339.

Myles, S., J. Peiffer, et al. (2009). "Association mapping: critical considerations shift from genotyping to experimental design." The Plant Cell Online 21(8): 2194-2202.

Oltmans, S. E., W. R. Fehr, et al. (2005). "Agronomic and Seed Traits of Soybean Lines with Low–Phytate Phosphorus." Crop Science 45(2): 593-598.

Openshaw, S. and H. Hadley (1978). "Maternal effects on sugar content in soybean seeds." Crop Science 18(4): 581-584.

Pierce, M. M. and R. L. Wehling (1994). "Comparison of sample handling and data treatment methods for determining moisture and fat in Cheddar cheese by near-infrared spectroscopy." Journal of agricultural and food chemistry 42(12): 2830-2835.

Psotka, J. and W. Shadow (1994). "NIR ANALYSIS IN THE WET CORN REFINING INDUSTRY-A TECHNOLOGY REVIEW OF METHODS IN USE." International Sugar Journal 96(1149): 358-360.

Raboy, V. (1997). Accumulation and storage of phosphate and minerals. Cellular and molecular biology of plant seed development, Springer: 441-477.

Raboy, V. (2001). "Seeds for a better future:'low phytate'grains help to overcome malnutrition and reduce pollution." Trends in Plant Science 6(10): 458-462.

Raboy, V. (2002). "Progress in breeding low phytate crops." The Journal of nutrition 132(3): 503S-505S.

Raboy, V. (2007). "Seed phosphorus and the development of low-phytate crops." Inositol phosphates: Linking agriculture and the environment: 111-132.

Raboy, V. (2009). "Approaches and challenges to engineering seed phytate and total phosphorus." Plant Science 177(4): 281-296.

Raboy, V., K. A. Young, et al. (2001). "Genetics and breeding of seed phosphorus and phytic acid." Journal of plant physiology 158(4): 489-497.

Risch, N. and K. Merikangas (1996). "The future of genetic studies of complex human diseases." Science-AAAS-Weekly Paper Edition 273(5281): 1516-1517.

Rodriguez-Otero, J. L., M. Hermida, et al. (1995). "Determination of fat, protein, and total solids in cheese by near-infrared reflectance spectroscopy." Journal of AOAC International 78(3): 802.

Scaboo, A., V. Pantalone, et al. (2009). "Confirmation of molecular markers and agronomic traits associated with seed phytate content in two soybean RIL populations." Crop Science 49(2): 426-432.

Semagn, K., Å. Bjørnstad, et al. (2010). "The genetic dissection of quantitative traits in crops." Electronic Journal of Biotechnology 13(5): 16-17.

Shadow, W. (1998). "Rapid analysis for the food industry using near-infrared spectroscopy. Perten Instruments North America." Inc., Springfield, IL.

Simpson, A. and J. Wilcox (1983). "Genetic and phenotypic associations of agronomic characteristics in four high protein soybean populations." Crop Science 23(6): 1077-1081.

Sneller, C., D. E. Mather, et al. (2009). "Analytical approaches and population types for finding and utilizing QTL in complex plant populations." Crop Science 49(2): 363-380.

Soto-Cerda, B. J. and S. Cloutier (2012). "Association mapping in plant genomes." Genetic diversity in plants. InTech, Rijeka: 29-54.

Stich, B. (2009). "Comparison of mating designs for establishing nested association mapping populations in maize and Arabidopsis thaliana." Genetics 183(4): 1525-1534.
Thompson, L. (1989). "Nutritional and physiological effects of phytic acid."

Vijayakumari, K., P. Siddhuraju, et al. (1996). "Effect of soaking, cooking and autoclaving on phytic acid and oligosaccharide contents of the tribal pulse,< i> Mucuna monosperma</i> DC. ex. Wight." Food Chemistry 55(2): 173-177.

Vinjamoori, D., J. Byrum, et al. (2004). "Challenges and opportunities in the analysis of raffinose oligosaccharides, pentosans, phytate, and glucosinolates." Journal of animal science 82(1): 319-328.

Walker, D., A. Scaboo, et al. (2006). "Genetic mapping of loci associated with seed phytic acid content in CX1834-1-2 soybean." Crop Science 46(1): 390-397.

Wehling, R., M. Pierce, et al. (1988). "Determination of Moisture, Fat and Protein in Spray- Dried Whole Egg by Near Infrared Reflectance Spectroscopy." Journal of Food Science 53(5): 1355-1359.

Wilcox, J. R. (1998). "Increasing seed protein in soybean with eight cycles of recurrent selection." Crop Science 38(6): 1536-1540.

Wilcox, J. R., G. S. Premachandra, et al. (2000). "Isolation of high seed inorganic P, low-phytate soybean mutants." Crop Science 40(6): 1601-1605.

Wilson, R. (1991). Designing value-added soybeans for markets of the future, American Oil Chemists.

Yamka, R., B. Hetzler, et al. (2005). "Evaluation of low-oligosaccharide, low-phytate whole soybeans and soybean meal in canine foods." Journal of animal science 83(2): 393-399.

Yamka, R. M. (2003). "Evaluation of low-oligosaccharide and low-oligosaccaride low-phytate whole soybeans and soybean meal in canine foods."

York, J. D., A. R. Odom, et al. (1999). "A phospholipase C-dependent inositol polyphosphate kinase pathway required for efficient messenger RNA export." Science 285(5424): 96-100.

Yuan, F.-J., H.-J. Zhao, et al. (2007). "Generation and characterization of two novel low phytate mutations in soybean (Glycine max L. Merr.)." Theoretical and Applied Genetics 115(7): 945-957.

Zhu, C., M. Gore, et al. (2008). "Status and prospects of association mapping in plants." The plant genome 1(1): 5-20.

Zuo, Y., G. Fahey, et al. (1996). "Digestion responses to low oligosaccharide soybean meal by ileally-cannulated dogs." Journal of animal science 74(10): 2441-2449.

Coward, L., N. C. Barnes, et al. (1993). "Genistein, daidzein, and their. beta.-glycoside conjugates: antitumor isoflavones in soybean foods from American and Asian diets." Journal of agricultural and food chemistry 41(11): 1961-1967.

Fageria, N. K., V. C. Baligar, et al. (2011). Growth and mineral nutrition of field crops, Taylor & Francis US.

Kennedy, A. R. (1995). "The evidence for soybean products as cancer preventive agents." The Journal of nutrition 125(3 Suppl): 733S-743S.

Kennedy, A. R. and B. F. Szuhaj (1994). Bowman-Birk inhibitor product for use as an anticarcinogenesis agent, Google Patents.

Keshun, L. (1997). Soybeans: chemistry, technology, and utilization, Chapman & Hall.
Kwon, S. (2009). "Nutrition & Education International Annual Report." 1-16.

Messina, M. J. (1999). "Legumes and soybeans: overview of their nutritional profiles and health effects." The American journal of clinical nutrition 70(3): 439s-450s.

Stats, S. (2001). "A reference guide to important soybean facts and figures." American Soybean Association.

Wang, H.-j. and P. A. Murphy (1994). "Isoflavone content in commercial soybean foods." Journal of agricultural and food chemistry 42(8): 1666-1673.

Association, I. S. (2010). "SOYBEAN HISTORY AT A GLANCE."

Baize, J. "2013 Global Market Outlook for Soybeans,Soymeal and Corn." PowerPoint.

Dwevedi, A. and A. M. Kayastha (2011). "Soybean: a multifaceted legume with enormous economic capabilities." Soyben-biochemistry, chemistry and physiology: 177-197.

Hymowitz, T. (1990). "Soybeans: The success story." Advances in new crops. Timber Press, Portland, OR: 159-163.

Shurtleff, W. (2010). History of Soybeans and Soyfoods in Southeast Asia (13th Century To 2010): Extensively Annotated Bibliography and Sourcebook, Soyinfo Center.

