GENETIC IMPROVEMENT OF SOYBEAN SEED COMPOSITION

by

ELIZABETH M. PRENGER

(Under the Direction of Zenglu Li)

ABSTRACT

Seed composition is crucial for the efficacy of soybeans [*Glycine max* (L.) Merr.] used in food, feed, and fuel. Soybean breeders strive to improve protein, oil, fatty acid, amino acid, and carbohydrate contents in soybean seeds while increasing yield. Negative relationships between protein content and both yield and oil content present challenges for breeders. Environment also affects the accumulation of seed components. Various approaches were taken to understand seed composition traits in relation to yield and environment and to determine the ability to manipulate these traits genetically to produce desirable soybean germplasm. Fast neutron mutants in elite backgrounds and a near-isogenic elite line with a high-protein introgression were utilized to further the understanding of seed composition and yield in southern soybean lines and to develop resources for improvement of seed composition in soybean.

INDEX WORDS: soybean, seed composition, protein, oil, sucrose, genetic, mutant, fast neutron, near-isogenic line, planting date, comparative genomic hybridization, whole genome sequencing, bulked segregant analysis

# GENETIC IMPROVEMENT OF SOYBEAN SEED COMPOSITION

by

ELIZABETH MARGARET PRENGER

BS, University of Missouri-Columbia, 2016

A Thesis Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

MASTER OF SCIENCE

ATHENS, GEORGIA

2018

GENETIC IMPROVEMENT OF SOYBEAN SEED COMPOSITION

by

ELIZABETH MARGARET PRENGER

| | |
|---|---|
| Major Professor: | Zenglu Li |
| Committee: | Peggy Ozias-Akins |
| | Wayne Parrott |

Electronic Version Approved:

Suzanne Barbour
Dean of the Graduate School
The University of Georgia
December 2018

# DEDICATION

I dedicate this thesis to my fiancé, Nick, my parents, Glenn and Jackie, my brothers, Jacob and Isaac, my grandparents, and entire extended family, who have supported me throughout this journey. Without your love, prayers, and guidance, I would not be where or who I am today. Thank you.

ACKNOWLEDGEMENTS

I have been fortunate to work with many remarkable people during the last few years. My advisor, Dr. Zenglu Li, has been a great guide and encourager throughout my graduate program, and my committee members, Dr. Wayne Parrott and Dr. Peggy Ozias-Akins, have given valuable time and input during this process. I am grateful for the direction and support of Drs. Melissa Mitchum and Andrew Scaboo and the many helpful people I worked with during my undergraduate degree. Thanks to all the members of the UGA Soybean Breeding and Genetics lab, past and present, for your help over the past several years: Dale Wood, Earl Baxter, Brice Wilson, Tatyana Nienow, Ricky Zoller, Jeremy Nation, Greg Gokalp, Clint Steketee, Ben Stewart-Brown, Silas Childs, Ivy Tran, Evan McCoy, Mary Campbell, Nicole Bachleda, Ethan Menke, Cecilia Giordano, Brooks Arnold, Sam McDonald, Mark Miller, Alexandra Ostezan, Dr. Rebecca Tashiro, Dr. Justin Vaughn, Dr. Jeff Boehm, and Dr. Miles Ingwers. I couldn't possibly name everyone who has helped me along the way, but I am truly grateful to all of you.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

# CHAPTER 1

# INTRODUCTION AND LITERATURE REVIEW

## Soybean and its uses

Soybean [*Glycine max* (L.) Merr.] is a dicot species in the Fabaceae native to Asia and a legume that fixes nitrogen in the atmosphere through symbiotic relationships with microbes in the soil. Soybean is a paleopolyploid crop that underwent two whole genome duplications about 59 and 13 million years ago, resulting in a highly duplicated genome. The genome consists of 20 chromosomes and is about 1.1 Gigabases in size (http://plants.ensembl.org; Schmutz et al., 2010).

Cultivated soybean is an annual plant with an erect, bushy growth habit that can be classified as indeterminate or determinate. Maturity Groups (MG) 000 to IV soybeans in the U.S. typically have indeterminate growth habit, while MG V and later tend to be determinate. These maturity group classifications result from sensitivity of soybean development to photoperiod. Morphology of the soybean flowers results in a highly self-pollinating plant (Kumudini, 2010). The flowers develop into pods, typically with at least two pods per inflorescence and up to five seeds per pod, although most pods contain two to three seeds (Carlson and Lersten, 2004).

Soybean is grown extensively in the United States and throughout the world with large amounts of production occurring in the U.S., Brazil, and Argentina. Smaller quantities are grown in China, Paraguay, India, Canada, and other countries. Total worldwide production was around 340.9 million metric tons in 2017. Soybean accounted for a third of crop area planted in the U.S. in 2017 and was valued at 41.0 billion U.S. dollars (http://soystats.com). Soybean meal is one of

the most popular sources of protein for animal feed, particularly poultry, swine, and dairy cattle, and is also prevalent in products for human use. Soybean oil is widely used in margarine, cooking oil, biodiesel, and plastics, and is also used as an emulsifier and lubricant. Soybeans account for 59% of oilseed production worldwide. Various soy protein products can be made after oil is extracted, such as soy flour and grits (http://soystats.com). Tofu, natto, miso, soymilk, edamame, and sprouts are some soy products commonly consumed by humans.

Modern soybeans are bred for good agronomic qualities, such as shattering and lodging resistance, yield, and disease resistance in addition to seed composition, such as protein and oil content. Soybean seeds consist of approximately 410 g $kg^{-1}$ protein and 210 g $kg^{-1}$ oil based on dry weight (Hartwig and Kilen, 1991) as well as approximately 120 g $kg^{-1}$ soluble carbohydrates (Hagely et al., 2013). The protein to oil energy ratio in most soybeans is about 2.0 (Yaklich, 2001). The desired meal protein content in soybean seeds is 48% or higher (Chung et al., 2003). Soybeans accounted for 70% of world protein meal consumption, and 42.0 million metric tons of soybean meal were produced in the United States in 2017. Poultry used the highest percentage of soybean meal in the livestock industry at 55%, followed by swine at 25%, beef at 8%, dairy at 7%, and petfood and other feed at 4% (http://soystats.com).

In order for soybean seeds to be utilized for end use, the seeds must be processed. Many modern soybean processing facilities use a solvent extraction process. The first step in this process is cleaning, drying the seeds to loosen the hull, and cracking. Hulls are removed, and the remaining cracked seed is heated and passed through a roller mill to produce extremely thin flakes. Flakes can then be sent to an expander for ease of oil extraction or sent directly to an extractor. Oil is removed from the flakes using a solvent and washed with hexane. The extracted oil is sent to an oil refinery. Resulting oil-free flakes are sent to a toaster to evaporate remaining

hexane and inactivate digestibility-inhibiting enzymes. The meal is then sent to a drier-cooler to achieve 13-14% moisture and is later screened and ground to produce uniform meal. Some processors will add hulls back to the meal to lower protein levels in the meal (Shurtleff and Aoyagi, 2007; Smith and Johnson, 2003).

**Importance of seed composition**

Soybean seed meal is crucial for the livestock industry as a source of protein and energy in feed. Advantages of soybean meal in livestock feed include high levels of amino acids, a good source of energy, vitamins and minerals, digestibility of amino acids, high availability, and reliability in a consistent product with fairly predictable prices for use in production (http://www.soymeal.org). In recent years, the protein content of U.S. soybeans, a major component of soy meal, has been decreasing (United Soybean Board, 2015). Farmers must produce high quality soybeans without sacrificing yield to meet the needs of the current livestock industry. A study done by the United Soybean Board indicates that a 10 g $kg^{-1}$ increase in soybean protein content could increase the amount that a farmer earns per hectare by $19 to $32 (United Soybean Board, 2015, http://unitedsoybean.org). In addition, sucrose in soybean meal is a source of metabolizable energy for monogastrics and makes up the majority of the 120 g $kg^{-1}$ of soybean seeds consisting of soluble carbohydrates (Hagely et al., 2013). The raffinose (RFO) family of oligosaccharides also makes up a portion of the soluble carbohydrates. The α-galactosidase enzyme is required to cleave these oligosaccharides. However, this enzyme is not present in the intestines of pigs, and chickens, and humans (monogastrics), so the oligosaccharides are degraded in the lower intestine by gut bacteria producing carbon dioxide,

hydrogen, and methane, resulting in diarrhea and flatulence in high doses (Knudsen and Li, 1991).

Although soybean is a good source of protein and energy for food and feed, it is not ideal for several reasons. First, it is low in the sulfur-containing essential amino acids methionine and cysteine, resulting in additional costs of supplementing the meal. Second, it is high in the oligosaccharides raffinose and stachyose. Additionally, the meal protein levels of many current Midwestern soybean cultivars are lower than the national standard, resulting in lower value soybeans (Patil et al., 2017). Soybeans are the largest source of protein meal and second largest source of vegetable oil worldwide (http://soystats.com). Therefore, improvement of soybean meal components and oil are crucial for the utility of soybean used in food, feed, and fuel.

The United Soybean Board and soybean researchers have set seed composition breeding goals for protein, meal protein, oil, oleic acid, linolenic acid, and sucrose (Table 1.1) to quantify seed composition standards for soybean breeders (United Soybean Board, personal communication).

## Factors affecting seed protein and composition

### *Relationships between seed protein, oil, carbohydrates, and yield*

Protein content of soybean seeds is negatively correlated with both oil content and yield (Burton, 1987; Hartwig and Hinson, 1972; Chung et al., 2003; Thorne and Fehr, 1970; Cober and Voldeng, 2000). It has been suggested that the negative correlation between protein and oil may be due to pleiotropic effects or tightly linked genes (Brummer et al., 1997; Chung et al., 2003). Opinions and supporting arguments differ on which of the two is likely causing the negative association. Brummer et al. (1997) indicated that the results of their study did not support the

pleiotropic effects theory, while Chung et al. (2003) suggested that QTL with pleiotropic effects is more probable and testing to determine whether protein and oil QTL are in tight repulsion phase would be costly and time-consuming. Several studies have shown that seed protein and oil, as well as maturity and yield QTL, all were mapped to a small interval of Linkage Group I (LG I) on Chromosome 20 (Chung et al., 2003; Nichols et al., 2006). The negative relationship often observed between seed protein and yield indicates that protein may be costly for the plants to produce (Chung et al., 2003). However, several studies have indicated that it may be possible to increase protein content without significantly adversely affecting yield. Wilcox and Zhang (1997) found that the regression of seed protein on seed yield for determinate (typically later maturity group) soybeans in their research was not significant. Progeny were identified with both high yield and elevated protein content from two crosses between high-protein indeterminate lines and average-protein determinate lines. An earlier study found that protein can be increased without significant effects on yield. In each of several backcross generations, lines with elevated protein content and yield similar to the low protein, high-yielding parent were identified (Wilcox and Cavins, 1995). A study done by Yin and Vyn (2005) indicated no significant relationship between protein concentration and yield in MG 0 and MG I lines planted in three years, although the relationship was negative. Interestingly, they also found that for every megagram per hectare increase in yield, seed oil concentration decreased 4.2 g $kg^{-1}$. Cober and Voldeng (2000) developed 886 outcross derived lines and 800 backcross derived lines using the parents 'AC Proteus' (high protein) and 'Maple Glen' (adapted, high-yielding). They selected six outcross lines and nine backcross lines and tested them at six locations. The authors did not observe any high protein lines with higher yield than the low protein parent in either backcross or outcross lines, but they observed very low association between seed protein and yield. The negative

relationship between protein and oil content is generally strong (Hurburgh et al., 1990). However, based on the results of two- and three-way crosses between high protein, exotic germplasm and adapted lines, Thorne and Fehr (1970) suggested that selection for high protein lines should be possible without drastically reducing oil content.

An additional important soybean breeding goal is to increase sucrose and decrease both raffinose and stachyose in soybean lines. A positive relationship between oil and sucrose content and a negative relationship between protein and sucrose content have been reported (Hymowitz et al., 1972; Jaureguy et al., 2011), further complicating simultaneous improvement of protein, oil, carbohydrates, and yield.

*Environmental effects on seed composition*

Seed composition components including protein, oil, and fatty acids change in developing seeds over time. In one study, protein, oil, and fatty acid components reached their peak percentages of total seed composition around 40 days after flowering (DAF), but 70% of total protein and oil were synthesized in the remaining 25 days of seed development (Rubel et al., 1972). Another study looked at the protein and oil accumulation of high and low protein lines at four different seed development stages in near-isogenic lines (NILs) for a chromosome 20 major protein content QTL identified in many bi-parental mapping and genome-wide association studies (GWAS) (https://soybase.org). Seed stages were designated by size, and differences in protein content between the high and low protein lines were apparent from the earliest stages. Differences in oil were stronger in stage four, the final stage studied (Bolon et al., 2010).

Environment, nutrient availability, and physiology all affect soybean seed composition. Various studies have investigated the role that temperature plays in seed composition. Seed

protein is positively correlated with temperature (Song et al., 2016; Kumar et al., 2006), while oil has been reported to be negatively correlated with temperature by some (Kumar et al., 2006) and positively correlated with temperature by others (Kane et al., 1997). Rainfall and water availability have been reported to have mixed relationships with protein (Rotundo and Westgate, 2009; Kumar et al., 2006), and protein content tends to be higher in southern soybeans, while oil may be higher in northern soybeans (Chung et al., 2003; Kumar et al., 2006).

Planting date effects on seed composition have mixed reports in literature. Some studies found that delayed planting date increased protein content in soybeans grown in North Dakota, the southeastern U.S., and Mississippi (Helms et al., 1990; Kane et al., 1997; Bellaloui et al., 2015). Others found that delayed planting decreased protein content in studies done in Arkansas, India, and Pakistan (Jaureguy et al., 2013; Billore et al., 2000; Muhammad et al., 2009). Most reports agree that oil tends to decrease with delayed planting date (Hu and Wiatrak, 2012). Reports of effects of delayed planting on sucrose accumulation are also mixed (Jaureguy et al., 2013; Bellaloui et al., 2015). With varied reports on the effects of planting date and environment on seed composition components, it may benefit growers to study how these conditions affect MG VII and VIII soybeans grown in Georgia.

*Genetic loci controlling for seed composition*

Genotype is a major factor contributing to soybean seed composition, and genotypes may interact with the environment to determine seed protein, oil, and carbohydrate contents (Herman, 2014). Numerous studies have been conducted to identify protein, oil, and carbohydrate content QTL in both biparental mapping populations and GWAS. Many of the results of these studies are recorded in the online database, Soybase (http://soybase.org). To date, there are 322 seed oil, five

seed oil plus protein, and 241 protein content QTL reported. Only sixteen of the oil QTL and sixteen of the protein QTL are designated "cq," or "confirmed QTL." Phansak et al. (2016) analyzed F-statistics and P-values of all reported protein and oil QTL to determine "highly likely" QTL. Protein QTL on chromosomes 1, 4, 6, 7, 11, 13, 15, and 20 as well as oil QTL on chromosomes 2, 6, 8, 14, 15, 19, and 20 were considered highly likely. In particular, two major protein QTL on chromosomes 15 and 20 have been detected repeatedly in many mapping studies (Diers et al., 1992; Sebolt et al., 2000; Chung et al., 2003; Nichols et al., 2006; Qi et al., 2011; Hwang et al., 2014; Vaughn et al., 2014; Bandillo et al., 2015; Warrington et al., 2015; Phansak et al., 2016).

The chromosome 15 protein content QTL was first reported by Diers et al. (1992), and the high protein allele from PI 468916 was associated with a 17 g $kg^{-1}$ increase in protein content and a decrease in oil content. A later study identified the chromosome 15 protein QTL in a Williams 82 × PI 407788A mapping population. The QTL was fine-mapped to a region between 3.59 and 4.12 Mb of chromosome 15, and the locus explained approximately 25% of phenotypic variation (Kim et al., 2016).

The chromosome 20 protein content QTL explains 7 to 65% of phenotypic variation in protein content (Phansak et al., 2016; Sebolt et al., 2000). Recent studies suggest that the chromosome 20 protein QTL is likely located somewhere in the 25 to 33 Mb region of chromosome 20 (Bolon et al., 2010; Hwang et al., 2014; Vaughn et al., 2014; Bandillo et al., 2015). Several candidate genes have been identified in this QTL region, including Glyma20g19620, Glyma20g19630, Glyma20g19680, Glyma20g21030, Glyma20g21040, and Glyma20g21080 (Bolon et al., 2010; Hwang et al., 2014). Another GWAS study narrowed the QTL to a region containing just three genes: Glyma20g21030, Glyma20g21040, and

Glyma20g21080. However, genes identified as the most plausible candidate genes in all QTL intervals included Glyma20g21030, Glyma20g21361, and Glyma20g21780 with a putative function relating to ammonium transport, a conserved oligomeric Golgi complex, and an ethylene receptor, respectively (Bandillo et al., 2015). Single nucleotide polymorphisms (SNPs) significantly associated with protein content detected by the Hwang et al. (2014) and Vaughn et al. (2014) GWA studies were not detected in all maturity groups of the Bandillo et al. (2015) study, but nearby significant SNPs were found in different subsets of maturity groups within the GWAS population. The Gm20_31610452 SNP was identified as the most significantly associated chromosome 20 SNP in a wide range of accessions (Vaughn et al., 2014).

To date, there are 37 seed sucrose and 15 seed oligosaccharide (sucrose, raffinose, and stachyose) content QTL listed in Soybase (https://soybase.org). Kim et al. (2006) identified putative oligosaccharide QTL on chromosomes 6, 12, 16, and 19. The QTL on chromosomes 12 and 16 were also associated with sucrose. Correlation coefficients from the same study for protein, oil, sucrose, and oligosaccharides implied the possibility of improving oligosaccharide content without greatly affecting protein or oil. Dierking and Bilyeu (2008) reported altered levels of the raffinose family of oligosaccharides resulting from a mutation in a raffinose synthase gene (*RS2*, Glyma06g18890 (W82.a1.v1)/Glyma.06g179200 (W82.a2.v1)). Hagely et al. (2013) studied carbohydrate profiles of various *RS2* mutants and functional *RS2* soybeans and observed weak RFO, low RFO, and ultra-low RFO lines with increased sucrose and decreased RFOs as a proportion of overall carbohydrates.

**Performance of major protein QTL in various genetic backgrounds**

In several studies involving the chromosome 15 and 20 protein QTL, the allele associated with an increase in protein also decreased both seed oil content and yield (Diers et al., 1992; Sebolt et al., 2000; Chung et al., 2003; Nichols et al., 2006). This result is consistent with the often-reported negative relationships between protein and oil contents and between protein content and yield (Burton, 1987; Hartwig and Hinson, 1972; Chung et al., 2003; Thorne and Fehr, 1970; Cober and Voldeng, 2000). The major QTL on chromosomes 15 and 20 have been introgressed into various backgrounds to study the effects of the high-protein alleles on protein, oil, and yield. In a recent experiment by Brzostowski and Diers (2017), the chromosome 15 major protein QTL from PI 407788A was introgressed into two MG II Midwestern lines. In each genetic background, the PI 407788A allele on chromosome 15 was associated with a significant increase in protein (11 g $kg^{-1}$) and decrease in oil (6 g $kg^{-1}$) on a 130 g $kg^{-1}$ moisture basis. The PI allele was also associated with a decrease in yield (57 to 109 kg $ha^{-1}$), though the decrease was not significant.

The chromosome 20 high-protein alleles from Danbaekkong and *Glycine soja* PI 468916 were crossed into various MG II and IV backgrounds (Brzostowski et al., 2017; Mian et al., 2017), and the effects on protein, oil, and yield were measured. A MG III soybean cultivar released in 2017 carries the high protein allele from Danbaekkong on chromosome 20. This cultivar has 57 g $kg^{-1}$ higher protein than the mean of four high-yielding check cultivars on a 130 g $kg^{-1}$ moisture basis. Despite the increase in protein, this cultivar had similar yields compared to the checks across 20 environments. However, this cultivar is not a NIL and was not directly comparable to the parents in yield trials (Mian et al., 2017). In another experiment, Danbaekkong was used as a source of the chromosome 20 high-protein allele, which was introgressed into two

MG II backgrounds, and the PI 468916 allele was introgressed into two MG II and two MG IV backgrounds. The Danbaekkong allele was consistently associated with a significant increase in protein (19 to 20 g $kg^{-1}$), decrease in oil (7 to 9 g $kg^{-1}$), and decrease in yield (363 to 455 kg $ha^{-1}$) in these MG-II backgrounds on a 130 g $kg^{-1}$ moisture basis. The PI 468916 allele was associated with increased protein (18 to 23 g $kg^{-1}$) and decreased oil (9 to 13 g $kg^{-1}$), but the decrease in yield (134 to 319 kg $ha^{-1}$) was not consistently significant across locations (Brzostowski et al., 2017). The Danbaekkong allele effects on yield in MG II and III backgrounds implicates a role of southern germplasm in reducing yield drag. Mian et al. (2017) used a southern (University of Georgia) breeding line containing the Danbaekkong allele in their crosses and did not see yield drag despite increased protein content. In contrast, Brzostowski et al. (2017) used only early-maturity germplasm as recurrent parents and either introgressed alleles directly from Danbaekkong or from an early-maturity $BC_3F_4$ plant carrying the PI 468916 allele at the chromosome 20 protein locus.

**Soybean genetics and genomics**

Soybean in North America is low in genetic diversity, with 86% of the parentage of modern U.S. cultivars contributed by fewer than 20 ancestral lines (Patil et al., 2017). Soybean is a diploid with 2n=40 chromosomes and a highly self-pollinating crop. Due in part to historical genetic bottlenecks, the soybean genome has a low frequency of single nucleotide polymorphisms (SNPs) with about 0.5 SNPs per 1 kilobase pair (kbp) of coding sequence (Song and Cregan, 2017). Additionally, soybean is considered a palaeopolyploid, with evidence of whole-genome duplication events that occurred about 59 and 13 million years ago. Many of the

genes duplicated in these events have been retained, resulting in multiple copies or paralogs of many genes (Schmutz et al., 2010).

Whole-genome sequencing and assembly of the MG III soybean cultivar Williams 82 was completed in 2010 using a whole genome shotgun sequencing approach. The completed reference genome (Glyma1.01 or Williams82.a1.v1) covered 85% of the estimated genome size with 955 Mb of 1.1 Gb covered (Schmutz et al., 2010). The Williams 82 version 1 reference genome contains more than 46,000 high-confidence protein-coding loci, 12.2% of which are putative transcription factors. An additional 20,000 low-confidence protein-coding loci are predicted (Schmutz et al., 2010). The Williams 82 genome assembly was revised using high-density linkage maps to develop version 2 (Wm82.a2.v1) (DOE Joint Genome Institute, https://phytozome.jgi.doe.gov/). The version 2 assembly covers 978.5 Mb (949.2 Mb excluding unplaced scaffolds) of the estimated 1.1 Gb genome. Over 56,000 protein-coding loci were predicted in Williams 82 version 2 (Valliyodan et al., 2017).

Due to the low genetic diversity of soybean, there has been an increase in efforts to broaden the genetic diversity of modern soybean cultivars. The soybean wild ancestor, *Glycine soja* [Sieb. & Zucc.] is often used as a source of genetic diversity and to introduce traits such as increased protein content or disease resistance in breeding programs (Singh, 2017; Patil et al., 2017). Induced mutations resulting from mutagens such as fast neutrons have also contributed to development of new phenotypic variation for traits of interest in soybeans (Bolon et al., 2011, 2014; Hwang et al., 2015; Stacey et al., 2016; Campbell et al., 2016; Dobbels et al., 2017).

**Effects of fast neutron bombardment on plant genomes**

Fast neutron (FN) ionizing radiation induces heritable genomic deletions, duplications, insertions, translocations, and single base pair changes in plants including soybean and Arabidopsis. Various levels of radiation have been used in former studies, with radiation levels ranging from 4 to 32 Gray Units (Gy) in soybean and 60 Gy in Arabidopsis (Bolon et al., 2011, 2014; Belfield et al., 2012). Mutations in irradiated Arabidopsis were mostly single base-pair substitutions, but insertions and deletions ranging from one base to greater than 1 kb were also detected (Belfield et al., 2012). Copy number variation (CNV) detected in 30 soybean FN mutants exposed to 4, 8, 16, or 32 Gy radiation consisted of 85.3% putative deletions and 14.8% putative duplications. Size of these mutations ranged from 986 bp to 3 Mb in size (Bolon et al., 2011). At these radiation levels, first generation mutant plant emergence peaked at 76% for the 4 Gy radiation dose and decreased to 61% for the 32 Gy dose. Only 49% of $M_1$ plants produced $M_2$ seed at 4 Gy radiation, and only 15% produced $M_2$ seeds at 32 Gy radiation (Bolon et al., 2011). Additional mutants were screened by Bolon et al. (2014), bringing the total number of mutants characterized to 264. An average of one duplication per mutant was discovered. Duplications ranged from 499 bp to over 50 Mb. An average of 2.5 homozygous deletions were detected per mutant, ranging in size from 493 bp to 8.1 Mb. An average of 1.1 hemizygous deletions was detected per mutant, and these deletions ranged in size from 4.9 kb to 9.3 Mb. Many of the hemizygous mutations were lost in later generations due to segregation of hemizygous loci, leading to retention of the wild type allele and inbreeding to homozygosity (Bolon et al., 2014).

**Comparative genomic hybridization and next-generation sequencing**

Array comparative genomic hybridization (CGH) is an effective way to identify CNV (Carter, 2007). A NimbleGen soybean microarray utilizing 696,139 unique probes based on the soybean reference genome, Williams 82 (Schmutz et al., 2010), was designed for CGH to detect differences in hybridization of each probe between mutants and a parent reference (Bolon et al., 2011). A normalized $\log_2$ ratio of mutant to control hybridization was calculated for each probe and used to detect CNV. SNP genotyping was used to exclude regions of intracultivar variation resulting in CGH peaks not induced by mutagenesis, and exome capture and resequencing was used to validate CGH-detected deletions (Bolon et al., 2011). The NimbleGen soybean microarray was updated to include nearly 1.4 million probes for greater characterization capacity in subsequent studies, and analyses were performed essentially as before (Bolon et al., 2014). The most recent CGH microarray from Agilent Technologies, Inc. contains 1 million unique probes enriched for genic regions. The new array requires less DNA per sample and less hybridization time. In several studies of FN mutants, whole genome sequencing (WGS) has been used to perform bulked segregant analysis using outcrossed populations or to further characterize mutations in the genomes of interest (Bolon et al., 2014; Campbell et al., 2016; Dobbels et al., 2017).

Whole genome sequencing was used by Hwang et al. (2015) to detect FN-induced genomic mutations and identify the causal mutation for a dwarf phenotype. An $M_5$ mutant in the Williams 82 background with a dwarf phenotype was characterized via Illumina HiSeq (Illumina, Inc.). A deletion was identified when sequencing reads were mapped to the reference genome (Williams 82 version 1) (Schmutz et al., 2010). Illumina HiSeq produced 100 bp paired-end sequence reads from 350 bp paired-end insert sizes with approximately 20x mapping depth

and 87% coverage of the reference genome. SNPs and indels were predicted using SAMTools software (Li et al., 2009), and large deletions were predicted using orientation and span size of reads mapped to the genome.

