

Assessment and Implementation of Value-Added Soybean Innovations for the Meal and Food Industry

Elizabeth Boadicea Fletcher

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

Doctor of Philosophy
In
Crop and Soil Environmental Sciences

Bo Zhang, Chair
A. Ozzie Abaye
David D. Kuhn
Gota Morota

April 24, 2025
Blacksburg, VA

Keywords: plant breeding, soybean, trypsin inhibitor, NIR, GWAS, CRISPR/Cas9, natto

Assessment and Implementation of Value-Added Soybean Innovations for the Meal and Food Industry

Elizabeth Boadicea Fletcher

ABSTRACT

Soybean (*Glycine max* [L.] Merr.) is a vital global crop for food and feed, but its value is constrained by consumer-driven quality traits and the presence of trypsin inhibitors (TIs) - antinutritional proteins that impair protein digestion in monogastric animals. This study employed a multifaceted approach combining phenotyping, genome editing, and association mapping to enhance soybean quality. First, we developed near-infrared reflectance spectroscopy (NIR) calibration models for the rapid estimation of total TI and Kunitz TI content in seeds and meals. Validated against HPLC data, these models demonstrated high predictive accuracy and offer a scalable tool for breeding and industry applications. Second, the CRISPR/Cas9-mediated knockout of *KTI1* and *KTI3* in the cultivar Williams 82 produced the low-TI line VTI5-26. Across multi-location field trials, VTI5-26 maintained reduced TI levels without compromising yield, maturity, or pest and disease resistance, and displayed significantly improved resistance to soybean cyst nematode. Third, a genome-wide association study (GWAS) of 146 natto soybean accessions identified novel SNPs associated with the key quality traits—water absorption, seed coat deficiency, cooked color, protein, and oil content—many of which were linked to stress-responsive gene networks. Collectively, these efforts illustrate how precision breeding and phenotyping tools can accelerate the development of soybean cultivars with enhanced nutritional and functional qualities.

Assessment and Implementation of Value-Added Soybean Innovations for the Meal and Food Industry

Elizabeth Boadicea Fletcher

GENERAL AUDIENCE ABSTRACT

Soybeans are an essential global crop for food and animal feed, but they naturally contain proteins called trypsin inhibitors (TIs) that can make it harder for animals to digest their food and grow properly. In this study, we explored new ways to improve soybean quality by reducing these harmful compounds and identifying traits important to food products like natto, a traditional Japanese fermented soybean dish. First, we developed a fast and cost-effective method to measure TI levels using near-infrared light instead of slower lab tests making it easier for farmers and food producers to check seed quality. We also used gene-editing technology (CRISPR/Cas9) to create a new soybean variety, VTI5-26, that significantly reduced Tis levels. This new variety grew just as well as other varieties in field trials and even showed better resistance to a common pest, the soybean cyst nematode. Finally, we evaluated over 100 natto-type soybeans and identified specific genes linked to desirable traits like protein content, oil levels, and seed appearance. These discoveries will support breeders to create soybeans that are healthier, easier to process, and better suited to different markets—benefiting both farmers and consumers.

Dedication

I dedicate this to my beloved parents – Mark Fletcher and Laura Fletcher (1968-2024).

Thank you for the continuous life-long outpouring of love and support.

Acknowledgements

I want to thank those who have made this long-awaited goal a reality. Thank you to my advisor Dr. Zhang and the Soybean Breeding team for their continued support academically and professionally. Thank you to the professors and mentors who advised, encouraged, and challenged me along the way. Finally, thank you to my family for their unwavering love and support.

Attributions

Chapter 1: Soybeans: The History of Genetic Understanding, Genetic Advancement and Soybean's Economic Importance

Elizabeth B. Fletcher: research and writing

Bo Zhang: editing

Chapter 2: Near-infrared reflectance spectroscopy calibration

Elizabeth B. Fletcher: data collection, analysis, writing

M. Luciana Rosso: conceptualization, data collection, editing

Troy Walker: sample preparation

Haibo Huang: sample preparation

Gota Morota: data analysis

Bo Zhang: conceptualization, editing

Chapter 3: Evaluation of Agronomic Performance, Biotic Stress Response, and Seed Quality in CRISPR-Edited Low Trypsin Inhibitor Soybean

Elizabeth B. Fletcher: data collection, analysis, writing

Luciana Rosso: data collection, editing

Usha Panta: data collection

Alejandro Rojas-Flechas: data collection

Emily Garant: data collection

Senyu Chen: data collection

Bo Zhang: conceptualization, editing

Chapter 4: Genome Wide Associate Study Uncovers Novel SNPs for Improving Natto-Specific Soybean Traits

Elizabeth B. Fletcher: data collection, analysis, writing

Jessica Wilbur: data collection, analysis

Zhibo Wang: analysis, writing

Jonathan Aims: data collection

Gota Morota: data analysis

Luciana Rosso: conceptualization, editing

Leandro Mozzoni: materials

Pengyin Chen: materials

Bo Zhang: conceptualization, materials, editing

Table of Contents

Dedication.....	iv
Acknowledgements.....	v
Attributions.....	vi
Table of Contents.....	vii
Chapter 1: Soybeans: The History of Genetic Understanding, Genetic Advancement and Soybean’s Economic Importance.....	1
1. History of Plant Breeding and Genetic Understanding.....	2
1.1 The Ancient Beginnings of Plant Breeding.....	2
1.2 Genetic Foundations.....	2
1.3 The Green Revolution.....	4
2. Advanced Techniques and Their Contribution to Improved Soybean.....	5
2.1 Genome Wide Association Study (GWAS)	6
2.2 Near Inferred Reflectance Spectroscopy	7
2.3 CRISPR/Cas9.....	8
3. History of Soybeans.....	9
4. Soybean Production in the U.S. and its Economic Importance.....	10
4.2 Soybean as Livestock Feed.....	12
4.2.1 Trypsin Inhibitor.....	12
4.3 Human consumption – Natto	14
5. Conclusion.....	15
Chapter 2: Near-infrared reflectance spectroscopy calibration	16
Abstract.....	17
1. Introduction	18
2. Materials and Methods	21
2.1 Plant Material	22
2.1.1 Whole Seed	22
2.1.2 Meal Preparation	23
2.2 Spectral Methodology	24
2.3 Trypsin Inhibitor quantification by HPLC	24
2.4 Model Creation, Cross-validation, and Statistical analysis	25
3. Results	25
3.1 Sample Concentration of Trypsin Inhibitor	25
3.2 Calibration Model Performance	26
4. Discussion	26
Tables	29
Figures	31
Chapter 3: Evaluation of Agronomic Performance, Biotic Stress Response, and Seed Quality in CRISPR-Edited Low Trypsin Inhibitor Soybean	34
Abstract.....	35
1. Introduction	36
2. Materials and Methods	38
2.1 Plant Materials	38
2.2 KTI Quantification	39
2.3 Evaluation of Agronomic Performance, Seed Quality	39

2.4 Evaluation of Pest Resistance	40
2.5 Evaluation of Soybean Cyst Nematode Resistance	41
2.6 Evaluation of Disease Resistance	42
2.7 Statistical Analysis	43
3. Results	44
3.1 KTI Concentrations	44
3.2 Agronomic Performance and Seed Quality.....	45
3.3 Pest Resistance	46
3.4 Soybean Cyst Nematode Resistance.....	46
3.5 Disease Resistance	46
4. Discussion	47
5. Conclusion	49
Tables.....	50
Chapter 4: Genome Wide Associate Study Uncovers Novel SNPs for Improving Natto-Specific Soybean Traits	53
Abstract	54
1. Introduction	55
2. Materials and Methods	56
2.1 Materials	56
2.2 Phenotypic Data Collection	57
2.3 Genotypic Data	57
2.4 Genome-wide Association Analysis and Candidate Gene Identification	58
3. Results	59
3.1 Phenotype Data	59
3.2 Significant SNPs and Candidate Genes for Natto Quality Traits	59
3.3 Significant SNPs and Candidate Genes for Seed Composition Traits	60
4. Discussion	60
4.1 Phenotype	61
4.2 Significant SNPs and Candidate Genes for Natto Quality Traits	62
4.3 Significant SNPs and Candidate Genes for Seed Composition Traits	63
5. Conclusion	65
Tables	67
Figures	70
References	73

Chapter 1: Foundations and Frontiers: The Evolution of Soybean Genetics, Breeding Technologies, and Economic Impact

Elizabeth B. Fletcher¹, Bo Zhang¹

¹ School of Plant and Environmental Sciences, Virginia Tech, Blacksburg, VA, United States

1. History of Plant Breeding and Genetic Understanding

1.1 The Ancient Beginnings of Plant Breeding

The domestication of plants was an integral part of human advancement and civilization. The natural process of plant selection evolved into a relationship of dependence between people and plants. Humans made a transition from being hunter-gatherers to becoming an agronomic society, which occurred about 10,000 years ago [1]. This age saw the transition from an ice age into more stable environments which enabled permanent settlements and the development of agricultural practices. Plant breeding started with choosing the most easily harvestable plant varieties [2]. A naturally occurring plant mutation that resulted in easier shattering would be more likely chosen for harvest and thus transmit its genetic material to successive generations. As selection processes became more intentional, people discovered the potential to control crop genetics.

1.2 Genetic Foundations

The establishment of the dedicated field of plant breeding started with the exploration of genetic principles during the 19th century. The advancement of heredity knowledge combined with selection and genetic modification techniques has dramatically enhanced agricultural productivity together with global food security and crop resilience [2].

Gregor Mendel earned his reputation as the Father of Genetics through his research on pea plants conducted in his monastery garden. In 1866 he published his findings which established the basis for inheritance laws [3]. Mendel observed how specific traits transferred between successive generations of pea plants after implementing specific

breeding techniques. He recorded characteristics of each parent and offspring, specifically focusing on stem length (tall or short), pod color (yellow/green) and shape (inflated/constricted), pea color (yellow/green) and shape (round/wrinkled), and flower color (purple/white) and flower position (axil or terminal). Mendel tracked these traits throughout successive generations which allowed him to propose three principles of inheritance in 1865 [4]:

- 1) The Law of Segregation – each parent separates their genes during gamete production and these genes randomly go to their offspring
- 2) The Law of Independent Assortment – Each gene functions independently when determining the inheritance of other genes
- 3) The Law of Dominance – The dominant allele covers up the expression of phenotypic traits from recessive alleles

The world forgot about Mendel's work until 1900 when three botanists Hugo de Vries, Carl Correns and Erich von Tschermak independently confirmed his findings. The three botanists performed their experiments on heredity without knowing about each other or Mendel's previous research. The discovery of Mendel's work after 1900 gave him credit for the findings while Vries, Correns and Tschermak used their work to prove and validate his Laws of Inheritance [5]. The rediscovery introduced plant breeding genetics which enabled scientists to choose beneficial traits through systematic selection.

During the early 1900s Wilhelm Johannsen used Mendel's laws to create the pure-line theory [6]. Through his bean research he proved how self-fertilized plants create genetically identical strains while demonstrating the essential role of selection in

agricultural improvement. According to Johannsen, the foundation of breeding programs today depends on genetic uniformity [7]. But pure line maintenance failed to deliver the production enhancements farmers and consumers needed for growing markets.

Plant breeding underwent a transformative shift with hybrid breeding techniques that emerged during the early 20th century. George Shull observed heterosis (hybrid vigor) by performing crosses between different inbred lines to create superior hybrids which surpassed their parent lines [8]. Hybrid corn varieties entered commercial markets during the 1920s which led to significant crop yield growth and established the advantages of controlled crossbreeding.