Association, I. S. (2010). "SOYBEAN HISTORY AT A GLANCE."

Shurtleff, W. (2010). History of Soybeans and Soyfoods in Southeast Asia (13th Century To 2010): Extensively Annotated Bibliography and Sourcebook, Soyinfo Center.

Abdurakhmonov, I. Y. and A. Abdukarimov (2008). "Application of association mapping to understanding the genetic diversity of plant germplasm resources." International journal of plant genomics 2008.

Abe, T., T. Ujiie, et al. (2004). "Varietal differences in free amino acid and sugar concentrations in immature seeds of soybean under raw and boiling treatments." JOURNAL-JAPANESE SOCIETY OF FOOD SCIENCE AND TECHNOLOGY 51(3): 172-176.

Agyeman, A., E. Parkes, et al. (2015). "AMMI and GGE biplot analyses of root yield performance of cassava genotypes in forest and coastal ecologies." International Journal of Agricultural Policy and Research 3(3): 122-132.

Akond, M., S. Liu, et al. (2014). "Identification of quantitative trait Loci (QTL) underlying protein, oil, and five major fatty acids' contents in soybean." American Journal of Plant Sciences 2014.

Aranzana, M. J., S. Kim, et al. (2005). "Genome-wide association mapping in Arabidopsis identifies previously known flowering time and pathogen resistance genes." PLoS Genetics 1(5): e60.

Association, I. S. (2010). "SOYBEAN HISTORY AT A GLANCE."

Baize, J. "2013 Global Market Outlook for Soybeans,Soymeal and Corn." PowerPoint.

Bates, D. M. (2010). "lme4: Mixed-effects modeling with R." URL http://lme4. r-forge. r-project. org/book.

Bentsink, L., C. Alonso-Blanco, et al. (2000). "Genetic analysis of seed-soluble oligosaccharides in relation to seed storability of Arabidopsis." Plant Physiology 124(4): 1595-1604.

Bernardo, R. (2002). Breeding for quantitative traits in plants, Stemma press Woodbury.

Buckler, E. S., J. B. Holland, et al. (2009). "The genetic architecture of maize flowering time." Science 325(5941): 714-718.

Carlson, J. B. and N. R. Lersten (2004). "Reproductive morphology." Soybeans: improvement, production, and uses(soybeansimprove): 59-95.

Center, C. (1995). "Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results." Nature genetics 11: 241.

Chung, J., H. Babka, et al. (2003). "The seed protein, oil, and yield QTL on soybean linkage group I." Crop Science 43(3): 1053-1067.

Cicek, M. (1997). Genetic analysis of quantitative trait loci associated with seed sucrose content using molecular markers in an interspecific Glycine cross, Citeseer.

Clarke, E. and J. Wiseman (2000). "Developments in plant breeding for improved nutritional quality of soya beans II. Anti-nutritional factors." The Journal of Agricultural Science 134(2): 125-136.

Clevinger, E. M. (2006). Mapping Quantitative Trait Loci for Soybean Quality Traits from Two Different Sources, Virginia Polytechnic Institute and State University.

Cober, E. and H. D Voldeng (2000). "Developing high-protein, high-yield soybean populations and lines." Crop Science 40(1): 39-42.

Collard, B., M. Jahufer, et al. (2005). "An introduction to markers, quantitative trait loci (QTL) mapping and marker-assisted selection for crop improvement: The basic concepts." Euphytica 142(1-2): 169-196.

Coward, L., N. C. Barnes, et al. (1993). "Genistein, daidzein, and their. beta.-glycoside conjugates: antitumor isoflavones in soybean foods from American and Asian diets." Journal of agricultural and food chemistry 41(11): 1961-1967.

Csanadi, G., J. Vollmann, et al. (2001). "Seed quality QTL identified in a molecular map of early maturing soybean." Theoretical and Applied Genetics 103(6-7): 912-919.

Dey, P. M. and R. Dixon (1985). Biochemistry of storage carbohydrates in green plants, Academic Press.

Diers, B., P. Keim, et al. (1992). "RFLP analysis of soybean seed protein and oil content." Theoretical and Applied Genetics 83(5): 608-612.

Ding, M., B. Tier, et al. (2007). Application of GGE biplot analysis to evaluate Genotype (G), Environment (E) and GxE interaction on P. radiata: a case study. Australasian Forest Genetics Conference Breeding for Wood Quality.

Dornbos Jr, D. and R. Mullen (1992). "Soybean seed protein and oil contents and fatty acid composition adjustments by drought and temperature." Journal of the American Oil Chemists Society 69(3): 228-231.

Dwevedi, A. and A. M. Kayastha (2011). "Soybean: a multifaceted legume with enormous economic capabilities." Soyben-biochemistry, chemistry and physiology: 177-197.

Fageria, N. K., V. C. Baligar, et al. (2011). Growth and mineral nutrition of field crops, Taylor & Francis US.

Farshadfar, E., M. M. Poursiahbidi, et al. (2012). "Evaluation of phenotypic stability in bread wheat genotypes using GGE-biplot." International Journal of Agriculture and Crop Sciences 4(13): 904-910.

Flint- Garcia, S. A., A. C. Thuillet, et al. (2005). "Maize association population: a high- resolution platform for quantitative trait locus dissection." The Plant Journal 44(6): 1054-1064.

Ghosh, J., P. Ghosh, et al. (2014). "An Assessment of Genetic Relatedness between Soybean [Glycine max (L.) Merrill] Cultivars Using SSR Markers." American Journal of Plant Sciences 5(20): 3089.

Gupta, P. K., S. Rustgi, et al. (2005). "Linkage disequilibrium and association studies in higher plants: present status and future prospects." Plant Molecular Biology 57(4): 461-485.

Hagos, H. G. and F. Abay (2013). "AMMI AND GGE biplot analysis of bread wheat genotypes in the northern part of Ethiopia." Journal of Plant Breeding and Genetics 1(1): 12-18.

Hall, D., C. Tegström, et al. (2010). "Using association mapping to dissect the genetic basis of complex traits in plants." Briefings in Functional Genomics 9(2): 157-165.

Hedley, C. L. (2001). Carbohydrates in grain legume seeds [electronic resource]: improving nutritional quality and agronomic characteristics, CABI.

Holland, J. B. (2007). "Genetic architecture of complex traits in plants." Current opinion in plant biology 10(2): 156-161.

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

Hou, A., P. Chen, et al. (2009). "Genetic variability of seed sugar content in worldwide soybean germplasm collections." Crop Science 49(3): 903-912.

Hou, A., P. Chen, et al. (2008). "Sugar variation in soybean seed assessed with a rapid extraction and quantification method." International Journal of Agronomy 2009.

Hu, G., C. Liu, et al. (2011). Integration of Major QTL of Important Agronomic Traits in Soybean, INTECH Open Access Publisher.

Huhn, M. R. (2003). Inheritance of soluble oligosaccharide content of soybean seeds, Virginia Polytechnic Institute and State University.

Hwang, E.-Y., Q. Song, et al. (2014). "A genome-wide association study of seed protein and oil content in soybean." BMC genomics 15(1): 1.

Hymowitz, T. (1990). "Soybeans: The success story." Advances in new crops. Timber Press, Portland, OR: 159-163.

Hymowitz, T. and C. Newell (1981). "Taxonomy of the genusGlycine, domestication and uses of soybeans." Economic Botany 35(3): 272-288.

Hyten, D., V. Pantalone, et al. (2004). "Seed quality QTL in a prominent soybean population." Theoretical and Applied Genetics 109(3): 552-561.

Hyten, D. L., I.-Y. Choi, et al. (2007). "Highly variable patterns of linkage disequilibrium in multiple soybean populations." Genetics 175(4): 1937-1944.