**Confirmation of fast neutron mutations and functional genomics**

PCR primers have been used in many cases to confirm the presence of candidate mutations in fast neutron mutants. A known *FAD2-1A* deletion mutant was used as a proof of concept by Bolon et al. (2011). Amplification of the target region and sequencing revealed the location of the deletion. Hwang et al. (2015) also used PCR to confirm several deletions in Williams 82 fast neutron mutated populations. Small deletions were validated by observing size differences in the amplicons between wild type and mutant, and larger deletions were validated by absence of a product in the wild type and presence of a product in the mutant. Additional methods have also been used to confirm the presence of multiple deletions in mutant lines. For example, exome capture and resequencing, RNA-seq, and RT-PCR have been used to confirm absence of genes of interest (Bolon et al., 2011; Campbell et al., 2016; Hwang et al., 2015). Outcross and backcross populations have been developed to map putative causal mutations using methods such as whole genome sequencing and subsequent QTL-seq (Bolon et al., 2014; Campbell et al., 2016; Dobbels et al., 2017). Transgenics have been used in several studies to validate the role of mutated genes in mutant phenotypes (Stacey et al., 2016; Campbell et al., 2016).

**Summary and objectives**

Both mutants and PI's may be used as sources of desirable traits in breeding programs for improvement of a variety of traits including seed protein, oil, and carbohydrate content. Decreasing costs of array and sequencing technologies enable the discovery of causal genomic regions of traits of interest, further expediting the process of developing competitively yielding soybean varieties with improved seed composition. The objectives of this research were to develop and characterize fast neutron mutants in elite MG VII backgrounds for improvement of soybean seed composition and to identify causal mutations for traits of interest and to describe the effects of a high protein allele in a near-isogenic line on yield, seed composition, and agronomic traits.

**References**

Bandillo, N., D. Jarquin, Q. Song, R. Nelson, P. Cregan, J. Specht, et al. 2015. A population structure and genome-wide association analysis on the USDA soybean germplasm collection. Plant Genome 8. doi:10.3835/plantgenome2015.04.0024.

Belfield, E., X. Gan, A. Mithani, C. Brown, C. Jiang, K. Franklin, et al. 2012. Genome-wide analysis of mutations in mutant lineages selected following fast-neutron irradiation mutagenesis of arabidopsis thaliana. Genome Research 22: 1306-1315.

Bellaloui, N., H.A. Bruns, H.K. Abbas, A. Mengistu, D.K. Fisher and K.N. Reddy. 2015. Agricultural practices altered soybean seed protein, oil, fatty acids, sugars, and minerals in the midsouth USA. Frontiers in Plant Science 6: 31. doi:10.3389/fpls.2015.00031.

Billore, S.D., O.P. Joshi and A. Ramesh. 2000. Performance of soybean (*Glycine max*) genotypes on different sowing dates and row spacings in vertisols. Indian J. Agric. Sci. 70: 577-580.

Bolon, Y.-T., A.O. Stec, J.-M. Michno, J. Roessler, P.B. Bhaskar, L. Ries, et al. 2014. Genome resilience and prevalence of segmental duplications following fast neutron irradiation of soybean. Genetics 198: 967-981. doi:10.1534/genetics.114.170340.

Bolon, Y.-T., B. Joseph, S.B. Cannon, M.A. Graham, B.W. Diers, A.D. Farmer, et al. 2010. Complementary genetic and genomic approaches help characterize the linkage group I seed protein QTL in soybean. BMC Plant Biology 10: 41-64. doi:10.1186/1471-2229-10-41.

Bolon, Y.-T., W.J. Haun, W.W. Xu, D. Grant, M.G. Stacey, R.T. Nelson, et al. 2011. Phenotypic and genomic analyses of a fast neutron mutant population resource in soybean. Plant Physiol. 156: 240-253. doi: doi: http://dx.doi.org/10.1104/pp.110.170811.

Brummer, E.C., G.L. Graef, J.H. Orf, J.R. Wilcox, and R.C. Shoemaker. 1997. Mapping QTL for seed protein and oil content in eight soybean populations. Crop Sci. 37(2): 370-378.

Brzostowski, L.F. and B.W. Diers. 2017. Agronomic evaluation of a high protein allele from PI407788A on chromosome 15 across two soybean backgrounds. Crop Sci. 57: 2972-2978. Doi: 10.2135/cropsci2017.02.0083.

Brzostowski, L.F., T. Pruski, J.E. Specht, and B.W. Diers. 2017. Impact of seed protein alleles from three soybean sources on seed composition and agronomic traits. Theor. Appl. Genet. 130(11): 2315-2326.

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

Campbell, B.W., A.N. Hofstad, S. Sreekanta, F. Fu, T.J.Y. Kono, J.A. O'Rourke, et al. 2016. Fast neutron-induced structural rearrangements at a soybean *NAP1* locus result in gnarled trichomes. Theor. Appl. Genet. 129: 1725-1738. doi:10.1007/s00122-016-2735-x.

Carlson, J.B. and N.R. Lersten. 2004. Reproductive morphology. *In:* H.R. Boerma and J.E. Specht, eds. Soybeans: Improvement, production and Uses. 3rd ed. ASA and CSSA, Madison, WI. Pages 59-93.

Carter, N.P. 2007. Methods and strategies for analyzing copy number variation using DNA microarrays. Nat. Genet. 39: S16-S21.

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

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

Dierking, E.C. and K.D. Bilyeu. 2008. Association of a soybean raffinose synthase gene with low raffinose and stachyose seed phenotype. Plant Genome 1: 135-145. doi:10.3835/plantgenome2008.06.0321.

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

Dobbels, A.A., J.-M. Michno, B.W. Campbell, K.S. Virdi, A.O. Stec, G.J. Muehlbauer, et al. 2017. An induced chromosomal translocation in soybean disrupts a KASI ortholog and is associated with a high-sucrose and low-oil seed phenotype. G3: Genes|Genomes|Genetics 7: 1215-1223. doi:10.1534/g3.116.038596.

George, A.A. and B.O. De Lumen. 1991. A novel methionine-rich protein in soybean seed: Identification, amino acid composition, and N-terminal sequence. J. Agric. Food Chem. 39: 224-227. doi:10.1021/jf00001a046.

Grant, D., R.T. Nelson, S.B. Cannon, and R.C. Shoemaker. 2010. SoyBase, the USDA-ARS soybean genetics and genomics database. Nucl. Acids Res. (2010) 38 (suppl 1): D843-D846. doi: 10.1093/nar/gkp798.

Hagely, K.B., D. Palmquist and K.D. Bilyeu. 2013. Classification of distinct seed carbohydrate profiles in soybean. J. Agric. Food Chem. 61: 1105-1111. doi:10.1021/jf303985q.

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

Hartwig, E.E. and T.C. Kilen. 1991. Yield and composition of soybean seed from parents with different protein, similar yield. Crop Sci. 31: 290-292.

Helms, T.C., C.R. Hurburgh, R.L. Lussenden and D.A. Whited. 1990. Economic-analysis of increased protein and decreased yield due to delayed planting of soybean. J. Prod. Agric. 3: 367-371.

Herman, E.M. 2014. Soybean seed proteome rebalancing. Frontiers in Plant Science 5(437):1-8.

Hu, M. and P. Wiatrak. 2012. Effect of planting date on soybean growth, yield, and grain quality: Review. Agron. J. 104: 785-790.

Hurburgh, Jr., C.R., T.J. Brumm, J.M. Guinn, and R.A. Hartwig. 1990. Protein and oil patterns in U.S. and world soybean markets. J. Am. Oil Chem. Soc. 67:966-973.

Hwang, E.-Y., S. Qijian, G. Jia, J.E. Specht, D.L. Hyten, J. Costa, et al. 2014. A genome-wide association study of seed protein and oil content in soybean. BMC Genomics 15: 1-25. doi:10.1186/1471-2164-15-1.

Hwang, W., M. Kim, Y. Kang, S. Shim, M. Stacey, G. Stacey, et al. 2015. Genome-wide analysis of mutations in a dwarf soybean mutant induced by fast neutron bombardment. Euphytica 203: 399-408. doi:10.1007/s10681-014-1295-x.

Hymowitz, T., F.I. Collins, W.M. Walker and J. Panczner. 1972. Relationship between content of oil, protein, and sugar in soybean seed. Agron. J. 64: 613-615.

Jaureguy, L., F. Ledesma Rodriguez, L. Zhang, P. Chen and K. Brye. 2013. Planting date and delayed harvest effects on soybean seed composition. Crop Sci. 53: 2162-2175.

Jaureguy, L.M., P. Chen and A.M. Scaboo. 2011. Heritability and correlations among food-grade traits in soybean. Plant Breed. 130: 647-652. doi:10.1111/j.1439-0523.2011.01887.x.

Kane, M.V., C.C. Steele, L.J. Grabau, C.T. MacKown and D.F. Hildebrand. 1997. Early-maturing soybean cropping system .3. Protein and oil contents and oil composition. Agron. J. 89: 464-469.

Kim, H.K., S.T. Kang and K.W. Oh. 2006. Mapping of putative quantitative trait loci controlling the total oligosaccharide and sucrose content of *Glycine max* seeds. J. Plant Res. 119: 533-538. doi:10.1007/s10265-006-0004-9.

Kim, M., S. Schultz, R. Nelson, and B.W. Diers. 2016. Identification and fine mapping of a soybean seed protein QTL from PI 407788A on chromosome 15. Crop Sci. 56: 219-225. Doi:10.2135/cropsci2015.06.0340.

Knudsen, K.E.B. and B.W. Li. 1991. Determination of oligosaccharides in protein-rich feedstuffs by gas-liquid chromatography and high-performance liquid chromatography. J. Agric. Food Chem. 39: 689-694. doi:10.1021/jf00004a013.

Kumar, V., A. Rani, S. Solanki and S.M. Hussain. 2006. Influence of growing environment on the biochemical composition and physical characteristics of soybean seed. J. Food Composition and Analysis 19: 188-195. doi:http://dx.doi.org/10.1016/j.jfca.2005.06.005.

Kumudini, S. 2010. Soybean Growth and Development. p. 48–69. *In* Singh, G. (ed.), Soybean: Botany, Production, and Uses. CAB International.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, et al. 2009. The Sequence alignment/map (SAM) format and SAMtools. Bioinformatics, 25, 2078-2079.

Mian, R., L. McHale, Z. Li, and A.E. Dorrance. 2017. Registration of 'HighPro1' soybean with high protein and high yield developed from a north x south cross. J. Plant Reg. 11: 51-54.

Muhammad, A., S.K. Khalil, K. Marwat, A.Z. Khan and I.H. Khalil. 2009. Nutritional quality and production of soybean land races and improved varieties as affected by planting dates. Pakistan J. Botany 41: 683-689.

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

Patil, G., R. Mian, T. Vuong, V. Pantalone, Q. Song, P. Chen, et al. 2017. Molecular mapping and genomics of soybean seed protein: a review and perspective for the future. Theor. Appl. Genet. 130: 1975-1991. Doi: 10.1007/s00122-017-2955-8.

Phansak, P., W. Soonsuwon, D.L. Hyten, Q. Song, P.B. Cregan, G.L. Graef, et al. 2016. Multi-population selective genotyping to identify soybean [*Glycine max* (L.) Merr.] seed protein and oil QTLs. G3: Genes|Genomes|Genetics 6: 1635-1648. doi:10.1534/g3.116.027656.

Qi, Z., S. Ya-nan, W. Qiong, L. Chun-yan, H. Guo-hua and C. Qing-shan. 2011. A meta-analysis of seed protein concentration QTL in soybean. Can. J. Plant Sci. 91: 221-230. doi:10.4141/cjps09193.

Rotundo, J.L. and M.E. Westgate. 2009. Meta-analysis of environmental effects on soybean seed composition. Field Crops Res. 110: 147-156. doi: dx.doi.org/10.1016/j.fcr.2008.07.012.

Rubel, A., R.W. Rinne, and D.T. Canvin. 1972. Protein, oil, and fatty acid in developing soybean seeds. Crop Sci. 12: 739-741.

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

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

Shurtleff, W., and A. Aoyagi. History of Soybean Crushing: Soy Oil and Soybean Meal - Part 1. 2007. *In*: History of Soybeans and Soyfoods, 1100 B.C. to the 1980s. Soyinfo Center. Web. http://www.soyinfocenter.com/HSS/soybean_crushing1.php

Singh, R.J. 2017. Botany and cytogenetics of soybean. In: H.T. Nguyen and M.K. Bhattacharyya, eds. *The Soybean Genome*, Compendium of Plant Genomes. Springer International Publishing. Pages 11-40. DOI 10.1007/978-3-319-64198-0_2.

Smith, K., and L. Johnson, comps. *Fact Sheet: Soybean Processing*. 2003. Soybean Meal Information Center. Web. http://www.soymeal.org/FactSheets/processing3.pdf

Song, Q. and P.B. Cregan. 2017. Classical and molecular genetic mapping. In: H.T. Nguyen and M.K. Bhattacharyya, eds. *The Soybean Genome*, Compendium of Plant Genomes. Springer International Publishing. Pages 41-56. DOI 10.1007/978-3-319-64198-0_2.

Song, W., R. Yang, T. Wu, C. Wu, S. Sun, S. Zhang, et al. 2016. Analyzing the effects of climate factors on soybean protein, oil contents, and composition by extensive and high-density sampling in China. J. Agric. Food Chem. 64: 4121-4130. doi:10.1021/acs.jafc.6b00008.

Soybean Meal Info Center. Gordon Denny LLC, Cornerstone Resources LLC, and Integrative Nutrition, Inc.. *Advantages of U.S. Soybean Meal in Domestic Feed Rations* [Pamphlet]. UnitedSoybean.org. http://www.soymeal.org/factsheets.html

Stacey, M.G., R.E. Cahoon, H.T. Nguyen, Y. Cui, S. Sato, C.T. Nguyen, et al. 2016. Identification of homogentisate dioxygenase as a target for vitamin E biofortification in oilseeds. Plant Physiol. 172: 1506.

Thorne, J.C. and W.R. Fehr. 1970. Incorporation of high-protein, exotic germplasm into soybean populations by 2- and 3-way crosses. Crop Sci. 10: 652-655.

United Soybean Board. (2015, November 19). Animal Agriculture and Soybean Quality. Retrieved February 1, 2017, from http://unitedsoybean.org/media-center/issue-briefs/animal-agriculture/

Valliyodan, B., S.-H. Lee, and H. Nguyen. 2017. Sequencing, assembly, and annotation of the soybean genome. In: H.T. Nguyen and M.K. Bhattacharyya, eds. *The Soybean Genome*, Compendium of Plant Genomes. Springer International Publishing. Pages 73-82. DOI 10.1007/978-3-319-64198-0_2.

Vaughn, J.N., R.L. Nelson, Q. Song, P.B. Cregan and Z. Li. 2014. The genetic architecture of seed composition in soybean is refined by genome-wide association scans across multiple populations. G3: Genes|Genomes|Genetics 4: 2283-2294. doi:10.1534/g3.114.013433.

Warrington, C.V., H. Abdel-Haleem, D.L. Hyten, P.B. Cregan, J.H. Orf, A.S. Killam, et al. 2015. QTL for seed protein and amino acids in the Benning × Danbaekkong soybean population. Theor. Appl. Genet. 128: 839-850. doi:10.1007/s00122-015-2474-4.

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

Wilcox, J.R. and G.D. Zhang. 1997. Relationships between seed yield and seed protein in determinate and indeterminate soybean populations. Crop Sci. 37: 361-364.

Yaklich, R.W. 2001. Beta-conglycinin and glycinin in high-protein soybean seeds. J. Agric. Food Chem. 49: 729-735.

Yin, X.H. and T.J. Vyn. 2005. Relationships of isoflavone, oil, and protein in seed with yield of soybean. Agron. J. 97: 1314-1321.

**Table 1.1:** Seed composition goals for soybean breeders

| Seed composition component | Goal (13% moisture basis) | Goal (dry matter basis) |
|---|---|---|
| Protein | ≥35% (North) ≥36% (South) | ≥40.2% (N) 41.4% (S) |
| Meal protein | ≥48% | |
| Oil | ≥19% (North) ≥18% (South) | ≥21.8% (N) 20.7% (S) |
| Oleic acid | ≥75% (as percent of total fatty acids) | |
| Linolenic acid | ≤3% (as percent of total fatty acids) | |
| Sucrose | | ≥7% |
| Cysteine + Methionine | 3.5% of total protein | George and de Lumen 1991 |

# CHAPTER 2

# IDENTIFICATION AND CHARACTERIZATION OF FAST-NEUTRON INDUCED MUTANTS WITH ELEVATED SEED PROTEIN CONTENT IN SOYBEAN [1]

[1] Prenger, E.M., R. Mian, R. Stupar, T. Glenn, and Z. Li. To be submitted to *Theoretical and Applied Genetics*.

**Abstract**

Soybean seed composition affects the utility of soybean. Improving soybean seed composition is an essential goal for soybean breeders. However, negative relationships between protein content and both yield and oil content present challenges. Fast neutron radiation introduces genomic mutations resulting in novel variation for traits of interest. Two elite soybean lines were irradiated with fast neutrons and screened for altered seed composition. Twenty-three lines with altered protein, oil, or sucrose content were selected based on near-infrared spectroscopy data from five environments over two years and yield tested at five locations. Mutants with significantly increased protein averaged 14 to 29 g $kg^{-1}$ more protein than the parents. Comparative genomic hybridization (CGH) identified putative mutations in a mutant, G15FN-12, that has 24 g $kg^{-1}$ higher protein than the parent genotype, and whole genome sequencing (WGS) of the mutant has confirmed these mutations. An $F_{2:3}$ population was developed from G15FN-12 to determine association between genomic changes and increased protein content. Bulked segregant analysis of the population using the SoySNP50K BeadChip identified a CGH- and WGS-confirmed deletion on chromosome 12 responsible for elevated protein content. The population was genotyped using a KASP marker designed at the mutation region based on the whole genome sequence. Results indicated that $F_2$ individuals homozygous for the deletion on chromosome 12 averaged 27 g $kg^{-1}$ higher seed protein and 8.4 g $kg^{-1}$ lower oil than homozygous wild type individuals. Mutants with altered seed composition are a new resource for gene function studies and provide elite materials for development of varieties with improved seed composition.

## Introduction

Soymeal is a major source of protein consisting of essential amino acids, and energy in animal feed. Soybeans account for 70% of worldwide protein meal consumption, indicating the utility of this commodity in feed (http://soystats.com). Protein in soybean meal is crucial for animal growth and development, and meal protein content must be at least 48% (Chung et al. 2003) in order to meet U.S. standards for feed and ensure the competitiveness of soybean. However, the protein content of recently released U.S. soybean cultivars has been decreasing (unitedsoybean.org). Farmers need to produce high quality soybeans without sacrificing yield to meet the current needs of the livestock industry. Amino acid content, available energy, and vitamins and minerals also contribute to the usefulness of soybean meal (http://soymeal.org). Although soybean meal contains the essential amino acids, it is relatively low in methionine and cysteine, resulting in the need to supplement amino acids. Soymeal also contains the oligosaccharides raffinose and stachyose (RFOs), which cause flatulence in monogastrics such as swine, chickens, and humans (Patil et al. 2017). In addition to the prevalence of soybean meal in world protein meal consumption, soybeans are also the second largest source of vegetable oil worldwide, and the oil fraction of soybeans is an important economic component (http://soystats.com). Complicating the need to improve seed composition and yield simultaneously are negative relationships between protein content and yield, oil content (Burton, 1987; Hartwig and Hinson, 1972; Chung et al. 2003; Thorne and Fehr, 1970; Cober and Voldeng, 2000), and sucrose content (Hymowitz et al. 1972; Jaureguy et al. 2011).

Various quantitative trait loci (QTL) mapping and genome wide association studies (GWAS) have identified QTL for seed protein, oil, and carbohydrate content (Diers et al. 1992; Brummer et al. 1997; Kim et al. 2006; Qi et al. 2011a, 2011b; Hwang et al. 2014; Vaughn et al.

2014; Bandillo et al. 2015; Warrington et al. 2015; Phansak et al. 2016). To date, there are about 240 seed protein, 322 seed oil, 5 seed protein plus oil, 2 seed protein to oil ratio, 37 seed sucrose, and 15 seed oligosaccharide (sucrose, raffinose, and stachyose) QTL reported (http://soybase.org). QTL for protein and oil contents have been reported on every chromosome (Chr) in the soybean genome. Phansak et al. (2016) reported that most of the QTL for protein and oil listed in Soybase have not been confirmed, and none have been cloned. In fact, only 16 of the seed protein QTL and 16 seed oil QTL listed in Soybase are designated as "cq," or "confirmed QTL." Two of these confirmed protein QTL are on Chrs 15 and 20 and have been detected in multiple studies (Diers et al. 1992; Sebolt et al., 2000; Chung et al., 2003; Nichols et al., 2006; Qi et al. 2011b; Hwang et al. 2014; Vaughn et al. 2014; Bandillo et al. 2015; Warrington et al., 2015; Kim et al., 2016; Phansak et al. 2016). The high-protein alleles at these loci are typically associated with decreased oil and yield. Sources of the high-protein alleles at these loci include PI 468916, PI 407788A, PI 437088A, and Danbaekkong (Sebolt et al. 2000; Chung et al. 2003; Nichols et al. 2006; Warrington et al., 2015; Kim et al., 2016; Brzostowski et al. 2017; Brzostowski and Diers, 2017). The protein QTL at the Chr 15 can increase protein content 5 to 18 g $kg^{-1}$, and this QTL typically accounts for 5 to 26% of phenotypic variation (Bandillo et al. 2015; Diers et al. 1992; Qi et al. 2011b; Vaughn et al. 2014; Warrington et al. 2015; Kim et al., 2016). Individuals with homozygous high-protein alleles on Chr 15 from PI 407788A were also associated with 57 to 109 kg $ha^{-1}$ decreases in yield in two genetic backgrounds from Maturity Group II (MG II) (Brzostowski and Diers, 2017). The Chr 20 protein QTL accounts for 7 to 65% of phenotypic variance for protein content, and individuals homozygous for the high-protein allele may have an increase in protein of 8 to 30 g $kg^{-1}$ and a decrease in oil of 4 to 17 g $kg^{-1}$ relative to individuals homozygous for the wild type allele (Diers

et al. 1992; Sebolt et al. 2000; Qi et al. 2011b; Hwang et al. 2014; Vaughn et al. 2014; Bandillo et al. 2015; Phansak et al. 2016; Warrington et al. 2015). Lines homozygous for the high-protein allele were also shown to have yield decreases of 134 to 455 kg $ha^{-1}$ across environments in MG II, III, and IV backgrounds and in crosses with a MG III cultivar (Chung et al. 2003; Nichols et al. 2006; Brzostowski et al. 2017). A high-protein NIL in a MG VII genetic background was recently shown to have yield parity with the recurrent parent, despite a 41 g $kg^{-1}$ increase in protein content due to homozygous Danbaekkong alleles at the Chr 20 locus. This result indicates the potential for increasing protein content without consequently decreasing yield in some genetic backgrounds (Prenger et al., 2018).

Two studies have reported protein content QTL in the 6 to 15 Mb region of Chr 12. This region of Chr 12 is near a deletion of interest identified in this study. Lu et al. (2013) mapped a protein content QTL in the 6-12 Mb region of Chr 12 using a population of 212 recombinant inbred lines (RILs). The QTL was identified in all six testing environments and explained 4.5% of phenotypic variance. Teng et al. (2017) mapped a protein content QTL in the 9-15 Mb region of Chr 12 using 129 RILs. The QTL explained 3.3 to 11.3% of phenotypic variation and was detected in seven out of 10 environments.

Many sources of high-protein alleles are from *Glycine max* plant introductions (PIs) or landraces (Patil et al. 2017). The soybean wild ancestor, *Glycine soja* [Sieb. & Zucc.] may be used as a source of genetic diversity in breeding programs to introduce important traits such as high protein content (Diers et al. 1992; Brummer et al. 1997). However, the use of *G. soja* or *G. max* landraces to introduce traits of interest in breeding programs may result in poor agronomic traits and yield and require substantial effort to break linkage drags. Sources of variation for seed

composition traits of interest in elite soybean backgrounds have the potential to expedite the development of soybeans with improved seed composition and yield parity.

Fast neutron (FN) radiation is a type of ionizing radiation that has been shown to induce heritable genomic deletions, duplications, insertions, translocations, and single base pair substitutions in soybean and Arabidopsis (Bolon et al. 2011, 2014; Belfield et al. 2012). The number and size of mutations as well as survival of mutant plants depends on the radiation dosage used. Copy number variations (CNV) detected by Bolon et al. (2014) in 264 fast neutron mutants exposed to 4, 8, 16, or 32 Gray Units (Gy) radiation included duplications ranging from 499 bp to 50 Mb in size. Hemizygous deletions in these mutants ranged from 4.9 kb to 9.3 Mb, and homozygous deletions ranged from 493 bp to 8.1 Mb. Anderson et al. (2016) studied the same mutants and determined the number of genes in duplicated regions ranged from 0 to 2,312 per plant, while the number of genes in deletion regions ranged from 0 to 290 per plant.

The genome sequence of soybean MG III cultivar 'Williams 82' was released in 2010 (Schmutz et al. 2010). The assembled genome covers approximately 955 Mb of an estimated 1.1 Gb total genome size. The Williams 82 genome was revised to develop version 2, which is 978.5 Mb in size (https://phytozome.jgi.doe.gov/). The Williams 82 genome sequence was used to develop an array Comparative Genomic Hybridization platform for copy number variation (CNV) detection (Bolon et al. 2011). The original NimbleGen soybean microarray utilized nearly 700,000 unique probes spaced throughout the genome to detect differences in hybridization between a mutant and a reference control. A normalized $\log_2$ ratio of control to mutant hybridization identified regions of the mutant genome containing duplications and deletions. However, due to gaps between probes spaced throughout the genome, structural variation events such as small deletions and duplications may not be detected (Bolon et al. 2011). An updated

array with 1.4 million probes was utilized in a later study with minor modifications. The expanded NimbleGen CGH array consisted of unique probe sequences designed based on the Williams 82 version 1 (W82.a1.v1.1) reference genome sequence (Schmutz et al. 2010) and spaced approximately 0.5 kb apart, on average (Bolon et al. 2014). The most recent CGH array from Agilent Technologies, Inc. consists of one million unique probes enriched for genic regions throughout the soybean genome. This array requires less genomic DNA and a shorter hybridization time than previous arrays, and CNV are identified essentially as in previous studies (Dobbels et al. 2017). CGH has been established as an effective way to detect large CNV throughout the soybean genome (Bolon et al. 2011, 2014; Stacey et al. 2016; Campbell et al. 2016; Dobbels et al. 2017).