1.3 The Green Revolution

The mid-20th century represented a transformative period in agricultural history worldwide. The period from 1940 to 1960 represented the Green Revolution which transformed crop development and worldwide food production into a major success story [9]. Traditional farming systems in Mexico, India and Pakistan operated at low productivity levels during that time because of losses due to pests, disease, and unpredictable weather. The restrictions in agricultural production created regular food deficits while the increasing population numbers intensified the concern [10]. Dr. Norman Borlaug emerged as a leading figure of this era through his work to support Mexican wheat farmers. Through Rockefeller Foundation support Borlaug created wheat breeds that yielded more and maintained short compact growth which reduced their susceptibility to field breakdown. His innovative shuttle breeding technique served as the main differentiator because it both accelerated crop breeding time and enhanced environmental adaptability [9].

Shuttle breeding requires breeding lines to undergo simultaneous growth in two completely different environments throughout the year. Borlaug tested his wheat varieties in both the hot, lowland Sonora Valley and the cool, high-altitude Toluca Valley in Mexico. The method involved moving wheat plants between two environments for one year, which enabled Borlaug to find suitable varieties that could adapt to different conditions while shortening the time needed to create stable new strains [11].

The results were game-changing. Food production levels in struggling countries increased substantially because of these innovations, which enhanced the availability of a stable food supply. Through his work Borlaug brought transformative changes to the whole field of plant breeding. Current plant breeding practices including accelerated breeding cycles and testing across multiple locations and climate resilience selection directly stem from the foundational work of Borlaug.

2. Advanced Techniques Driving Soybean Improvement

Scientists from the 19th and 20th centuries established the basis of our modern understanding of genetics along with advanced genetic techniques. Molecular genetics emerged as a scientific field in the closing years of the twentieth century to enable researchers to achieve precise gene alterations, which benefit crop improvement. Genome-Wide Association Studies (GWAS), Near Infrared Reflectance Spectroscopy (NIR), and Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR/Cas9) represent some of these techniques among others.

2.1 Genome Wide Association Study (GWAS)

GWAS has emerged as a fundamental method to explore the genetic roots of complex crop traits. GWAS evaluates naturally occurring plant genetic variation between different populations to detect genome locations affecting essential traits, which enables breeders to select those traits precisely and effectively [12].

GWAS relies fundamentally on the principle of linkage disequilibrium which describes how particular genetic markers tend to pass through inheritance together. GWAS examines tens of thousands of genetic markers (SNPs) to determine associations between genome segments and observable traits. This tool has allowed researchers to analyze population genetic diversity to provide detailed information about trait variation [13].

In plant breeding, GWAS is a powerful tool to discover genes responsible for traits including drought tolerance, disease resistance, seed quality and other traits. GWAS has proven useful in soybeans for identifying genes that influence seed protein content together with oil composition and resistance to soybean cyst nematode, which significantly impacts yield.

identification of genetic markers through research enables breeders to speed up variety improvement with increased focus on their program objectives [14]. By providing data on traits and associated SNPs, genomic selection models are improved, increasing their ability to predict which progeny will show promising field performance. This technology cuts down the requirement for thorough field testing across numerous lines which reduces costs while saving time and land [15].

2.2 Near Infrared Reflectance Spectroscopy

Scientists and breeders have transformed their plant material analysis with the fast and non-invasive Near-infrared reflectance spectroscopy (NIR). Researchers use near-infrared reflectance spectroscopy (NIR) to measure light absorption and reflection within the 780-2500 nm range which enables them to analyze seed and tissue compositions without sample destruction [16].

The fast analysis of protein, oil, moisture, starch and fiber content in plant samples becomes possible through the use of NIR [17]. Soybean breeders utilize NIR technology for routine measurements of seed protein and oil content because these traits determine both nutritional value and market price. Traditional lab-based analysis of these compounds requires time-consuming methods that destroy samples but NIR technology enables fast analysis of numerous samples without damaging them and keeps seeds viable for planting or additional research [18].

Modern breeding programs have widely adopted NIR technology as a fundamental component for fast and large-scale high-throughput data collection. New portable NIR devices help users perform field and harvest site analysis by advancing their on-site analytical capabilities [19]. Researchers have now linked NIR data with genomic information using machine learning and chemometric modeling to improve selection decisions for nutritional components [20]. This adoption of NIR can simplify the evaluation of complex traits. As a result the process becomes streamlined and requires minimal time and resources while enabling informed breeder selections, which speeds promising lines through the breeding pipeline.

2.3 CRISPR/Cas9

Plant genetic improvement techniques have experienced a fundamental shift because of CRISPR/Cas9. The bacterial immune system gene-editing tool enables researchers to make precise DNA modifications at specific locations within plant genetic material [21]. The DNA editing tool CRISPR/Cas9 outperforms earlier gene editing systems such as zinc finger nucleases and TALENs, in both speed and accuracy [22].

CRISPR provides researchers with the ability to perform specific gene modifications which avoids adding foreign genetic material. Developers can create non-transgenic crops through this method which reduces both regulatory challenges and public resistance. CRISPR allows multiplex editing to modify various genetic regions simultaneously which proves beneficial for traits with multiple genetic regulators.

The potential of CRISPR in soybean breeding has been clearly demonstrated in recent studies. Scientists have used CRISPR technology to create high-oleic soybean varieties which produce healthier fats with better oxidative stability, benefiting food processing [23]. Researchers have also employed CRISPR technology to build resistance against soybean mosaic virus and to lower trypsin inhibitors [24, 25]. These genetic modifications can lead to improved yield levels, better product quality, and more environmentally friendly practices.

The future development of DNA-free editing together with nanoparticle delivery, speed breeding and genomic prediction tools will enhance CRISPR's capabilities. The improvement of these systems will increase their usage in creating soybean varieties that become more resilient while providing better nutrition and productivity levels to fulfill changing global demands.

3. History of Soybeans

The modern soybean species (*Glycine max* [L.] Merr.) that serves as a global food and crop originated from the natural species *Glycine soja* (Siebold & Zucc.), which inhabits East Asian regions. The transition of wild plants into essential crops started when humans first domesticated plants between 6,000 and 9,000 years ago [26]. This process started at a fundamental period in human development when early agricultural societies began to actively cultivate and enhance their plant species. During the Eastern Zhou Dynasty, China made substantial agricultural developments approximately 2510 BCE[27].

Studies combining archaeological evidence with genetic analysis have revealed more accurate details about soybean origin and the development of its domestication. Recent studies indicate that the Yellow River basin in China presents a primary location for the origins of cultivated soybean. Genetic analysis reveals that local soybean varieties from this area exhibit higher diversity levels than other strains, indicating possible early cultivation and improvement of wild soybeans [28]. The discovery of burned soybean plant remains near ancient fire pits, further supports the timeline of prehistoric people consuming this crop [29].

The semi-wild soybean species *Glycine gracilis* Skvortz. holds a mysterious place in the story of soybean domestication. Fukuda suggested in the 1930s that *Glycine gracilis* represented a phase between *Glycine soja* and *G. max* [30], leaving scientists to debate the evolutionary significance of the soybean variety for several decades, although the correct position in the timeline remained unknown.

Genomic research advancements have made it possible to clarify this picture in recent years. Han et al. conducted an extensive analysis of 500 soybean samples including their

wild, semi-wild and domesticated varieties [31]. The research revealed surprising results that *G. gracilis* possessed genetic traits which differentiated it from its close relatives and thus disproved the previous theory proposed by Hymowitz in 1970 that *G. gracilis* was a hybrid [30]. The current data indicates that *G. gracilis* occupies a separate and individual position in the domestication sequence, which provides new insights into the developmental history of soybean.

The history of soybean domestication shows the deep bond between human beings and plants. Human selection of a wild species over thousands of years resulted in the development of one of the most globally cultivated crops. Understanding the origin of soybean and its evolutionary development provides valuable lessons for future crop improvement to address present-day challenges such as climate change and global food security.

4. Soybean Production in the U.S. and its Economic Importance

Soybeans (*Glycine max* [L.]) are the most widely grown oilseed crop globally, which demonstrates their fundamental importance in agriculture and commerce [32]. Current data shows soybeans make up about 59% of total worldwide oilseed production [32]. The many uses of soybeans and their growing market demand have established their crucial role in numerous industries, ranging from livestock feed, human consumption to biofuels and related products.

Soybeans function as the foundation of American agricultural operations. The total soybean plantings reached 87.5 million acres in the United States during 2022, while the average yield per acre reached 49.5 bushels [33]. A total production figure of 4.28 billion

bushels resulted from 2022 crop, despite a 4% reduction from the previous year. The United States remains one of the leading soybean-producing nations together with Brazil and Argentina [33].

Soybean cultivation within the United States generates a considerable amount to the economy. According to market data for 2023, the U.S. soybean crop generated \$60.7 billion in total value [34]. The United States relies heavily on soybeans both for local production and international trade purposes. During the 2022/2023 marketing year U.S. soybean exports reached 2.21 billion bushels, which represented 52% of the total production [34]. China leads the way as the primary destination for U.S. soybeans followed by the European Union, Mexico and Japan.

Soybean production represents an essential part of Virginia's agricultural industry as farmers in Virginia grew soybeans on approximately 570,000 acres during 2023. The projected harvested area stood at 560,000 acres and had an estimated soybean production at 21.1 million bushels, at an average yield of 41.0 bushels [35].

The uses of soybeans extend into multiple areas. Approximately 80% of processed soybeans are used for animal feed production, mainly serving poultry, pigs and cattle operations. Industrial applications and human nutrition account for the remaining portion of the global soybean production including biofuels, lubricants, and other industrial products.

4.2 Soybean as Livestock Feed

Soybeans (*Glycine max*) function as a vital animal feed source across the globe because they contain abundant protein and beneficial amino acids. Current worldwide

statistics show that between 70 and 80% of produced soybeans become soybean meal, which serves primarily as animal feed [33]. Soybean meal is an essential component in feed formulations for poultry and pigs, and also serves ruminant animals because of its digestible nutrient profile [36]. Soybean meal is widely used in animal feed because of its reliability, adaptability, and affordability compared with other protein sources. Additionally, soybean-based feeds support optimal growth and, thus, are fundamental to large-scale animal agriculture.

Soybean meal mainly goes to poultry production, followed by swine and then dairy and beef cattle. Soy cakes or flakes are the standard soybean meal processing forms before they get added to compound feed products. Soybeans that are not processed into meal and are instead fed to livestock directly only make up a very small portion of the total soybean consumption because processing enhances nutritional value and reduces the anti-nutritional factor - trypsin inhibitor [36].

4.2.1 Trypsin Inhibitor

Although soybeans are well-suited for livestock feed due to their protein content and digestibility, they are negatively affected by naturally occurring antinutritional factors, particularly trypsin inhibitors (TIs). These compounds hinder digestion by covalently binding to trypsin, a digestive enzyme, thus inhibiting protein digestion and absorption in animals [37 - 39]

High levels of TIs in animal diets can cause several adverse physiological effects. At low doses, they may impair protein metabolism and growth and feed conversion. In

higher doses, animals may develop abnormal organ enlargement, including the pancreas, liver, and intestine, and in some cases, may suffer from pancreatitis [40 - 42].

Soybeans contain two major forms of trypsin inhibitors: Kunitz type (KTI) and Bowman-Birk type (BBTI). Kunitz is a larger 21.5 kDa protein and is a very stable, and in many cases, irreversible trypsin inhibitor [18,19]. In contrast, BBTI is much smaller, 7 to 8 kDa, and has many isoforms. It can inhibit both trypsin and chymotrypsin and thus has a higher inhibitory action [45, 46].