Jones, D., M. DuPont, et al. (1999). "The discovery of compositional variation for the raffinose family of oligosaccharides in pea seeds." Seed Science Research 9(04): 305-310.

Karner, U., T. Peterbauer, et al. (2004). "myo-Inositol and sucrose concentrations affect the accumulation of raffinose family oligosaccharides in seeds." Journal of experimental botany 55(405): 1981-1987.

Kennedy, A. R. (1995). "The evidence for soybean products as cancer preventive agents." The Journal of nutrition 125(3 Suppl): 733S-743S.

Kennedy, A. R. and B. F. Szuhaj (1994). Bowman-Birk inhibitor product for use as an anticarcinogenesis agent, Google Patents.

Keshun, L. (1997). Soybeans: chemistry, technology, and utilization, Chapman & Hall.
Kloth, K. J., M. P. Thoen, et al. (2012). "Association mapping of plant resistance to insects." Trends in plant science 17(5): 311-319.

Kwon, S. (2009). "Nutrition & Education International Annual Report." 1-16.

Lander, E. S., P. Green, et al. (1987). "MAPMAKER: an interactive computer package for constructing primary genetic linkage maps of experimental and natural populations." Genomics 1(2): 174-181.

Latham, M. C. (1997). Human nutrition in the developing world, Food & Agriculture Org.

Lee, S., M. Bailey, et al. (1996). "RFLP loci associated with soybean seed protein and oil content across populations and locations." Theoretical and Applied Genetics 93(5-6): 649-657.

Lowell, C. A. and T. M. Kuo (1989). "Oligosaccharide metabolism and accumulation in developing soybean seeds." Crop Science 29(2): 459-465.

Mansur, L., K. Lark, et al. (1993). "Interval mapping of quantitative trait loci for reproductive, morphological, and seed traits of soybean (Glycine max L.)." Theoretical and Applied Genetics 86(8): 907-913.

Messina, M. J. (1999). "Legumes and soybeans: overview of their nutritional profiles and health effects." The American journal of clinical nutrition 70(3): 439s-450s.

Middelbos and Fahey Jr, Ed. (2008). Soybean Carbohydrates: Chemistry, Production Processing, and Utilization, AOSC Press, Urbana, Illinois.

Moongkanna, J., S. Nakasathien, et al. (2011). "SSR markers linking to seed traits and total oil content in soybean." Thai Journal of Agricultural Science 44(4): 233-241.

Murphy, P. A. (2008). Soybeans: Chemistry, Production Processing, and Utilization. AOCS Press, Urbana, Illinois.: 229-268.

Myles, S., J. Peiffer, et al. (2009). "Association mapping: critical considerations shift from genotyping to experimental design." The Plant Cell Online 21(8): 2194-2202.

Obendorf, R. L. (1997). "Oligosaccharides and galactosyl cyclitols in seed desiccation tolerance." Seed Science Research 7(2): 63-74.

Panthee, D., V. Pantalone, et al. (2005). "Quantitative trait loci for seed protein and oil concentration, and seed size in soybean." Crop Science 45(5): 2015-2022.

Panthee, D. R. (2005). "Genetic Mapping of Quantitative Trait Loci Conditioning Protein Concentration and Quality, and Other Seed Characteristics in Soybean [Glycine max (L.) Merrill]."

Pedersen, P. and B. Elbert (2004). Soybean growth and development, Iowa State University, University Extension Ames, IA.

Pennycooke, J. C., M. L. Jones, et al. (2003). "Down-regulating α-galactosidase enhances freezing tolerance in transgenic petunia." Plant Physiology 133(2): 901-909.

Phansak, P. (2010). "Detection of soybean seed protein QTL using selective genotyping."

Piper, E. L. and K. I. Boote (1999). "Temperature and cultivar effects on soybean seed oil and protein concentrations." Journal of the American Oil Chemists' Society 76(10): 1233-1241.

Qiu, B., P. Arelli, et al. (1999). "RFLP markers associated with soybean cyst nematode resistance and seed composition in a 'Peking'×'Essex'population." Theoretical and Applied Genetics 98(3-4): 356-364.

Risch, N. and K. Merikangas (1996). "The future of genetic studies of complex human diseases." Science-AAAS-Weekly Paper Edition 273(5281): 1516-1517.

Semagn, K., Å. Bjørnstad, et al. (2010). "The genetic dissection of quantitative traits in crops." Electronic Journal of Biotechnology 13(5): 16-17.

Shafii, B. and W. J. Price (1998). "Analysis of genotype-by-environment interaction using the additive main effects and multiplicative interaction model and stability estimates." Journal of Agricultural, Biological, and Environmental Statistics: 335-345.

Shoemaker, R. and J. Specht (1995). "Integration of the soybean molecular and classical genetic linkage groups." Crop Science 35(2): 436-446.

Shurtleff, W. (2010). History of Soybeans and Soyfoods in Southeast Asia (13th Century To 2010): Extensively Annotated Bibliography and Sourcebook, Soyinfo Center.

Singh, R. and T. Hymowitz (1999). "Soybean genetic resources and crop improvement." Genome 42(4): 605-616.

Skoneczka, J., M. Maroof, et al. (2009). "Identification of candidate gene mutation associated with low stachyose phenotype in soybean line PI200508." Crop Science 49(1): 247-255.

Sneller, C., D. E. Mather, et al. (2009). "Analytical approaches and population types for finding and utilizing QTL in complex plant populations." Crop Science 49(2): 363-380.

Sorrells, M. E. and J. Yu (2009). Linkage disequilibrium and association mapping in the Triticeae. Genetics and genomics of the Triticeae, Springer: 655-683.

Soto-Cerda, B. J. and S. Cloutier (2012). "Association mapping in plant genomes." Genetic diversity in plants. InTech, Rijeka: 29-54.

Specht, J., K. Chase, et al. (2001). "Soybean response to water." Crop Science 41(2): 493-509.

Sprenger, N. and F. Keller (2000). "Allocation of raffinose family oligosaccharides to transport and storage pools in Ajuga reptans: the roles of two distinct galactinol synthases." The Plant Journal 21(3): 249-258.

Stats, S. (2001). "A reference guide to important soybean facts and figures." American Soybean Association.

Stich, B. (2009). "Comparison of mating designs for establishing nested association mapping populations in maize and Arabidopsis thaliana." Genetics 183(4): 1525-1534.

Tahir, M., M. Båga, et al. (2012). "An Assessment of Raffinose Family Oligosaccharides and Sucrose Concentration in Genus." Crop Science 52(4): 1713-1720.

van Kleunen, M. and K. Ritland (2005). "Estimating heritabilities and genetic correlations with marker-based methods: an experimental test in Mimulus guttatus." Journal of Heredity 96(4): 368-375.

Vinjamoori, D., J. Byrum, et al. (2004). "Challenges and opportunities in the analysis of raffinose oligosaccharides, pentosans, phytate, and glucosinolates." Journal of animal science 82(1): 319-328.

Wang, H.-j. and P. A. Murphy (1994). "Isoflavone content in commercial soybean foods." Journal of agricultural and food chemistry 42(8): 1666-1673.

Wilcox, J. R. and J. F. Cavins (1995). "Backcrossing high seed protein to a soybean cultivar." Crop Science 35(4): 1036-1041.

Xavier, A. (2015). Mixed Model Approach for Genotypic Imputation. Plant and Animal Genome XXIII Conference, Plant and Animal Genome.

Zhang, J., A. Singh, et al. (2015). "Genome- wide association and epistasis studies unravel the genetic architecture of sudden death syndrome resistance in soybean." The Plant Journal 84(6): 1124-1136.

Zhu, C., M. Gore, et al. (2008). "Status and prospects of association mapping in plants." The plant genome 1(1): 5-20.

Schnebly, S. and W. Fehr (1993). "Effect of years and planting dates on fatty acid composition of soybean genotypes." Crop Science 33(4): 716-719.