CGH can be combined with next-generation sequencing to detect mutations in fast neutron mutants of soybean and to help determine associations between mutations and phenotypes of interest. Bolon et al. (2014) associated a deletion on Chr 10 with a high oil/low protein phenotype using a combination of CGH, bulked segregant analysis (BSA), and high-throughput paired-end sequencing. The association was validated in segregating backcross progeny. A short petiole mutant phenotype in this same study was found to be associated with a tandem duplication on Chr 17 based on CGH data from multiple $M_4$ plants, paired-end high-throughput sequencing, and PCR confirmation in an outcross population.

Stacey et al. (2016) found an association between a brown seed mutant phenotype and deletion of the gene *GmHGO1*. CGH identified three candidate deletions, and a putative causal gene was determined based on predicted gene function. The role of the gene was confirmed through Southern blot analysis, chemical tissue analysis, and transgenic complementation (Stacey et al. 2016). Campbell et al. (2016) identified the causal mutation for a gnarled trichome

mutant phenotype using CGH and BSA through whole genome sequencing of bulked plants. BSA identified a deletion on Chr 20 containing a candidate gene homologous to the Arabidopsis gene *NAP1* with a reported role in trichome development. PCR, RNA-seq, and transgenic complementation were used to validate the association of the deletion of the *NAP1* homolog with gnarled trichomes. Dobbels et al. (2017) associated a high sucrose/low oil mutant phenotype with a reciprocal translocation between Chrs 8 and 13 interrupting a *KAS1* ortholog. CGH detected putative CNV, and BSA through whole genome sequencing of bulks identified the candidate causal mutation. PCR and sequencing confirmed the presence of the mutation, and a backcross population was used to validate the role of the translocation in the high sucrose/low oil phenotype. Hwang et al. (2015) identified a fast neutron mutant in the 'Williams 82' background showing a dwarf phenotype. Whole genome sequencing identified three deletions that were validated using PCR, and reverse-transcription PCR (RT-PCR) was utilized to determine loss of expression of a peroxidase superfamily protein in one of the large deletions detected.

FN mutagenesis of elite germplasm could facilitate quicker development of new varieties with improved seed composition phenotypes. Additionally, the mutants will broaden the scope of FN-induced mutations in soybean and increase the possibility of finding novel mutations in genomic studies for application in plant breeding programs. The objective of this work was to develop, identify, and characterize fast neutron mutant lines in elite genetic backgrounds from late maturity groups for improvement of seed composition and to identify and confirm the genomic regions associated with increased protein content in these mutants.

## Materials and Methods

*Population development and advancement*

Fast neutron mutant lines were developed by irradiating 2.3 kg seed of each of two elite genotypes, G00-3213 and G00-3880. Both G00-3213 and G00-3880 are high-yielding MG VII breeding lines developed at the University of Georgia. The mutant line development and advancement scheme is shown in Figure 2.1. Irradiation of seeds was performed at the McClellan Nuclear Radiation Center in McClellan, CA in 2013. A radiation dosage of 25 Gray Units (Gy) was used based on the results of a study by Bolon et al. (2011), who studied the effects of 4, 8, 16, and 32 Gy of fast neutron radiation on emergence, seed production, and frequency of mutant phenotypes of soybean.

The $M_1$ seeds were grown in Watkinsville, GA in 2013 and advanced to the $M_2$ by single pod descent. About 1,200 $M_2$ plants from each genetic background, G00-3213 and G00-3880, were grown in Watkinsville, GA in 2014. Seed composition analysis of the $M_{2:3}$ seed was conducted using Near Infrared (NIR) spectroscopy on a Perten DA 7250 analyzer (PerkinElmer Inc., Stockholm, Sweden). NIR provides a seed composition profile with percent protein, oil, amino acids, and carbohydrates reported on a dry matter basis and seed fatty acid content reported as a percent of the total oil.

NIR data were factored into the selection of 231 $M_{2:3}$ lines to grow in plant rows in 2015 in Watkinsville, GA. Four reps of G00-3213 and six reps of G00-3880 were included in the test. The $M_{2:4}$ bulked seed from each mutant line and parent row were analyzed via NIR after harvesting by row. The $M_{2:4}$ seed was grown in plant rows in 2016 at four locations (Watkinsville, GA; Plains, GA; Caswell, NC; and Clayton, NC) with six reps each of G00-3213 and G00-3880. Each mutant line was grown in one rep per location in an augmented design. In

Watkinsville, seeds were planted at a density of approximately 43 seeds meter$^{-1}$ in one-row plots with 2.1 m length and 76 cm row spacing. In Plains, seed were planted at a density of approximately 37 seeds meter$^{-1}$ in two-row plots with a length of 4.9 m and 76 cm row spacing. In North Carolina, single rows with 3.7-m length and 97-cm row spacing were planted at a density of 27 seeds meter$^{-1}$. At maturity, three individual plants were harvested from each of 98 mutant lines in Watkinsville that were selected for altered protein, oil, or sucrose to determine within-plot seed composition variation. Entire plots were then individually harvested as a bulk. Both individual plants and entire rows were analyzed using NIR spectroscopy.

*Seed composition analysis (NIR and wet chemistry)*

NIR analysis was performed on approximately 180-250 whole seeds in a white dish using a Perten DA 7250 analyzer. Each seed sample was analyzed based on a calibration curve developed using hundreds of samples with known seed composition values (Soybean NIR Consortium). Based on the results of an in-house experiment (unpublished data), NIR values for samples large enough to cover the bottom of the small dish (about 90-110 seeds) are accurate for protein and oil. The same experiment revealed that smaller samples with around 38 seeds measured in a mirror cup return similar protein and oil results as a fully covered small white dish. NIR data from four locations in 2016 were used to confirm stably altered seed composition traits. NIR data from 2015 were not directly included in analyses due to poor seed quality from late harvest, although 2015 NIR values were noted to either agree or disagree with the 2016 increase or decrease in seed composition values relative to the parent genotype.

Seed samples from 18 selected mutant lines and parent seed samples were also analyzed for protein and oil contents at the University of Missouri Agricultural Experiment Station

Chemical Laboratories using a wet chemistry method in 2016 to confirm the NIR results. The proximate analysis package using LECO protein extraction was performed. Percent protein, moisture, crude fat, ash, and fiber were reported on an as-is basis and converted to dry basis for comparison to NIR values. Correlations between NIR and wet chemistry values were determined using JMP version 13.2 software (SAS Institute, 2017).

*Comparative genomic hybridization and whole genome sequencing analysis*

Four $M_4$ mutant lines (G15FN-12, G15FN-23, G15FN-54, and G15FN-109) were selected for genomic mutation characterization via comparative genomic hybridization (CGH). Mutant lines were selected based on both NIR and wet chemistry data. NIR data of CGH mutants with wet chemistry confirmation are shown in Table 2.1a. Protein and oil NIR values from the single plant used as the source for CGH, and both NIR and wet chemistry values on a whole plot basis for high protein mutant G15FN-12 are shown in Table 2.1b.

CGH analyses were performed on a single $M_{4:5}$ plant for each mutant line along with the parent reference. Genomic DNA was extracted using a Qiagen DNeasy plant kit. CGH was performed in the Soybean Genomics Lab at the University of Minnesota, St. Paul, MN using the procedure described in Dobbels et al. (2017). Briefly, the most recent CGH microarray by Agilent Technologies, Inc. utilizes one million unique probes designed based on the soybean reference genome version Glyma.Wm82.a2.v1 (https://phytozome.jgi.doe.gov/). The following procedures were performed to manufacturer specifications. A total of 500 ng DNA of each mutant of interest as well as a parent reference were labeled with dye (Cy3 for the mutants and Cy5 for the parent reference). Labeled DNA was hybridized to the microarray for 66 hours at 67˚. A $\log_2$ ratio of mutant to control hybridization was calculated for each probe. Software

including Agilent Genomic Workbench (version 7.0.4.0) and the Agilent feature extraction (version 12.0.0.7) were used to identify significant anomalies for each CGH run from raw data. CGH results were visualized using JMP version 13.0 software and Microsoft Excel. Significant probes underlying potential CNV were determined by setting a $\log_2$ threshold of three standard deviations above the average or three standard deviations below the average $\log_2$ ratio for each mutant (Bolon et al. 2011). Significant regions found in more than one mutant from the same background indicated potential intra-cultivar variation.

Whole genome sequencing (WGS) was performed for two of the CGH mutants (G15FN-12 and G15FN-54) along with the parent, G00-3213. Three individual $M_{4:5}$ plants grown from seed of each selected mutant were tissue sampled from the greenhouse in 2017, and seed composition phenotypes were verified using NIR after harvest. Tissues were lyophilized using a VirTis Freezemobile 35L freeze dryer and ground with a Spex Sample Prep Geno/Grinder 2010 machine using BB's. Genomic DNA was extracted using a CTAB protocol modified from Keim et al. (1988). Genomic DNA was sheared to an average size of 500 bp using a Bioruptor-UCD200 (Diagenode Inc., Deville, NJ, USA) for all samples except those already degraded to that approximate size. Library prep was performed using a KAPA Hyper Kit with iTru adaptors and primers (Roche Sequencing, USA; Illumina, Inc.; Glenn et al., 2016). Each product was run on a 1% gel at a volume of 3 µl for size verification, cleanup, quantification with Qubit (Thermo Fisher Scientific, Waltham, MA, USA), and pooling for sequencing. Sequencing was performed on Illumina HiSeq (Illumina Inc., 2017) at the Oklahoma Medical Research Foundation core facility, and produced 150-bp paired-end reads with an approximate insert size of 500 bp. FastQ file preparation, sequence quality control, and cleanup were performed using CyVerse apps (www.cyverse.org). Apps used include Uncompress files with gunzip (Vaughn, 2012),

Concatenate multiple files (Walls, 2018), HTProcess-prepare_directories-and-run_fastqc (Andrews, 2014; Barthelson, 2014b), and HTProcess_trimmomatic_0.32 (Bolger et al. 2014; Barthelson, 2014a). Paired reads of G15FN-12 and G00-3213 were aligned to the 'Williams 82' reference genome version 2.0 (Schmutz et al. 2010; https://phytozome.jgi.doe.gov/). Paired reads of G15FN-12, G15FN-54, and G00-3213 were aligned to the newly released 'Lee' reference genome (https://soybase.org/projects/SoyBase.B2018.01.php). Alignment of sequences was performed using either BWA-mem_0.7.15 or BWA-mem_longreads-0.7.15 (Li, 2013; Devisetty, 2016a; Cooksey, 2017). SAM files were converted to indexed BAM files using the SAM to Sorted BAM and SAM to sorted BAM-0.1.19_app_for_workflows Apps (Li et al. 2009; Li, 2011; Barthelson, 2016; Devisetty, 2016b). Most of the programs mentioned previously are currently available through CyVerse (www.cyverse.org).

Alignments of G15FN-12, G15FN-54, and G00-3213 sequences to the Williams 82 (W82.a2.v1) reference genome were also performed using the University of Georgia Sapelo cluster, which is a Linux cluster running a 64-bit CentOS operating system. Quality control and read cleanup were performed with FastQC (Andrews, 2014), cutadapt (Martin, 2011), and trimmomatic (Bolger et al. 2014). Alignment to the reference genome was performed using SAMTools (Li et al. 2009; Li, 2011) and the Burrows-Wheeler Aligner (BWA) (Li and Durbin, 2009). Integrative Genomics Viewer (IGV) (Robinson et al. 2011; Thorvaldsdóttir et al. 2013) was used to view each mutant genome alignment and compare it to the G00-3213 parent using either 'Williams 82' or 'Lee' as the reference genome.

Initial PCR primers were designed for candidate deletions on Chrs 10, 12, and 17 to test for these deletions in G15FN-12. The reverse primer for each candidate deletion was designed within the deletion region so that no DNA product would be amplified in plants homozygous for

the mutation of interest and a product would be amplified in the wild type. Apex Taq 2X Master Mix (Genesee Scientific, San Diego, CA, USA) was used to perform PCR reactions. For each 10-µl reaction, 5 µl Apex Taq 2X Master Mix and 0.3 µl of each primer (Sigma-Aldrich, St. Louis, MO, USA) was added to result in final concentrations of 1X and 0.3 µm, respectively. DNA was added in volumes of 1-2 µl to reach a total of 10 to 50 ng, and PCR-grade water was added to each reaction to reach a volume of 10 µl. These initial primers were used due to large gaps between CGH probes and unknown breakpoints for deletions of interest.

*Mapping mutant genomic regions responsible for elevated protein content using a bi-parental population*

Mutant line G15FN-12 with elevated protein was crossed with the elite variety 'Benning' (Boerma et al. 1997) to develop populations for validation of causal mutations for traits of interest. The cross between Benning and G15FN-12 was made in summer 2016 in the UGA Soybean Breeding Program crossing block near Watkinsville, GA. $F_1$ plants were advanced in the UGA greenhouse in the winter of 2016-2017 in Watkinsville, GA. In addition, 10 individual plants were harvested from the G15FN-12 crossing block rows in the fall of 2016, and the $M_{4:5}$ seed of each plant was grown as a single plant row in summer 2017 to study variation in protein content among plants within the mutant line and to determine the protein content inheritance patterns.

The $F_2$ seeds from each $F_1$ plant were harvested in the spring of 2017 and were grown at the UGA Iron Horse Plant Science Farm, Watkinsville, GA in the summer of 2017. Phenotypic data in the $F_2$ generation was used to verify a successful outcross at $F_1$ development. After identifying successful Benning × G15FN-12 plants, 276 $F_2$ plants and six Benning and G15FN-

12 plants were individually tagged and tissue sampled in 96-well plates in the field. Tissue samples were lyophilized using a VirTis Freezemobile 35L freeze dryer, and DNA was extracted using a CTAB protocol modified from Keim et al. (1988). Each plant was harvested and threshed individually at maturity, and $F_{2:3}$ seed was analyzed using NIR spectroscopy with a Perten DA 7250 machine.

Based on NIR analysis results, two bulks were formed using genomic DNA of 20 $F_2$ progeny with the highest protein and 20 with the lowest protein from the Benning × G15FN-12 population. Bulk samples were genotyped using the SoySNP50K Infinium BeadChip (Song et al. 2013) in the Soybean Genomics and Improvement Laboratory, USDA-ARS, Beltsville, MD.

Allele calls were made using Genome Studio software (Illumina Inc., 2017), and regions of interest were manually checked for accuracy. SNPs polymorphic between Benning and G00-3213 were identified first. Next, genomic regions with markers polymorphic between the two bulks were identified. Polymorphic regions where the low protein bulk matched Benning and the high protein bulk matched G00-3213 or was heterozygous, were identified as regions putatively associated with the trait of interest. G00-3213 SNP data were used in place of the actual mutant since it is the parent background for G15FN-12.

SNPs identified using BSA were used to locate putative mutation regions for Kompetitive Allele Specific PCR (KASP) marker development, and a KASP marker was developed for the putative deletion on Chr 12. The KASP marker was designed to have a FAM or mutant forward primer bridging the putative deletion, a VIC or wild type forward primer sitting within the 3' end of the deletion, and a common reverse primer in the flanking region at the 3' end of the deletion. $F_2$ progeny were assayed using the developed KASP marker to test associations between the genotypes and NIR data.

The KASP (LGC Genomics, Middlesex, UK) assays were performed by a similar protocol to Pham et al. (2013). Briefly, in each reaction, 2.0 µl of KASP 2X assay mix, 0.055 µl 5X primer assay mix (Sigma-Aldrich, St. Louis, MO, USA), and 2.0 µl DNA at an approximate concentration of ~20 ng $\mu l^{-1}$ were mixed. Reactions were performed in a 384-well PCR plate, and a KASP PCR program was used to amplify the region of interest for SNP genotyping. After the PCR reaction, the plate was read and genotypes for each reaction were recorded using a Tecan Infinite® M1000 Pro (Tecan Group Ltd., Männedorf, Switzerland). Genotypes and NIR phenotype data were tested for associations in JMP Pro version 13.2. The same methods will be used to confirm associations between genotypes and phenotypes in the $F_{3:4}$ generation.

*Yield trials*

NIR data from five total environments over two years were used to select 23 stable seed composition mutant lines. These were used for yield trials in 2017 to study the effects of mutations and altered seed composition on yield and agronomic traits. Both parents, G00-3213 and G00-3880, were included as checks. Mutant lines were selected based on stably altered protein, oil, or sucrose levels. Each $M_{2:5}$ mutant line was grown in five locations (Watkinsville, GA ; Plains, GA; Plymouth, NC; Caswell, NC; and Bossier, LA) in a randomized complete block design. Three reps per location were used, except in the case of insufficient seed for four mutant lines, for which the third rep or an entire location was dropped. Each plot consisted of three or four rows with the inner row(s) harvested. Plots in Watkinsville and Plains, GA and Bossier, LA consisted of four rows with 4.9 m in length and 76 cm between rows. Seeds were planted at a density of about 27 seeds $meter^{-1}$. Plots in Plymouth and Caswell, NC consisted of three rows with 6.4 m in length and 97 cm between rows. Seeds were planted at a density of about 31 seeds

meter$^{-1}$. In 2017, additional lines were included in the plant rows for single plant selection to study within-row variation and to increase seed.

Yield, protein, oil, and sucrose content, maturity, plant height, lodging, and seed size were measured for each plot. Protein, oil, and sucrose contents were measured using NIR and reported as g kg$^{-1}$ of seed on a dry matter basis for all reps at all locations. Maturity was recorded as the number of days past August 31 required for 95% of a plot to reach physiological maturity (R8) (Fehr and Caviness 1977) and was noted for all reps at four locations except for one rep at Caswell, NC. Plant height was measured in centimeters for all reps at three locations. Plant lodging was measured on a scale from 1 to 5, where one indicates plants were erect and five indicates plants were prostrate on the ground. Lodging was measured for all reps at four locations except one rep at Caswell, NC. Seed size was measured in grams 100-seeds$^{-1}$ and was measured for all reps at all locations.

*Statistical Analysis*

Analyses of seed composition across all four locations in 2016 were performed to identify stable seed composition mutants for further study. Augmented design analysis of mutant lines across all four locations in 2016 was performed using the lmerTest package in R software (Kuznetsova et al. 2016; R Core Team, 2013). Analysis returned a Best Linear Unbiased Predictor (BLUP) value for each mutant line. The following formula was used: model=lmer (Trait ~ check + (1|Location) + (1|Name), data=data). NIR data from 2015 were used to determine consistency of altered seed composition between years.

Data quality control of yield trials was performed using Agrobase (Agronomix Inc., 2012) to identify extreme outliers. These extreme outliers were replaced with values imputed by

Agrobase software based on surrounding plots and the other reps of the same genotype. Unusual seed composition values were subjected to NIR analysis a second time to minimize machine error.

Data of yield trials were analyzed in SAS version 9.4 (SAS Institute, 2017). A PROC MIXED procedure was utilized, in which genotype was treated as a fixed effect and environment, genotype by environment interaction, and rep within environment were random effects. Tukey's honestly significant difference (Tukey's HSD, $\alpha$=0.05) was used to determine significant differences for traits including yield, protein, oil, sucrose, maturity, height, lodging, and seed weight in mutant lines relative to the parent and check genotypes. Correlations among seed composition traits and yield were calculated using parents, commercial checks, and mutant lines except G15FN-54, due to extremely low oil and high sucrose contents. Data from all reps and locations were analyzed using JMP version 13.2 software.

## Results

### *Identification and performance of seed composition mutants*

NIR data from 2016 were used to develop a seed composition profile for each mutant line (Table 2.2). NIR data for 2015 were used to identify mutants with an increase or decrease in seed composition components, but were not included in analyses due to poor seed quality resulting from a late harvest. Augmented design analyses of 2016 NIR data confirmed the altered seed composition in mutants of interest. In the G00-3213 background, BLUPs of 90 mutant lines for protein content ranged from 408.5 to 431.2 g $kg^{-1}$, oil content ranged from 112.2 to 201.8 g $kg^{-1}$, and sucrose content was between 55.4 to 94.3 g $kg^{-1}$. In 141 mutant lines from the G00-3880 background, protein content ranged from 369.8 to 426.4 g $kg^{-1}$, oil content ranged from 174.4 to

210.8 g $kg^{-1}$, and sucrose content was between 52.9 and 72.3 g $kg^{-1}$. Mutant lines of interest had higher protein, lower oil, or increased sucrose content compared to the parent BLUPs in augmented analyses. These included G15FN-12 with 18.3 g $kg^{-1}$ higher protein, G15FN-23 with 8.5 g $kg^{-1}$ higher protein, G15FN-54 with 83.7 g $kg^{-1}$ lower oil and 27.1 g $kg^{-1}$ higher sucrose, and G15FN-109 with 32.0 g $kg^{-1}$ higher protein than the parents.

Wet chemistry analysis of eighteen mutant lines for protein and oil contents revealed a correlation coefficient between wet chemistry values and NIR values of r = 0.92 for protein and r = 0.86 for oil. The $R^2$ value of the line of best fit was 84.6% for protein and 73.8% for oil. Each regression line was significant with a P-value of <0.0001. Results of wet chemistry data showed that NIR is a reasonably accurate high-throughput method of analyzing seed composition.

Seeds from ten individual G15FN-12 plants harvested from the 2016 crossing block were grown as ten different plant rows in 2017, and NIR data collected after harvest revealed fairly consistent protein content among $M_{4:5}$ rows for the G15FN-12 selections. Protein content in these rows ranged from 432.9 to 466.0 g $kg^{-1}$, oil ranged from 168.5 to 186.8 g $kg^{-1}$, and sucrose ranged from 58.3 to 67.8 g $kg^{-1}$. Nine out of ten rows had greater than 447 g $kg^{-1}$ protein and less than 186 g $kg^{-1}$ oil. In comparison, the G00-3213 parent contained about 432.3 g $kg^{-1}$ protein, 182.3 g $kg^{-1}$ oil, and 73.9 g $kg^{-1}$ sucrose. This consistency among rows derived from individual plants indicates that the mutations causing phenotypes of interest are fixed and supports the idea that the increase in protein content in G15FN-12 is mainly due to a single mutation.

*Characterization of mutant lines using comparative genomic hybridization and sequencing analysis*

Comparative genomic hybridization was performed on four mutants (G15FN-12, G15FN-23, G15FN-54, and G15FN-109). The results revealed two to three large deletions per mutant. For each mutant, individual probes with significant hybridization ratios potentially coinciding with smaller deletions or duplications were also identified. Significant association of single probes was determined by a method similar to that of Bolon et al. (2011), in which probes with a $\log_2$ hybridization ratio three standard deviations greater or smaller than the average $\log_2$ hybridization ratio for each mutant were considered significant. Probes that were significantly associated across all mutants from the same background were considered to represent potential intracultivar variation.

Based on CGH data, there are two large deletions in G15FN-12, a high protein mutant, on Chrs 12 and 17 (Figure 2.2). The largest deletion was located from 15.7 to 21.1 Mb of Chr 12, is approximately 5.4 Mb in size, and contains 137 gene models (Figure 2.3). The other large deletion observed in G15FN-12 is located in the 30.5 Mb region of Chr 17 and is approximately 4 kb in size. The Chr 17 mutation contains no gene models, based on the Williams 82 reference genome. Several significant individual probes were located on Chrs 6 and 10. The Chr 6 probe has a positive $\log_2$ hybridization ratio, indicating a possible duplication. The Chr 10 probe has a negative $\log_2$ hybridization ratio and is located in the 51.2 Mb region. The Chr 10 significant probe is located in the intron of one gene, Glyma.10G295100, a putative Callose Synthase 9 gene in the 1,3-β-D-glucan biosynthesis pathway (https://phytozome.jgi.doe.gov/phytomine/). A putative intra-cultivar variation peak on Chr 2 in the G15FN-12 mutant also appears in the G15FN-23-3 mutant.

G15FN-23, a high protein mutant, has two large putative deletions plus the same putative intra-cultivar variation found on Chr 2 in G15FN-12. One of the deletions is on Chr 18 in the 5.7 to 5.8 Mb region, is approximately 50 kb in size, and contains up to seven gene models. The other large deletion is located on Chr 19 in the 16.9 to 20.7 Mb region. The most significant probes are in a 1 kb region around 17.5 Mb, flanked on both sides by significant probes with much smaller $\log_2$ ratios. This distribution and magnitude of significant probes could indicate a homozygous deletion 1 kb in size or a hemizygous deletion up to 3.7 Mb in size containing from zero to 28 gene models. A single significant probe with a high positive $\log_2$ hybridization ratio is also found on Chr 14 in the 9.8 Mb region that may involve one gene model.

G15FN-54, a high sucrose/low oil mutant, contains three large putative deletions and several significant individual probes with highly negative or positive $\log_2$ hybridization ratios based on CGH data. One of the large deletions is found on Chr 10 in the 18.5 to 20.8 Mb region, is approximately 2.3 Mb in size, and contains 14 gene models. Another large deletion is found on Chr 13 in the 13.9 Mb region, is about 17 kb in size, and interrupts one gene model. The final large deletion is located in the 7.7 to 8.0 Mb region of Chr 16, is approximately 300 kb in size, and contains 21 gene models. Two significant individual probes on Chrs 8 (19.4 Mb) and 19 (4.1 Mb) have significant negative $\log_2$ hybridization ratios. Two additional probes on Chrs 16 (4.0 Mb) and 19 (1.7 Mb) had significant positive hybridization ratios, potentially indicating duplications. A single gene model overlaps with the Chr 19 probe with a significant negative hybridization ratio, and another gene model is located between the Chr 19 probe with a significant positive hybridization ratio and the two adjacent probes.

The G15FN-109 mutant has more than 20 large peaks based on CGH data. It is likely that many of these are naturally occurring intracultivar variation, but it is difficult to tell without

other G00-3880-derived mutants for comparison. CGH of additional mutants in the G00-3880 background and additional sequencing could help to characterize these regions.

CGH results were further defined in the mutants G15FN-12 (Table 2.3) and G15FN-54 using whole genome sequencing data and alignment to both the 'Williams 82' version 2.0 and 'Lee' genomes. Williams 82 is a MG III cultivar, and Lee is MG VI. Since Lee is closer in maturity to our MG VII lines, typically it is more genetically similar. This is evidenced by similarity analysis using 50k whole genome SNP makers, which show that G00-3213 has 77.8% and 63.8% similarity to Lee and Williams 82, respectively. Williams 82 is still a valuable reference genome because of extensive research on the genome and characterization of putative gene transcripts. The parent of both mutants, G00-3213, was also sequenced. The G15FN-12 final alignment had an approximate average read depth of 26x, G15FN-54 was covered at an approximate depth of 20x, and the parent G00-3213 was covered at an approximate depth of 4x after trimming and alignment to the Lee genome.

Sequencing confirmed the Chr 12 deletion in G15FN-12 based on both Williams 82 and Lee. However, the Chr 17 deletion was found only in the alignment to Williams 82 and was not found in the alignment to Lee in regions of homology based on BLAST results. A single probe on Chr 10 was found to overlap with a 17-bp deletion based on both the Williams 82 and Lee alignments. For G15FN-54, the deletions on Chr 10, 13, and 16 were confirmed in both the Williams 82 and Lee alignments. Physical positions of deletions based on the Williams 82 and Lee references varied slightly. Individual significant probes from each mutant were investigated, and no obvious mutations were found in these regions based on sequencing data, although probes with significant positive $\log_2$ ratios may be difficult to identify in WGS data due to variable

sequencing depth throughout the genome and effort to establish whether sequencing variation is due to a duplication or sequencing depth.