Soybeans are usually heated to reduce the antinutritional effects of TIs before being used in animal feed. This practice has been employed for over a century, with benefits observed even before its scientific rationale was fully understood. Today, conventional industrial processing of soybeans involves heating them to 121°C for 15 minutes to inactivate TIs [47]. Nonetheless, there are drawbacks to this method. It can be quite costly and may result in the loss of 5-20% of protein available because of heat destruction [48].

Although undesirable in animal diets, TIs play a crucial role in protecting soybeans from insect pests. The problem of breeding to lower the levels of TIs is complicated by the fact that such breeding may result in the loss of pest resistance and reduced yield [49].

4.3 Human consumption - Natto

Although a large part of the world's soybean production goes towards livestock feed and biofuel sectors, a growing percentage—13.91 million tons in recent data—goes towards human consumption [50]. Global soy product consumption has steadily increased, with per capita consumption increasing from 1.29 kg in 2010 to 1.77 kg in 2021 [51]. Much

of this food-grade soy is consumed through traditional soy-based foods like tofu, soy milk, and natto.

Natto, a traditional Japanese food product, is produced through the fermentation of cooked soybeans (*Glycine max* [L.]) with the bacterium *Bacillus subtilis* [52]. It is known for its sticky texture, pungent smell, and chewiness, and has been consumed in Japan for over five centuries. Although it has an unorthodox appearance and taste, it is valued for its health-promoting properties, which include lowering cholesterol, strengthening bones, and boosting the immune system [53]. A 100-gram serving of natto provides about 211 calories and 19 grams of protein, making it a nutritious and filling food [54]. Apart from protein, natto is a good source of bioactive compounds and other nutrients including vitamin K, isoflavones, biogenic amines, and the enzyme nattokinase [54]. Nattokinase has been identified as a potential drug candidate with properties to help lower blood pressure, act as an anticoagulant, improve vision, and reduce inflammation [55-56].

Recently, the health benefits of natto have been recognized in the broader world, and as such, it is becoming a globally sought-after item in the health food industry outside Japan. The growing market demand for natto has resulted in increased production and highlighted the need for continued soybean breeding to meet evolving consumer preferences and processing requirements [57].

5. Conclusion

Soybean functions as a fundamental crop worldwide because of its dual value as a protein-rich food source and versatile animal feed and a growing importance in human nutrition. The evolution of Plant breeding is evident in the domestication and improvement

of soybean over thousands of years, progressing from early selection methods to current genetic tools. The advancement of soybean cultivars to their current high-performance state resulted from Mendelian genetics and the Green Revolution, which established a foundation for modern breeding practices. Soybean yield potential alongside environmental resistance and adaptability have been directly affected by innovations in hybridization and shuttle breeding.

GWAS, together with NIR and CRISPR-Cas9 genome editing have accelerated soybean improvement through modern breeding techniques. These tools enable scientists to identify and select precise, valuable traits including seed composition and stress tolerance and reduced antinutritional factors such as trypsin inhibitors. Soybean breeding programs will need to use advanced technologies to create new, nutritious, high-yielding varieties if soybean demand keeps expanding in animal feed and health-conscious human markets. The future of soybean breeding will rely on a foundation of classical knowledge integrated with molecular innovations to overcome complex global food and feed system challenges.

Chapter 2: Near-infrared reflectance spectroscopy calibration

Elizabeth B. Fletcher¹, **M. Luciana Rosso**¹, **Troy Walker**², **Haibo Huang**², **Gota Morota**³ and **Bo Zhang**¹

¹ School of Plant and Environmental Sciences, Virginia Tech, Blacksburg, VA, United States

² Department of Food Science and Technology, Virginia Tech, Blacksburg, VA, United States

³ Department of Agricultural and Environmental Biology, University of Tokyo, Tokyo, Japan

Abstract

Trypsin inhibitors (TI) are naturally occurring antinutritional factors found in soybean seeds [*Glycine max.* (L.)], that decrease the growth rate of livestock, causing malnutrition and digestion troubles. The current accurate method to quantify TI levels in soybean seeds or meals is by High Performance Liquid Chromatography (HPLC); however, it is time-consuming, creating bottlenecks in industrial processing. Establishing a near-infrared reflectance spectroscopy (NIR) model for estimating TI in seeds and meals would provide a more efficient and cost-effective method for breeding programs and feed producers. In this study, 300 soybean lines, both seeds and meals, were analyzed for TI content using HPLC, and calibration models were created based on spectral data collected from a Pertenda 7250 NIR instrument. The resulting models demonstrated robust validation, achieving accuracy rates of 97% for seed total TI, 97% for seed Kunitz TI, and 89% for meal total TI. The findings of this study are significant as no NIR calibration models had previously been developed for TI estimation in soybean seed and meal. These models can be used by breeding programs to efficiently assess their lines and by industry to quickly evaluate their soybean meal quality.

1. Introduction

Soybean [*Glycine max* (L.)] is a highly valued row crop, primarily due to its high protein and oil content. The high protein content, balanced amino acid profile, and various vitamins and minerals contained in a soybean has made it a staple in animal feed production [1,2]. In the United States, 70% of soybeans produced are destined for animal feed – totaling 33.12 million tons in 2021 [59]. Despite its nutritional benefits, soybean contains antinutritional factors such as trypsin inhibitors (TI), that inhibit the absorption and digestibility of nutrients in livestock [2-6]. When livestock consume significant amounts of TI, it can result in severe health implications, such as limited growth because of the limited protein and nutrient absorption. In more severe cases, overconsumption of TI may lead to the enlargement of the pancreas, liver, intestines, and even result in pancreatitis [7-10]

TI is a naturally occurring protein found in legume species that binds strongly to trypsin, a digestive enzyme in animals, effectively blocking its active site and hindering the digestion process [9,11]. Soybean has two types of TI– Kunitz trypsin inhibitor (KTI) [43] and Bowman-Birk trypsin inhibitor (BBTI) [11,13,14]. The KTI is monomeric protein in soybean seeds, consisting of 181 amino acids and 21.5 kDa (Kunitz, 1947). It is more abundant than BBTI and forms a strong, irreversible complex with trypsin [44]. In contrast, BBTI is a smaller protein, 7-8 kDa, with various isoforms. Present in lower quantities, BBTI inhibits both trypsin and chymotrypsin [16-18].

To negate the antinutritional effects of TI in soybeans, the raw seeds are heated and processed into a meal. The benefits of heating soybeans were first observed in 1917, and the practice was widely adopted even though the reasons behind its improvement of the

feed's nutritional value were not yet understood [48]. In recent years, heating raw soybean to 121°C for 15 minutes is the industry's standard practice to deactivate TI within soy meal for livestock consumption [47]. However, this process requires careful control - overheating the soy meal can destroy beneficial nutrients – while insufficient heating may leave TI concentrations too high for animal consumption. As such, TI concentration and nutritional value of the meal is regularly checked to maintain a quality product.

To date, numerous methods for detecting TI in soybeans have been published [17, 21-24]. However, these existing techniques for measuring TI concentrations are often time-consuming. High performance liquid chromatography (HPLC), an analytical chemistry technique that separates compounds within a mixture for quantification, is a widely used and efficient method at the time of this study [22, 24]. Although HPLC is known for its accuracy and sensitivity, it has several disadvantages. The HPLC method is time-consuming, requires extensive sample preparation, and demands costly maintenance, including expensive reagents, specialized equipment, and skilled personnel, particularly when analyzing complex protein mixtures. The initial equipment cost of a high throughput system, such as the one used in this study, can range from \$40,000 - \$120,000 [69]. The estimated cost in 2025 is \$1.89 per sample [18]. Approximately 13 minutes per sample is required for HPLC analysis, not including the time needed to grind and prepare samples. These factors make HPLC less practical for high-throughput or routine analysis. Given these limitations, there is a growing need in the animal feed industry for more efficient and practical methods to measure TIs. The creation and adoption of a near infrared reflectance spectroscopy (NIR) calibration for quantifying TIs on an industrial scale would allow for a more time-efficient evaluation of a soybean seed and meal.

Near-infrared reflectance spectroscopy (NIR) offers a promising alternative for industrial-scale quantification of TIs. Benchtop models, such as the one used in this study, range from \$15,000 - \$60,000 [70]. NIR is a secondary analytical technique that quickly identifies and evaluates chemical compositions of a substance [71]. This is achieved by emitting near-infrared light (700 -2500nm) onto a sample, where the wavelengths are absorbed based on the C-O, C-H, and C-N chemical bonds present, and the resulting spectral data are then used to determine composition of various chemical compounds based on pre-existing calibrations [72]. NIR spectroscopy is a non-destructive, fast, and cost-effective method for analyzing agricultural products quality, requiring minimal sample preparation and providing results in approximately 20 seconds, thereby reducing the need for costly and time-consuming laboratory testing.

It was first developed in 1946 as part of a project funded by the US Department of Agriculture to develop a method for grading eggs [73]. Since then, it has become a staple instrument used in breeding programs to evaluate lines for advancement based on desired traits and in the food processing industry to evaluate quality of a product [74]. NIR's application in agriculture has expanded with successful calibrations for various compounds including moisture, protein, and oil content, demonstrating its versatility and efficiency [75]. For soybean quality assessment, AACC International (formerly the American Association of Cereal Chemists) currently recommends the NIR method for analyzing protein, crude fat, and moisture content in soybean seeds [76].

The specific wavelengths utilized and available calibrations for NIR analysis vary depending on the manufacturer and the instrument used. In this study, raw spectra data were collected using a DA7250 NIR instrument manufactured by Perten Instruments. The

DA7250 NIR can predict the concentration of a sample in under 10 seconds by utilizing a samples absorption within the 950-1650 nm wavelength range. To our knowledge, no calibration specifically designed for the determination of TI in whole soybean seeds and soybean meal has been developed for the DA7250 or similar NIR models. Developing accurate calibrations for this instrument would significantly reduce the time required to measure TI concentrations in whole seeds and processed meals, thereby leading to a more efficient production process.

As such, the objectives of this study were to (a) to develop a DA7250 NIR calibration for the accurate and rapid prediction of TI concentration in whole soybean seeds to be used by breeding programs for evaluating lines potential, and (b) to develop a DA7250 NIR calibration for the accurate and time-saving prediction of TI concentration in soybean meal, enhancing efficiency for industrial feed producers.

2. Materials and Methods

2.1 Plant Material

2.1.1 Whole Seed

A total of 300 soybean plant introductions (PI) were selected from a diverse USDA soybean germplasm collection (Singer et al., 2022). The 300 samples consisted of soybean in maturity groups 4 and 5, representing diverse origins (China, Japan, USA, Russia, Nepal, Korea, Vietnam, and Morocco) [77] They were grown in 3 m two-row plots with 76 cm row spacing and harvested in Blacksburg, Virginia in 2020 and 2021. The 300 samples were selected to represent a diverse range of trypsin inhibitor (TI) concentrations, including low, mid, and high concentrations of KTI, BBTI, and Total TI (TTI) based on data obtained by HPLC analysis following Rosso et al., 2018. TTI was calculated as the sum of KTI and

BBTI concentrations per sample. The inclusion of these samples allowed for a model creation that represented naturally occurring TI ranges in soybeans.