Yan, W., L. Hunt, et al. (2000). "Cultivar evaluation and mega-environment investigation based on the GGE biplot." Crop Science 40(3): 597-605.

Zhe, Y., J. G. Lauer, et al. (2010). "Effects of genotype× environment interaction on agronomic traits in soybean." Crop Science 50(2): 696-702.

Abdurakhmonov, I. Y. and A. Abdukarimov (2008). "Application of association mapping to understanding the genetic diversity of plant germplasm resources." International journal of plant genomics **2008**.

Abe, T., T. Ujiie, et al. (2004). "Varietal differences in free amino acid and sugar concentrations in immature seeds of soybean under raw and boiling treatments." JOURNAL-JAPANESE SOCIETY OF FOOD SCIENCE AND TECHNOLOGY **51**(3): 172-176.

Agyeman, A., E. Parkes, et al. (2015). "AMMI and GGE biplot analyses of root yield performance of cassava genotypes in forest and coastal ecologies." International Journal of Agricultural Policy and Research **3**(3): 122-132.

Akond, M., S. Liu, et al. (2014). "Identification of quantitative trait Loci (QTL) underlying protein, oil, and five major fatty acids' contents in soybean." American Journal of Plant Sciences **2014**.

Akond, M., S. Liu, et al. (2015). "Quantitative Trait Loci Underlying Seed Sugars Content in" MD96-5722" by" Spencer" Recombinant Inbred Line Population of Soybean." Food and Nutrition Sciences **6**(11): 964.

Aranzana, M. J., S. Kim, et al. (2005). "Genome-wide association mapping in Arabidopsis identifies previously known flowering time and pathogen resistance genes." PLoS Genetics **1**(5): e60.

Baize, J. (2013). "Global Market Outlook for Soybeans, Soymeal and Corn."

Baker, R. (1988). "Tests for crossover genotype-environmental interactions." Canadian journal of plant science **68**(2): 405-410.

Balestre, M., V. B. d. Santos, et al. (2010). "Stability and adaptability of upland rice genotypes." Crop Breeding and Applied Biotechnology **10**(4): 357-363.

Bandillo, N., D. Jarquin, et al. (2015). "A population structure and genome-wide association analysis on the USDA soybean germplasm collection." The plant genome **8**(3).

Bates, D., M. Maechler, et al. (2014). "lme4: Linear mixed-effects models using Eigen and S4." R package version **1**(7).

Bates, D. M. (2010). "lme4: Mixed-effects modeling with R." URL http://lme4. r-forge. r-project. org/book.

Bentsink, L., C. Alonso-Blanco, et al. (2000). "Genetic analysis of seed-soluble oligosaccharides in relation to seed storability of Arabidopsis." Plant Physiology **124**(4): 1595-1604.

Bernardo, R. (2002). Breeding for quantitative traits in plants, Stemma press Woodbury.

Broman, K. W., H. Wu, et al. (2003). "R/qtl: QTL mapping in experimental crosses." Bioinformatics **19**(7): 889-890.

Brummer, E., G. Graef, et al. (1997). "Mapping QTL for seed protein and oil content in eight soybean populations." Crop Science **37**(2): 370-378.

Buckler, E. S., J. B. Holland, et al. (2009). "The genetic architecture of maize flowering time." Science **325**(5941): 714-718.

Carlson, J. B. and N. R. Lersten (2004). "Reproductive morphology." Soybeans: improvement, production, and uses(soybeansimprove): 59-95.

Center, C. (1995). "Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results." Nature genetics **11**: 241.

Choung, M.-G. (2010). "Determination of sucrose content in soybean using near-infrared reflectance spectroscopy." Journal of the Korean Society for Applied Biological Chemistry **53**(4): 478-484.

Chung, J., H. Babka, et al. (2003). "The seed protein, oil, and yield QTL on soybean linkage group I." Crop Science **43**(3): 1053-1067.

Churchill, G. A. and R. W. Doerge (1994). "Empirical threshold values for quantitative trait mapping." Genetics **138**(3): 963-971.

Cicek, M. (1997). Genetic analysis of quantitative trait loci associated with seed sucrose content using molecular markers in an interspecific Glycine cross, Citeseer.

Clarke, E. and J. Wiseman (2000). "Developments in plant breeding for improved nutritional quality of soya beans II. Anti-nutritional factors." The Journal of Agricultural Science **134**(2): 125-136.

Clevinger, E. M. (2006). "Mapping quantitative trait loci for soybean quality traits from two different sources."

Clevinger, E. M. (2006). Mapping Quantitative Trait Loci for Soybean Quality Traits from Two Different Sources, Virginia Polytechnic Institute and State University.

Cober, E. and H. D Voldeng (2000). "Developing high-protein, high-yield soybean populations and lines." Crop Science **40**(1): 39-42.

Collard, B., M. Jahufer, et al. (2005). "An introduction to markers, quantitative trait loci (QTL) mapping and marker-assisted selection for crop improvement: The basic concepts." Euphytica **142**(1-2): 169-196.

Cook, J. P., M. D. McMullen, et al. (2012). "Genetic architecture of maize kernel composition in the nested association mapping and inbred association panels." Plant physiology **158**(2): 824-834.

Coward, L., N. C. Barnes, et al. (1993). "Genistein, daidzein, and their. beta.-glycoside conjugates: antitumor isoflavones in soybean foods from American and Asian diets." Journal of agricultural and food chemistry **41**(11): 1961-1967.

Csanadi, G., J. Vollmann, et al. (2001). "Seed quality QTLs identified in a molecular map of early maturing soybean." Theoretical and Applied Genetics **103**(6-7): 912-919.

Dei, H. (2011). Soybean as a Feed Ingredient for Livestock and Poultry, INTECH Open Access Publisher.

Dey, P. M. and R. Dixon (1985). Biochemistry of storage carbohydrates in green plants, Academic Press.

Dierking, E. (2009). Characterization of raffinose synthase genes in soybean, University of Missouri--Columbia.

Diers, B., P. Keim, et al. (1992). "RFLP analysis of soybean seed protein and oil content." Theoretical and Applied Genetics **83**(5): 608-612.

Ding, M., B. Tier, et al. (2007). Application of GGE biplot analysis to evaluate Genotype (G), Environment (E) and GxE interaction on P. radiata: a case study. Australasian Forest Genetics Conference Breeding for Wood Quality.

Dornbos Jr, D. and R. Mullen (1992). "Soybean seed protein and oil contents and fatty acid composition adjustments by drought and temperature." Journal of the American Oil Chemists Society **69**(3): 228-231.

Dwevedi, A. and A. Kayastha (2011). "Soybean: a multifaceted legume with enormous economic capabilities, soybean-biochemistry, chemistry and physiology." In Tech.

Ersoz, E. S., J. Yu, et al. (2009). Applications of linkage disequilibrium and association mapping in maize. Molecular Genetic Approaches to Maize Improvement, Springer**:** 173-195.

Eskandari, M., E. R. Cober, et al. (2013). "Genetic control of soybean seed oil: II. QTL and genes that increase oil concentration without decreasing protein or with increased seed yield." Theoretical and Applied Genetics **126**(6): 1677-1687.

Fageria, N. K., V. C. Baligar, et al. (2011). Growth and mineral nutrition of field crops, Taylor & Francis US.

Farshadfar, E., M. M. Poursiahbidi, et al. (2012). "Evaluation of phenotypic stability in bread wheat genotypes using GGE-biplot." International Journal of Agriculture and Crop Sciences **4**(13): 904-910.

Flint-Garcia, S. A., A. C. Thuillet, et al. (2005). "Maize association population: a high-resolution platform for quantitative trait locus dissection." The Plant Journal **44**(6): 1054-1064.