Based on CGH and whole genome sequencing results, PCR primers were designed to target the deletions in G15FN-12 on Chrs 10, 12, and 17. These primers were designed to amplify a product in both homozygous wild type-individuals and heterozygous individuals and to amplify no product in individuals homozygous for the mutation (Figure 2.4). Six mutant plants of G15FN-12, wild type G00-3213, and Benning were genotyped using these markers. The results indicated that all six mutants genotyped were homozygous for the Chr 12 and 17 deletions. However, the Chr 10 deletion was present in only half of the selected plants genotyped from the mutant, G15FN-12. Interestingly, genotyping of $F_2$ individuals in the Benning × G15FN-12 population with the Chr 17 deletion primers revealed putative homozygous deletions in the exact same individuals with a homozygous Chr 12 deletion. Only one Chr 12 wild-type individual contained the putative homozygous Chr 17 deletion. Based on this information and the absence of the Chr 17 deletion in the Lee alignments, a BLAST search was performed checking the putative Chr 17 deletion sequence and primer sequences against both the Lee reference Chr 17 and Lee Chr 12 FASTA files. Primers were previously checked and specificity was confirmed based on the Williams 82 genome. In the Lee genome, the Chr 17 deletion sequence and primers more closely match a region near 18 Mb of Chr 12 than any region of Chr 17. This would indicate that the putative Chr 17 deletion is not a separate deletion, but was detected possibly due to presence of the Chr 12 deletion, which extends from 15.9 to 21.6 Mb of Chr 12 in the Lee genome alignment, or due to a translocation in the Lee genome. These results along with inability to find the Chr 17 deletion in the G15FN-12 alignment to the Lee genome indicated the

Chr 17 deletion was possibly falsely identified due to presence of the Chr 12 deletion, although further investigation must be done to determine if a translocation is involved.

*Bulked segregant analysis of an $F_{2:3}$ population*

Based on NIR analysis results, high protein and low protein bulks of the $F_{2:3}$ population derived from Benning × G15FN-12 were formed and genotyped using the SoySNP 50K Infinium Beadchip (Song et al. 2013). Over 9,000 SNPs were polymorphic between the parents, Benning and G15FN-12, where genotypes of G00-3213 were used in place of G15FN-12. Since the BSA was performed using $F_2$ plants, large linkage blocks were expected. Thus, multiple polymorphic SNPs matching the expected parents in putative causal mutation regions for the phenotype of interest were expected.

Comparison of SNPs between the high protein and low protein bulks from the Benning × G15FN-12 population revealed two regions which contained four or more polymorphic SNPs in close proximity, for which the low protein bulk genotype matched the Benning SNP and the high protein bulk was either heterozgyous or matched G00-3213. One region was located from approximately 40.8 to 46.1 Mb of Chr 3 based on the Williams 82 genome, and contains only four markers at 40.81, 41.10, 46.06, and 46.07 Mb. Whole genome sequencing did not reveal obvious mutations in this region, although a more in-depth analysis of the region should be performed. The second region is from 1.7 to 32.1 Mb of Chr 12. This region contains 30 markers, with 29 of the SNPs located from 1.7 to 14.8 Mb and a single SNP at 32.1 Mb. These SNPs surround a large deletion revealed by both CGH and WGS. The Chr 12 deletion was chosen as a target region for study based on both BSA and results of PCR primers used to genotype the mutant, parent, and Benning.

*Mapping genomic deletions responsible for elevated protein content*

CGH data identified putative causal mutations for traits of interest and facilitated the design of PCR primers to test for deletions, and WGS provided a detailed view of the sequences surrounding these mutations for high-throughput Kompetitive Allele Specific PCR (KASP) marker development. Based on BSA results, a KASP marker was designed for the deletion on Chr 12. The KASP marker was tested on all 276 $F_2$ progeny derived from the Benning × G15FN-12 population (Figure 2.5). Both genotype and phenotypic NIR data were available for a total of 254 $F_2$ plants. Of the 254 plants, 34 were homozygous for the chromosome 12 deletion, 129 plants were heterozygous, and 91 plants were homozygous wild type. This is different from the expected 1:2:1 segregation ratio expected for a single mutation. Lower than expected rates of homozygous mutation plants may be due to a fitness cost of harboring the deletion or differences in segregation due to the large size of the deletion and chromosomal pairing issues during meiosis. Analysis of genotype and protein contents of these 254 $F_2$ progeny derived from the Benning × G15FN-12 revealed significant differences among all genotypic groups in relation to protein and some differences in other seed components (Table 2.4). Average protein of the plants homozygous for the Chr 12 deletion was 449.2 g $kg^{-1}$, while average protein for wild type plants was 422.1 g $kg^{-1}$, and protein of the plants homozygous for the Chr 12 deletion was significantly higher ($P<0.0001$) than the average protein of plants with the homozygous wild-type allele. The heterozygous plants have a protein content of 429.9 g $kg^{-1}$, which behaves as a partial dominant effect (Figure 2.6). This deletion explains approximately 20.5% of phenotypic variation for protein in this population.

*Impact of mutations on yield and agronomic traits*

NIR data from 2016 and 2015 were analyzed to select 23 mutant lines for inclusion in 2017 Georgia advanced yield trials. The 23 mutants were selected for either increased protein or sucrose content. In these two years, nearly all of the mutant lines (22 out of 23) showed an increase in protein compared to the parent across five environments, and 2 out of 23 mutant lines showed an increase in sucrose compared to the parent. One of the lines, G15FN-54, had both increased protein and increased sucrose, although it also exhibited at least 36 g $kg^{-1}$ lower oil than the parent.

Results of 2017 yield trials from five locations revealed significant decreases in yield compared to the parent check for 3 out of 23 mutant lines tested, based on a PROC MIXED analysis in SAS. Yield is reported on a Least Squares Means basis and significance is based on Tukey's HSD with $\alpha$=0.05. Yield of mutants in the G00-3213 background ranged from 2,743.8 to 3,207.8 kg $ha^{-1}$, which is below the yield of G00-3213 (3,389.4 kg $ha^{-1}$). Yield of mutants in the G00-3880 background ranged from 2,582.4 kg $ha^{-1}$ to 3,631.5 kg $ha^{-1}$, and G00-3880 yielded 3,564.3 kg $ha^{-1}$. In the G00-3213 mutants, protein measured as g $kg^{-1}$ of seed on a dry weight basis ranged from 430.2 to 454.8 g $kg^{-1}$, oil from 108.6 to 187.2 g kg-1, and sucrose from 62.4 to 108.9 g kg-1. In comparison, G00-3213 had 430.5 g $kg^{-1}$ protein, 187.7 g $kg^{-1}$ oil, and 71.8 g $kg^{-1}$ sucrose. G00-3880 mutants contained 395.1 to 440.3 g $kg^{-1}$ protein, 169.7 to 188.8 g $kg^{-1}$ oil, and 59.9 to 76.5 g $kg^{-1}$ sucrose, while G00-3880 contained 411.4 g $kg^{-1}$ protein, 188.0 g $kg^{-1}$ oil, and 70.4 g $kg^{-1}$ sucrose.

Overall, the protein content of 10 out of 23 mutant lines was significantly higher than that of the respective parent check and one was significantly decreased. The oil content of five mutant lines was significantly decreased, sucrose content of three lines was significantly

decreased, and one line had significantly increased sucrose. Another mutant line had a 6 g $kg^{-1}$ higher sucrose than the parent, but the increase was not significant. Seven lines showed significantly ($P \leq 0.05$) increased protein content (14 to 29 g $kg^{-1}$) without a significant decrease in yield. However, an overall decrease in yield (278 to 648 kg $ha^{-1}$) in these mutants relative to the parents was observed. Most lines did not have significantly altered height, lodging, or seed size relative to the parent check. Maturity was significantly delayed in two mutant lines, G15FN-189 and G15FN-216. Two lines, G15FN-12 (from the G00-3213 background) and G15FN-189 (from the G00-3880 background), had significantly larger seed size than the respective parents (Table 2.5).

The four CGH mutants (G15FN-12, G15FN-23, G15FN-54, and G15FN-109) were among the 23 mutants included in the 2017 UGA Advanced Yield Trials. G15FN-12 had significant increases in protein (24.3 g $kg^{-1}$) and seed size (1.9 g 100-$seed^{-1}$), and significantly decreased sucrose (9.4 g $kg^{-1}$) compared to the parent genotype, G00-3213. Yield of G15FN-12 was lower than the parent, but not significantly lower (391.4 kg $ha^{-1}$). G15FN-23 had no significant changes relative to G00-3213, although yield (378.8 kg $ha^{-1}$), oil (3.8 g $kg^{-1}$), and sucrose (4.7 g $kg^{-1}$) were decreased and protein was increased (13.2 g $kg^{-1}$) relative to the parent. G15FN-54 had significantly decreased oil (79.1 g $kg^{-1}$) and increased sucrose (37.1 g $kg^{-1}$) compared to the parent genotype, G00-3213. Yield (565.3 kg $ha^{-1}$) and protein (0.3 g $kg^{-1}$) were also decreased in this mutant compared to G00-3213, though the differences were not considered significant. G15FN-109 had significantly decreased yield (-694.9 kg $ha^{-1}$), oil (-18.3 g $kg^{-1}$), and sucrose (-10.4 g $kg^{-1}$), and increased protein (+28.9 g $kg^{-1}$) relative to the parent, G00-3880. Yield trial data for the CGH mutants can also be found in Table 2.5.

Correlations among seed composition components were calculated using all parents, checks, and mutant lines except G15FN-54 due to the extremely high sucrose and low oil. Protein and yield (r = -0.27), protein and oil (r = -0.37), and protein and sucrose (r = -0.54) all had significant (P<0.0001) negative relationships. Oil and sucrose (r = -0.07) and oil and yield (r = -0.02) had non-significant negative relationships. Only sucrose and yield had a significant positive correlation (r = 0.36, P<0.0001) (Table 2.6).

## Discussion

### *Challenges in breeding high protein soybean cultivars*

Negative correlations among seed components and yield have been well-documented. Protein content, especially, has a consistent negative relationship with oil content and yield (Burton, 1987; Hartwig and Hinson, 1972; Chung et al. 2003; Thorne and Fehr, 1970; Cober and Voldeng, 2000). Negative relationships have been reported between protein and sucrose as well (Hymowitz et al. 1972; Jaureguy et al. 2011). These negative correlations complicate the simultaneous improvement of soybean protein, oil, sucrose, and yield. Attempts have been made to break these negative relationships, with some studies indicating the negative relationship between protein and yield may be mitigated (Wilcox and Cavins, 1995; Wilcox and Zhang, 1997; Yin and Vyn, 2005; Cober and Voldeng, 2000). High-protein soybean germplasm has been used as parents in numerous bi-parental mapping studies and in GWA studies. However, most of these high-protein sources are unadapted plant introductions (PIs) and landraces and could require significant time and resource investment to develop improved varieties containing the desired high-protein alleles (https://soybase.org; Patil et al. 2017). Some sources of major protein QTL include the Korean tofu cultivar Danbaekkong (Warrington et al. 2015), *Glycine soja* PI

468916 (Diers et al. 1992; Sebolt et al. 2000; Nichols et al. 2006), PI 437088A (Chung et al. 2003), and PI 407788A (Kim et al. 2016). These QTL have been studied in various genetic backgrounds to determine the effects on oil, yield, and other agronomic components (Brzostowski and Diers, 2017; Brzostowski et al. 2017; Mian et al. 2017). The high protein allele from each source was associated with an increase in protein and a decrease in oil. Effects on yield were mixed, although the high-protein allele was typically associated with a significant decrease in yield. In our study, fast neutron mutants were developed in elite backgrounds to attempt to mitigate the yield drag associated with high protein.

*Fast neutron technology*

Fast neutron radiation has been used to induce heritable genetic mutations such as deletions, duplications, inversions, and translocations in soybeans resulting in novel phenotypic variation for traits of interest. These mutants have been utilized for gene function studies (Bolon et al. 2011, 2014; Hwang et al. 2015; Campbell et al. 2016; Stacey et al. 2016; Dobbels et al. 2017). Advances in microarray and sequencing technology have enabled detection of mutations leading to association of genomic mutations with mutant phenotypes, and CGH is an established method of detecting large copy number variations within the soybean genome (Bolon et al. 2011, 2014). Next-generation sequencing technologies are also increasingly affordable and informative in discovering mutations within the genome through methods including QTL-seq and whole genome sequencing (Bolon et al. 2011, 2014; Hwang et al. 2015; Campbell et al. 2016; Dobbels et al. 2017). Specific mutations and genes have been associated with mutant phenotypes of interest in each of the previous studies using a variety of methods.

Previous studies of fast neutron mutants combine various array technologies, next generation sequencing, and mapping approaches to identify candidate mutations and genes for the mutant phenotype of interest. Some of these methods, such as QTL-seq, are both expensive and computationally intensive and require significant time and expertise to complete the analysis. In our study, we used an approach combining CGH, minimal whole genome resequencing of only the mutant and parent, and BSA using the SoySNP50K Infinium BeadChip for a simplified and less expensive way to map putative causal mutations for the mutant phenotype. Previous studies established a clear route in soybean for using mutants in gene function discovery. These studies were largely performed to identify causal genes for mutant phenotypes and did not put much emphasis on the direct use of mutants as sources of desirable traits for use in breeding programs.

In this study, mutant phenotypes were identified that are useful for both gene discovery and as sources of traits for the development of improved soybean varieties. Previous studies used early maturity FN mutants, so development of mutants in MG VII backgrounds provides more resources for development of soybean germplasm with improved seed composition and competitive yield. Fast neutron mutants were developed in two elite backgrounds, G00-3213 and G00-3880, which have high yield and approximately 410 to 430 g $kg^{-1}$ protein, 180 to 190 g $kg^{-1}$ oil, and 7 g $kg^{-1}$ sucrose. Variation for seed composition traits such as protein, oil, and sucrose contents existed among the mutants from each genetic background. Due to the nutritional and economic importance of protein, oil, and sucrose, these components were the focus of this mutant screen. The discovery of causal genes for these traits of interest could lead to more precise introgression of the trait into elite backgrounds and direct development of improved

plants through CRISPR gene editing, if development and regulation allow such a product to be profitable.

Correlations among seed composition traits and yield based on 2017 yield trial data show significant negative relationships between protein and yield, protein and oil, and protein and sucrose. A significant positive relationship exists between sucrose and yield. Non-significant negative relationships exist in these mutant lines between oil and sucrose and between oil and yield. Oil and yield are often positively correlated (Hartwig and Hinson, 1972), but in this experiment, the correlation was nearly zero. The strength and direction of correlations in this set of lines may have been affected by several factors, including the genotypes used as parents for mutant development, effects of mutations within the lines on agronomic performance, an overall decrease in yield in the mutant lines, generally higher protein content, and slightly decreased oil content. The mutants showed an overall decrease in yield relative to the parents. This is expected due to the increase in protein in many of the mutants relative to the parents. However, it is possible that background mutations are contributing to the yield decrease, since fast neutrons induce multiple random mutations per plant with unknown effects (Bolon et al. 2011). A decrease in yield is expected with an increase in protein content, but the negative relationship may potentially be mitigated with the removal of background mutations in a backcrossing or outcrossing scheme.

*Approaches for mutation identification and association with phenotypes of interest*

Comparative genomic hybridization (CGH) is a well-established method for detecting large CNV within the soybean genome (Bolon et al. 2011, 2014; Campbell et al. 2016; Dobbels et al. 2017). Despite the use of over one million probes spaced throughout the genome, it is

possible that CGH analysis does not detect smaller mutations that fall between probes. Additionally, the space between probes is not consistent, making PCR primer design more complicated around mutations detected in regions with large gaps between probes. In these cases, next generation sequencing can be used to provide more detailed information about mutations across the entire genome. Whole genome sequencing can be used to confirm mutations found in CGH, provide in-depth coverage of the entire genome, show exact breakpoints of mutations, and reveal smaller mutations missed by CGH.

Methods such as BSA can be used to identify the strongest candidate regions or mutations resulting in the phenotype of interest. Use of $F_2$ plants in the bulks should lead to large linkage blocks and simple detection of associations using the SoySNP 50K Infinium BeadChip. Once these candidate regions are identified through BSA, they can be confirmed through CGH and WGS. The sequencing data can also be used to design high-throughput genotyping markers. In this study, BSA results pointed to a large deletion on chromosome 12 that had an association with increased protein. The deletion was also identified in CGH and WGS, and the details in these analyses allowed for development of a KASP marker to confirm an association with protein in the $F_{2:3}$ population from a cross between Benning and the high-protein mutant, G15FN-12.

Initial genotyping of the Benning × G15FN-12 (high protein) $F_2$ population using regular PCR primers designed for the deletions revealed almost perfect cosegregation of the deletions on Chrs 12 and 17 revealed by CGH. The CGH probes were designed based on the Williams 82 reference genome (Bolon et al. 2011). The deletion detected on Chr 17 in CGH is highly homologous to a region of Chr 12 based on the newer Lee genome assembly that falls within the

confirmed Chr 12 deletion. Based on this information and genotyping results for the putative deletion, the chromosome 17 deletion was considered a false peak.

*Candidate genes for elevated protein content*

The Soybase database (Grant et al. 2010) lists several protein QTL on Chr 12. One is located in the 4-6 Mb region (Brummer et al. 1997; Liang et al. 2010). Another region around 35-39 Mb was identified by Lee et al. (1996), Qiu et al. (1999), Kabelka et al. (2004), and Eskandari et al. (2013). Lu et al. (2013) identified a QTL in the 6-12 Mb region. This QTL was identified in a population of 212 RILs grown in six environments. The Chr 12 QTL was found in all six environments with a LOD score of 3.9 and explained 4.5% of phenotypic variation. Teng et al. (2017) identified another protein QTL in the 9-15 Mb region of Chr 12 that was detected in seven environments and explained 3.3-11.3% of phenotypic variation across environments. This is the closest QTL to the deletion region found to be associated with altered protein content in G15FN-12. These QTL do not clearly overlap with the deletion in G15FN-12; however, they are in close proximity, and the QTL locations could be slightly different from the exact location due to gaps between markers used to construct linkage maps or phenotyping error.

The large Chr 12 deletion in G15FN-12 contains a total of 137 putative genes (Table 2.S1). Annotations for these genes including GO functions were from Soybase (https://soybase.org) and Phytozome (Goodstein et al. 2012). Additional data provided through the RNA-Seq Atlas provides quantified gene expression at young leaf, flowering, pod development, and post-flowering intervals, and in roots and nodules (Severin et al. 2010). Genes of interest were identified using a combination of putative function, GO annotation, and level of expression during flowering and seed development.

Genes of interest for high protein content within this deletion have putative functions including transcription factors, transporter activity, nucleotide binding, cell growth and auxin response, stress response, enzymatic activity (catabolic process, proteasome complex, and [endo] peptidase activity), cell growth/ATP binding/post-embryonic development, response to wounding, ubiquitin-related activity and embryo development, and organic solute transport. These candidate genes were identified using gene expression data provided by Severin et al. (2010) by identifying genes with high levels of expression post-flowering during seed development. Other candidate genes are putatively involved in protein phosphorylation, protein kinase activity, ATP binding, and carbohydrate binding. These genes are candidates based on their involvement in lectin, a moderately abundant protein in soybeans, or in protein kinase functions (Herman, 2014). Additional candidate genes are putatively involved in glycerol metabolism (Wilson, 2004), based on their role in fatty acid metabolism and the negative relationship between oil and protein. Due to the large number of genes within the Chr 12 deletion and the complexity of protein accumulation in seeds, narrowing down candidate genes for the elevated protein content is very difficult and will require substantial time and resource investment to pinpoint the causal gene(s). Filtering to a handful of candidate causal genes within the deletion could result in the use of gene editing to validate the role of the gene in plant phenotype.

*Implications of using these FN mutants in breeding*

The mutant lines developed in this program can be used as sources of desirable seed composition traits in conventional breeding schemes. Use of these mutants in a backcrossing system would remove random background mutations and move the causal mutation into an elite

background. The removal of background mutations could result in mitigation of the yield drag observed in mutant line yield data. However, if improved seed composition traits are caused by large deletions, these deletions may contain genes important for yield and agronomic traits, lessening the functionality of mutants as sources of desirable traits for breeding. In addition, yield will likely still be affected if breeding for increased protein due to the biologically expensive nature of seed protein production.

Validation of a candidate gene function and advancement of gene-editing capabilities in soybean could lead to direct deletion of the causal gene in elite lines. Conventional mutagenesis induces multiple mutations at random locations throughout the genome, whereas gene editing technologies are much more targeted to the region of interest with far fewer off-target effects (Van de Wiel et al. 2017). Many mutations found in fast neutron mutants are large and contain multiple genes. Loss of function of these genes may contribute to poor performance of the mutant line, making it an unsuitable breeding line. Gene editing would be a much more targeted method of achieving the mutant phenotype without the concomitant yield drag or poor agronomic traits associated with the original mutation. Gene editing could also result in shortened development time of soybean varieties with increased protein in elite backgrounds, depending on regulation of gene-editing techniques.

The techniques used in these studies are applicable in other crops with high quality genomics data available. Array CGH has been used in Arabidopsis (DeBolt, 2010), maize (Springer et al. 2009), rice (Yu et al. 2011), and barley (Muñoz-Amatriaín et al. 2013). Advancement of array-based and sequencing technologies will continue to enable the association of mutations with altered plant phenotypes, potentially leading to the use of gene-editing technologies for replication of phenotypes of interest. Based on the results from this study, gene

editing technology (CRISPR-Cas) may be used in the future to create high protein lines in other soybean maturity groups, and identification of mutants with desirable characteristics will allow for their use in breeding programs for development of improved varieties.

**Acknowledgements**

We thank the United Soybean Board for funding this research. Thanks to Dale Wood, Earl Baxter, Brice Wilson, Jeremy Nation, Greg Gokalp, Ricky Zoller, and Tatyana Nienow for technical support. Thanks to Blair Buckley for growing the Bossier, LA yield trial.

**References:**

Anderson JE, Michno J-M, Kono TJY, Stec AO, Campbell BW, Curtin SJ, Stupar RM (2016) Genomic variation and DNA repair associated with soybean transgenesis: a comparison to cultivars and mutagenized plants. BMC Biotechnol 16:41-41

Andrews S (2014) FastQC A Quality Control tool for High Throughput Sequence Data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

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

Barthelson R (2014a) HTProcess_trimmomatic_0.32. rogerab. CyVerse https://de.cyverse.org/de/?type=apps&app-id=93820714-89b7-4008-a9d1-98cb35ae3b09&system-id=de

Barthelson R (2014b) HTProcess-prepare_directories-and-run_fastqc-0.1. rogerab. CyVerse. https://de.cyverse.org/de/?type=apps&app-id=b4d15ab5-81b6-47e1-8637-26824e1f1679&system-id=de

Barthelson R (2016) SAM-to-sorted-BAM. rogerab. CyVerse. https://pods.iplantcollaborative.org/wiki/display/DEapps/Convert+SAM-to-sorted-BAM

Belfield E, Gan X, Mithani A, Brown C, Jiang C, Franklin K, Alvey E, Wibowo A, Jung M, Bailey K, Kalwani S, Ragoussis J, Mott R, Harberd N (2012) Genome-wide analysis of mutations in mutant lineages selected following fast-neutron irradiation mutagenesis of *Arabidopsis thaliana*. Genome Res 22:1306-1315

Boerma HR, Hussey RS, Phillips DV, Wood ED, Rowan GB, Finnerty SL (1997) Registration of 'Benning' soybean. Crop Sci 37:1982

Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinform 30(15):2114-2120

Bolon Y-T, Haun WJ, Xu WW, Grant D, Stacey MG, Nelson RT, Gerhardt DJ, Jeddeloh JA, Stacey G, Muehlbauer GJ, Orf JH, Naeve SL, Stupar RM, Vance CP (2011) Phenotypic and genomic analyses of a fast neutron mutant population resource in soybean. Plant Physiol 156:240-253

Bolon Y-T, Stec AO, Michno J-M, Roessler J, Bhaskar PB, Ries L, Dobbels AA, Campbell BW, Young NP, Anderson JE, Grant DM, Orf JH, Naeve SL, Muehlbauer GJ, Vance CP, Stupar RM (2014) Genome resilience and prevalence of segmental duplications following fast neutron irradiation of soybean. Genetics 198:967-981

Brummer EC, Graef GL, Orf J, Wilcox JR, Shoemaker RC (1997) Mapping QTL for seed protein and oil content in eight soybean populations. Crop Sci 37:370-378

Brzostowski LF, Diers BW (2017) Agronomic evaluation of a high protein allele from PI407788A on chromosome 15 across two soybean backgrounds. Crop Sci 57:2972-2978

Brzostowski LF, Pruski T, Specht JE, Diers BW (2017) Impact of seed protein alleles from three soybean sources on seed composition and agronomic traits. Theor Appl Genet 130:2315-2326

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

Campbell BW, Hofstad AN, Sreekanta S, Fu F, Kono TJY, O'Rourke JA, Vance CP, Muehlbauer GJ, Stupar RM (2016) Fast neutron-induced structural rearrangements at a soybean *NAP1* locus result in gnarled trichomes. Theor Appl Genet 129:1725-1738

Chung J, Babka HL, Graef GL, Staswick PE, Lee DJ (2003) The seed protein, oil, and yield QTL on soybean linkage group I. Crop Sci 43:1053-1067

Cober ER, Voldeng HD (2000) Developing high-protein, high-yield soybean populations and lines. Crop Science 40:39-42

Cooksey A (2017) BWA mem_longreads-0.7.15. Upendra Devisetty. CyVerse. https://de.cyverse.org/de/?type=apps&app-id=e813971a-7d2d-11e7-9247-008cfa5ae621&system-id=de

DeBolt S (2010) Copy number variation shapes genome diversity in Arabidopsis over immediate family generational scales. Genome Biol Evol 2:441–453

Devisetty UK (2016a) BWA mem 0.7.15. Upendra Devisetty. CyVerse. https://de.cyverse.org/de/?type=apps&app-id=4c7e942e-6408-11e6-a1f5-008cfa5ae621&system-id=de

Devisetty UK (2016b) SAM_to_Sorted_BAM-0.1.19 (app for workflows). Upendra Kumar Devisetty. CyVerse. https://de.cyverse.org/de/?type=apps&app-id=9b837c54-122f-11e6-98c7-dfa2ad9d9ddd&system-id=de

Diers BW, Keim P, Fehr WR, Shoemaker RC (1992) RFLP analysis of soybean seed protein and oil content. Theor Appl Genet 83:608-612

Diers BW, Shoemaker RC (1992) Restriction-fragment-length-polymorphism analysis of soybean fatty-acid content. J Am Oil Chem Soc 69:1242-1244

Dobbels AA, Michno J-M, Campbell BW, Virdi KS, Stec AO, Muehlbauer GJ, Naeve SL, Stupar RM (2017) An induced chromosomal translocation in soybean disrupts a KASI ortholog and is associated with a high-sucrose and low-oil seed phenotype. G3 Genes Genomes Genet 7:1215-1223

Eskandari M, Cober E, Rajcan I (2013) Genetic control of soybean seed oil: II. QTL and genes that increase oil concentration without decreasing protein or with increased seed yield. Theor Appl Genet 126(6): 1677-1687

Fehr WR, Caviness CE (1977) Stages of soybean development. Special Report. 87. https://lib.dr.iastate.edu/specialreports/87/

Goodstein DM., Shu S, Howson R, Neupane R, Hayes RD, Fazo J, et al (2012) Phytozome: a comparative platform for green plant genomics. Nucleic Acids Res 40(Database issue): D1178–D1186. http://doi.org/10.1093/nar/gkr944

Grant D, Nelson RT, Cannon SB, Shoemaker RC (2010) SoyBase, the USDA-ARS soybean genetics and genomics database. Nucleic Acids Res 38 (suppl 1): D843-D846. doi: 10.1093/nar/gkp798.