2.1.2 Meal Preparation

For analysis of soybean meal, meal samples were prepared from the 300 PI lines. First, the starting moisture content of each sample was determined by placing 3 g of whole seed in an oven set to 103 °C for 72 hours. After each sample cooled, the final mass was recorded, and moisture content calculated. The whole seeds were then cracked and dehulled by placing 20 g of each sample in a hopper with a roller mill. The moisture content for the remaining soybean meat was readjusted to 15% using the following equation:

$$0.176 * (X) - 1.176 * (Y) * (X) = Z$$

X = Dehulled mass (grams)

Y = Moisture content in decimal form (not percentage)

Z = DI water to be added (milliliters)

Following the addition of the determined amount of water, the samples were then set in an incubator at 65 °C for 15 minutes (Okedigba et al., 2023). The samples were then added to a roller mill for the creation of flakes. Solvent extraction was performed on the resulting flakes. Each sample was run on a Dionex ASE 350, set to the following method: 65 °C, static time 15 minutes, 3 cycles, solvent B-Hexane. After the completion of the run, the resulting samples were set under a fume hood overnight to allow for evaporation of the solvent. Each sample was then stored in an airtight bag.

2.2 Spectral Methodology

Approximately 50 g of whole soybean seeds were selected from the 300 lines. Each of the 300 samples were individually placed in the small seed breeding tray and scanned

on the Perten DA7250 NIRS instrument (Perten Instruments). The spectral data recorded for each sample consisted of 141 datapoints collected across a wavelength range of 950 – 1,650 nm (Operation and Handling - DA 7250TM NIR | PerkinElmer, n.d.). A complete spectral dataset from the 300 samples was then exported from the instrument to be used for model creation. The same method was repeated with approximately 50 g of soybean meal and the spectra data exported.

2.3 Trypsin inhibitor quantification by HPLC

The HPLC method used to quantify TI in both the untreated seed and the meal was conducted following the procedure previously developed by Rosso et al. (2018). Briefly, 10 mg of finely ground soybean seed powder was mixed with 1.5 mL of 0.1 M sodium acetate buffer (pH 4.5). Samples were vortexed and shaken for 1 h at room temperature. The sample was centrifuged at 12,000 rpm for 15 min. One-mL of the supernatant was filtered through a syringe with an IC Millex-LG 13-mm mounted 0.2-mm low protein binding hydrophilic millipore (polytetrafluoroethylene [PTFE]) membrane filter (Millipore Ireland). The TI in solution was separated on an Agilent 1260 Infinity series (Agilent Technologies) equipped with a guard column (4.6 x 5 mm) packed with POROS R2 10-mm Self Pack Media and a Poros R2/H perfusion analytical column (2.1 x 100 mm, 10 μ m). The mobile Phase A consisted of 0.01% (v/v) trifluoroacetic acid in Milli-Q water, and the mobile Phase B was 0.085% (v/v) trifluoroacetic acid in acetonitrile. The injection volume was 10 μ L and the detection wavelength was 220 nm.

2.4 Model Creation, Cross-validation, and Statistical analysis

The method for model creation, cross-validation, and statistical analysis followed the procedure reported by Lord et al., 2021 [78]. The CAMO Unscrambler X software

(CAMO Analytics AS) was used for the spectroscopic data pretreatment, model creation, and internal cross-validation of model. A representative subsample of 124 seed and 112 meal samples were selected using R (R Core Team, 2024), ensuring an equal distribution of low, mid, and high concentrations of each of the six TI values (mg/g): STTI: seed total trypsin inhibitor, SKTI: seed Kunitz trypsin inhibitor, SBBTI: seed Bowman-Birk trypsin inhibitor, MTTI: meal total trypsin inhibitor, MKTI: meal Kunitz trypsin inhibitor, and MBBTI: meal Bowman-Birk trypsin inhibitor. The spectroscopic data for each of the six datasets was then pretreated first with standard normal variation, followed by detrending which corrected for light scatter and particle size. Models were created using the transformed data and partial least squares regression (PLSR) based on previous studies [34, 35]. The number of PLSR components, 10, was determined based on the number of factors that minimized the predicted residual error sum of squares. To perform a 10-fold cross-validation, the samples were randomly divided into 10 equal segments, each consisting of 11 or 12 samples. Unscrambler performed cross-validation by holding out the samples randomly placed in a segment. The model was then recalibrated without the selected samples and the recalibrated model was used to predict the values of the withheld samples. This was repeated for each segment, until all samples had been withheld. Each model resulted in an R^2 and a root-mean-square-error (RMSE) value for the calibration and cross-validation of each model. The R^2 demonstrates statistically how well the TI concentrations determined by HPLC match the concentrations predicted from the spectral model. The RMSE values represent the average error present within the model[81].

3. Results

3.1. Sample Concentration of Trypsin Inhibitor

As expected, the seed total trypsin inhibitor (STTI) subsample had the highest concentration of TI, as its seeds were raw and without any heat treatment. The total TI in seeds ranged from 2.62 – 13.13% of the total seed content (Table 1). TI concentration in soybean varieties have been reported to range from 0.07-18.7%, with anything less than 6% being considered low and greater than 10% high concentration [82]. Therefore, the broad range of TI concentrations in our subsample set effectively captured the diversity found in soybean germplasm, providing a robust basis for calibration model development. However, the preparation of soybean meal resulted in a notable decrease in total TI concentration by at least 2 mg/g across all samples, reflecting the effectiveness of processing in reducing antinutritional factors. The reduction in TI was expected, given the heat treatment applied during meal preparation, which is known to denature these proteins.

3.2. Calibration Model Performance

The calibration and validation statistics for each model are presented in Table 2, and the model fit is shown in Figure 1. The STTI and SKTI validation models both resulted in a 97% rate of accuracy, based on the R^2 values. Given that the STTI range was 2.62 – 13.13% and the model's RMSE of 1.579, indicating a 13.7% error rate in the validation model. This is considered a moderate range of error and may be acceptable depending on the intended application of the model. The SKTI validation model had a more acceptable error rate of 10.3% calculated from the reported RMSE value of 0.74. The meal total trypsin inhibitor (MTTI) resulted in a moderately successful validation model as well with an 86% accuracy rate and a moderate error rate of 15%.

In contrast, the seed and meal Bowman-Birk trypsin inhibitor (SBBTI and MBBTI) models performed poorly, with validation R^2 values of only 0.017 and 0.016, respectively. As shown in Figure 2, BBTI averaged 47% of STTI and 27% for MTTI composition. Additionally, the low performance for the MKTI model may be due to the reduced concentration of TI by the applied heat treatment, reducing the levels of KTI present to values undeterminable by the NIR instrument used in this study.

The summary of the spectral data's contribution to the successful models is shown in Figure 3. In the three graphs, upward peaks represent a portion of wavelength that had a positive correlation to the model creation. The downward peaks have a negative correlation and interference with the model creation. In each of the successful models, the wavelength range of 1450-1470 nm was significant, indicating that the chemical structure of TI may absorb this range of infrared wavelengths. The wavelength range 1410-1430nm had a negative correlation with the model creation for both STTI and SKTI. This close range between the negative and positive peak may also be a reason for the error rate in the validation models. The STTI model had a high peak at 1390 – 1410nm, not present in the other models.

4. Discussion

The model statistics shown in Table 2 suggest that the NIR calibration models for STTI, SKTI, and MTTI can be reliably used for rapid TI quantification in soybean breeding and industrial applications. In contrast each attempt at a BBTI model was unsuccessful, suggesting that the chemical composition of BBTI may not be suitable for accurate NIR prediction of BBTI concentrations. The molecular size of BBTI in soybeans is approximately 8 kDa [32-34], which is much smaller than that of KTI (approximately

20 kDa). In addition to its smaller molecular size, BBTI consists of seven disulfide bonds, while KTI has only two [84].

Other organic compounds consisting of multiple disulfide bonds, similar to BBTI, have been reported to correspond peaks around 1700 nm [86], which lies beyond the spectral range of the DA7250 instrument (900 nm – 1650 nm). Additionally, the low performance of BBTI may have negatively impacted the overall accuracy of the STTI calibration model, contributing to the 4% decrease between STTI and SKTI. Despite the poor performance of the MKTI and MBBTI models alone, the MTTI models may have been successful in part due to the higher combined concentration of the total TIs compared to the lower concentrations of the individual TI samples.

The wavelength range identified in Figure 3, does correlate to a wavelength range previously reported in a study that created an NIR calibration model on another instrument for soy cakes [87]. In this study by Hoffmann, they identified two wavelength ranges related to TI that are beyond the range analyzed by the Perten DA7250 (1640-1830nm and 2100 – 2300nm). A study performed in 2009 also identified peaks in wavelengths greater than 1650nm for heat-treated trypsin inhibitor activity [88]. This suggests that the error rates and the unsuccessful attempts to create a model for SBBTI, MKTI, and MBBTI may once again be due to the limitations of the instrument used in this study. However, given the popularity of the Perten DA7250 instrument, our models are expected to be broadly adopted by soybean breeders and meal processors.

In conclusion, three models – STTI, SKTI, and MTTI were successfully developed and validated. The novelty of this study's findings offer a rapid and cost-effective alternative to traditional HPLC methods that was not previously available. Although

methods such as HPLC will still remain necessary for precise quantification of TI within a seed or meal sample, the Perten DA7250 instrument provides a valuable tool for quick evaluation of a soybean seed for a breeding program and the meal for an industrial feed producer. Future directions include expanding model development to other commonly used benchtop NIRs for further accessibility. Additionally, further research focused on developing accurate models for BBTI quantification using NIR with spectral ranges accommodating BBTI's molecular structure would be beneficial. Soybean breeding programs will benefit from the rapid assessment of TI content, enabling them to efficiently screen and focus resources on varieties that meet their objectives, thereby avoiding investment in non-viable lines. Within the industry sector, the adoption of an NIR model for estimating total TI in meal samples will allow for a more time-efficient and cost-effective evaluation of their production line and final product quality control.

Tables

Table 1. TI concentration determined by HPLC in the six models; seed total trypsin inhibitor (STTI), seed Kunitz trypsin inhibitor (SKTI), seed Bowman-Birk trypsin inhibitor (SBBTI), meal total trypsin inhibitor (MTTI), meal Kunitz trypsin inhibitor (MKTI), and meal Bowman-Birk trypsin inhibitor (MBBTI).

TI Model	Mean %¹	Range %¹	SD²	CV³
STTI	8.39	2.62-13.13	1.87	0.22
SKTI	4.4	0.49-7.78	1.06	0.24
SBBTI	3.99	0.34-7.79	1.35	0.33
MTTI	4.02	0.38-10.25	1.55	0.37
MKTI	2.98	0.24-10.76	1.26	0.472
MBBTI	1.14	0.0-4.27	0.72	0.63

¹ Mean and range reported in mg/g of sample (%)

² SD: standard deviation

³ CV: coefficient of variance

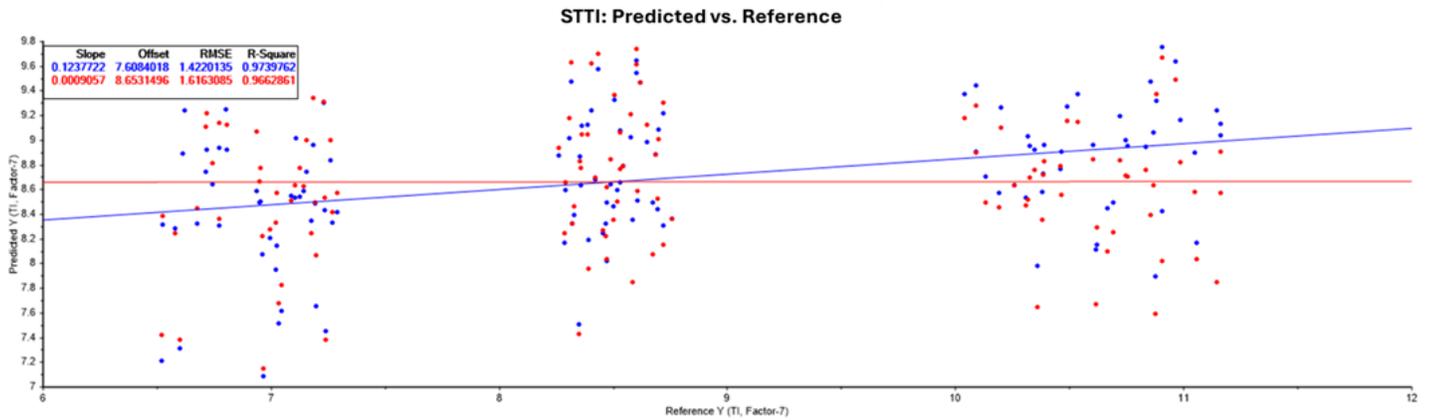
Table 2. Calibration and validation for the TI models, with R2 representing model accuracy and RMSE the average error rate.