Frutos, E., M. P. Galindo, et al. (2014). "An interactive biplot implementation in R for modeling genotype-by-environment interaction." Stochastic Environmental Research and Risk Assessment **28**(7): 1629-1641.

Gauch, H. G. (2006). "Statistical analysis of yield trials by AMMI and GGE." Crop Science **46**(4): 1488-1500.

Ghosh, J., P. Ghosh, et al. (2014). "An Assessment of Genetic Relatedness between Soybean [Glycine max (L.) Merrill] Cultivars Using SSR Markers." American Journal of Plant Sciences **5**(20): 3089.

Greenacre, M. J. (2010). Biplots in practice, Fundacion BBVA.

Gupta, P. K., S. Rustgi, et al. (2005). "Linkage disequilibrium and association studies in higher plants: present status and future prospects." Plant Molecular Biology **57**(4): 461-485.

Gurmu, F., H. Mohammed, et al. (2009). "Genotype x Environment interactions and stability of soybean for grain yield and nutrition quality." African Crop Science Journal **17**(2).

Hagos, H. G. and F. Abay (2013). "AMMI AND GGE biplot analysis of bread wheat genotypes in the northern part of Ethiopia." Journal of Plant Breeding and Genetics **1**(1): 12-18.

Hall, D., C. Tegström, et al. (2010). "Using association mapping to dissect the genetic basis of complex traits in plants." Briefings in Functional Genomics **9**(2): 157-165.

Hedley, C. L. (2000). Carbohydrates in grain legume seeds: Improving nutritional quality and agronomic characteristics, CABI.

Hedley, C. L. (2001). Carbohydrates in grain legume seeds [electronic resource]: improving nutritional quality and agronomic characteristics, CABI.

Holland, J. B. (2007). "Genetic architecture of complex traits in plants." Current opinion in plant biology **10**(2): 156-161.

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

Hothorn, T. and B. S. Everitt (2014). A handbook of statistical analyses using R, CRC press.

Hou, A., P. Chen, et al. (2009). "Genetic variability of seed sugar content in worldwide soybean germplasm collections." Crop Science **49**(3): 903-912.

Hou, A., P. Chen, et al. (2008). "Sugar variation in soybean seed assessed with a rapid extraction and quantification method." International Journal of Agronomy **2009**.

Hu, G., C. Liu, et al. (2011). Integration of Major QTLs of Important Agronomic Traits in Soybean, INTECH Open Access Publisher.

Huehn, M. (2011). "On the bias of recombination fractions, Kosambi's and Haldane's distances based on frequencies of gametes." Genome **54**(3): 196-201.

Huhn, M. R. (2003). Inheritance of soluble oligosaccharide content of soybean seeds, Virginia Polytechnic Institute and State University.

Hwang, E.-Y., Q. Song, et al. (2014). "A genome-wide association study of seed protein and oil content in soybean." BMC genomics **15**(1): 1.

Hymowitz, T. (1990). "Soybeans: The success story." Advances in new crops: 159-163.

Hymowitz, T. and C. Newell (1981). "Taxonomy of the genusGlycine, domestication and uses of soybeans." Economic Botany **35**(3): 272-288.

Hyten, D., V. Pantalone, et al. (2004). "Seed quality QTL in a prominent soybean population." Theoretical and Applied Genetics **109**(3): 552-561.

Hyten, D. L., I.-Y. Choi, et al. (2007). "Highly variable patterns of linkage disequilibrium in multiple soybean populations." Genetics **175**(4): 1937-1944.

Jaureguy, L. M., P. Chen, et al. (2011). "Heritability and correlations among food-grade traits in soybean." Plant Breeding **130**(6): 647-652.

Johnson, H. W., H. Robinson, et al. (1955). "Estimates of genetic and environmental variability in soybeans." Agronomy Journal **47**(7): 314-318.

Jones, D., M. DuPont, et al. (1999). "The discovery of compositional variation for the raffinose family of oligosaccharides in pea seeds." Seed Science Research **9**(04): 305-310.

Kang, M. S. and H. G. Gauch (1996). "Genotype-by-environment interaction."

Karner, U., T. Peterbauer, et al. (2004). "myo-Inositol and sucrose concentrations affect the accumulation of raffinose family oligosaccharides in seeds." Journal of experimental botany **55**(405): 1981-1987.

Kaya, Y., M. Akçura, et al. (2006). "GGE-biplot analysis of multi-environment yield trials in bread wheat." Turkish Journal of Agriculture and Forestry **30**(5): 325-337.

Kennedy, A. R. (1995). "The evidence for soybean products as cancer preventive agent." The Journal of nutrition **125**(3): 733S.

Kennedy, A. R. and B. F. Szuhaj (1994). Bowman-Birk inhibitor product for use as an anticarcinogenesis agent, Google Patents.

Keshun, L. (1997). Soybeans: chemistry, technology, and utilization, Chapman & Hall.

Kim, H. K., S. T. Kang, et al. (2006). "Mapping of putative quantitative trait loci controlling the total oligosaccharide and sucrose content of Glycine max seeds." Journal of plant research **119**(5): 533-538.

Kim, J.-S., P. E. Klein, et al. (2005). "Chromosome Identification and Nomenclature of Sorghum bicolor." Genetics **169**(2): 1169-1173.

> Linkage group identities and homologies were determined for metaphase chromosomes of Sorghum bicolor (2n = 20) by FISH of landed BACs. Relative lengths of chromosomes in FISH-karyotyped metaphase spreads of the elite inbred BTx623 were used to estimate the molecular size of each chromosome and to establish a size-based nomenclature for sorghum chromosomes (SBI-01-SBI-10) and linkage groups (LG-01 to LG-10). Lengths of arms were determined to orient linkage groups relative to a standard karyotypic layout (short arms at top). The size-based nomenclature for BTx623 represents a reasonable choice as the standard for a unified chromosome nomenclature for use by the sorghum research community.

Kloth, K. J., M. P. Thoen, et al. (2012). "Association mapping of plant resistance to insects." Trends in plant science **17**(5): 311-319.

Kwon, J. M. and A. M. Goate (2000). "The candidate gene approach." Alcohol research and health **24**(3): 164-168.

Kwon, S. (2009). "Nutrition & Education International Annual Report." 1-16.

Lander, E. S., P. Green, et al. (1987). "MAPMAKER: an interactive computer package for constructing primary genetic linkage maps of experimental and natural populations." Genomics **1**(2): 174-181.

Latham, M. C. (1997). Human nutrition in the developing world, Food & Agriculture Org.

Lee, S., M. Bailey, et al. (1996). "RFLP loci associated with soybean seed protein and oil content across populations and locations." Theoretical and Applied Genetics **93**(5-6): 649-657.

Lian, L. and G. de los Campos (2015). "FW: An R package for Finlay-Wilkinson Regression That Incorporates Genomic/Pedigree Information and Covariance Structures Between Environments." G3: Genes| Genomes| Genetics: g3. 115.026328.

Lowell, C. A. and T. M. Kuo (1989). "Oligosaccharide metabolism and accumulation in developing soybean seeds." Crop Science **29**(2): 459-465.

Lu, W., Z. Wen, et al. (2013). "Identification of the quantitative trait loci (QTL) underlying water soluble protein content in soybean." Theoretical and Applied Genetics **126**(2): 425-433.

Lyimo, H., R. Pratt, et al. (2012). "Variation in aggressiveness among isolates of Cercospora zeae-maydis inlow-, medium-and high-altitude maize agro-ecologies of Tanzania." Archives of Phytopathology and Plant Protection **45**(9): 1076-1086.

Lynch, M. and B. Walsh (1998). "Genetics and analysis of quantitative traits."

Maughan, P., M. S. Maroof, et al. (2000). "Identification of quantitative trait loci controlling sucrose content in soybean (Glycine max)." Molecular Breeding **6**(1): 105-111.