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

Herman EM (2014) Soybean seed proteome rebalancing. Front Plant Sci 5(437):1-8

Hwang E-Y, Qijian S, Jia G, Specht JE, Hyten DL, Costa J, Cregan PB (2014) A genome-wide association study of seed protein and oil content in soybean. BMC Genom 15:1-25

Hwang W, Kim M, Kang Y, Shim S, Stacey M, Stacey G, Lee S-H (2015) Genome-wide analysis of mutations in a dwarf soybean mutant induced by fast neutron bombardment. Euphytica 203:399-408

Hymowitz T, Collins FI, Walker WM, Panczner J (1972) Relationship between content of oil, protein, and sugar in soybean seed. Agron J 64:613

Jaureguy LM, Chen P, Scaboo AM (2011) Heritability and correlations among food-grade traits in soybean. Plant Breed 130:647-652

Kabelka EA, Diers BW, Fehr WR, LeRoy AR, Baianu IC, You T, Neece DJ, Nelson RL (2004) Putative alleles for increased yield from soybean plant introductions. Crop Sci 44(3): 784-791

Keim P, Olson TC, Shoemaker RC (1988) A rapid protocol for isolating soybean DNA. Soybean Genet Newsl 18: 150–152

Kim HK, Kang ST, Oh KW (2006) Mapping of putative quantitative trait loci controlling the total oligosaccharide and sucrose content of Glycine max seeds. J Plant Res 119:533-538

Kim M, Schultz S, Nelson RL, Diers BW (2016) Identification and fine mapping of a soybean seed protein QTL from PI 407788A on chromosome 15. Crop Sci 56(1):219-225. DOI: 10.2135/cropsci2015.06.0340

Kuznetsova A, Brockhoff PB Christensen RHB (2016). lmerTest: Tests in linear mixed effects models. R package version 2.0-33.https://CRAN.R-project.org/package=lmerTest

Kwanyuen P, Pantalone VR, Burton JW, Wilson RF (1997) A new approach to genetic alteration of soybean protein composition and quality. J Am Oil Chem Soc 74:983-987

Lee SH, Bailey MA, Mian MAR, Carter, Jr. TE, Shipe ER, Ashley DA, Parrott WA, Hussey RS, Boerma HR (1996) RFLP loci associated with soybean seed protein and oil content across populations and locations. Theor Appl Genet 93(5-6): 649-657

Li H (2011) A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinform 27(21); 2987-2993

Li H (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:1303.3997v2 (q-bio.GN).

Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler Transform. Bioinform 25:1754-60

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup (2009) The Sequence alignment/map (SAM) format and SAMtools. Bioinform 25:2078-2079

Liang H, Yu Y, Wang S, Lian Y, Wang T, Wei Y, Gong P, Liu X, Fang X, Zhang M (2010) QTL Mapping of isoflavone, oil and protein contents in soybean (*Glycine max* L. Merr.). Ag Sci China 9(8): 1108-1116

Lu W, Wen Z, Li H, Yuan D, Li J, Zhang H, Huang Z, Cui S, Du W (2013) Identification of the quantitative trait loci (QTL) underlying water soluble protein content in soybean. Theor Appl Genet 126:425-433

Martin M (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal, 17(1):10-12. DOI: http://dx.doi.org/10.14806/ej.17.1.200

Mian R, McHale L, Li Z, Dorrance AE (2017) Registration of 'HighPro1' soybean with high protein and high yield developed from a north x south cross. J Plant Reg 11: 51-54

Muñoz-Amatriaín M, Eichten SR, Wicker T, Richmond TA, Mascher M, Steuernagel B, Scholz U, Ariyadasa R, Spannagl M, Nussbaumer T, Mayer KF, Taudien S, Platzer M, Jeddeloh JA, Springer JM, Muehlbauer GJ, Stein N (2013) Distribution, functional impact, and origin mechanisms of copy number variation in the barley genome. Genome Biol 14(6), R58. http://doi.org/10.1186/gb-2013-14-6-r58

Nichols DM, Glover KD, Carlson SR, Specht JE, Diers BW (2006) Fine mapping of a seed protein QTL on soybean linkage group I and its correlated effects on agronomic traits. Crop Sci 46:834-839

Patil G, Mian R, Vuong T, Pantalone V, Song Q, Chen P, Shannon G, Carter TE, Nguyen H (2017) Molecular mapping and genomics of soybean seed protein: a review and perspective for the future. Theor Appl Genet 130:1975-1991

Pham AT, McNally K, Abdel-Haleem H, Boerma HR, Li Z (2013) Fine mapping and identification of candidate genes controlling the resistance to southern root-knot nematode in PI 96354. Theor Appl Genet 126(7):1825-1838

Phansak P, Soonsuwon W, Hyten DL, Song Q, Cregan PB, Graef GL, Specht JE (2016) Multi-Population Selective Genotyping to Identify Soybean [Glycine max (L.) Merr.] Seed Protein and Oil QTLs. G3 Genes Genomes Genet 6:1635-1648

Prenger EM, Mian R, Buckley B, Boerma HR, Li Z (2018). Introgression of a high protein allele into an elite soybean variety results in a high-protein near-isogenic line with yield parity. Chapter 3. MS Thesis, University of Georgia.

Qi Z, Wu Q, Han X, Sun Y-n, Du X-y, Liu C-y, Jiang H-w, Hu G-h, Chen Q-s (2011a) Soybean oil content QTL mapping and integrating with meta-analysis method for mining genes. Euphytica 179:499-514

Qi Z, Ya-nan S, Qiong W, Chun-yan L, Guo-hua H, Qing-shan C (2011b) A meta-analysis of seed protein concentration QTL in soybean. Can J Plant Sci 91:221-230

Qiu BX, Arelli PR, Sleper DA (1999) RFLP markers associated with soybean cyst nematode resistance and seed composition in a 'Peking' x 'Essex' population. Theor Appl Genet 98(3-4): 356-364.

R Core Team (2013) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/

Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP (2011) Integrative Genomics Viewer. Nat Biotechnol 29, 24–26

Schmutz et al. (2010) Genome sequence of the palaeopolyploid soybean. Nat 463:178-183

Sebolt AM, Shoemaker RC, Diers BW (2000) Analysis of a quantitative trait locus allele from wild soybean that increases seed protein concentration in soybean. Crop Sci 40:1438-1444

Severin AJ, Woody JL, Bolon Y-T, Joseph B, Diers BW, Farmer AD, Muehlbauer GJ, Nelson RT, Grant D, Specht JE, Graham MA, Cannon SB, May GD, Vance CP, Shoemaker RC (2010) RNA-Seq Atlas of Glycine max: A guide to the soybean transcriptome. BMC Plant Biol 10:160

Song Q, Hyten DL, Jia G, Quigley CV, Fickus EW, Nelson RL, Cregan PB (2013) Development and Evaluation of SoySNP50K, a High-Density Genotyping Array for Soybean. PLoS ONE 8:e54985

Springer NM, Ying K, Fu Y, Ji T, Yeh C-T, Jia Y, Wu W, Richmond T, Kitzman J, Rosenbaum H, Iniguez AL, Barbazuk WB, Jeddeloh JA, Nettleton D, Schnable PS (2009) Maize inbreds exhibit high levels of copy number variation (CNV) and presence/absence variation (PAV) in genome content. PLoS Genet 5(11): e1000734. https://doi.org/10.1371/journal.pgen.1000734.

Stacey MG, Cahoon RE, Nguyen HT, Cui Y, Sato S, Nguyen CT, Phoka N, Clark KM, Liang Y, Forrester J, Batek J, Do PT, Sleper DA, Clemente TE, Cahoon EB, Stacey G (2016) Identification of Homogentisate Dioxygenase as a Target for Vitamin E Biofortification in Oilseeds. Plant Physiol 172:1506

Teng, W, Li W, Zhang Q, Wu D, Zhao X, Li H, Han Y, Li W (2017) Identification of quantitative trait loci underlying seed protein content of soybean including main, epistatic, and QTL × environment effects in different regions of Northeast China. Genome 60(8): 649-655. Doi: 10.1139/gen-2016-0189.

Thorne JC, Fehr WR (1970) Incorporation of High-Protein, Exotic Germplasm into Soybean Populations by 2- and 3-way Crosses1. Crop Sci 10:652-655

Thorvaldsdóttir H, Robinson JT, Mesirov JP (2013) Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief in Bioinform 14:178-192

van de Wiel CCM, Schaart JG, Lotz LAP, Smulders MJM (2017) New traits in crops produced by genome editing techniques based on deletions. Plant Biotechnol Rep 11:1-8. Doi 10.1007/s11816-017-0425-z

Vaughn JN, Nelson RL, Song Q, Cregan PB, Li Z (2014) The Genetic Architecture of Seed Composition in Soybean Is Refined by Genome-Wide Association Scans Across Multiple Populations. G3 Genes Genomes Genet 4:2283-2294

Vaughn M (2012) Uncompress files with gunzip. CyVerse http://www.gnu.org/software/gzip/

Walls R (2018) Concatenate multiple files. BusyBox. CyVerse. https://de.cyverse.org/de/?type=apps&app-id=77830f32-084a-11e8-a871-008cfa5ae621&system-id=de

Warrington CV, Abdel-Haleem H, Hyten DL, Cregan PB, Orf JH, Killam AS, Bajjalieh N, Li Z, Boerma HR (2015) QTL for seed protein and amino acids in the Benning × Danbaekkong soybean population. Theor Appl Genet 128:839-850

Wilcox JR, Cavins JF (1995) Backcrossing high seed protein to a soybean cultivar. Crop Sci 35:1036-1041

Wilcox JR, Zhang GD (1997) Relationships between seed yield and seed protein in determinate and indeterminate soybean populations. Crop Sci 37:361-364

Wilson RF (2004) Seed composition. In: Boerma HR, Specht JE (eds) Soybean: Improvement, production and uses, 3rd edn. ASA and CSSA, Madison, WI, pp 671-677

Yin XH, Vyn TJ (2005) Relationships of isoflavone, oil, and protein in seed with yield of soybean. Agron J 97:1314-1321

Yu P, Wang C, Xu Q, Feng Y, Yuan X, Yu H, Wang Y, Tang S, Wei X (2011) Detection of copy numer variations in rice using array-based comparative genomic hybridization. BMC Genom 12:372. https://doi.org/10.1186/1471-2164-12-372

**Table 2.1:** Protein and oil of the selected mutant lines for comparative genomic hybridization and their parents. A) Protein and oil values on a whole-plot basis. B) Protein and oil values on a single plant selected as the CGH seed source, and on a whole-plot basis for the high protein mutant, G15FN-12.

A)

| Name | Line Type | Location | Near-Infrared (NIR) | | Wet chemistry | |
|---|---|---|---|---|---|---|
| | | | Protein | Oil | Protein | Oil |
| | | | ----------g kg$^{-1}$---------- | | | |
| G15FN-12 | Mutant | Watkinsville, GA | 460.9 | 181.1 | 447.6 | 203.0 |
| G15FN-23 | Mutant | Watkinsville, GA | 461.1 | 180.8 | 456.2 | 198.9 |
| G00-3213* | Parent | Watkinsville, GA | 409.2 | 205.9 | 385.2 | 200.1 |
| G15FN-54 | Mutant | Plains, GA | 431.5 | 118.8 | 421.7 | 145.2 |
| G00-3213* | Parent | Plains, GA | 421.9 | 191.1 | 395.3 | 180.8 |
| G15FN-109 | Mutant | Watkinsville, GA | 435.1 | 179.3 | 440.1 | 203.3 |
| G00-3880* | Parent | Watkinsville, GA | 395.4 | 204.5 | 370.0 | 200.1 |

Note: An asterisk (*) indicates a parent control. Protein and oil are reported on a dry matter basis.

B)

| Sample Type | Name | Line Type | Location | NIR | | Wet chemistry | |
|---|---|---|---|---|---|---|---|
| | | | | Protein | Oil | Protein | Oil |
| | | | | ----------g kg$^{-1}$---------- | | | |
| Single plant | G15FN-12 | Mutant | Watkinsville, GA | 470.7 | 186.1 | - | - |
| | G00-3213 | Parent | Watkinsville, GA | 410.3 | 205.8 | - | - |
| Whole plot | G15FN-12 | Mutant | Watkinsville, GA | 460.9 | 181.1 | 447.6 | 203.0 |
| | G00-3213 | Parent | Watkinsville, GA | 409.2 | 205.9 | 385.2 | 200.1 |

Note: The single plant values for G15FN-12 are from the CGH analysis seed source.

**Table 2.2:** Summary of 231 mutant lines derived from G00-3213 and G00-3880 based on augmented design analysis across four locations in 2016.

| Background | Trait | Number of lines | Range of mutant lines ----------g $kg^{-1}$---------- | Parent estimate ------g $kg^{-1}$------ |
|---|---|---|---|---|
| G00-3213 | Protein | 90 | 408.5 - 431.2 | 412.9 |
| | Oil | | 112.2 - 201.8 | 195.9 |
| | Sucrose | | 55.4 - 94.3 | 67.2 |
| G00-3880 | Protein | 141 | 369.8 - 426.4 | 394.4 |
| | Oil | | 174.4 - 210.8 | 197.6 |
| | Sucrose | | 52.9 - 72.3 | 65.3 |

Note: Protein, oil, and sucrose contents are reported in g $kg^{-1}$ on a dry matter basis.

**Table 2.3:** Size and location of large deletions found in both CGH and whole genome sequencing of mutant line G15FN-12.

| Mutant | Chromosome | Williams 82 | | Lee | |
|---|---|---|---|---|---|
| | | Location | Size | Location | Size |
| G15FN-12 High protein | Gm10 | 51,204,213-51,204,229 | 17 bp | 53,362,526-53,362,542 | 17 bp |
| | Gm12 | 15,730,575-21,156,836 | 5.4 Mb | 15,971,575-21,637,281 | 5.7 Mb |
| | Gm17 | 30,588,869-30,592,911 | 4 Kb | Not found | |

**Table 2.4:** Average seed composition values for chromosome 12 deletion genotypes in the $F_2$ Benning × G15FN-12 population.

| Trait | Homozygous deletion | Heterozygous | Homozygous wild type | *P*-value |
|---|---|---|---|---|
| Protein (g $kg^{-1}$) | 449.2 | 429.9 | 422.1 | <0.0001 |
| Oil (g $kg^{-1}$) | 185.1 | 189.4 | 193.5 | 0.0001 |
| Number of $F_2$ families | 34 | 129 | 91 | - |
| Sucrose (g $kg^{-1}$) | 55.6 | 61.5 | 62.3 | 0.0004 |
| Number of $F_2$ families | 26 | 110 | 76 | - |

Note: Sucrose values were excluded for smaller samples due to potential NIR inaccuracy.

**Table 2.5:** Performance of mutants in yield and agronomic traits across five locations in 2017.

| Genotype | Description | Yield kg ha$^{-1}$ | Protein g kg$^{-1}$[a] | Oil g kg$^{-1}$[a] | Sucrose g kg$^{-1}$ | Maturity d past Aug. 31 | Height cm | Lodging score[b] | Seed Size g 100-seed$^{-1}$ | Seed Quality score[c] |
|---|---|---|---|---|---|---|---|---|---|---|
| G15FN-5 | mutant | 3080.9 | 439.9 | 184.3 | 69.3 | 51.2 | 98.2 | 2.6 | 15.3 | 1.3 |
| G15FN-12[d] | mutant | 3000.7 | 454.8*** | 180.9 | 62.4*** | 51.1 | 87.2 | 1.9 | 17.3** | 1.4 |
| G15FN-14 | mutant | 2954.9 | 437.1 | 186.0 | 66.9 | 51.2 | 93.6 | 2.1 | 15.0 | 1.3 |
| G15FN-16 | mutant | 2912.6 | 430.3 | 186.7 | 69.5 | 50.9 | 90.8 | 2.1 | 15.9 | 1.4 |
| G15FN-23[d] | mutant | 3013.3 | 443.7 | 183.9 | 67.1 | 52.1 | 93.4 | 2.3 | 16.6 | 1.4 |
| G15FN-35 | mutant | 3074.7 | 440.5 | 186.5 | 65.5 | 52.1 | 99.9 | 2.0 | 15.9 | 1.4 |
| G15FN-37 | mutant | 3028.5 | 431.1 | 187.2 | 70.7 | 52.6 | 99.6 | 2.4 | 16.0 | 1.3 |
| G15FN-52 | mutant | 3184.5 | 438.0 | 186.2 | 69.1 | 50.4 | 90.6 | 2.2 | 16.1 | 1.4 |
| G15FN-54[d] | mutant | 2826.8 | 430.2 | 108.6*** | 108.9*** | 53.6 | 87.8 | 2.3 | 15.6 | 1.8 |
| G15FN-67 | mutant | 2744.3 | 444.8* | 180.5 | 68.2 | 50.9 | 94.3 | 2.9 | 14.8 | 1.4 |
| G15FN-75 | mutant | 3184.5 | 430.9 | 185.9 | 71.1 | 50.7 | 94.8 | 2.1 | 15.8 | 1.3 |
| G15FN-80 | mutant | 3210.5 | 441.6 | 182.3 | 67.3 | 51.4 | 88.3 | 2.6 | 15.3 | 1.3 |
| **G00-3213** | **parent** | **3392.1** | **430.5** | **187.7** | **71.8** | **50.0** | **98.5** | **1.9** | **15.4** | **1.4** |
| G15FN-104 | mutant | 2899.4 | 421.8 | 181.2 | 70.3 | 53.8 | 96.2 | 2.2 | 14.1 | 1.4 |
| G15FN-109[d] | mutant | 2871.6* | 440.3*** | 169.7*** | 59.9*** | 56.6 | 99.6 | 2.6 | 14.3 | 1.9 |
| G15FN-122 | mutant | 3632.8 | 395.1** | 188.8 | 76.5 | 50.8 | 99.6 | 2.4 | 14.2 | 1.4 |
| G15FN-124 | mutant | 3288.1 | 427.6** | 183.0 | 67.2 | 49.8 | 107.8 | 3.0 | 15.8 | 1.6 |
| G15FN-141 | mutant | 3266.6 | 425.7* | 180.6 | 68.7 | 56.2 | 107.8 | 2.7 | 16.5 | 1.6 |
| G15FN-165 | mutant | 3343.7 | 421.4 | 174.1*** | 71.2 | 54.2 | 88.9 | 2.1 | 14.7 | 1.7 |
| G15FN-176 | mutant | 3233.4 | 432.9*** | 176.3** | 69.7 | 55.2 | 103.3 | 2.4 | 15.1 | 1.4 |
| G15FN-189 | mutant | 2793.3* | 433.8*** | 180.0 | 65.0 | 58.9** | 93.6 | 2.3 | 17.4** | 2.1 |
| G15FN-215 | mutant | 3146.4 | 439.9*** | 178.3 | 67.6 | 53.0 | 101.0 | 2.0 | 15.6 | 1.4 |
| G15FN-216 | mutant | 2584.6*** | 429.4*** | 176.6** | 65.5 | 57.5* | 92.9 | 1.9 | 16.6 | 1.5 |
| G15FN-244 | mutant | 3096.2 | 437.9*** | 181.3 | 60.5*** | 52.8 | 103.9 | 2.5 | 14.9 | 1.4 |
| **G00-3880** | **parent** | **3566.5** | **411.4** | **188.0** | **70.4** | **51.9** | **101.3** | **2.3** | **15.3** | **1.3** |
| AG7733 | commercial check | 3598.3 | 412.7 | 183.6 | 79.1 | 53.5 | 102.7 | 2.1 | 15.6 | 1.4 |
| AG7934 | commercial check | 3533.8 | 410.8 | 196.8 | 64.6 | 54.6 | 106.7 | 1.8 | 14.5 | 1.2 |

Note: Significance is based on Tukey's HSD ($\alpha$=0.05) with *=0.05, **=0.01, and ***≤0.001.

[a] Protein and oil contents are reported on a dry matter basis.

[b] Lodging score is from 1 to 5, with 1 indicating erect plants and 5 indicating plants were prostrate on the ground.

[c] Seed quality score is from 1 to 5, with 1 indicating few cracked, damaged, or diseased seeds and 5 indicating many poor quality seeds.

[d] Indicates CGH mutants

**Table 2.6:** Correlations among yield and seed composition traits in 2017 yield trials.

| | Yield | Protein | Oil |
|---|---|---|---|
| Protein | -0.27*** | | |
| Oil | -0.02$^{NS}$ | -0.37*** | |
| Sucrose | 0.36*** | -0.54*** | -0.07$^{NS}$ |

Note: *, **, and *** indicate significance at $P$=0.05, 0.01, and 0.001, respectively.

**Table 2.S1:** Glyma.12G gene models located within the chromosome 12 deletion.

| Gene model | Gene model length (bp) | Putative Function |
|---|---|---|
| Glyma.12g135400 | 2685 | PPR repeat family (PPR_2) |
| Glyma.12g135500 | 4664 | PPR repeat (PPR) // Pentatricopeptide repeat domain (PPR_3) |
| Glyma.12g135600 | 4922 | Leucine-rich repeat-containing protein |
| Glyma.12g135700 | 22698 | Protein C44H4.4 |
| Glyma.12g135800 | 165 | |
| Glyma.12g135900 | 547 | Gag-pol-related retrotransposon |
| Glyma.12g136000 | 7244 | Autophagy-related protein 18F |
| Glyma.12g136100 | 3313 | Galacturonosyltransferase 12-related |
| Glyma.12g136200 | 1311 | Domain of unknown function (DUF4283) |
| Glyma.12g136300* | 4071 | Transcription factor ILR3-related |
| Glyma.12g136400* | 1943 | Thioredoxin-like protein CXXS1 |
| Glyma.12g136500 | 301 | |
| Glyma.12g136600 | 6322 | Kinesin family member C1 (KIFC1) |
| Glyma.12g136700 | 9543 | Leucine-rich repeat-containing protein |
| Glyma.12g136800 | 7790 | Formin-like protein 12-related |
| Glyma.12g136900 | 1884 | |
| Glyma.12g137000 | 1558 | NAD dependent epimerase/dehydratase |
| Glyma.12g137100 | 16289 | Calcium/calmodulin-regulated receptor-like kinase |
| Glyma.12g137200 | 2729 | Kinase interacting (KIP1-like) family protein |
| Glyma.12g137300 | 554 | Domain of unknown function (DUF4283) |
| Glyma.12g137400 | 2094 | AAA ATPase |
| Glyma.12g137500 | 825 | Unconventional prefoldin RPB5 interactor 1 (URI1) |
| Glyma.12g137600 | 815 | Protein suppressor of PHYA-105 1 |
| Glyma.12g137700 | 3598 | GRAS domain family (GRAS) |
| Glyma.12g137800 | 2307 | Leucine-rich repeat-containing protein |
| Glyma.12g137900 | 9445 | Leucine Rich Repeat (LRR_3) // TIR domain (TIR_2) |
| Glyma.12g138000 | 597 | Androgen induced inhibitor of proliferation AS3 / PDS5-related |
| Glyma.12g138100 | 8119 | (-)-germacrene D synthase (GERD) |
| Glyma.12g138200 | 1680 | |
| Glyma.12g138300 | 243 | Auxin responsive protein (Auxin_inducible) |
| Glyma.12g138400 | 3426 | Leucine-rich repeat-containing protein |
| Glyma.12g138500 | 1147 | Leucine-rich repeat-containing protein |
| Glyma.12g138600 | 4829 | (-)-germacrene D synthase (GERD) |
| Glyma.12g138700 | 2229 | Protein NRT1/ PTR family 5.1 |
| Glyma.12g138800 | 8351 | Alpha-copaene synthase |
| Glyma.12g138900 | 7575 | |
| Glyma.12g139000* | 5392 | MFS transporter OCT family solute carrier family 22 member 4/5 |
| Glyma.12g139100 | 6082 | |
| Glyma.12g139200* | 3890 | RNA recognition motif. (a.k.a. RRM RBD or RNP domain) (RRM_1) |

| | | |
|---|---|---|
| Glyma.12g139300 | 1140 | Beta-glucosidase 41-related |
| Glyma.12g139400 | 347 | 60S ribosomal protein L7 |
| Glyma.12g139500 | 429 | Beta-glucosidase 41-related |
| Glyma.12g139600 | 5449 | mRNA capping enzyme (Pox_MCEL) |
| Glyma.12g139700* | 4515 | |
| Glyma.12g139800 | 3235 | |
| Glyma.12g139900* | 839 | 2-C-methyl-D-erythritol 2,4-cyclodiphosphate synthase / MECDP-synthase |
| Glyma.12g140000 | 648 | RNA helicase |
| Glyma.12g140100* | 2179 | Protein of unknown function DUF538 |
| Glyma.12g140200* | 3621 | S-locus glycoprotein domain (S_locus_glycop) // D-mannose binding lectin (B_lectin) // Protein tyrosine kinase (Pkinase_Tyr) // PAN-like domain (PAN_2) |
| Glyma.12g140300* | 3493 | S-locus glycoprotein domain (S_locus_glycop) // D-mannose binding lectin (B_lectin) // Protein tyrosine kinase (Pkinase_Tyr) // PAN-like domain (PAN_2) |
| Glyma.12g140400* | 2962 | S-locus glycoprotein domain (S_locus_glycop) // Protein tyrosine kinase (Pkinase_Tyr) // PAN-like domain (PAN_2) |
| Glyma.12g140500* | 1951 | D-mannose binding lectin (B_lectin) |
| Glyma.12g140600 | 7493 | (-)-germacrene D synthase (GERD) |
| Glyma.12g140700 | 894 | Peptidase C1-like family (Peptidase_C1_2) |
| Glyma.12g140800* | 5971 | S-locus glycoprotein domain (S_locus_glycop) // D-mannose binding lectin (B_lectin) // Protein tyrosine kinase (Pkinase_Tyr) // PAN-like domain (PAN_2) |
| Glyma.12g140900 | 4279 | PIF1-like helicase (PIF1) // Helitron helicase-like domain at N-terminus (Helitron_like_N) |
| Glyma.12g141000* | 3139 | Indole-3-acetic acid-amido synthetase GH3.5-related |
| Glyma.12g141100 | 8781 | WD40 repeat protein |
| Glyma.12g141200 | 692 | |
| Glyma.12g141300 | 439 | DnaJ domain (DnaJ) |
| Glyma.12g141400 | 3387 | Protein DRE2 required for cell viability |
| Glyma.12g141500 | 1129 | Flavodoxin-like quinone reductase 1 |
| Glyma.12g141600 | 3581 | Aldo/keto reductase |
| Glyma.12g141700* | 4511 | Transcription factor BHLH123 |
| Glyma.12g141800 | 884 | Uracil phosphoribosyltransferase / UMP pyrophosphorylase // Uridine kinase / Uridine monophosphokinase |
| Glyma.12g141900 | 2177 | NAD(P)H dehydrogenase (quinone) (wrbA) |
| Glyma.12g142000 | 3771 | |
| Glyma.12g142100* | 4356 | S-locus glycoprotein domain (S_locus_glycop) // D-mannose binding lectin (B_lectin) // Protein tyrosine kinase (Pkinase_Tyr) // PAN-like domain (PAN_2) |
| Glyma.12g142200* | 3823 | Inactive G-type lectin S-receptor-like serine/threonine-protein kinase SRK-related |
| Glyma.12g142300* | 3566 | Inactive G-type lectin S-receptor-like serine/threonine-protein kinase SRK-related |
| Glyma.12g142400* | 3515 | Inactive G-type lectin S-receptor-like serine/threonine-protein kinase SRK-related |
| Glyma.12g142500 | 447 | OS03G0366900 protein |
| Glyma.12g142600 | 1463 | Defensin-like protein 155-related |