TI Model	Sample Size	Calibration		Validation	
		R ² ¹	RMSE ²	R ²	RMSE
STTI	124	0.937	1.46	0.968	1.579
SKTI	124	0.979	0.676	0.975	0.741
SBBTI	124	0.027	1.269	0.017	1.287
MTTI	112	0.892	1.398	0.864	1.56
MKTI	112	0.059	1.116	0.052	1.126
MBBTI	112	0.021	0.641	0.016	0.648

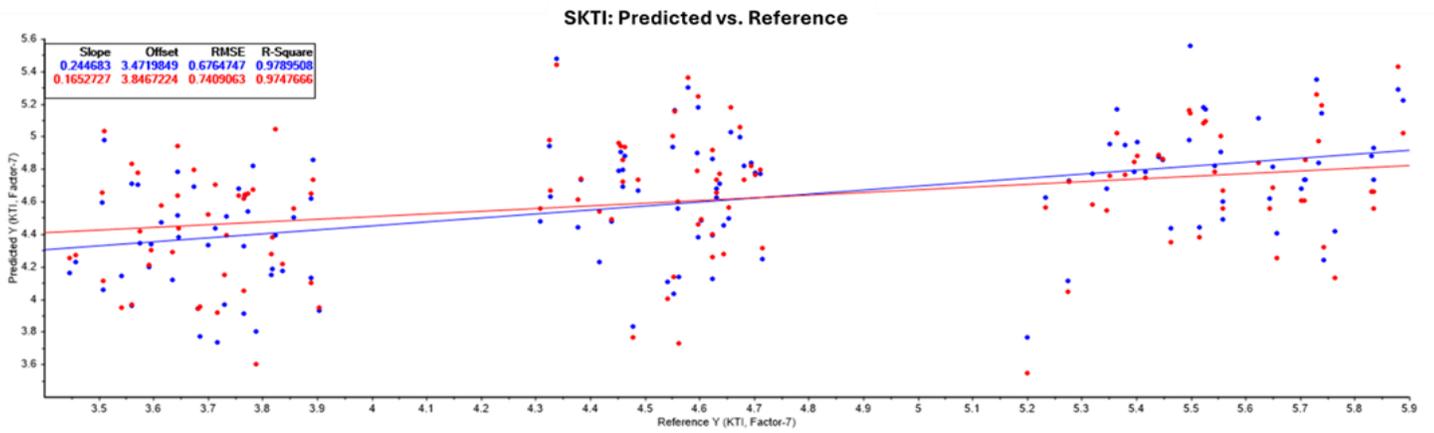
¹ R²: coefficient of determination

² RMSE: root-mean-square-error

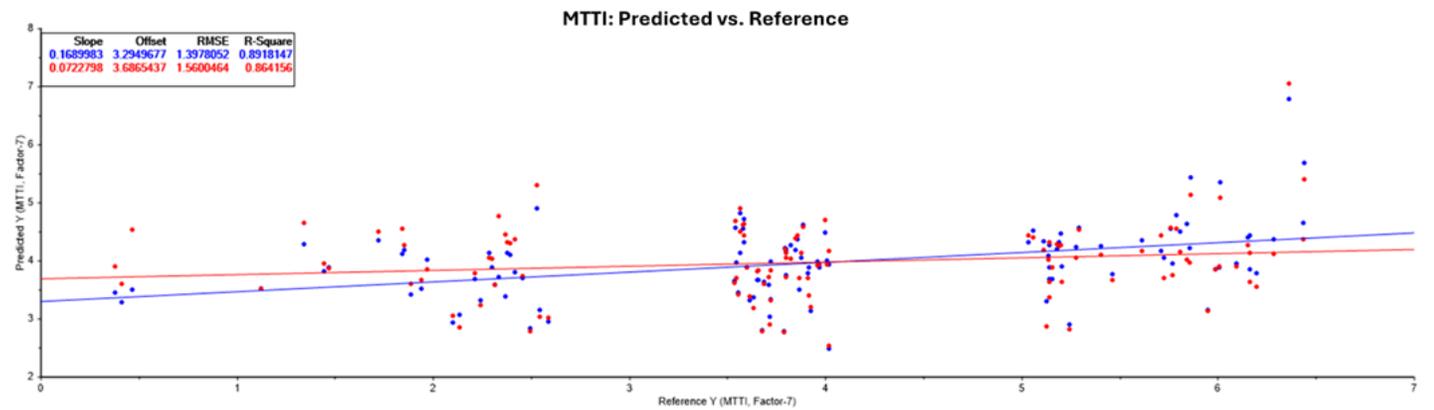
Figures



(a)

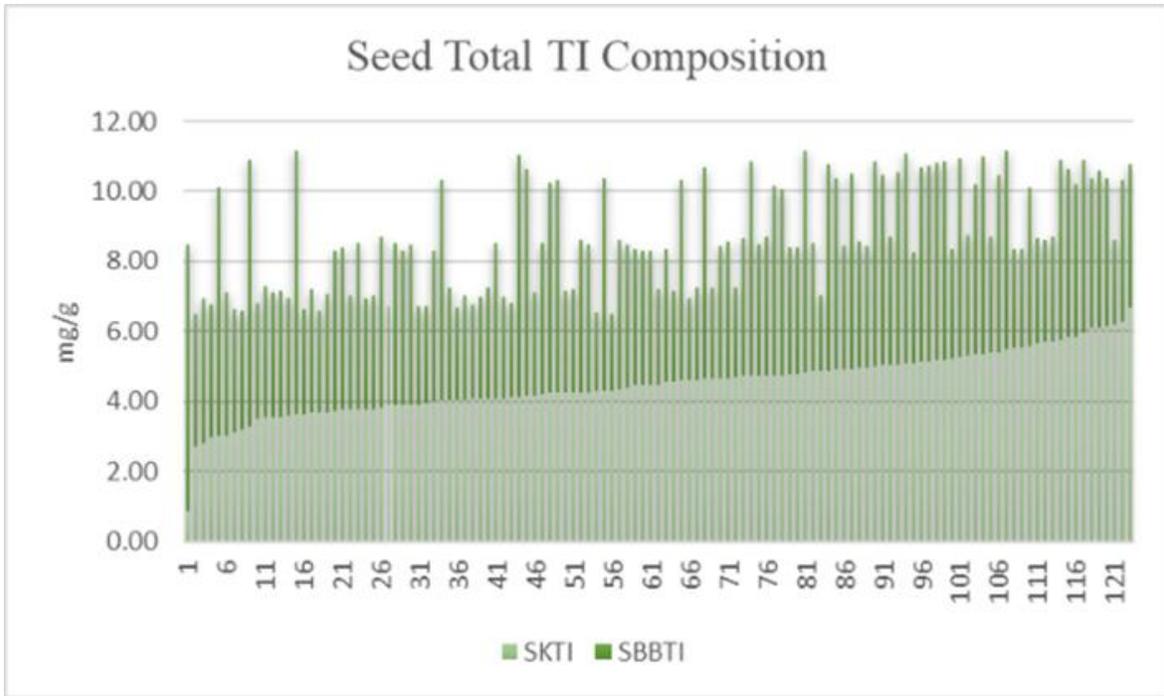


(b)

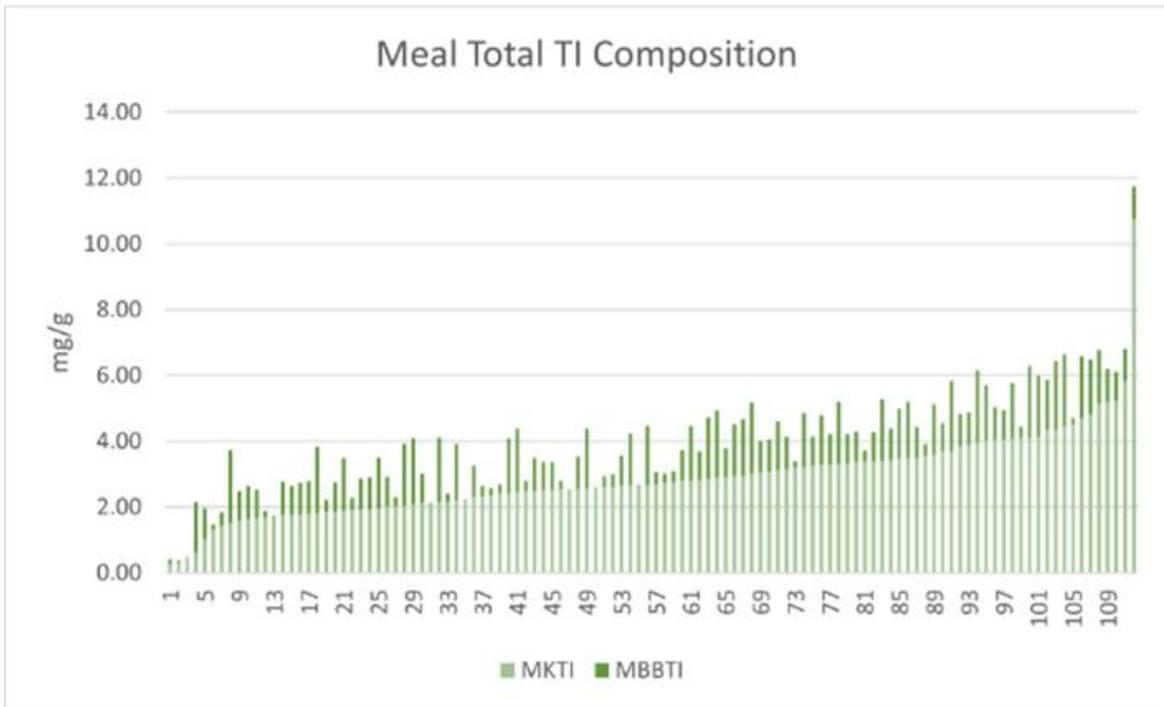


(c)

Figure 1. Relationship between the predicted (red) and the reference (blue) values and fit of (a) STTI, (B) SKTI, and (C) MTTI models.