Meade, K. A. (2012). "Genetic dissection of canonical models of maize kernel growth and development."

Messina, M. J. (1999). "Legumes and soybeans: overview of their nutritional profiles and health effects." The American journal of clinical nutrition **70**(3): 439s-450s.

Middelbos and Fahey Jr, Ed. (2008). Soybean Carbohydrates: Chemistry, Production Processing, and Utilization, AOSC Press, Urbana, Illinois.

MITROVIÃ, B., S. TRESKI, et al. (2012). "Evaluation of Experımental Maize Hybrids Tested in Multi-Location Trials Using Ammi and GGE Biplot Analyses." Turkish Journal of field crops **17**(1): 35-40.

Mohammadi, R. and A. Amri (2013). "Genotype× environment interaction and genetic improvement for yield and yield stability of rainfed durum wheat in Iran." Euphytica **192**(2): 227-249.

Moongkanna, J., S. Nakasathien, et al. (2011). "SSR markers linking to seed traits and total oil content in soybean." Thai Journal of Agricultural Science **44**(4): 233-241.

Murphy, P. A. (2008). Soybeans: Chemistry, Production Processing, and Utilization. AOCS Press, Urbana, Illinois.**:** 229-268.

Myles, S., J. Peiffer, et al. (2009). "Association mapping: critical considerations shift from genotyping to experimental design." The Plant Cell Online **21**(8): 2194-2202.

Nazareth, Z. M. (2009). "Compositional, functional and sensory properties of protein ingredients."

Nzuve, F., S. Githiri, et al. (2013). "Analysis of genotype x environment interaction for grain yield in Maize hybrids." Journal of Agricultural Science **5**(11): 75.

Obendorf, R. L. (1997). "Oligosaccharides and galactosyl cyclitols in seed desiccation tolerance." Seed Science Research **7**(2): 63-74.

Panthee, D., V. Pantalone, et al. (2005). "Quantitative trait loci for seed protein and oil concentration, and seed size in soybean." Crop Science **45**(5): 2015-2022.

Pedersen, P. and B. Elbert (2004). Soybean growth and development, Iowa State University, University Extension Ames, IA.

Pennycooke, J. C., M. L. Jones, et al. (2003). "Down-regulating α-galactosidase enhances freezing tolerance in transgenic petunia." Plant Physiology **133**(2): 901-909.

Phansak, P. (2010). "Detection of soybean seed protein QTLS using selective genotyping."

Piper, E. L. and K. I. Boote (1999). "Temperature and cultivar effects on soybean seed oil and protein concentrations." Journal of the American Oil Chemists' Society **76**(10): 1233-1241.

Prado, S. A., C. G. López, et al. (2014). "The genetic architecture of maize (Zea mays L.) kernel weight determination." G3: Genes| Genomes| Genetics **4**(9): 1611-1621.

Qiu, B., P. Arelli, et al. (1999). "RFLP markers associated with soybean cyst nematode resistance and seed composition in a 'Peking'×'Essex'population." Theoretical and Applied Genetics **98**(3-4): 356-364.

Rakshit, S., K. Ganapathy, et al. (2012). "GGE biplot analysis to evaluate genotype, environment and their interactions in sorghum multi-location data." Euphytica **185**(3): 465-479.

Risch, N. and K. Merikangas (1996). "The future of genetic studies of complex human diseases." Science-AAAS-Weekly Paper Edition **273**(5281): 1516-1517.

Salvi, S., S. Corneti, et al. (2011). "Genetic dissection of maize phenology using an intraspecific introgression library." BMC plant biology **11**(1): 1.

Semagn, K., Å. Bjørnstad, et al. (2010). "The genetic dissection of quantitative traits in crops." Electronic Journal of Biotechnology **13**(5): 16-17.

Shafii, B. and W. J. Price (1992). "STATISTICAL ANALYSIS OF GENOTYPE-BY-ENVIRONMENT INTERACTION USING THE AMMI MODEL AND STABILITY ESTIMATES."

Shafii, B. and W. J. Price (1998). "Analysis of genotype-by-environment interaction using the additive main effects and multiplicative interaction model and stability estimates." Journal of Agricultural, Biological, and Environmental Statistics: 335-345.

Shoemaker, R. and J. Specht (1995). "Integration of the soybean molecular and classical genetic linkage groups." Crop Science **35**(2): 436-446.

Shurtleff, W. and A. Aoyagi (2010). History of Soybeans and Soyfoods in Southeast Asia (13th Century to 2010): Extensively Annotated Bibliography and Sourcebook, Soyinfo Center.

Singh, R. and T. Hymowitz (1999). "Soybean genetic resources and crop improvement." Genome **42**(4): 605-616.

Skoneczka, J., M. Maroof, et al. (2009). "Identification of candidate gene mutation associated with low stachyose phenotype in soybean line PI200508." Crop Science **49**(1): 247-255.

Sneller, C., D. E. Mather, et al. (2009). "Analytical approaches and population types for finding and utilizing QTL in complex plant populations." Crop Science **49**(2): 363-380.

Sorrells, M. E. and J. Yu (2009). Linkage disequilibrium and association mapping in the Triticeae. Genetics and genomics of the Triticeae, Springer**:** 655-683.

Soto-Cerda, B. J. and S. Cloutier (2012). "Association mapping in plant genomes." Genetic diversity in plants. InTech, Rijeka: 29-54.

SoyStats (2012). "Soy stats 2012." Am. Soybean Assoc., St. Louis, MO.

SoyStats (2012). SoyStats: a reference guide to important soybean facts & figures; 2012.

Specht, J., K. Chase, et al. (2001). "Soybean response to water." Crop Science **41**(2): 493-509.

Sprenger, N. and F. Keller (2000). "Allocation of raffinose family oligosaccharides to transport and storage pools in Ajuga reptans: the roles of two distinct galactinol synthases." The Plant Journal **21**(3): 249-258.

Stats, S. (2001). "A reference guide to important soybean facts and figures." American Soybean Association.

Stich, B. (2009). "Comparison of mating designs for establishing nested association mapping populations in maize and Arabidopsis thaliana." Genetics **183**(4): 1525-1534.

Tahir, M., M. Båga, et al. (2012). "An Assessment of Raffinose Family Oligosaccharides and Sucrose Concentration in Genus." Crop Science **52**(4): 1713-1720.

Valliyodan, B., H. Shi, et al. (2015). "A simple analytical method for high-throughput screening of major sugars from soybean by normal-phase HPLC with evaporative light scattering detection." Chromatography Research International **2015**.

van Kleunen, M. and K. Ritland (2005). "Estimating heritabilities and genetic correlations with marker-based methods: an experimental test in Mimulus guttatus." Journal of Heredity **96**(4): 368-375.

Vaughn, J. N., R. L. Nelson, et al. (2014). "The genetic architecture of seed composition in soybean is refined by genome-wide association scans across multiple populations." G3: Genes| Genomes| Genetics **4**(11): 2283-2294.

Wang, S., C. Basten, et al. (2006). Windows QTL Cartographer Raleigh, Department of Statistics, North Carolina State University,
.

Wang, Y., P. Chen, et al. (2014). "Quantitative trait loci analysis of soluble sugar contents in soybean." Plant Breeding **133**(4): 493-498.

WeiKai, Y., L. Hunt, et al. (2001). "Biplot analysis of multi-environment trial data." Quantitative genetics, genomics and plant breeding: 289-303.

Wilcox, J. R. and J. F. Cavins (1995). "Backcrossing high seed protein to a soybean cultivar." Crop Science **35**(4): 1036-1041.

Wilson, L. M., S. R. Whitt, et al. (2004). "Dissection of maize kernel composition and starch production by candidate gene association." The Plant Cell **16**(10): 2719-2733.

Xavier, A. (2015). Mixed Model Approach for Genotypic Imputation. Plant and Animal Genome XXIII Conference, Plant and Animal Genome.