| | | |
|---|---|---|
| Glyma.12g142700 | 5201 | AMP deaminase |
| Glyma.12g142800 | 3684 | Cysteine-rich receptor-like protein kinase 27-related |
| Glyma.12g142900* | 1203 | Calcium binding protein |
| Glyma.12g143000* | 4042 | S-locus glycoprotein domain (S_locus_glycop) // D-mannose binding lectin (B_lectin) // Protein tyrosine kinase (Pkinase_Tyr) // PAN-like domain (PAN_2) |
| Glyma.12g143100 | 1090 | Beta-13-N-acetylglucosaminyltransferase |
| Glyma.12g143200 | 4536 | Cysteine-rich receptor-like protein kinase 27-related |
| Glyma.12g143300* | 2624 | Protein kinase domain (Pkinase) // PAN-like domain (PAN_2) |
| Glyma.12g143400 | 1185 | NAD dependent epimerase/dehydratase |
| Glyma.12g143500 | 3437 | |
| Glyma.12g143600 | 1507 | Bucentaur related |
| Glyma.12g143700 | 1277 | |
| Glyma.12g143800 | 4142 | U3 small nucleolar ribonucleoprotein protein IMP4 |
| Glyma.12g143900 | 3599 | Cysteine-rich receptor-like protein kinase 27-related |
| Glyma.12g144000 | 986 | |
| Glyma.12g144100 | 4126 | Cysteine-rich receptor-like protein kinase 27-related |
| Glyma.12g144200* | 4991 | Protein tyrosine kinase (Pkinase_Tyr) // PAN-like domain (PAN_2) |
| Glyma.12g144300 | 1657 | Cysteine-rich receptor-like protein kinase 27-related |
| Glyma.12g144400 | 3989 | Cysteine-rich receptor-like protein kinase 27-related |
| Glyma.12g144500 | 17604 | Cysteine-rich receptor-like protein kinase 27-related |
| Glyma.12g144600 | 1152 | EMB |
| Glyma.12g144700 | 1903 | |
| Glyma.12g144800 | 3321 | Cysteine-rich receptor-like protein kinase 27-related |
| Glyma.12g144900 | 1567 | Calcium-transporting ATPase / Calcium-translocating P-type ATPase |
| Glyma.12g145000 | 440 | |
| Glyma.12g145100 | 1778 | No apical meristem (NAM) protein (NAM) |
| Glyma.12g145200 | 576 | Camp-response element binding protein-related |
| Glyma.12g145300 | 3377 | Cysteine-rich receptor-like protein kinase 27-related |
| Glyma.12g145400* | 4289 | S-locus glycoprotein domain (S_locus_glycop) // D-mannose binding lectin (B_lectin) // Protein tyrosine kinase (Pkinase_Tyr) // PAN-like domain (PAN_2) |
| Glyma.12g145500 | 459 | Myb/SANT-like DNA-binding domain (Myb_DNA-bind_3) |
| Glyma.12g145600 | 261 | |
| Glyma.12g145700 | 877 | |
| Glyma.12g145800 | 1266 | |
| Glyma.12g145900 | 916 | EF-hand calcium-binding domain containing protein |
| Glyma.12g146000 | 512 | S locus-related glycoprotein 1 binding pollen coat protein (SLR1-BP) |
| Glyma.12g146100* | 3966 | Universal stress protein family (Usp) |
| Glyma.12g146200 | 603 | Trehalose-phosphate phosphatase E-related |
| Glyma.12g146300* | 1054 | Long chain acyl-COA synthetase 2 |
| Glyma.12g146400 | 2359 | (+)-abscisic acid 8'-hydroxylase / ABA 8'-hydroxylase |
| Glyma.12g146500 | 2278 | |
| Glyma.12g146600* | 15211 | Ubiquitin carboxyl-terminal hydrolase 5/13 [EC:3.4.19.12] (USP5_13 UBP14) |

| | | |
|---|---|---|
| Glyma.12g146700* | 7786 | Diacylglycerol kinase 5-related |
| Glyma.12g146800 | 288 | FBD |
| Glyma.12g146900 | 2960 | PPR repeat domain-containing protein |
| Glyma.12g147000 | 18221 | Myosin-2 |
| Glyma.12g147100 | 1482 | Serine/arginine-rich splicing factor 2 |
| Glyma.12g147200* | 11418 | 20S proteasome subunit alpha 1 (PSMA6) |
| Glyma.12g147300 | 5123 | Cytochrome C-type biogenesis CCDA-like chloroplastic protein |
| Glyma.12g147400 | 6580 | Protein W09D10.1 |
| Glyma.12g147500 | 2784 | |
| Glyma.12g147600* | 2137 | |
| Glyma.12g147700* | 9833 | Organic solute transporter-related |
| Glyma.12g147800 | 168 | |
| Glyma.12g147900 | 6016 | Trans-sulfuration enzyme family member |
| Glyma.12g148000 | 3058 | |
| Glyma.12g148100 | 3093 | cyclin-dependent kinase 2 (CDK2) |
| Glyma.12g148200* | 3506 | Interleukin-1 receptor-associated kinase 4 protein-related |
| Glyma.12g148300 | 1138 | |
| Glyma.12g148400 | 5746 | PAP-specific phosphatase HAL2-like |
| Glyma.12g148500 | 1584 | Expressed protein |
| Glyma.12g148600 | 7163 | Peptidyl-prolyl cis-trans isomerase CYP37 chloroplastic |
| Glyma.12g148700 | 2487 | Beta-galactosidase 8 |
| Glyma.12g148800 | 1300 | |
| Glyma.12g148900 | 2920 | NAC domain containing protein 97 |
| Glyma.12g149000 | 390 | |

Note: Genes of interest for high protein are indicated with an asterisk (*). Putative functions are from the Phytomine feature of Phytozome. The first returned function is listed.

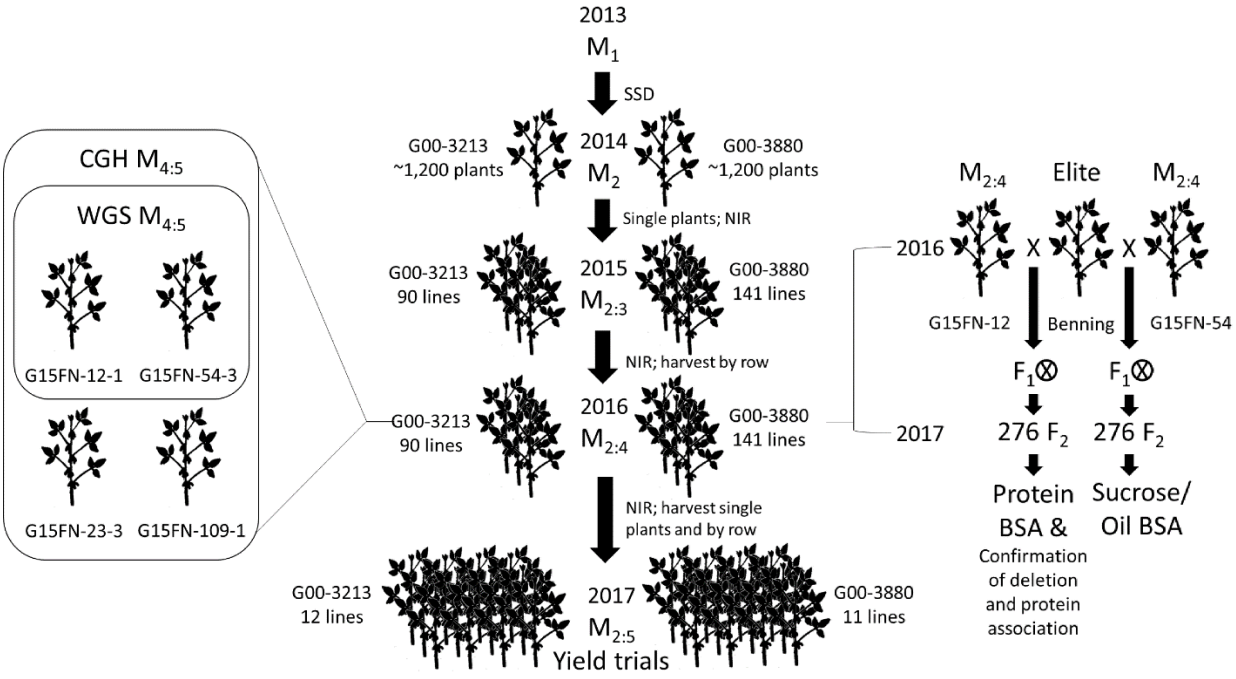


**Figure 2.1:** Scheme of selection and advancement of mutant lines

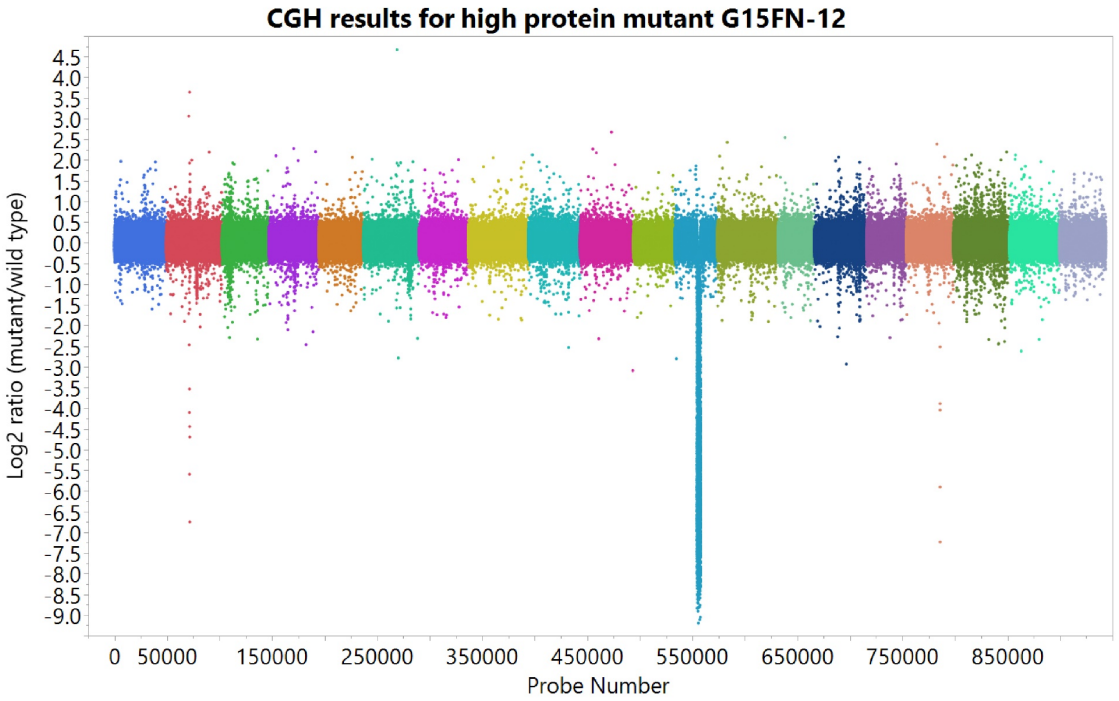


**Figure 2.2:** CGH results for mutant G15FN-12

Three negative peaks can be seen, indicating three putative large deletions. Each color indicates one of the 20 chromosomes of soybean. Chromosome 1 is on the left and 20 is on the right.

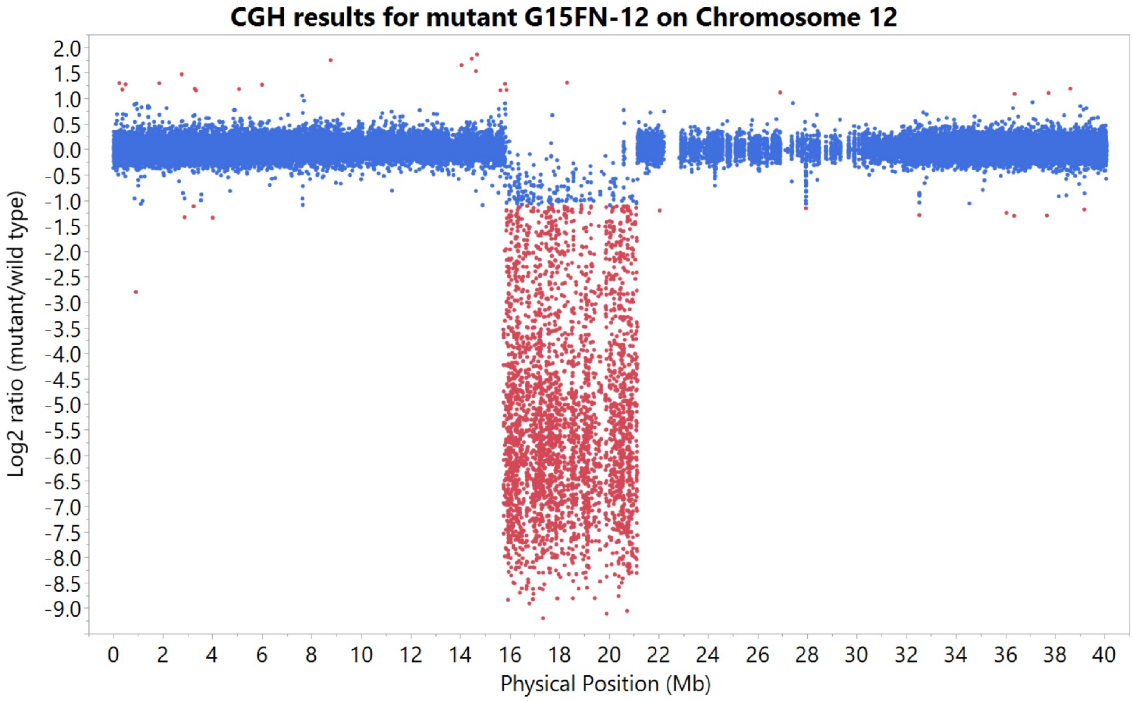


**Figure 2.3:** CGH results for mutant G15FN-12 on chromosome 12.

Each dot is a unique CGH probe. Red probes are significant, and blue probes are not significant. Significance is based on the average $\log_2$ hybridization ratio for mutant G15FN-12 plus or minus three standard deviations.

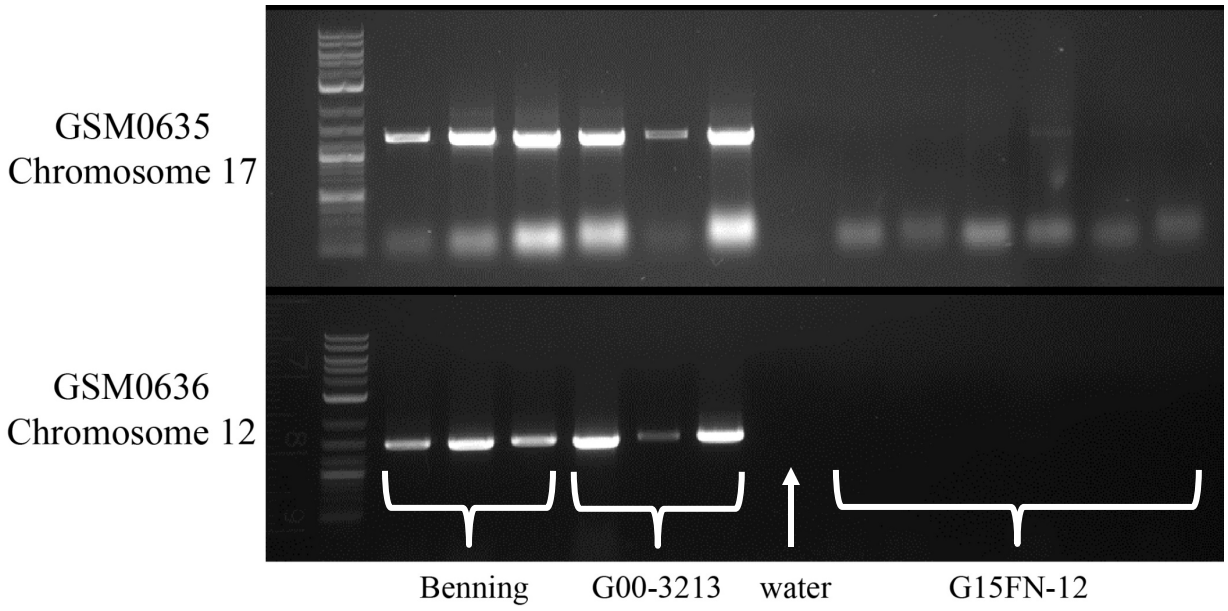


**Figure 2.4:** Confirmation of deletions with PCR. PCR primers were designed to target putative deletion regions. Primer sets were tested using wild type Benning, the mutant parent background G00-3213, and six G15FN-12 samples, for the deletions on chromosome 17 (top), and 12 (bottom). A 1 kb+ ladder was used for both GSM0635 and GSM0636.

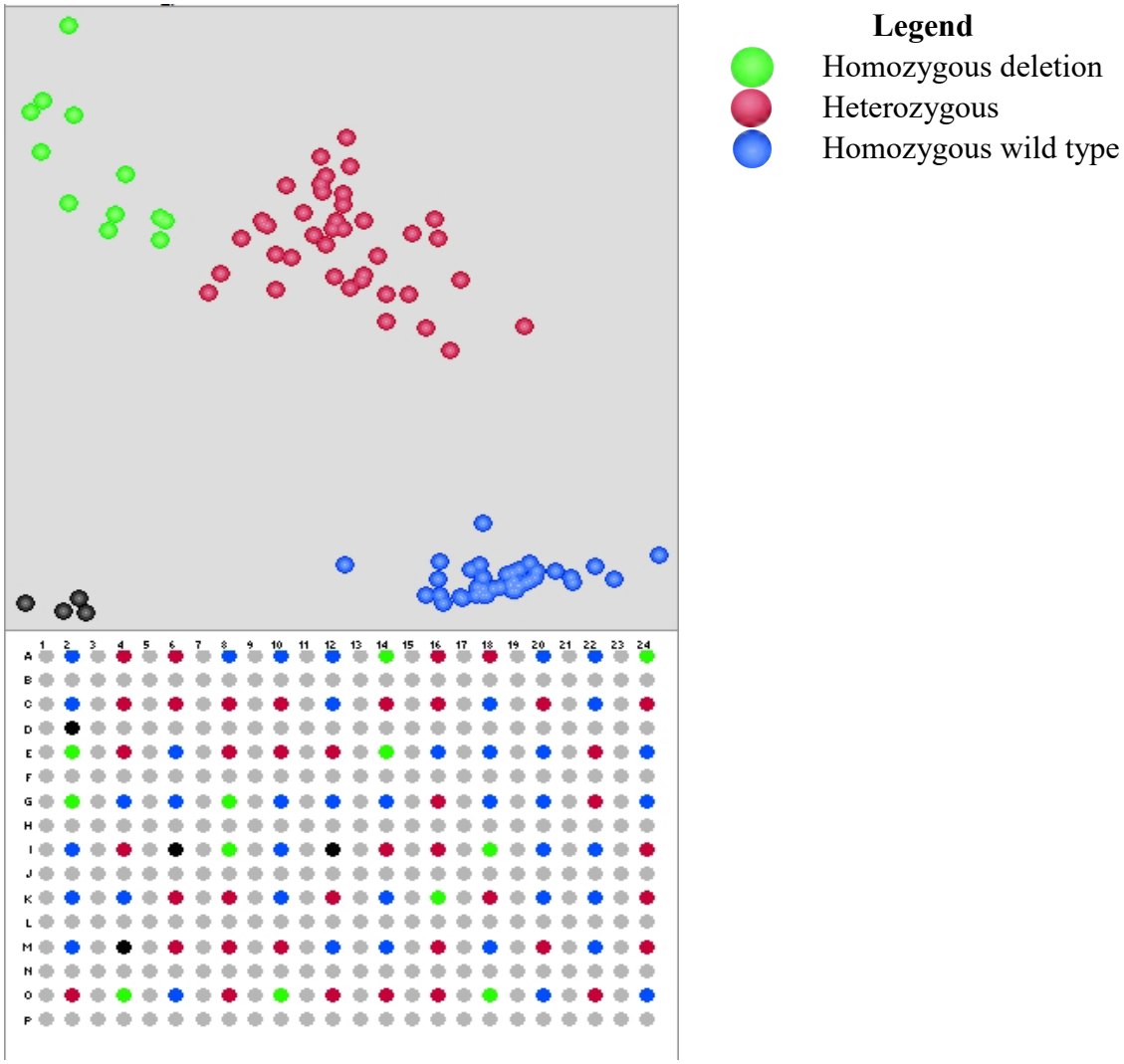


**Figure 2.5:** Genotyping of the Benning × G15FN-12 population using the KASP marker for the chromosome 12 deletion. Plates were read with a Tecan Infinite® M1000 Pro reader. Green dots indicate $F_2$ individuals homozygous for the deletion, red dots indicate heterozygotes, and blue dots indicate homozygous wild type individuals.

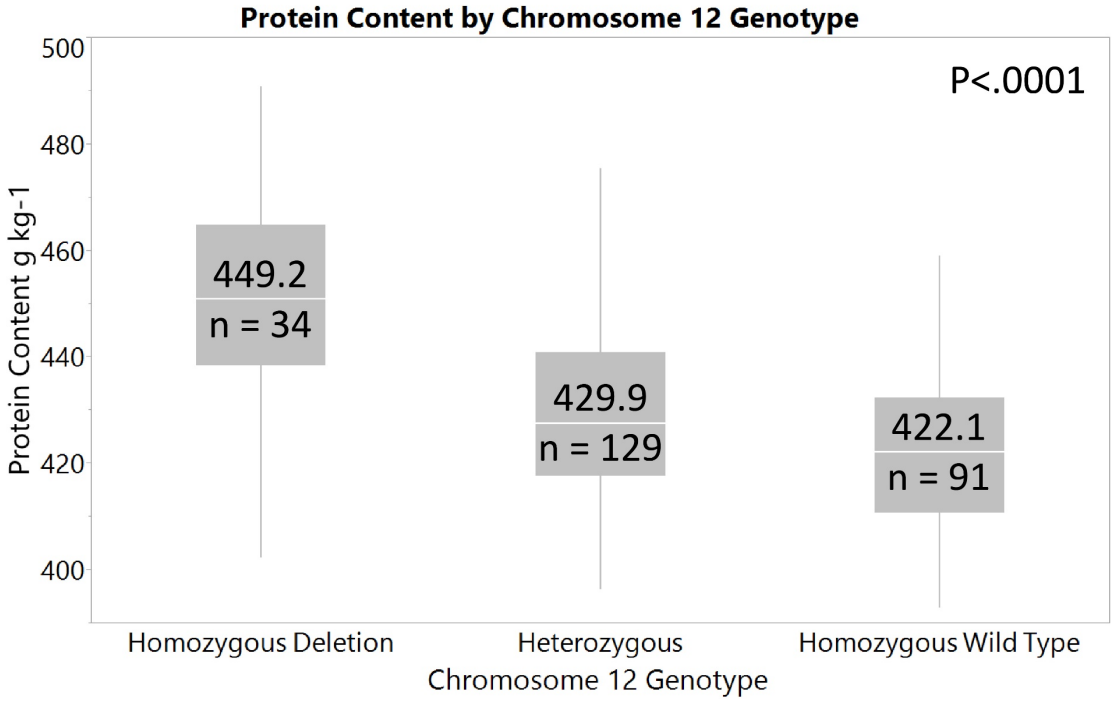


**Figure 2.6:** Protein differences among genotypes in the $F_2$ Benning × G15FN-12 population

Note: Protein content is reported in g $kg^{-1}$ on a dry matter basis. *P*<.0001. An "n" indicates the number of $F_2$ plants with the corresponding genotype.

# CHAPTER 3

# INTROGRESSION OF A HIGH PROTEIN ALLELE INTO AN ELITE SOYBEAN VARIETY RESULTS IN A HIGH-PROTEIN NEAR-ISOGENIC LINE WITH YIELD PARITY [1]

[1] Prenger, E.M., R. Mian, B. Buckley, H.R. Boerma, and Z. Li. To be submitted to *Crop Science*.

**Abstract**

Soybean [*Glycine max* (L.) Merr.] meal is the largest source of protein meal consumption worldwide. Meal is a high-protein product with all the essential amino acids for animal growth and development. Soybeans are valued for their suitability for use in feed, but the protein content of soybeans in the United States has been decreasing. Soybean seed protein content is negatively correlated with both yield and oil content, and these negative relationships complicate the simultaneous improvement of seed protein and selection for high-yielding varieties. A major protein QTL on chromosome (Chr) 20 has been identified in genome-wide association studies (GWAS) and bi-parental population mapping studies. The effects of various sources of Chr 20 high-protein alleles have been studied in several Northern and Midwestern genetic backgrounds of soybeans. The high-protein allele is often associated with a significant increase in protein and decreases in oil and yield. The Chr 20 high-protein allele from a Korean tofu cultivar, Danbaekkong (PI 619083), was introgressed into an elite MG VII cultivar, Benning (PI 595645), at the University of Georgia. A near-isogenic line (NIL), designated Benning HP, was developed through backcrossing and yield-tested in 2015 and 2017. Results from eight environments over two years revealed yield parity between the recurrent parent, Benning, and the NIL, despite a 41 g $kg^{-1}$ increase in protein in the NIL. The NIL also contained 19 g $kg^{-1}$ less oil than Benning. This NIL may be used to develop high protein, high-yielding soybean varieties for the Southern U.S.