(a)



(b)

Figure 2. Percentage of KTI and BBTI content for (A) whole seed and (B) meal samples

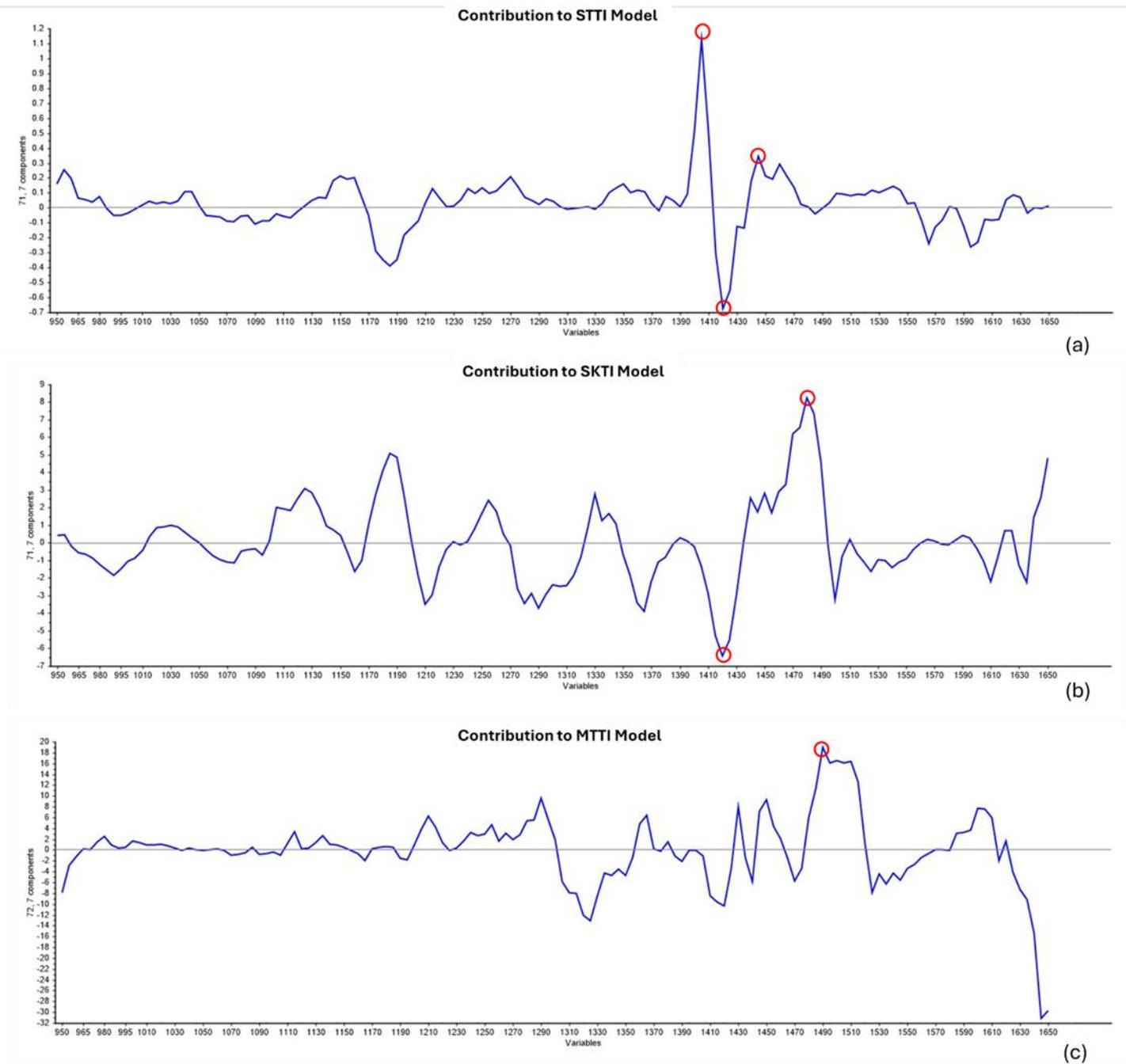


Figure 3. Absorption of each wavelength and its contribution to the model creation: (a) STTI (b) SKTI (c) MTTI. Red circles highlight wavelength ranges with significant model impact.

Chapter 3: Evaluation of Agronomic Performance, Biotic Stress Response, and Seed Quality in CRISPR-Edited Low Trypsin Inhibitor Soybean

Elizabeth B. Fletcher¹, Luciana Rosso¹, Usha Panta², Alejandro Rojas-Flechas³, Emily Garant³, Senyu Chen⁴, Bo Zhang^{1*}

¹ Virginia Tech, School of Plant and Environmental Sciences, Blacksburg, VA, USA

² Virginia Tech, Department of Entomology, Blacksburg, VA, USA

³ Michigan State University, Plant Genomics, East Lansing, MI, USA

⁴ University of Minnesota, Department of Plant Pathology, Waseca, MN, USA

Abstract

Soybean [*Glycine max* (L.) Merr.] is a key livestock feed, but the presence of trypsin inhibitors (TIs), an antinutritional factor essential for pest defense, can cause gastrointestinal issues and limit animal growth. To address this, we developed the low-TI soybean line VT5-26 by knocking out two genes, *KTII* and *KTI3*, in Willams82 (WM82) using CRISPR/Cas9 gene editing. This study aimed to confirm that, following propagation, 1) VT5-26 retained the intended gene edits and low TI levels and 2) these edits did not negatively impact agronomic traits, biotic stress responses or seed quality compared to WM82. VT5-26, WM82, and a commercial variety (AG 3803) were planted in Blacksburg and Warsaw, VA, in the growing seasons of 2023 and 2024. The results confirmed stable inheritance of the *KTII* and *KTI3* edits and consistently low TI levels in VT5-26. No significant differences were observed between VT5-26 and WM82 in plant height, lodging, maturity, or yield. VT5-26 also exhibited similar response to insect feeding and soil-born pathogens. Notably, VT5-26 demonstrated significantly improved nematode resistance to soybean cyst nematode (279 vs. 384 cysts/plant). A 1% decrease in seed protein content in VT5-26 compared to WM82, although statistically significant, was expected due to the reduction of TI, a proteinaceous compound. These findings confirm the successful knockout of *KTII* and *KTI3* without negatively affecting agronomic performance or biotic stress response. The results validate CRISPR/Cas9 as a precise and effective tool to reduce TI in soybean seed, paving the way to the development of crops better suited to meet global food production challenges.

1. Introduction

Soybean [*Glycine max* (L.) Merr.] is one of the most valuable row crops globally, widely cultivated for its high protein and oil content. Its nutritional profile includes a complete set of amino acids, combined with essential vitamins and minerals, making soybean a foundational ingredient staple in animal feed formulation [39], [58]. In the United States, approximately 70% of soybeans produced are destined for animal feed, amounting to 33.12 million tons in 2021 [59]. Despite these nutritional benefits, soybean contains antinutritional factors, such as trypsin inhibitors (TI) that inhibit the absorption and digestibility of nutrients in livestock [37], [39], [38], [60]. High intake of TI can cause severe health implications in animals such as decreased growth rates because of the limited protein and nutrient absorption. In severe cases, overconsumption of TI may lead to the enlargement of the pancreas, liver, intestines, and even result in pancreatitis [40], [41], [61], [42].

TI is a naturally occurring protein in legumes that strongly binds to trypsin, a digestive enzyme in animals, effectively blocking its active site and hindering the digestion process [42], [62]. High dietary intake of TI has been associated with reduced growth rates in animals due to impaired nutrient assimilation. In severe cases, chronic consumption can lead to hypertrophy of the pancreas, liver, and intestines, and has been linked to inflammation and pancreatitis [40], [42], [61], [89]. Soybean has two types of TI: Kunitz trypsin inhibitor (KTI) [43] and Bowman-Birk trypsin inhibitor (BBTI) [62], [45], [63]. KTI, a monomeric protein with 181 amino acids with a molecular weight of 21.5 kDa, is more abundant than BBTI and forms a strong, irreversible complex with trypsin [44]. In

contrast, BBTI, a smaller 7-8 kDa protein with various isoforms, is present in lower quantities, and inhibits both trypsin and chymotrypsin [65], [64], [46].

To negate the antinutritional effects of TI in soybeans, the raw soybean seeds are heated and processed into meal. The benefits of heating soybeans were first observed in 1917, and the practice was widely adopted to improve the nutritional value even though the mechanism were not initially understood [48]. In recent years, heating raw soybean to 121°C for 15 minutes is the industry's standard practice to deactivate TI in soy meal [47]. While effective, this process is energy-intensive, expensive, and can lead to the degradation of heat-sensitive nutrients, resulting in protein losses of up to 5–20%. Moreover, overprocessing can further reduce feed quality, highlighting the need for soybean varieties with inherently lower TI content.

Despite its antinutritional properties, TI plays a vital role in the soybean plants' natural defense against pests. and it is widely expressed throughout soybean tissues [49]. This dual role makes breeding for low-TI soybean particularly challenging, as reducing TI levels can compromise pest resistance and overall field performance. To address this issue, the Virginia Tech Soybean Breeding program developed a low-TI line using CRISPR/Cas9 technology. Using Williams82 (WM82) as the genetic background, two seed-specific genes *KTII* and *KTI3*, were knocked out to create a low-TI line, VTI5-26 (Wang et al., 2023) . Because *KTII* and *KTI3* are exclusively expressed in seeds, their removal is not expected to affect TI expression in vegetative tissues, thus preserving the plant's defensive integrity.

This targeted approach offers several key advantages: (1) it minimizes the need for post-harvest thermal processing, potentially reducing energy costs and nutrient loss; (2) it

enhances the nutritional quality and digestibility of soybean meal; and (3) it presents a scalable solution to improve feed efficiency across the livestock industry. In this context, the development of VTI5-26 represents a significant advancement in soybean improvement efforts, leveraging genome-editing tools to reconcile nutritional goals with agronomic resilience.

The current study aims to evaluate the stability and agronomic performance of VTI5-26 across multiple growing environments. Specifically, our objectives were to (1) confirm the stable inheritance of the *KTI1* and *KTI3* knockouts and the persistence of low TI levels following propagation, and (2) assess whether these genetic edits impact key agronomic traits, biotic stress responses, or seed quality in comparison to the wild-type WM82 and a commercial variety. Field trials were conducted in 2023 and 2024 at two contrasting locations in Virginia, Blacksburg and Warsaw, to capture environmental variability and ensure robust trait evaluation. Results from these trials will help determine the suitability of VTI5-26 as a commercially viable, low-TI soybean cultivar and validate CRISPR/Cas9 as a powerful tool in modern plant breeding to address complex nutritional and agronomic challenges.

2. Materials and Methods

2.1 Plant Materials

The low-trypsin inhibitor soybean line, VTI5-26, was developed by knocking out two seed-specific genes, *KTI1* and *KTI3*, in the Williams 82 (WM82) genetic background using CRISPR/Cas9 genome editing technology (Wang et al., 2023). WM82 is widely used

soybean cultivar released in 1972, developed from a cross of ‘Wayne’ × I57-0034 by the U.S. Regional Soybean Laboratory and the Illinois Experimental Agricultural Station [90].

2.2 KTI Quantification

The KTI concentration in VTI5-26 and W82 seeds harvested in 2023 and 2024 was quantified using the a high-performance liquid chromatography (HPLC) method developed by Rosso et al. (2018) [18]. Briefly, 10 mg of finely ground soybean seed powder was extracted with 1.5 mL of 0.1 M sodium acetate buffer (pH 4.5). by vortexing and shaking at room temperature for 1 hour. The mixture then was centrifuged at 12,000 rpm for 15 minutes. A 1 mL aliquot of the supernatant was filtered through 0.2- μ m low-protein binding hydrophilic Millipore (polytetrafluoroethylene [PTFE]) membrane (Millipore Ireland). KTI separation and quantification was performed using an Agilent 1260 Infinity series HPLC system (Agilent Technologies) fitted with a guard column (4.6 \times 5 mm) and a Poros R2/H perfusion analytical column (2.1 \times 100 mm, 10 μ m). The mobile phase consisted of Phase A (0.01% [v/v] trifluoroacetic acid in Milli-Q water) and Phase B (0.085% [v/v] trifluoroacetic acid in acetonitrile). The injection volume was 10 μ L, and detection was carried out at a wavelength of 220 nm.