Xiao, Y., H. Tong, et al. (2015). "Genome-wide dissection of the maize ear genetic architecture using multiple populations." New Phytologist.

Xu, S. and W. R. Atchley (1995). "A random model approach to interval mapping of quantitative trait loci." Genetics **141**(3): 1189-1197.

Yan, W. (2002). "Singular-value partitioning in biplot analysis of multienvironment trial data." Agronomy Journal **94**(5): 990-996.

Yan, W., L. Hunt, et al. (2000). "Cultivar evaluation and mega-environment investigation based on the GGE biplot." Crop Science **40**(3): 597-605.

Zeng, A., P. Chen, et al. (2014). "Identification of Quantitative Trait Loci for Sucrose Content in Soybean Seed." Crop Science **54**(2): 554-564.

Zeng, A., P. Chen, et al. (2015). "Identification and confirmation of quantitative trait loci for stachyose content in soybean seed." Plant Breeding **134**(2): 178-185.

Zeng, Z. B. (1994). "Precision Mapping of Quantitative Trait Loci." Genetics **136**(4): 1457-1468.

Zhang, J., A. Singh, et al. (2015). "Genome-wide association and epistasis studies unravel the genetic architecture of sudden death syndrome resistance in soybean." The Plant Journal **84**(6): 1124-1136.

Zhang, Z., X. Wu, et al. (2015). "Genetic dissection of the maize kernel development process via conditional QTL mapping for three developing kernel-related traits in an immortalized F2 population." Molecular Genetics and Genomics: 1-18.

Zhu, C., M. Gore, et al. (2008). "Status and prospects of association mapping in plants." The plant genome **1**(1): 5-20.