## Introduction

Soybeans [*Glycine max* (L.) Merr.] are the largest source of protein meal worldwide due to the high protein content containing amino acids essential for animal and human growth and development (http://soystats.com; Patil et al., 2017). In the United States alone, 42.0 million metric tons of soybean meal were produced in 2017. Of those, 31.1 million metric tons were fed to livestock, including poultry, swine, beef, dairy, pets, and other animals. In addition to being a crucial source of protein meal, soybeans are the second largest source of vegetable oil worldwide (http://soystats.com). A typical U.S. soybean contains approximately 410 g $kg^{-1}$ protein and 210 g $kg^{-1}$ oil on a dry matter basis (Hartwig and Kilen, 1991). Both soybean meal and oil are economically important components of soybeans.

Prevalence of soybean protein and oil products indicate the importance of soybean seed composition for end use. Payment to farmers on a weight basis and the need to meet global demand necessitates breeding for high yield in addition to improved seed composition. However, negative relationships between protein and oil contents and between protein content and yield complicate the simultaneous improvement of seed protein and yield (Burton, 1987; Hartwig and Hinson, 1972; Chung et al., 2003; Thorne and Fehr, 1970; Cober and Voldeng, 2000). Hanson et al. (1961) noted a high-protein and high-yield combination in soybean should be possible based on energy transformation models. Multiple studies have indicated the possibility of developing soybeans with increased protein content, without significantly decreasing yield (Wilcox and Cavins, 1995; Wilcox and Zhang, 1997; Yin and Vyn, 2005; Cober and Voldeng, 2000). Wilcox and Cavins (1995) identified lines with increased protein and high yield by crossing a high-protein maturity group 00 (MG 00) accession, 'Pando,' to a low-protein, high-yielding MG IV cultivar, 'Cutler 71'. $F_4$-derived lines were selected as high protein donor parents for

backcrossing based on protein content and agronomic similarity to 'Cutler 71,' and in each backcross generation studied, lines with elevated protein and yield similar to 'Cutler 71' were identified. 'Cutler 71' averaged 2461 kg $ha^{-1}$ and 408 g $kg^{-1}$ protein, while 'Pando' yielded 800 kg $ha^{-1}$ and averaged 498 g $kg^{-1}$ protein. In the $BC_2$ and $BC_3$ generations grown in four-row plots in two to three replications, progeny were identified that contained more than 460 g $kg^{-1}$ protein and yielded equal to or greater than 2400 kg $ha^{-1}$. Wilcox and Zhang (1997) indicated that the regression of seed protein on yield was not significant in 61 determinate progeny from two crosses between high-protein indeterminate lines and average-protein determinate lines. Three determinate entries with the highest yield from one cross contained 449 to 478 g $kg^{-1}$ seed protein. Yield was similar to that of the average-protein parent, which contained 414 g $kg^{-1}$ protein. Two out of three highest-yielding entries from the other cross followed a similar pattern. Yin and Vyn (2005) found no significant relationship between seed protein concentration and yield in MG 0 and MG I lines planted from 1998-2000. Protein (g $kg^{-1}$) and yield (Mg $ha^{-1}$) were negatively correlated, and the regression line had a slope of -2.11 and an $R^2$ value of 0.01. Cober and Voldeng (2000) used outcross and backcross populations to develop lines with high protein and yield. They tested six outcross-derived lines and nine backcross-derived lines at six locations. Selected lines (n=42) from each crossing scheme exhibited low association between protein content and seed yield in one year tested, with correlation coefficients of -0.06 for backcross lines and -0.21 for outcross lines. A second year of correlations between protein and yield resulted in correlation coefficients of -0.07 for outcross lines (n=886) and -0.08 for backcrossed lines (n=800). Although the correlations are close to zero, they are significant, possibly due to the large number of observations leading to false positives. Both crossing schemes resulted in lines with significantly higher seed protein (456 to 501 g $kg^{-1}$) than the

recurrent parent, 'Maple Glen' (427 g $kg^{-1}$). Several of these individuals have yield similar to that of 'Maple Glen,' despite the increase in protein content. Results such as these indicate the possibility of breeding high-protein lines without decreasing yield.

Numerous protein QTL have been identified using bi-parental populations and genome-wide association studies (GWAS). QTL for protein content have been reported on each of 20 chromosomes (https://soybase.org). Two major protein QTL on Chrs 15 and 20 have been detected in many studies (Diers et al., 1992; Sebolt et al., 2000; Chung et al., 2003; Nichols et al., 2006; Qi et al., 2011; Bolon et al., 2010; Hwang et al, 2014; Vaughn et al., 2014; Bandillo et al., 2015; Warrington et al., 2015; Kim et al., 2016; Phansak et al., 2016).

Diers et al. (1992) originally mapped both the Chr 15 and Chr 20 protein and oil QTL in an A81-356022 x PI 468916 population. Sebolt et al. (2000) introgressed the high-protein alleles on Chrs 15 and 20 from the *Glycine soja* PI 468916 into a MG III breeding line. Lines homozygous for the Chr 20 high protein allele had significantly increased protein content (12 to 30 g $kg^{-1}$) compared to lines homozygous for the average protein cultivar allele, and the QTL indicated with the linked markers accounted for 41-65% of phenotypic variation in protein content in several populations across years. Lines homozygous for the PI 468916 allele on Chr 20 also had reduced oil content (9 g $kg^{-1}$) and yield. However, the PI 468916 allele on Chr 15 did not significantly affect protein content in this study.

Chung et al. (2003) developed recombinant inbred lines (RILs) from a cross of the high-protein PI 437088A and cultivar Asgrow 3733. The results indicated that homozygous PI alleles on Chr 20 were associated with an 18 g $kg^{-1}$ increase in protein content, an 11 g $kg^{-1}$ decrease in oil content, and up to a 268 kg $ha^{-1}$ decrease in yield. The $R^2$ value for the most highly associated marker ranged from 27.7 to 45.3%. In this study, protein showed strongly negative correlations

with both oil content and yield in multiple years. Nichols et al. (2006) introgressed the Chr 20 high protein allele from PI 468916 into a MG III genetic background and found an association between markers on Chr 20 with increased protein content (11.6 to 21.1 g $kg^{-1}$) and decreased oil content (7.7 to 12.4 g $kg^{-1}$) and yield (4.7 to 308.9 kg $ha^{-1}$), although the decrease in yield was not always significant. Warrington et al. (2015) mapped a major QTL to Chr 20 in a Benning × Danbaekkong RIL population. The Chr 20 QTL explained 55% of phenotypic variation in protein content. The additive effect of the Danbaekkong allele amounted to a 13.6 g $kg^{-1}$ increase in protein.

Kim et al. (2016) fine-mapped the Chr 15 protein QTL to a region between 3.59 and 4.12 Mb, based on the Williams 82 version 2 reference genome. PI 407788A was the source of the high protein allele at the Chr 15 locus, and Williams 82 was the normal protein parent. Lines homozygous for the PI allele had 9 g $kg^{-1}$ higher protein and 5 g $kg^{-1}$ lower oil than homozygous Williams 82 alleles, and the locus explained approximately 26% of phenotypic variation for protein. Recently, Brzostowski and Diers (2017) introgressed the Chr 15 high-protein QTL from PI 407788A into two elite MG II genetic backgrounds, AR09-192019 and LD02-4485. The PI allele was associated with an 11 g $kg^{-1}$ increase in protein and a 6 g $kg^{-1}$ decrease in oil content on a 130 g $kg^{-1}$ moisture basis. The PI allele was also associated with a non-significant decrease in yield, amounting to a 109 kg $ha^{-1}$ yield decrease in the AR09-192019 background and a 57 kg $ha^{-1}$ yield decrease in the LD02-4485 background.

The effects of Chr 20 protein QTL alleles in various backgrounds have been characterized in several recent studies. Brzostowski et al. (2017) introgressed high-protein alleles from two sources, Danbaekkong (PI 619083) and *Glycine soja* PI 468916, into various MG II and MG IV genetic backgrounds. The Danbaekkong allele, designated “CHR20-D,” was

introgressed into two MG II genetic backgrounds and was associated with significantly increased protein content (19 to 20 g $kg^{-1}$) and decreased oil content (7 to 9 g $kg^{-1}$) in both populations across environments on a 130 g $kg^{-1}$ moisture basis. The CHR20-D allele was also associated with a significant yield decrease of greater than 300 kg $ha^{-1}$ across environments. The PI allele from PI 468916, designated "CHR20-PI," was also introgressed into two MG II and two MG IV genetic backgrounds. The CHR20-PI allele was associated with significantly increased protein content (18 to 23 g $kg^{-1}$) and decreased oil content (9 to 13 g $kg^{-1}$) across environments, but the effects on yield were variable. The PI allele was associated with a decrease in yield (134 to 270 kg $ha^{-1}$) across environments for each population, although the decrease was not always significant.

A recent MG III germplasm release, 'Highpro1,' was developed with the Danbaekkong Chr 20 allele conferring increased protein content (Mian et al., 2017). The line contained 57 g $kg^{-1}$ more protein than the check cultivars on a 130 g $kg^{-1}$ moisture basis and also yielded the same as the mean of the check cultivars across 20 environments. The source of the Danbaekkong allele was the University of Georgia breeding line, GASF98-114, which was developed from a QTL mapping population used by Warrington et al. (2015). GASF98-114 is an $F_5$-derived RIL from a cross of the elite MG VII cultivar 'Benning' (PI595645) (Boerma et al., 1997) and Korean MG V cultivar 'Danbaekkong' (PI 619083) (Kim et al., 1996). GASF98-114 is also the high-protein allele donor for Benning HP NIL in this study.

Effects of the Danbaekkong Chr 20 allele on protein, oil, and yield have not been described in southern U.S. germplasm. Based on the effect of this allele described by Harris (2001), the Benning HP NIL was developed to study the performance of this allele in southern

germplasm, to investigate and characterize yield parity between Benning and the high-protein Benning NIL, and to compare protein and oil contents and agronomic performance.

## Materials and Methods

### *Development of Benning HP*

Benning HP is a NIL developed by backcrossing a high-protein $F_5$-derived RIL from a cross between Benning (PI 595645) (Boerma et al., 1997) and Danbaekkong (PI 619083) (Kim et al., 1996) to the recurrent parent, Benning. Benning is an elite MG VII cultivar released in 1996 from the University of Georgia with purple flowers, tawny pubescence, tan pod walls, determinate growth habit, brown hilum color, and a yellow seed coat. The donor parent, Danbaekkong, is a late MG IV South Korean accession with high seed protein content. An $F_5$-derived RIL with high seed protein content, GASF98-114, was selected and backcrossed to Benning RR starting in 2000. The first backcross between Benning RR and G98SF-114 took place at the UGA Plant Sciences Farm in Watkinsville, GA in the summer of 2000. Additional backcrosses of the progeny containing the high-protein allele were made to advance to the $BC_4F_1$ generation. Backcrosses were conducted either at the UGA Plant Science Farm or in the greenhouse. The backcrossing process was facilitated by marker-assisted selection (MAS). During the winter of 2002-2003, the $BC_4F_1$ generation was self-fertilized in the greenhouse to produce the $BC_4F_2$ generation. Benning HP is derived from the $BC_4F_2$ generation. Plants homozygous for the Chr 20 Danbaekkong allele were designated G03MG-97. G03MG-97 was used as the seed source for 2015 and 2017 yield trials. Benning HP was not entered into yield trials in 2016 because of poor seed quality of the 2015 seed. Due to backcrossing to Benning RR,

the Benning HP seed used for genetic analyses and yield trials is a Roundup Ready® version of the NIL.

The recurrent parent used in backcrossing, Benning RR, also called G99-G3438, was developed by backcrossing a Resnik-RR plant to Benning (Boerma et al., 1997). Resnik-RR was provided as a source of the Roundup Ready® trait through an agreement between Monsanto and the University of Georgia Research Foundation, Inc. The Georgia Agricultural Experiment Stations released Benning as a conventional variety in 1996. Benning is resistant to pests such as southern stem canker (*Diaporthe aspalathi*), bacterial pustule (*Xanthomonas campestris* pv. *glycines*), southern root knot nematode (*Meloidogyne incognita*), and soybean cyst nematode (*Heterodera glycines*) race 3, and is moderately resistant to prevalent races of frogeye leaf spot (*Cercospora sojina* K. Hara) (Boerma et al., 1997).

*Characterization of the Danbaekkong introgression fragment on chromosome 20*

Leaf tissue was collected from 17 Benning HP plants in the crossing block in September 2016. Tissues were lyophilized for 24 hours and ground to a fine powder using a Spex Sample Prep Geno/Grinder 2010 machine and BB's. Genomic DNA was extracted using a CTAB protocol modified from Keim et al. (1988) and optimized for genotyping with the SoySNP50K Infinium BeadChip (Song et al., 2013). Approximately 100 µl of genomic DNA diluted to 50 ng $\mu l^{-1}$ were used for genotyping with the SoySNP50K Infinium BeadChip in the Soybean Genomics and Improvement Lab, USDA-ARS, Beltsville, Maryland. Genome Studio V2011.1 (Illumina, Inc., 2011) was used to perform SNP calls, and manual editing was performed in regions of interest. SoySNP 50K genotype data for Benning (PI 595645) and Danbaekkong (PI 619083) were available through Soybase (https://soybase.org) and the UGA Soybean Breeding

and Genetics Lab database. The Chr 20 introgression location and size were determined by matching the Benning HP SNP genotypes to either Benning or Danbaekkong for each SNP marker polymorphic between Benning and Danbaekkong. The introgression was visualized using Flapjack (Milne et al., 2010) and JMP Pro version 13 (SAS Institute, 2017) software. A total of 518 markers polymorphic between 'Benning' and 'Danbaekkong' was used to estimate the size and location of the Chr 20 introgression.

*Yield test and seed composition analysis*

Benning and Benning HP were grown in Georgia Advanced Yield Trials in 2015 at three locations with three reps per location in a randomized complete block design. Yield trials included the elite lines 'Boggs' (Boerma et al., 2000) and 'Woodruff' (Boerma et al., 2012) as checks along with other entries. Trials were grown in Watkinsville, GA; Plains, GA; and Florence, SC. Plots in Watkinsville and Plains, GA consisted of four rows with 76.2 cm row spacing and 4.9 m row length. Plots were planted at a density of about 27 seeds meter$^{-1}$. At harvest, the two middle rows were harvested with Almaco plot combines. Plots consisted of four rows 6.1 m in length. Seeds were planted at a density of approximately 33 seeds m$^{-1}$ in the Florence, SC location. Best management practices were applied in these yield trials. Plots in Georgia were end-trimmed in September before harvest. A total of 61 cm was removed from both ends of each plot, resulting in a 3.7 m length for yield measurements.

Benning and Benning HP were grown in yield trials for a second year in the 2017 UGA Advanced Yield Trials at five locations with three reps per location in a randomized complete block design. Elite checks included AG 7733 and AG 7934, both MG VII varieties developed and released by Monsanto Company. Trials were grown in Watkinsville, GA; Plains, GA; Clayton

(Plymouth), NC; Caswell, NC; and Bossier, LA. Plots grown in Watkinsville and Plains, GA and Bossier, LA consisted of four rows on 76.2 cm row spacing and 4.9 m in length. Approximately 27 seeds meter$^{-1}$ were planted in these locations. At harvest, the middle two rows were harvested. All plots in Watkinsville and Plains, GA were end-trimmed before harvest. Both ends of each plot were trimmed 61 cm from the end, leaving a plot 3.7 m in length for harvest. Plots in Clayton (Plymouth) and Caswell, NC consisted of three rows on 96.5 cm row spacing with 6.4 m row length. Seeds were planted at a density of approximately 31 seeds meter$^{-1}$ in these plots. Best management practices were used throughout the growing season.

Traits including flower color, pubescence color, pod wall color, and hilum color were recorded. These traits were used as quality control measures. Maturity, plant height, lodging, seed size, seed quality, seed protein content, and seed oil content were also recorded each year. Maturity was measured as the number of days past August 31 required for at least 95% of a plot to reach physiological maturity (R8). Plant height was measured in centimeters. Lodging was evaluated on a scale from 1 to 5, with 1 indicating the plants in the plot were upright and 5 meaning plants were prostrate on the ground. Seed size was measured as grams 100-seed$^{-1}$. Seed quality was rated on a scale from 1 to 5, with a rating of 1 indicating good seed quality with very few cracked, wrinkled, or diseased seeds and a rating of 5 indicating very poor quality with many cracked, wrinkled, or diseased seeds. Yield was measured in kg ha$^{-1}$. Seed protein and oil contents were measured as a percent of the total seed on a dry seed basis. Seed composition was analyzed via NIR using a Perten DA 7250 machine (Perten, Peoria, IL). The calibration curve used to quantify seed protein and oil content was developed from hundreds of soybean samples with known composition values (Soybean NIR Consortium).

In 2015, yield and lodging were measured for all three reps at three locations. Maturity and height were recorded for all three reps in Watkinsville, GA and Florence, SC. Seed size, seed quality, seed protein content, and seed oil content were recorded for all three reps in Plains, GA. The yield trial in Watkinsville, GA was not harvested until December 9, 2015, making seed composition and quality data inaccurate. Due to poor seed quality in 2015, Benning HP was increased in the summer of 2016 in Watkinsville, GA. In 2017, yield, seed size, seed quality, and protein and oil content were measured for all available reps at all five locations. Maturity and lodging were recorded for all three reps at three locations and two reps at Caswell, NC. Plant height was recorded at three locations.

Quality control was performed using JMP Pro version 13.2 to identify extreme outliers. Seed composition outliers were subjected to NIR analysis a second time to minimize machine error. Extreme yield outliers were identified using a coefficient of variance (CV) in the Agrobase database (Agronomix Inc., 2012), and a corrected yield value was imputed for individual outlier plots based on surrounding plots and additional reps of the same line.

*Statistical Analyses*

Data analyses were performed in JMP Pro version 13.2 using a Mixed Fit Model function. Line (Genotype, G) was treated as a fixed effect, and location (Environment, E), genotype by location interaction (G x E), and rep within location were treated as random effects. Values for each trait are reported as Least Squares Means (LS Means), and significant differences were determined using Tukey's Honestly Significant Difference (Tukey's HSD) where $\alpha=0.05$. Additional analyses across both years and all locations were performed as stated previously in SAS version 9.4 software (SAS Institute, 2017) using the PROC GLM procedure to

estimate the effect of each variable. Analyses of each individual environment were performed in JMP Pro version 13.2 using a fixed ANOVA with Genotype and Rep (or block) as factors to determine significant differences based on Tukey's HSD ($\alpha$=0.05).

*Evaluation of protein and oil accumulation in different planting dates*

Benning and Benning HP were included in 2017 USDA Uniform Test MG VII (UT7) in Watkinsville, GA to evaluate differences in accumulation of protein and oil between early- and late-planted plots. The "early" planting date is typically when Georgia farmers plant soybean, and the "late" date is planted about a month after farmers prefer to have soybean in the ground. Each genotype was grown in three replicates in a randomized complete block design. Early plots were planted on May 18, 2017 and harvested Nov. 20, 2017 with a total time in the field of 26.5 weeks. Late plots were planted on June 19, 2017 and harvested Nov. 20, 2017 with a total time in the field of 22 weeks. Data for protein and oil accumulation were evaluated with SAS version 9.4 software using the PROC GLM procedure and a fixed model. Genotype, environment, genotype by environment interaction, and rep within environment [rep(environment)] were treated as fixed effects. Least Squares Means were estimated for each genotype by environment interaction, and significant differences were based on Tukey's HSD ($\alpha$=0.05). Benning and Benning HP were both also planted in the 2016 UGA Soybean Breeding Program's crossing block, which consists of an early and late planting date. Benning was planted in five rows in each planting date, and Benning HP was planted in three rows per planting date. Average protein and oil content for each genotype were compared between planting dates. Statistical analysis could not be performed due to insufficient experimental design, but trends were noted.

## Results

*Genetic characterization of the Danbaekkong high protein allele introgression*

A total of 518 polymorphic SNPs on Chr 20 revealed the approximate size of the high-protein introgression. SNP markers for Chr 20 begin at 26,325 bp and end at 46,763,584 bp of Chr 20 based on the Williams 82 version 2 reference genome. The introgression begins around 422 kb of Chr 20 and ends around 35.6 Mb of Chr 20. (Figure 3.1). The approximate introgression size was determined using 518 SNP markers polymorphic between Benning and Danbaekkong.

The Chr 20 high-protein introgression in Benning HP is approximately 35.2 Mb in total size. This makes up about 74% of the physical chromosome and contains up to 1,152 gene models, based on the Williams 82 version 2.0 genome sequence from JBrowse (Skinner et al., 2009). Previous studies have identified the 24.5 to 32.9 Mb region of Chr 20 as the most likely location of the major protein QTL, based on the Williams 82 version 1.0 reference genome (Schmutz et al., 2010). Bolon et al. (2010) identified a protein QTL between 24.5 and 32.9 Mb of Chr 20. Several GWA studies also identified and refined the QTL region. Hwang et al. (2014) narrowed the QTL region to between 27.6 and 30.0 Mb. Vaughn et al. (2014) identified the probable QTL region about 1 Mb upstream of that identified by Hwang et al. (2014) in the 30.9 to 31.9 Mb region, and Bandillo et al. (2015) identified the 29.6 to 31.9 Mb region as the most likely location of the high protein locus. Gm20_31610452 (Gm20_32752214 W82.a2.v1) was identified as the most significantly associated SNP (Vaughn et al., 2014).

*Yield and agronomic trait performance of Benning and the Benning HP NIL*

Yield trials in 2015 revealed yield parity between Benning (3,476.8 kg $ha^{-1}$) and Benning HP (3,490.3 kg $ha^{-1}$), with a non-significant yield increase of 13.5 kg $ha^{-1}$ in Benning HP. Benning HP yielded 103.6% of Boggs and 89.8% of Woodruff in 2015. Maturity of Benning and Benning HP were very similar, 63.2 and 64.3 days past August 31, respectively. Benning HP grew to 83.4 cm height and had a lodging score of 2.4. Seed size of Benning HP was 12.1 g 100-$seed^{-1}$, which was significantly smaller than Benning at 14.2 g 100-$seed^{-1}$. Seed quality of Benning HP did not differ significantly from Benning. Protein content of Benning HP was 464.2 g $kg^{-1}$ and oil content was 201.3 g $kg^{-1}$ on a dry seed basis. Benning seed had 429.1 g $kg^{-1}$ protein and 221.1 g $kg^{-1}$ oil. Compared to Benning, Benning HP had a 35.1 g $kg^{-1}$ increase in protein and a 19.8 g $kg^{-1}$ decrease in oil. Both protein and oil values are significantly different. Table 3.1 contains the results of 2015 yield trials.

Yield trials in 2017 confirmed yield parity between Benning and Benning HP. Benning yielded 3,154.0 kg $ha^{-1}$, and Benning HP yielded 3,107.0 kg $ha^{-1}$, amounting to a non-significant yield decrease in the NIL of 47.0 kg $ha^{-1}$. Benning HP yielded 86.4% and 88.0% of commercial checks AG 7733 and AG 7934, respectively. Benning HP matured four days earlier than Benning in 2017 (significant at $\alpha$=0.05). Benning HP reached a height of 97.0 cm and had a lodging score of 3.0. Neither of these values differed significantly from Benning (96.5 cm height and a lodging score of 3.3). Seed size of Benning HP was significantly lower than that of Benning again in 2017, with seed sizes of 14.1 and 15.2 g 100-$seed^{-1}$, respectively. Seed quality did not differ significantly between Benning HP and Benning. Protein content of Benning HP was 462.2 g $kg^{-1}$, and oil content was 179.1 g $kg^{-1}$. These values are significantly different from Benning, which

had 419.8 g $kg^{-1}$ protein and 198.3 g $kg^{-1}$ oil. Compared to Benning, it amounts to a 42.4 g $kg^{-1}$ increase in protein content and a 19.2 g $kg^{-1}$ decrease in oil content in Benning HP (Table 3.2).

Analysis of Benning HP and Benning across years and environments revealed yield parity between the two lines with Benning HP yielding 99.2% of Benning (a non-significant yield decrease of 26.9 kg $ha^{-1}$) across a total of eight environments. Benning HP had 41.2 g $kg^{-1}$ higher protein content and 19.3 g $kg^{-1}$ lower oil content than Benning across six environments, which were both significantly different. Across six environments, Benning HP matured approximately two days earlier and was 1.2 g 100-$seed^{-1}$ lighter than Benning. Plant height, lodging, and seed quality were not significantly different between Benning and Benning HP across five to seven environments (Table 3.3).

Evaluation of yield, protein, and oil in each environment in 2015 and 2017 revealed a significant genotype by environment interaction for yield in the 2017 Bossier, LA environment (Table 3.4). Yield of Benning HP was significantly lower than that of Benning at the Bossier location in 2017, with a yield difference of 627.7 kg $ha^{-1}$. In four of the seven remaining locations, Benning HP had higher yield than Benning, while in the other three locations, Benning HP had lower yield than Benning, although none of these differences were significant. Protein content of Benning HP was significantly higher than Benning in all six environments, and oil content of Benning HP was significantly lower. Analysis of yield, protein, and oil across all years and locations did not reveal significant G x E effects for these traits (P=0.48, 0.10, and 0.83, respectively) (Table 3.3b).

*Comparison of protein and oil content of Benning HP between planting dates*

Benning and Benning HP were planted in the 2017 USDA Uniform Test MG VII (UT7) in Watkinsville, GA at both an early (typical) and a late planting date. Benning HP acquired 473.3 g $kg^{-1}$ seed protein and 174.0 g $kg^{-1}$ oil in the early planting and 473.1 g $kg^{-1}$ protein and 178.8 g $kg^{-1}$ oil in the late planting. These protein and oil values are nearly equal between early and late planting and are not significantly different. In contrast, Benning had 429.4 g $kg^{-1}$ seed protein and 190.8 g $kg^{-1}$ oil in the early planting and 412.3 g $kg^{-1}$ protein and 209.5 g $kg^{-1}$ oil in the late planting. Protein content is not significantly different between the two planting dates, but oil content is significantly different between dates ($P$=0.01).

In the 2016 crossing block, Benning HP averaged 467.8 g $kg^{-1}$ protein and 177.7 g $kg^{-1}$ oil in the early planting date and averaged 450.5 g $kg^{-1}$ protein and 182.1 g $kg^{-1}$ oil in the late planting date. Delayed planting resulted in 17.3 g $kg^{-1}$ lower protein and 4.5 g $kg^{-1}$ higher oil. In comparison, Benning averaged 411.9 g $kg^{-1}$ protein and 199.8 g $kg^{-1}$ oil in the early planting date and 400.0 g $kg^{-1}$ protein and 205.4 g $kg^{-1}$ oil in the late planting date. Similar to Benning HP, delayed planting of Benning resulted in a 12.0 g $kg^{-1}$ decrease in protein and a 5.7 g $kg^{-1}$ increase in oil.

**Discussion**

A well-documented negative relationship between protein and yield in soybeans (Burton, 1987; Hartwig and Hinson, 1972; Chung et al., 2003; Thorne and Fehr, 1970; Cober and Voldeng, 2000) has led soybean breeders to attempt to break or mitigate this negative relationship using several approaches. As a result of this negative relationship, soybean breeders selected for higher yield, leading to an overall decrease in the protein content of U.S. soybeans in

recent years (http://unitedsoybean.org). Past studies have indicated it should be possible to select for increased yield without significantly decreasing protein (Wilcox and Zhang, 1997; Wilcox and Cavins, 1995; Yin and Vyn, 2005; Cober and Voldeng, 2000), and soybeans with increased protein would benefit end users of soybean meal as well as farmers who could earn an additional $19 to $32 per hectare if soybean seed protein is increased by just 10 g $kg^{-1}$ (http://unitedsoybean.org).