2.3 Evaluation of Agronomic Performance, Seed Quality

Field evaluations of VTI5-26 were conducted during the 2023 and 2024 growing seasons at two locations in Virginia: Blacksburg and Warsaw. Trials were conducted in a randomized complete block design (RCBD) with three replications. The experimental lines included VTI5-26, WM82 (wild-type), and AG3803 (a commercial check variety). Each plot consisted of four rows (76.2 cm spacing), measuring 3.6 m long in Warsaw and 4.8 m

in Blacksburg. The inner two rows were designated for agronomic traits and seed harvest, and the outer rows were used for pest resistance assessments.

Agronomic traits included:

- **Maturity (MAT):** Defined as the date when 95% of pods exhibited their mature pod color (Fehr, Caviness, and Vorst, 1977).
- **Plant Height (HT):** Measured in centimeters from the soil surface to the apex of the plant at maturity (Shapiro & Flowerday, 1987).
- **Lodging (LOD):** Scored visually on a 1–5 scale, with 1 indicating fully erect plants and 5 indicating plants completely lodged.

Seed quality (QUAL) was evaluated using a 1–5 rating scale, where 1 indicated excellent quality and 5 indicated poor quality, based on visual inspection for seed development, wrinkling, brightness, and damage.

Protein and oil content were analyzed using a Perten DA7250 near-infrared reflectance spectroscopy (NIRS) instrument, with values reported on a dry weight basis. Seed yield (YLD) was calculated based on harvested seed weight, adjusted to 13% moisture, and converted to bushels per acre (60 lbs/bu) based on plot dimensions. Seed size was expressed as the average weight (g) of a 100-seed sample.

2.4 Evaluation of Pest Resistance

Pest resistance evaluations focused on three key insect pests: stink bugs (*Halyomorpha halys*), Japanese beetles (*Popillia japonica*), and Mexican bean beetles (*Epilachna varivestis*).

- **Leaf feeding damage (LFD):** assessed at the R1 growth stage. Five leaves were randomly selected from the interior two rows of each plot and visually rated for damage using a standardized scale that estimated missing leaf tissue. An average LFD score was then calculated for each plot.
- **Pod feeding damage (PFD):** evaluated at R6 stage, while pods/beans were still green. A total of 100 pods were randomly collected from each plot, opened, and inspected for feeding damage, which was identified by a black mark on the pod corresponding to misshapen bean. The percentage of damaged pods was recorded for each plot.

2.5 Evaluation of Soybean Cyst Nematode Resistance

The soybean lines were evaluated for resistance/susceptibility to the soybean cyst nematode (SCN), *Heterodera glycines*, specifically HG Type 7 (race 3), originally collected from a field at the University of Minnesota Southern Research and Outreach Center (UMN SROC), Waseca, Minnesota. Before the experiment, the nematode population was cultured on the susceptible soybean cultivar ‘Sturdy’ for approximately 45 days. SCN eggs were extracted from the roots and used to prepare an egg suspension for inoculation.

The SCN development assay was conducted on soybean plants grown in containers (4 cm diameter × 13.5 cm height) in a growth room at the UMN SROC. Each container was filled halfway with autoclaved soil (80% sand, 20% field clay loam), inoculated

with 1,500 eggs in 2.5 mL of water, and then topped with additional soil to about 2 cm below the rim. A second inoculation of 1,500 eggs in 2.5 mL of water was applied to the soil surface, followed by sowing one soybean seed per cone-tainer and covering it with about 1 cm of soil. Twelve replicates were used per soybean line, with ‘Williams 82’ included as the susceptible control.

The cone-tainers were arranged randomly on a rack in the growth room, maintained at 25 °C with a 16-hour photoperiod, and watered using a sprinkler irrigation system. After 35 days, cysts (females) were extracted from the roots and soil using standard protocols. Briefly, the soil and roots were transferred to a beaker with water. The roots were removed, and females were washed onto 850- μ m-aperture sieve nested on 250- μ m-aperture sieve. The soil suspension was also poured over the sieves to recover cysts. After rinsing, cysts retained on the 250- μ m-aperture sieve were collected and separated from debris by flotation-centrifugation in 76% sucrose solution. Cysts per plant were counted under a dissecting microscope. The Female Index (FI) for each plant was calculated as: $FI = (\text{Number of females on a plant} \times 100) / (\text{Mean number of females on Williams 82})$ [91].

2.6 Evaluation of Disease Resistance

To test the tolerance of soybean cultivars VT5-26, AG 3803, and Williams82 (susceptible check), a variety of soybean pathogens were selected for screening, including *Fusarium virguliforme*, *Pythium sylvaticum*, *Phytophthora sansomeana*, *Rhizoctonia solani* AG4, *Macrophomina phaseolicola* and *Diaporthe longicolla*. Pathogen Inoculum for all isolates, except *Diaporthe*, was prepared by growing the corresponding isolates on sterile millet in a 500 mL flask. Briefly, millet was soaked in water overnight in distilled water, strained, and transferred to flasks. Flasks were covered with foil and autoclaved for

35 minutes. Flasks were allowed to cool overnight and then autoclaved a second time. Five 8 mm plugs were cut from each culture from their respective isolate and aseptically transferred into the flasks containing the millet. The millet inoculum was incubated at room temperature for 14 days while being mixed regularly to establish complete colonization and facilitate the separation of grains.

Seedling assays were prepared in 473 mL capacity foam cups with 8-mm drainage holes at the bottom. The cups were filled from the bottom with 200 mL of vermiculite, 4 g of inoculated millet, 150 mL of vermiculite, five seeds of the same cultivar (AG3803, VT5-26, or WM82), then 75 mL of vermiculite. Plants were watered every other day and maintained in a growth chamber with a light regimen of 16 h light (300 mE) and 8 h dark, at a temperature of approximately 25° C and 80% humidity for 16 days. Every isolate had five replicates per experiment, per cultivar, along with two types of controls: non-inoculated sterile millet and non-millet, to monitor any effects of the millet on the seedlings. At the end of the experiment, the germination rate was recorded, and plant roots were carefully rinsed with tap water to remove any debris. The shoot mass, and root mass was then measured in grams using a weighing scale. Root length was collected using APS Assess 2.0 (American Phytopathological Society, St. Paul, MN), an image analysis software. Root samples were then collected and plated to confirm the identity of the pathogen.

In the case of *Diaporthe longicolla*, a toothpick inoculation as described in Ghimire et al. (2019) [92] was conducted by plating an isolate of *Diaporthe longicolla*, along with sterile toothpicks on potato dextrose agar. Plates were stored at room temperature for 18 days to allow for the isolation to completely colonize the toothpicks. Soybean seeds of the

same three cultivars (AG3803, VT5-26, and WM82) were maintained in non-millet vermiculite for 18 days before inoculation until the formation of the first trifoliolate. The toothpicks were then inserted into the stem of the soybean plants (V2-V3) between the unifoliolate and trifoliolate and sealed with petroleum jelly to prevent the stem from drying out. Plants were placed back in the growth chambers in the same growth conditions to the experiment above for 4 weeks before each plant's lesion size was measured in millimeters using a caliper. Stem samples were then collected and plated on water agar to confirm the isolate.

2.7 Statistical Analysis

The data collected in this study were statistically analyzed using RStudio (version 4.4). An analysis of variance (ANOVA) was performed to assess differences among genotypes for all traits. least square means (LSD) and coefficient of variation (CV) were calculated to measure the relative variability of the data relative to the grand mean value. Tukey's Honest Significant Difference (HSD) test was used for multiple comparisons, with significance determined at $\alpha = 0.05$.

3. Results

3.1 KTI Concentrations

KTI concentrations were quantified from composite seed samples of VT5-26 and W82 harvested in 2023 and 2024. VT5-26 had KTI levels were 0.158 mg/g in 2023 and 0.156 mg/g in 2024. In contrast, W82 had significantly higher KTI concentrations of 5.817

mg/g in 2023 and 6.935 mg/g in 2024. These results confirm the successful and stable knock-out of *KTI1* and *KTI3* genes in VT5-26 across two growing seasons.

3.2 Agronomic Performance and Seed Quality

Field trials conducted in 2023 and 2024 assessed agronomic performance and seed quality of VT5-26 in comparison with WM82 and the commercial cultivar AG3803 (Table 1). VT5-26 matured at the same rate as WM82 and about 15.0 days later than the check AG3803. VT5-26 was approximately 2.54–5.08 cm taller than both checks. Lodging scores remained low and consistent across all genotypes, with VT5-26 averaging around 2.0, similar to the test mean and reference lines.

Seed yield performance varied by year. In 2023, VT5-26 outperformed WM82 by 4 bu/ac, 2023 however, in 2024, yielded 5 bu/ac less. Across both years, no statistically significant differences in yield were observed between VT5-26, WM82, and AG3803. Seed quality ratings remained consistent across all three lines in both years, with each genotype averaging a two-year score of 1.75, indicating visually high-quality seed with minimal wrinkling or discoloration.

In terms of seed composition, VT5-26 had a protein content of 40.3% in 2023, which was similar to WM82's 40.8%, and slightly higher than AG3803 (39.6%). In 2024, protein content in VT5-26 was slightly reduced to 39.6%, while WM82 remained higher at 40.7%. Although the two-year average protein content for VT5-26 (39.95%) was only 0.8% lower than WM82, this difference was statistically significant. Oil content was consistent across both years and all genotypes, with VT5-26 averaging 20%, similar to both checks.

3.3 Pest Resistance

Pest pressure was higher in 2023 than in 2024, resulting in increased PFD during the R6 stage (Table 1). While in both years VT5-26 exhibited more pod and leaf feed damage compared to both checks., However, these differences were not statistically significant.

3.4 Soybean Cyst Nematode Resistance

Significant improvements were observed in soybean cyst nematode (SCN) resistance. VT5-26 showed an average of 279 cysts per plant, significantly fewer than WM82 (384 cysts per plant), representing a reduction of 105 cysts across 12 replications (Table 2).

3.5 Disease Resistance

Under *Macrophomina phaseolina* inoculation, VT5-26 showed increased shoot mass (6.66 g) compared to WM82 (5.80 g) and AG3803 (5.26 g), but reduced root mass (0.97 g) relative to WM82 (1.59 g) and AG3803 (1.46 g). Germination rate was also lower in VT5-26 (76%) compared to 92% for both WM82 and AG3803. Despite these observed differences, no statistically significant differences were detected for any of these traits (Table 3).

Under *Rhizoctonia solani* screening, shoot mass for VT5-26 (7.81 g) was comparable to WM82 (7.89 g), while AG3803 showed higher biomass (10.24 g). Root mass was lower in VT5-26 (3.29 g) than in WM82 (4.01 g) and AG3803 (5.32 g). Interestingly, VT5-26 showed a higher germination rate (84%) than WM82 (76%) but was slightly lower than

AG3803 (88%). As with *Macrophomina*, none of these differences were statistically significant across the five replications.

4. Discussion

The results of the KTI quantification and gene sequencing confirm the stability and effectiveness of the CRISPR/Cas9-induced deletions of *KTI1* and *KTI3* in the VT5-26 soybean line. The deletion of *KTI1* and *KTI3* resulted in an approximately 98% reduction of KTI expression in the seed of VT5-26 compared to wild type WM82. The findings support the conclusion that the edition of these two seed-specific genes did not significantly impact the agronomic performance or seed quality of VT5-26 compared to WM82. While there was a statistically significant reduction in protein content, less than 1% lower than W82, this minor decrease was expected due to the removal of KTI, a seed storage protein. This minor decrease in protein is negligible when compared to the estimated 15% loss of digestible protein that occurs during industrial heat treatment to deactivate TI [93].