APPENDIX

APPENDIX

R codes for chapter 3

```
P&O<-read.table("FILEANOVAPOF46Removed.csv ",fill=T,header=T,sep=",", )
head(P&O)
library(HSAUR2)
str(PO)

# Basic summary
summary (P&O[8:12])
par(mfrow=c(1,1))
boxplot(P&O$Protein~P&O$Env,main="Distribution of Protein by
Environment",cex=0.9,cex.main=2,cex.lab=1.5,cex.axis=1.3,ylab="%Protein",ylim=c(29
,39),las=1, xlab="Environment",col="gold",notch=TRUE,outline = FALSE)

boxplot(P&O$Protein~P&O$FamNo,main="Distribution of Protein by
Population",cex=0.9,cex.main=2,cex.lab=1.5,cex.axis=0.8,ylab="%Protein",ylim=c(29,3
9),las=1, xlab="Family",col="gold",notch=TRUE,outline = FALSE)

boxplot(P&O$Oil~P&O$Env,main="Distribution of Oil by
Environment",cex=0.9,cex.main=2,cex.lab=1.5,cex.axis=1.4,ylab="%Oil",ylim=c(16,23),
las=1, xlab="Environment",col="gold",notch=TRUE,outline = FALSE)

boxplot(P&O$Oil~P&O$FamNo,main="Distribution of Oil by
Population",cex=0.9,cex.main=2,cex.lab=1.5,cex.axis=0.8,ylab="%Oil",ylim=c(16,23),la
s=1, xlab="Family",col="gold",notch=TRUE,outline = FALSE)

par(mfrow=c(2,3))
hist(P&O$Protein,main="Distribution of Individual Plot
Data",cex=1,cex.main=1.7,cex.lab=1.6,cex.axis=1,las=1,
xlab="Protein%",col="gold",notch=TRUE,use="pairwise.complete.obs")
mx <- mean(33.79)
abline(v = mx, col = "red", lwd = 2)
```

```
hist(P&O$Oil,main="Distribution of Individual Plot
Data",cex=1,cex.main=1.7,cex.lab=1.6,cex.axis=1,las=1,
xlab="Oil%",col="Gold",notch=TRUE,use="pairwise.complete.obs")
mx <- mean(19.70)
abline(v = mx, col = "blue", lwd = 2)


# Correlation analysis
Overall correlation between protein and oil
library (psych)
pairs.panels(P&O[8:12])


# Correlation on family basis
require(plyr)
func <- function(xx)
{
  return(data.frame(COR = cor(xx$Protein, xx$FamNO,use="pairwise.complete.obs" )))
}
ddply(P&O, .(FamNo), func)

func <- function(xx)
{
  return(data.frame(COR = cor(xx$Oil, xx$FamNO,use="pairwise.complete.obs" )))
}
ddply(P&O, .(FamNo), func)

# Variance estimates
P&O = read.csv("", header=T, sep="," )
attach(P&O)
# Rename variables for ease of use
Protein = as.numeric(Protein)
Oil = as.numeric(Oil)
LINE = as.factor(Line)
LOC = as.factor(Loc)
YEAR = as.factor(Year)
REP = as.factor(Rep)
Family = as.factor(Family)


## Calculate variance components
```

```
# requires lme4 package
library(lme4)
# Linear Model with random effects for variance components
y = lmer(Protein ~ (1|LINE) + (1|YEAR) + (1|LOC) + (1|LINE:YEAR))
# Extract variance components
Summary (y)

# calculate coefficient of variation (CV)
CV=sqrt(residual)/(grand mean)*100


#R code for GWAS using the SoyNAM and NAM R packages
install.packages("NAM", repos=c("http://rstudio.org/_packages",
"http://cran.rstudio.com"))
library(SoyNAM)
data(soynam)
head(ENV(trait = "protein/oil"))
P=BLUP(trait = "oil",family="all",env = c(1,2,3,4,5,6,7,13,14),
MAF=0.1,use.check=F,impute="FM",rm.rep=TRUE)

# all the required files for GWAS was extracted from SoyNAM package and then the
NAM package was used to conduct GWAS.

#Load NAM package
library(NAM)
# Set folder
setwd("C:/Users/Wali Salari/Desktop/GWAS/GWAS-NEWGENODATA")
# loading phenotypes
Protein = as.vector (data.matrix (read.delim ("NAM-Protein/NEWPhenoProtein.csv",
header=F)))
Oil = as.vector (data.matrix (read.delim ("NAM-Oil/NEWPhenoOil.csv", header=F)))
Pheno = cbind(Protein,Oil)

# Loading chromosome, family and genotype
chr = as.vector(data.matrix(read.delim("NAM-Protein/chrNEWGENO.csv",header=F)))
fam = as.vector(data.matrix(read.delim("NAM-Protein/NEWDATAfam.csv",header=F)))
gen = read.delim("NAM-Protein/NewGENOTYPIC.csv",header=T,sep=",");
gen=data.matrix(gen)
```

```
# Remove replicated observations
cleaned = cleanREP(y = Pheno,fam = fam,gen = gen)

# Quality control
gen=snpQC(gen=cleaned$gen,psy=1,MAF=0.2,remove=TRUE,impute=FALSE)
Prot = cleaned$y[,1]
Oil = cleaned$y[,2]
fam = cleaned$fam

# GWAS
testP=gwas2(Prot,gen,fam,chr)
testO=gwas2(Oil,gen,fam,chr)

#plot GWAS result
plot(x=testP,colA=2,colB=3,pch=20,alpha=0.05/4118,main="Oil/Protein",cex.main=1.8,
cex.lab=1.3,cex.axis=1.3, cex = 1,lwd=6)
#Find the lrt threshold
optim(1,fn=function(x)abs(-dchisq(x,df=0.5,log=T)+log(0.05/4119)),method="CG")$par
#Identify marker significant at lrt 15.59
w = which( testP$PolyTest$lrt > 15.59 )
colnames(gen)[w]
as.data.frame(colnames(gen)[w])
#phenotypic variation explained by the significant markers
j = lm(Protein~gen[,w])
aov(j)
plot(j)
Return
summary(j)$r.squared

# Saving GWAS result
write.csv(TestP,'output.csv')

# Statistical Analysis System (SAS) code
ods rtf file="P&O.csv.rtf" style= minimal bodytitle;
title 'Summary statistics FAM46 removed P&O';

data P&O;
  infile 'P&O.csv' dsd firstobs=2 missover;
```

```
 length Env $ 10 Strain $20 ; max=60000;
 input  Protein Oil Proteingrkg Oilgrkg;

 if Env='' then delete;
 run;
proc sort data= AllPomostupdated;
 by Env;
proc means data= AllPomostupdated noprint;
 class Env Family;
 var Protein Oil Proteingrkg Oilgrkg;
 output out=Means(rename=(_type_=Type)) N= Mean= StdDev= Min= Max= /autoname;
 run;


proc print data=Means noobs uniform split='_';
 format HtIn_Mean--Oil_Max 7.2;
 where Type in (2,3);
 var Env Family Type Proteingrkg_N Proteingrkg_Mean Proteingrkg_StdDev
Proteingrkg_Min Proteingrkg_Max
Oilgrkg_N Oilgrkg_Mean Oilgrkg_StdDev Oilgrkg_Min Oilgrkg_Max;
 run;

proc print data=Means noobs uniform split='_';
 where Type in (2,3);
 format HtIn_Mean--Oil_Max 7.2;
 var Env Family Type   Protein_N Protein_Mean Protein_StdDev Protein_Min
Protein_Max
 Oil_N Oil_Mean Oil_StdDev Oil_Min Oil_Max;
 run;

ods rtf close;

# R code for chapter 4
Carbo<-read.table("Soycabohydrates.csv",header=TRUE,sep=",",)
summary(Carbo)
library(multcomp)

#ANOVA and distribution
aov.out1 <- aov(Carbo$Sucrose ~ Carbo$Strain + Carbo$Year, data=Carbo)
options(show.signif.stars=T)
```

```
summary(aov.out1)

boxplot(Carbo$Sucrose~Carbo$Year,main="Distribution of %Sucrose by
Year",cex=0.5,cex.main=2,cex.lab=1.7,cex.axis=1.4,ylab="%Sucrose",ylim=c(4,9),las=1,
xlab="Year",col="gold",notch=TRUE, outline = T)

hist(Carbo$Sucrose,main="Sucroes",cex=1,cex.main=1.7,cex.lab=1.6,cex.axis=1,las=1,x
lab="Sucrose%",col="cyan",notch=TRUE,use="pairwise.complete.obs",breaks=25)
mx <- mean(6.873)
abline(v = mx, col = "red", lwd = 2)

aov.out2 = aov(Carbo$Stachyose ~ Carbo$Strain * Carbo$Year, data=Carbo)
options(show.signif.stars=T)
summary(aov.out2)

boxplot(Carbo$Stachyose~Carbo$Year,main="Distribution of %Stachyose by
Year",cex=0.5,cex.main=2,cex.lab=1.7,cex.axis=1.4,ylab="%Stachyose",ylim=c(1.5,5.5),
las=1, xlab="Year",col="gold",notch=TRUE, outline = T)

hist(Carbo$Stachyose,main="Stachyose",cex=1,cex.main=1.7,cex.lab=1.6,cex.axis=1,las
=1,xlab="Stachyose%",col="yellow",notch=TRUE,use="pairwise.complete.obs",breaks=
25)
mx <- mean(4.05)
abline(v = mx, col = "black", lwd = 2)

aov.out3 = aov(Carbo$Raffinose ~ Carbo$Strain * Carbo$Year, data=Carbo)
options(show.signif.stars=T)
summary(aov.out3)

boxplot(Carbo$Raffinose~Carbo$Year,main="Distribution of %Raffinose by
Year",cex=0.5,cex.main=2,cex.lab=1.7,cex.axis=1.4,ylab="%Raffinose",ylim=c(0.5,1.6),
las=1, xlab="Year",col="gold",notch=TRUE, outline = T)
hist(Carbo$Raffinose,main="Raffinose",cex=1,cex.main=1.7,cex.lab=1.6,cex.axis=1,las=
1,xlab="Raffinose%",col="gold",notch=TRUE,use="pairwise.complete.obs",breaks=25)
mx <- mean(0.9723)
abline(v = mx, col = "black", lwd = 2)

# R code for chapter 5
```

```
POParent = read.csv("ProandOil.csv", header=T)
## Check to ensure data imported correctly
str(POParent)
head(POParent)
tail(POParent)
## Attach dataset
attach(POParent)

boxplot(Protein~Loc, xlab="%Protein", ", main=" Protein ", col="gray")
# Rename variables for ease of use
Protein=as.numeric(Protein)
Oil = as.numeric(Oil)
Strain= as.factor(Strain)
LOC = as.factor(Location)
YEAR = as.factor(Year)
REP = as.factor(Rep)
# Distribution
hist(Protein, col="gray")
hist(Oil, col="gray")


##gge bioplots
bioplot<-read.table("GGEProtParent1.txt", header=T)
bioplot<-read.table("GGEOilParent1.txt", header=T)
head(bioplot)
#GGE
library(GGEBiplotGUI)
GGEBiplot(bioplot)

# SAS code
title 'MIXED analysis of variance for Soybean Protein and Oil';
options ps=73 ls=120 nocenter nonumber;
data one;
  infile 'POParent.GGE-biplot.csv' dsd firstobs=2 missover;
  Length Loc $ 12 Env $ 10 Strain $ 16 Family $ 8;
  input Loc Year Env Strain Family Block Environment Location Protein Oil;
  if Loc =" then delete;
  run;
```

```
proc means data=one noprint;
  class Loc Year Strain ;* Block ;
  var Protein Oil;
  format Protein Oil 6.2;
  output out=Means(rename=(_type_=Type)) N= Mean= Std= Min= Max= / autoname;
  run;
title2 'SImple statistics for Main effects';
proc print data=Means noobs split='_';
  where Type in (0,1,2,4);
  var Loc Year Strain Type Protein_N Protein_Mean  Protein_StdDev  Protein_Min
Protein_Max
                    Oil_N Oil_Mean  Oil_StdDev  Oil_Min  Oil_Max;
  run;
title1 'GLM analysis of variance for Soybean Oil and Protein';
Year Loc*Year Block Year*Strain Loc*Year*Strain / test;
  quit;


title1 'GLM analysis of variance for Soybean Oil and Protein';
proc glm data=one;
  class Loc Year Strain Block;
  model Protein Oil = Loc Year Block Strain Loc*Strain Year*Strain Loc*Year*Strain /
ss3;* / ddfm=Satterth;
  random Year Block Year*Strain Loc*Year*Strain / test;
  quit;
```

VITA

VITA

Mohammad Wali Salari was born to a typical Afghan family in Kabul, Afghanistan. In 2005, he graduated from the Agronomy Department of the College of Agriculture at Kabul University. Following graduation with Bachelor of Science degree in Agriculture, he became an assistant lecturer in the agronomy department. In 2006, he was awarded an Afghan Merit Scholar Fellowship to work toward a Master of Science degree in agronomy at Purdue University. He completed six months of English language course at the American University of Afghanistan followed by five months non-degree program in English Language and Cultural Exchange Training at Indiana University-Purdue University Indianapolis. He successfully completed his MS degree at Purdue University in 2010 in Plant Genetics and Breeding. His MS thesis research focused on wheat stem rust disease which is one of the most devastating diseases of wheat in Afghanistan. After returning to Kabul, he began teaching in the Agronomy department at Kabul University and worked for the Ministry of Agriculture, Irrigation and Livestock (MAIL) first as an Organization Development Manger and then as Project Director for the period of February 2011 to August 2012. He completed his Ph.D. degree in August 2016. His Ph.D. dissertation research focused on identifying QTL controlling seed composition traits in soybean.