Previous attempts to produce a high protein line with yield parity when using the Chr 20 protein QTL have produced mixed results. Use of the Danbaekkong Chr 20 allele in two midwestern genetic backgrounds (MG II) resulted in backcross lines with significantly increased protein, but with decreased oil content and yield (Brzostowski et al., 2017). Use of the *G. soja* PI 468916 as the source of the high protein allele on Chr 20 resulted in increased protein, decreased oil, and variable effects on yield in four MG II and IV backcross populations. Although effects on yield associated with the PI 468916 allele were not always significant, yield of the lines associated with the PI allele was less than that with the recurrent parent allele across environments (Brzostowski et al., 2017). A recent MG III soybean release, 'Highpro1', has increased protein and decreased oil content as a result of introgression of the Danbaekkong high protein allele on Chr 20 and also has yield parity with checks, although this cultivar is not a NIL and therefore is not directly comparable to the parents (Mian et al., 2017). The donor of the Chr 20 protein allele for 'Highpro1' was the University of Georgia breeding line that also was used to develop Benning HP. The results of the Benning HP yield trials indicate that the negative association between protein and yield may be mitigated using the Danbaekkong allele in some genetic backgrounds.

The United Soybean Board has set goals of 350 g $kg^{-1}$ or higher protein, 480 g $kg^{-1}$ or higher meal protein, and 190 g $kg^{-1}$ or higher oil content on a 130 g $kg^{-1}$ (13%) moisture basis (United Soybean Board, 2017, personal communication). Although Benning HP falls short of the oil content goal with an oil content of 159.0 g $kg^{-1}$ (on a 130 g $kg^{-1}$ moisture basis) across six environments, it exceeds the United Soybean Board's goals for seed protein and meal protein content with 402.4 g $kg^{-1}$ protein over two years and more than 500 g $kg^{-1}$ meal protein.

Benning HP yielded 100% of the recurrent parent, Benning, across environments in 2015. In 2017, Benning HP yielded 98.5% of Benning, and the difference was not considered significantly different (Tukey's HSD, $\alpha$=0.05). Analyses of yield, protein, and oil across years and locations revealed the G x E factor was not significant for these traits. However, analysis of the same traits in individual environments revealed a significant decrease in yield of Benning HP compared to Benning only in the 2017 Bossier, LA environment. Detection of a significant (P=0.01) decrease in yield in individual analysis of the Bossier, LA environment when G x E was not significant in the ANOVA may be due to simplification of the model and a small number of replications when performing analysis in only one environment with a genotype and a block, or rep, variable. The difference in yield between Benning and Benning HP in the Bossier, LA environment may not have been strong enough to appear in analysis across all years and locations. This decrease in yield of Benning HP in the Bossier, LA environment may have been due to several factors including environmental and growing conditions, or variability in plots due to placement within blocks in the field. The increase in protein and decrease in oil content of Benning HP compared to Benning were significant in all environments. Soybean seed composition is affected by factors such as temperature and rainfall (Song et al., 2016; Kumar et

al., 2006), which may have played a role in the magnitude of differences in seed composition between Benning HP and Benning in various environments.

Similarity of yields between Benning and Benning HP across environments indicates it is possible to significantly increase protein content in the Benning background without a corresponding decrease in yield, although the oil content is decreased. Since the increase in protein in Benning HP is due to a single locus on Chr 20 reported by Warrington et al. (2015), introgression into other genetic backgrounds can be done relatively quickly using a marker-assisted backcrossing scheme. Benning was released in 1996 (Boerma et al., 1997), so the yield of this variety is not equal to the varieties developed more recently. Nevertheless, Benning provides an elite background for Benning HP and results in a high-protein, elite source of breeding material for developing varieties with competitive yield and increased protein. However, the conclusions of this study are limited due to the characterization of the Danbaekkong Chr 20 high protein alleles in only one genetic background. In order to make broader conclusions, the high protein alleles must be studied in additional elite southern germplasm backgrounds.

The introgression from Danbaekkong is large, consisting of approximately 74% of the physical Chr 20 as indicated in JBrowse (Skinner et al., 2009). Benning HP was developed by backcrossing the donor and subsequent offspring carrying the Chr 20 high-protein allele for four generations to Benning RR, and the yield trial seed source was derived from the $BC_4F_2$. Danbaekkong is expected to contribute less than 3.1% of Benning HP's genomic sequence. With an approximate genome size of 1.1 Gb, this means up to 34 Mb of sequence are expected to come from Danbaekkong. Based on characterization of the introgression on chromosome 20, the estimated introgression size accounts for the entirety of the estimated genetic material from

Danbaekkong. The large introgression size of the Danbaekkong high protein allele could be a result of using flanking SSR markers for selection during backcrossing or structural variation. Therefore, using gene specific markers or tightly linked markers as well as flanking markers to select the numerous individuals with the trait of interest as well as desired recombinants will help reduce the size of the introgression region to avoid potential linkage drag, unless structural variation is present.

Results of yield trials indicate that Benning HP can be used in breeding programs to develop additional elite, high-yielding lines with increased protein content, though the increase in protein content is concurrent with a decrease in oil. The ability to significantly increase protein without decreasing yield may be dependent on the use of southern germplasm. The scope of this study is limited to the Chr 20 Danbaekkong allele effects in a single MG VII genetic background. Performance of the high-protein allele in additional elite southern backgrounds must be evaluated to determine whether the introgression results in yield parity despite an increase in seed protein content across a variety of southern germplasm.

**Acknowledgements**

We thank the United Soybean Board for funding this research. Thanks to Dale Wood, Earl Baxter, Brice Wilson, Jeremy Nation, Greg Gokalp, Ricky Zoller, and Tatyana Nienow for technical support.

**References**

Bandillo, N., D. Jarquin, Q. Song, R. Nelson, P. Cregan, J. Specht, et al. 2015. A population structure and genome-wide association analysis on the USDA soybean germplasm collection. Plant Genome 8. doi:10.3835/plantgenome2015.04.0024.

Boerma, H.R., R.S. Hussey, D.V. Phillips, E.D. Wood, G.B. Rowan, and S.L. Finnerty. 1997. Registration of 'Benning' soybean. Crop Sci. 37:1982.

Boerma, H.R., R.S. Hussey, D.V. Phillips, E.D. Wood, G.B. Rowan, and S.L. Finnerty. 2000. Registration of 'Boggs' soybean. Crop Sci. 40:294-295.

Boerma, H.R., R.S. Hussey, D.V. Phillips, and E.D. Wood. 2012. Soybean variety G00-3209. U.S. Patent No. US 8,304,616 B2.

Bolon, Y.-T., B. Joseph, S.B. Cannon, M.A. Graham, B.W. Diers, A.D. Farmer, et al. 2010. Complementary genetic and genomic approaches help characterize the linkage group I seed protein QTL in soybean. BMC Plant Biology 10: 41-64. doi:10.1186/1471-2229-10-41.

Brim, C.A. 1966. Dare soybeans. Crop Sci. 6:95.

Brzostowski, L.F. and B.W. Diers. 2017. Agronomic evaluation of a high protein allele from PI407788A on chromosome 15 across two soybean backgrounds. Crop Sci. 57: 2972-2978. Doi: 10.2135/cropsci2017.02.0083.

Brzostowski, L.F., T. Pruski, J.E. Specht, and B.W. Diers. 2017. Impact of seed protein alleles from three soybean sources on seed composition and agronomic traits. Theor. Appl. Genet. 130: 2315-2326.

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

Buss, G.R., H.M. Camper, Jr., and C.W. Roane. 1988. Registration of 'Hutcheson' soybean. Crop Sci. 28:1024-1025.

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

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

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

Grant, D., R.T. Nelson, S.B. Cannon, and R.C. Shoemaker. 2010. SoyBase, the USDA-ARS soybean genetics and genomics database. Nucl. Acids Res. 38 (suppl 1): D843-D846. doi: 10.1093/nar/gkp798.

Hanson, W.D., R.C. Leffel, and R.W. Howell. 1961. Genetic analysis of energy production in the soybean. Crop Sci. 1:121-126.

Harris, D.K. 2001. Genetic improvement of soybean seed traits and resistance to bud blight and root-knot nematodes. MS thesis, University of Georgia.

Hartwig, 1958. Registration of 'Lee' Soybean. Agron. J. 50:690-691.

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

Hartwig, E.E., and D.J. Gray. 1991. The uniform soybean tests, southern region, 1990. USDA-ARS, Stoneville, MS.

Hartwig, E.E. and T.C. Kilen. 1991. Yield and composition of soybean seed from parents with different protein, similar yield. Crop Sci. 31: 290-292.

Hinson, K., R.A. Kinlock, H.A. Peacock, W.H. Chapman, and W.T. Scudder. 1981. `Braxton' soybean. Florida Agric. Exp. Stn. Circular S-276.

Hwang, E.-Y., S. Qijian, G. Jia, J.E. Specht, D.L. Hyten, J. Costa, et al. 2014. A genome-wide association study of seed protein and oil content in soybean. BMC Genomics 15: 1-25. doi:10.1186/1471-2164-15-1.

Keim, P., T. C. Olson, and R. C. Shoemaker. 1988. A rapid protocol for isolating soybean DNA. Soybean Genet. Newsl. 18: 150–152.

Kim, M., S. Schultz, R. Nelson, and B. Diers. 2016. Identification. 2016 Identification and fine mapping of a soybean seed protein QTL from PI 407788A on chromosome 15. Crop Sci 56(1):219-225. DOI: 10.2135/cropsci2015.06.0340

Kim, S.D., E.H. Hong, Y.H. Kim, S.H. Lee, Y.K. Seong, K.Y. Park, et al. 1996. A new high protein and good seed quality soybean variety “Danbaekkong.” RDA J. Agr. Sci. Upl. Ind. Crops 38:228-232.

Kumar, V., A. Rani, S. Solanki and S.M. Hussain. 2006. Influence of growing environment on the biochemical composition and physical characteristics of soybean seed. J. Food Comp. Anal. 19: 188-195. doi:http://dx.doi.org/10.1016/j.jfca.2005.06.005.

Mian, R., L. McHale, Z. Li, and A.E. Dorrance. 2017. Registration of ‘HighPro1’ soybean with high protein and high yield developed from a north x south cross. J. Plant Reg. 11: 51-54.

Milne I., P. Shaw, G. Stephen, M. Bayer, L. Cardle, W.T.B. Thomas, A.J. Flavell, and D. Marshall. 2010. Flapjack – graphical genotype visualization. Bioinformatics 26: 3133-3134.

Mississippi State University Agricultural and Forestry Experiment Station. 1976. Information Sheet no. 1331.

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

Patil, G., R. Mian, T. Vuong, V. Pantalone, Q. Song, P. Chen, et al. 2017. Molecular mapping and genomics of soybean seed protein: a review and perspective for the future. Theor. Appl. Genet. 130: 1975-1991. Doi: 10.1007/s00122-017-2955-8.

Phansak, P., W. Soonsuwon, D.L. Hyten, Q. Song, P.B. Cregan, G.L. Graef, et al. 2016. Multi-population selective genotyping to identify soybean [*Glycine max* (L.) Merr.] seed protein and oil QTLs. G3: Genes|Genomes|Genetics 6: 1635-1648. doi:10.1534/g3.116.027656.

Qi, Z., S. Ya-nan, W. Qiong, L. Chun-yan, H. Guo-hua and C. Qing-shan. 2011b. A meta-analysis of seed protein concentration QTL in soybean. Can. J. Plant Sci. 91: 221-230. doi:10.4141/cjps09193.

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

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

Skinner, M.E., A.V. Uzilov, L.D. Stein, C.J. Mungall, and I.H. Holmes. 2009. JBrowse: A next-generation genome browser. Genome Res. 19: 1630-1638.

Smith, T.J. 1968. Registration of York soybeans. Crop Sci. 8:776.

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

Song, Q., D. Hyten, G. Jia, C. Quigley, E. Fickus, R. Nelson, and P. Cregan. 2013. Development and evaluation of SoySNP50K, a high-density genotyping array for soybean. PLoS One 8(1):e54985.

Song, W., R. Yang, T. Wu, C. Wu, S. Sun, S. Zhang, et al. 2016. Analyzing the effects of climate factors on soybean protein, oil contents, and composition by extensive and high-density sampling in china. J. Agric. Food Chem. 64: 4121-4130. doi:10.1021/acs.jafc.6b00008.

Thorne, J.C. and W.R. Fehr. 1970. Incorporation of high-protein, exotic germplasm into soybean populations by 2- and 3-way crosses1. Crop Sci. 10: 652-655.

Vaughn, J.N., R.L. Nelson, Q. Song, P.B. Cregan and Z. Li. 2014. The genetic architecture of seed composition in soybean is refined by genome-wide association scans across multiple populations. G3: Genes|Genomes|Genetics 4: 2283-2294. doi:10.1534/g3.114.013433.

Warrington, C.V., H. Abdel-Haleem, D.L. Hyten, P.B. Cregan, J.H. Orf, A.S. Killam, et al. 2015. QTL for seed protein and amino acids in the Benning × Danbaekkong soybean population. Theor. and Appl. Genet. 128: 839-850. doi:10.1007/s00122-015-2474-4.

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

Wilcox, J.R. and G.D. Zhang. 1997. Relationships between seed yield and seed protein in determinate and indeterminate soybean populations. Crop Sci. 37: 361-364.

Yin, X.H. and T.J. Vyn. 2005. Relationships of isoflavone, oil, and protein in seed with yield of soybean. Agron. J. 97: 1314-1321.

**Table 3.1:** Performance of Benning HP compared to checks in 2015 UGA Advanced Yield Trials

| Line | Description | Seed yield † kg ha$^{-1}$ | Maturity ‡ d after 31 Aug | Plant height ‡ cm | Lodging ¶ score | Seed size# g 100-sd$^{-1}$ | Seed quality †† score | Protein content ‡‡ g kg$^{-1}$ | Oil content ‡‡ g kg$^{-1}$ |
|---|---|---|---|---|---|---|---|---|---|
| Benning HP | NIL | 3490.3$^{ab}$ | 64.3$^{a}$ | 83.4$^{a}$ | 2.4$^{a}$ | 12.1$^{c}$ | 4.0$^{a}$ | 464.2$^{a}$ | 201.3$^{b}$ |
| Benning | RP | 3476.8$^{ab}$ | 63.2$^{a}$ | 82.6$^{a}$ | 2.6$^{a}$ | 14.2$^{ab}$ | 3.7$^{a}$ | 429.1$^{c}$ | 221.1$^{a}$ |
| Boggs | check | 3369.2$^{b}$ | 63.5$^{a}$ | 73.2$^{a}$ | 2.4$^{a}$ | 12.7$^{bc}$ | 3.6$^{a}$ | 444.3$^{b}$ | 213.3$^{a}$ |
| Woodruff | check | 3887.1$^{a}$ | 65.3$^{a}$ | 77.7$^{a}$ | 2.6$^{a}$ | 14.9$^{a}$ | 2.5$^{b}$ | 426.6$^{c}$ | 203.3$^{b}$ |

Note: Values followed by a different letter are significantly different based on Tukey's HSD (0.05). Least Squares Means are reported for seed yield, maturity, plant height, and lodging. The mean of three reps from one location are reported for seed size, seed quality, protein content, and oil.
† Yield evaluations were performed in three locations in 2015 (Watkinsville, GA; Plains, GA; and Florence, SC) with 3 replications per location.
‡ Maturity and plant height were recorded in two locations.
¶ Lodging: 1= erect plants and 5= plants parallel with the ground; recorded in all three locations
# Seed size was recorded in Plains, GA.
†† Seed quality ratings were recorded in Plains, GA. The quality rating scale is based on the presence of cracked, split, wrinkled, discolored, or diseased seeds in a sample where 1=low presence and 5=high presence of defects.
‡‡ Protein and oil were recorded in Plains, GA. Values are reported on a dry weight basis.
RP = recurrent parent
NIL = Near-isogenic line

**Table 3.2:** Performance of Benning HP compared to checks in 2017 UGA Advanced Yield Trials

| Line | Description | Seed yield † kg ha$^{-1}$ | Maturity ‡ d after 31 Aug | Plant height § cm | Lodging ¶ score | Seed size # g 100-sd$^{-1}$ | Seed quality # score | Protein content # g kg$^{-1}$ | Oil content # g kg$^{-1}$ |
|---|---|---|---|---|---|---|---|---|---|
| Benning HP | NIL | 3107.0$^{b}$ | 46.3$^{c}$ | 97.0$^{a}$ | 3.0$^{ab}$ | 14.1$^{b}$ | 1.4$^{a}$ | 462.2$^{a}$ | 179.1$^{b}$ |
| Benning | RP | 3154.0$^{b}$ | 50.4$^{b}$ | 96.5$^{a}$ | 3.3$^{a}$ | 15.2$^{a}$ | 1.6$^{a}$ | 419.8$^{b}$ | 198.3$^{a}$ |
| AG 7733 | check | 3597.9$^{a}$ | 53.4$^{a}$ | 102.6$^{a}$ | 2.1$^{ab}$ | 15.5$^{a}$ | 1.5$^{a}$ | 410.0$^{b}$ | 187.2$^{b}$ |
| AG 7934 | check | 3530.6$^{ab}$ | 54.5$^{a}$ | 106.7$^{a}$ | 1.8$^{b}$ | 14.5$^{ab}$ | 1.4$^{a}$ | 410.0$^{b}$ | 199.4$^{a}$ |

Note: Values followed by a different letter are significantly different based on Tukey's HSD (0.05). All values reported are Least Squares Means.
† Yield evaluations performed in five locations in 2017 (Watkinsville, GA; Plains, GA; Plymouth, NC; Caswell, NC; and Bossier, LA) with 3 replications per location.
‡ Maturity recorded in four locations.
§ Plant height recorded in three locations.
¶ Lodging: 1= erect plants and 5= prostrate within a plot. Lodging recorded in four locations.
# Seed size, seed quality, and protein and oil were recorded from all five locations. The quality rating scale is based on the presence of cracked, split, wrinkled, discolored, or diseased seeds in a sample where 1=low presence and 5=high presence of defects. Protein and oil content are reported on a dry weight basis.
RP = recurrent parent
NIL = Near-isogenic line

**Table 3.3:** a) Yield and agronomic trait performance of Benning and Benning HP across years and locations. b) ANOVA for yield, protein, and oil across years and locations.

3.3a)

| Line | Description | Seed yield † kg ha$^{-1}$ | Maturity ‡ d after 31 Aug | Plant height § cm | Lodging ¶ score | Seed size # g 100-sd$^{-1}$ | Seed quality †† score | Protein content ‡‡ g kg$^{-1}$ | Oil content ‡‡ g kg$^{-1}$ |
|---|---|---|---|---|---|---|---|---|---|
| Benning HP | NIL | 3248.2$^{a}$ | 52.5$^{a}$ | 91.7$^{a}$ | 2.8$^{a}$ | 13.8$^{b}$ | 1.8$^{a}$ | 462.5$^{a}$ | 182.8$^{b}$ |
| Benning | RP | 3275.1$^{a}$ | 54.6$^{a}$ | 90.9$^{a}$ | 2.9$^{a}$ | 15.0$^{a}$ | 1.9$^{a}$ | 421.3$^{b}$ | 202.1$^{a}$ |

Note: Values followed by a different letter are significantly different based on Tukey's HSD (0.05). All values reported are Least Squares Means.
† Yield evaluations performed in eight total environments across two years with 3 replications per location.
‡ Maturity recorded in six environments.
§ Plant height recorded in five environments.
¶ Lodging: 1= erect plants and 5= prostrate within a plot. Lodging recorded in seven environments.
# Seed size recorded in six environments.
†† Seed quality ratings were recorded in six environments. The quality rating scale is based on the presence of cracked, split, wrinkled, discolored, or diseased seeds in a sample where 1=low presence and 5=high presence of defects.
‡‡ Protein and oil were recorded in six environments, and values are reported on a dry weight basis.
RP = recurrent parent
NIL = Near-isogenic line

3.3b)

| Trait | Source | DF | Sum of Squares | Mean Square | F Value | *Pr* > F |
|---|---|---|---|---|---|---|
| Yield | Genotype | 1 | 1.69 | 1.69 | 0.05 | 0.83 |
| | Environment | 7 | 354.72 | 50.67 | 1.37 | 0.26 |
| | G x E | 7 | 250.10 | 35.73 | 0.97 | 0.48 |
| | Rep(environment) | 8 | 238.30 | 29.79 | 0.81 | 0.60 |
| | Error | 24 | 885.20 | 36.88 | | |
| Protein | Genotype | 1 | 152.73 | 152.73 | 122.82 | 0.0001 |
| | Environment | 5 | 7.14 | 1.43 | 2.53 | 0.07 |
| | G x E | 5 | 6.22 | 1.24 | 2.20 | 0.10 |
| | Rep(environment) | 6 | 6.35 | 1.06 | 1.87 | 0.14 |
| | Error | 18 | 10.17 | 0.56 | | |
| Oil | Genotype | 1 | 33.60 | 33.60 | 245.23 | <.0001 |
| | Environment | 5 | 3.29 | 0.66 | 2.04 | 0.12 |
| | G x E | 5 | 0.69 | 0.14 | 0.42 | 0.83 |
| | Rep(environment) | 6 | 0.92 | 0.15 | 0.48 | 0.82 |
| | Error | 18 | 5.80 | 0.32 | | |

Note: Analyses performed using SAS program PROC GLM to estimate effect of each variable in a mixed model.

**Table 3.4:** Comparison of yield, protein content, and oil content between Benning HP and Benning at individual environments in 2015 and 2017.

| Location | 2017 | | | | | | 2015 | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Yield kg ha$^{-1}$ | | Protein g kg$^{-1}$ | | Oil g kg$^{-1}$ | | Yield kg ha$^{-1}$ | | Protein g kg$^{-1}$ | | Oil g kg$^{-1}$ | |
| | Benning | Benn. HP | Benning | Benn. HP | Benning | Benn. HP | Benning | Benn. HP | Benning | Benn. HP | Benning | Benn. HP |
| Watkinsville, GA | 3371.5 | 3844.5$^{NS}$ | 417.0 | 461.5**** | 200.2 | 178.7** | 2678.8 | 2425.5$^{NS}$ | - | - | - | - |
| Plains, GA | 3104.7 | 3194.4$^{NS}$ | 417.0 | 468.0**** | 204.8 | 182.4*** | 3295.3 | 3418.5 $^{NS}$ | 464.2 | 429.1*** | 201.3 | 221.1** |
| Bossier, LA | 3219.0 | 2591.4** | 416.4 | 456.1*** | 200.8 | 181.4** | - | - | - | - | - | - |
| Caswell, NC | 2972.5 | 2815.5$^{NS}$ | 435.8 | 463.0* | 189.2 | 175.5* | - | - | - | - | - | - |
| Plymouth, NC | 3107.0 | 3086.8$^{NS}$ | 412.7 | 462.4**** | 196.4 | 177.4*** | - | - | - | - | - | - |
| Florence, SC | - | - | - | - | - | - | 4458.7 | 4629.0$^{NS}$ | - | - | - | - |

Note: *, **, *** indicate significance at the level of *P*=0.05, 0.01, and 0.001, respectively. Protein and oil were only measured in the Plains, GA environment in 2015, due to late harvest resulting in poor seed quality in Watkinsville, GA and Florence, SC.

NS = not significant

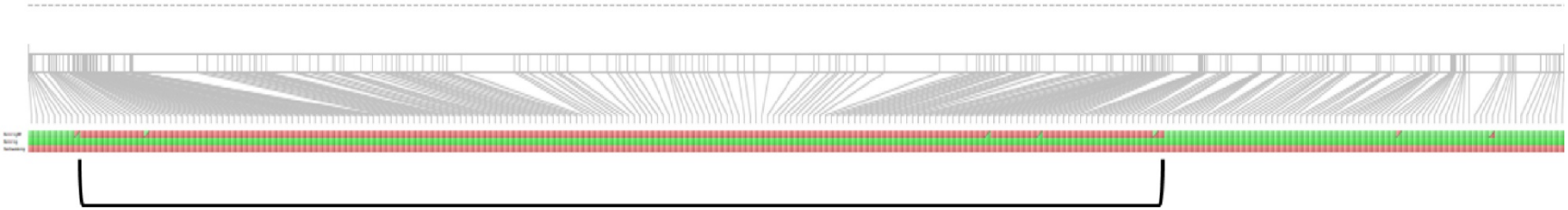

**Figure 3.1:** Graph displaying the Danbaekkong introgression into chromosome 20 of Benning HP.

The top, middle, and bottom rows show the Benning HP, Benning, and Danbaekkong genotypes, respectively. Green indicates SNPs matching the Benning genotype, and red indicates SNPs not matching the Benning genotype. The black bracket indicates the Danbaekkong introgression in the Benning HP NIL

# CHAPTER 5

## SUMMARY

Seed composition is a critical breeding goal due to the importance of protein, oil, and carbohydrates for the utility of soybean. Soybean is the largest source of vegetable protein meal consumption and second largest vegetable oil source worldwide. Advantages of the use of soybean meal include high levels of amino acids, good digestibility, and a reliable supply. However, protein content of recent releases has been decreasing, and the carbohydrate content of soybean is not ideal as a source of metabolizable energy. Negative relationships between seed components and yield complicate the improvement of seed composition while simultaneously breeding for high yield.

This research identified fast neutron mutants with increased protein or sucrose content in elite genetic backgrounds to be used as sources of improved seed composition traits in breeding programs, and the mutants also are a source of material for gene-function studies. Mutant G15FN-12 in particular contains a large deletion in the 15 to 21 Mb region of chromosome 12 that was identified through a combination of comparative genomic hybridization (CGH) and whole genome sequencing (WGS). Bulked segregant analysis of an $F_2$ elite × mutant population identified the Chr 12 deletion region as putatively associated with protein content. This deletion is approximately 5.5 Mb in size and contains 137 genes. Subsequent genotyping of the population using a deletion-specific marker indicated this deletion is associated with 27 g $kg^{-1}$ higher protein content than homozygous wild-type alleles.

Another goal was to characterize yield, seed composition, and agronomic performance of a near-isogenic line (NIL) with a high protein introgression across eight environments. The Chr 20 high-protein allele from a Korean tofu cultivar, Danbaekkong (PI 619083), was introgressed into an elite MG VII cultivar, Benning (PI 595645), at the University of Georgia. A near-isogenic line (NIL), designated Benning HP, was developed through backcrossing and yield-tested in 2015 and 2017. Results from eight environments over two years revealed yield parity between the recurrent parent, Benning, and the NIL, despite a 41 g $kg^{-1}$ increase in protein in the NIL. The NIL also contained 19 g $kg^{-1}$ less oil than Benning. This NIL may be used as a high yielding source of high-protein alleles for breeding improved soybean varieties.