Moreover, the protein eliminated through gene editing is largely non-digestible by livestock, further reinforcing the nutritional advantage of the low-TI trait. A slight, though not statistically significant, increase in leaf and pod feeding damage was observed in VT5-26 compared to the control lines. Trypsin inhibitors are known to contribute to plant defense by inhibiting insect digestive proteases [94]. While the gene deletions specifically targeted the KTI expression in the seed, it is possible that they may have subtly influenced gene expression in other tissues, particularly in leaves. However, because the observed increase in feeding damage was not statistically significant, the low-TI trait appears to

preserve overall pest resistance, indicating that seed-specific gene deletions do not compromise the plant's above-ground defenses.

Interestingly, there was a statistically significant difference in resistance to SCN, with VT5-26 having an average of 105 less cysts per plant compared to WM82. Although all lines in the study exceeded the threshold for SCN susceptibility (20 cysts per 100–200 cc of soil [95]), the marked reduction in VT5-26 suggests a possible link between SCN response and the deleted *KTI1* and *KTI3* genes or their downstream regulatory effects. While previous studies have shown that overexpression of TI in roots can enhance SCN resistance, the role of reduced seed TI expression in nematode resistance remains unclear. These findings warrant further investigation into the potential involvement of seed-derived TIs in root defense mechanisms.

VT5-26 also showed modest, though not statistically significant, differences in response to fungal pathogens *Macrophomina phaseolina* and *Rhizoctonia solani*, which cause seedling blight, root rot, and stem rot (Marquez et al., 2021; Crop Protection Network, 2019). While variations were observed in shoot mass, root mass, and germination rate, these differences did not reach significance. Previous studies have suggested a potential relationship in TI expression, specifically Bowman-Burke Trypsin Inhibitor, and the plants' immune defense against fungal pathogens [96], [97]. While the function of protease inhibitors such as TI in the immune system remains unclear, our findings raise the possibility of a relationship between KTI and fungal disease resistance.

5 Conclusion

The industrial-scale denaturation of trypsin inhibitors (TI) to improve soybean meal digestibility is both costly and time-consuming, that can compromise protein quality. VT5-26 demonstrated no statistically significant difference in agronomic performance, pest and disease resistance compared to the wild-type WM82, highlighting its potential as a viable low-TI alternative. By reducing TI specifically in the seed without affecting TI levels in other tissues, VT5-26 offer growers and processors a promising solution that could enhance soybean meal protein quality by minimizing the need for heat denaturation without compromising field performance.

Additionally, the development of VT5-26 using CRISPR/Cas9 and the results of this study demonstrated the effectiveness and potential of gene-editing technology for crop varietal improvement. This study provides proof of concept for using gene editing to improve seed quality traits while preserving agronomic and defensive traits. Future efforts will focus on introducing the low-TI trait into elite, high-yielding germplasm to expand its use in commercial production and continue meeting the evolving needs of soybean producers and processors.

Tables

Table 1. Performance of VT15-26 in the Virginia Tech field trials (2023-2024)										
YEAR	NAME	MAT [†] (days)	HT [‡] (cm)	LOD [§]	YIELD (bu/ac)	QUAL [‡]	PROTEIN	OIL	LFD	PFD
							(%)	(%)	(%)	(%)
2023 (2 LOC*)	VT15-26	17.3	101.6	2.1	40.3	1.5	40.3	21.1	9.6	55.3
	WM82 (Check)	16.6	102.3	2.1	36.6	1.5	40.8	20.6	9.7	47.3
	AG 3803 (Check)	26.8	101.6	2.5	45.8	1.5	39.4	20.4	6.9	42.6
	LSD**	3.9	9.9	0.8	18.7	0.0	0.6	0.9	3.9	10.5
	CV***	15.3	20.0	32.1	37.2	0.0	1.3	3.8	36.9	17.7
	GRAND MEAN	20.3	101.8	2.25	40.9	1.5	40.1	20.7	8.7	48.4
2024 (2 LOC)	VT15-26	16.3	104.9	2.0	50.2	2.0	39.6	20.3	8.1	27.3
	WM82 (Check)	16.2	102.3	1.3	55.9	2.0	40.7	20.4	5.4	22.5
	AG 3803 (Check)	31.2	100.3	1.8	62.9	2.0	39.7	20.5	6.5	22.3
	LSD**	1.4	5.3	0.8	11.0	0.0	0.8	0.7	5.1	6.6
	CV***	5.5	10.6	37.2	15.9	0.0	1.6	2.8	63.0	56.2
	GRAND MEAN	21.2	102.3	1.7	56.4	2.0	40.0	20.3	6.7	5.0
2023 & 2024	VT15-26	16.8 <i>b</i>	103.1 <i>a</i>	2.0 <i>a</i>	45.2 <i>a</i>	1.75 <i>a</i>	39.9 <i>b</i>	20.6 <i>a</i>	8.8 <i>a</i>	41.3 <i>a</i>
	WM82 (Check)	16.4 <i>b</i>	102.3 <i>a</i>	1.7 <i>a</i>	46.3 <i>a</i>	1.75 <i>a</i>	40.7 <i>a</i>	20.4 <i>a</i>	7.6 <i>a</i>	34.9 <i>ab</i>
	AG3803 (Check)	29.0 <i>a</i>	100.8 <i>a</i>	2.2 <i>a</i>	54.3 <i>a</i>	1.75 <i>a</i>	39.9 <i>b</i>	20.4 <i>a</i>	6.7 <i>a</i>	32.5 <i>b</i>
	LSD**	2.6	6.3	0.6	9.5	0	0.63	0.6	3.4	7.2
	CV***	12.7	16.0	31.8	19.4	0	1.5	3.3	44.0	19.7
	GRAND MEAN	20.7	102.1	1.9	48.6	1.75	40.0	20.5	7.7	36.2
LOC*: Location LSD**: Fisher's least significant difference with the significant difference at the 0.05 probability level CV***: Coefficient of variation MAT [†] : Maturity index HT [‡] : Height, cm LOD [§] : Lodging QUAL [‡] : Seed quality LFD: Leaf feeding damage PFD: pod feeding damage Statistical Significance: indicated by <i>a, b</i>										

Table 2. Performance of VT15-26 in Soybean Cyst Nematode Screening (2024)	
NAME	NCC
WM82	384
VT15-26	279
AG3803	129
LSD**	40.0
CV***	19.1
GRAND MEAN	263.6
LSD**: Fisher's least significant difference with the significant difference at the 0.05 probability level CV***: Coefficient of variation NCC: Cyst count per plant	

Table 3: Performance of VT15-26 in Disease Resistance Screening				
PATHOGEN	NAME	SHOOT MASS (g)	ROOT MASS (g)	GERMINATION (%)
<i>Macrophomina</i>	VT15-26	6.66 <i>a</i>	0.97 <i>a</i>	76.00 <i>a</i>
	WM82	5.80 <i>a</i>	1.59 <i>a</i>	92.00 <i>a</i>
	AG3803	5.26 <i>a</i>	1.46 <i>a</i>	92.00 <i>a</i>
	LSD**	2.95	0.89	25.00
	CV***	27.68	36.74	16.04
	GRAND MEAN	5.91	1.35	86.00
<i>Rhizoctonia</i>	VT15-26	7.81 <i>a</i>	3.29 <i>a</i>	84.00 <i>a</i>
	WM82	7.89 <i>a</i>	4.01 <i>a</i>	76.00 <i>a</i>
	AG3803	10.24 <i>a</i>	5.32 <i>a</i>	88.00 <i>a</i>
	LSD**	5.84	2.96	46.00
	CV***	37.36	38.96	30.91
	GRAND MEAN	8.65	4.21	82.00
LSD**: Fisher's least significant difference with the significant difference at the 0.05 probability level CV***: Coefficient of variation Statistical Significance: indicated by <i>a, b</i>				

Chapter 4: Genome Wide Associate Study Uncovers Novel SNPs for Improving Natto-Specific Soybean Traits

Elizabeth B. Fletcher¹, Jessica Wilbur², Zhibo Wang³, Jonathan Aims¹, Gota Morota⁴, Luciana Rosso¹, Leandro Mozzoni⁵, Pengyin Chen⁶, and Bo Zhang¹

¹School of Plant and Environmental Sciences, Virginia Tech, Blacksburg, VA, USA

²Nufarm Americas Inc., Woodland, CA, USA

³Danforth Center, St Louis, MO, USA

⁴Department of Agricultural and Environmental Biology, University of Tokyo, Tokyo, Japan

⁵Bayer Crop Science, Lincoln, NE, USA

⁶Fisher Delta Research Center, University of Missouri, Portageville, MO, USA

Abstract

Natto is a traditional Japanese fermented soybean dish, requiring varieties with unique traits such as water absorption (WA), seed coat deficiency (SCD), light cooked color (CC) and optimum protein and oil balance. However, genetic control of these traits remains largely unexplored. This study aimed to identify genetic loci associated with natto quality-specific traits in soybean to guide breeding efforts for improved varieties. We evaluated 146 natto soybean accessions across three locations over two years. The genotypic data included 26,927 SNPs from publicly available sources. A genome-wide association study was performed, significant SNPs were identified for all five traits, WA, CC, SCD, protein and oil. We uncovered four, six, and two novel SNPs for WA, CC, and SCD respectively. Additionally, nine novel SNPs associated with protein and four SNPs for oil were identified. The candidate genes for the natto quality traits were largely associated with stress responses. These findings provide valuable markers for developing high quality natto soybeans to meet producer and consumer demands.

1 Introduction

Soybean (*Glycine max.* [L] Merr.) is highly valued for its protein content, making it one of the world's largest row crops with global production reaching approximately 350 million metric tons annually [51]. While the majority of this production is destined for animal feed and biofuel, 13.91 million tons are consumed by humans as an increase of nearly 5 million tons from 2010 [50]. In addition, individual human consumption of soy was increased from 1.29 kg per capita in 2010 to 1.77 kg per capita in 2021 [51]. A large portion of food-grade soy is consumed in the form of traditional soyfood such as tofu, soy milk, and natto.

Natto is a traditional Japanese dish made by steaming soaked soybeans (*Glycine max.*[L]) and fermenting them with the bacteria, *Bacillus subtilis* [52]. Notorious for its stringy, slimy texture and strong ammonia smell, natto has been a staple in the Japanese diet for at least 500 years. Despite its unusual appearance, natto is well known for its numerous health benefits including prevention of heart disease and the strengthening of bones and immune systems [53], [98]. A 100 g serving provides 211 calories and 19 g of protein, making it a nutritious and filling source of energy [54]. Beyond its high protein content, natto provides many bioactive compounds and essential nutrients such as vitamin K, isoflavones, biogenic amines and nattokinase [54]. Nattokinase, a naturally occurring enzyme and active ingredient extracted from natto [55], has been touted as a natural alternative for lowering blood pressure, acting as an anticoagulant, improving retina performance and reducing inflammation [56].

In recent years, natto has gained popularity beyond Japan's borders, entering the health food scene in other countries, driving increased demand and production [57]. This growing interest

has highlighted the need for continued advancement in soybean variety development to meet evolving consumer preferences and production requirements.

The breeding and selection of soybean lines for natto production requires the evaluation of many phenotypic traits that are not typically the focus of breeding programs. These traits are primarily related to seed composition and natto appearance, aligning with the preference of Japanese consumers and the natto manufacture [99], [100]. In terms of appearance, Japanese consumers prefer small seeds with a light color after steaming. To ensure proper fermentation, the soybean seed must have high water absorption for optimum steaming, and a low protein-high oil content. Additionally, a low seed coat deficiency is also important to maintain the seed integrity after soaking, cooking and fermentation.

This study is novel, as no GWAS has been previously conducted for the natto quality-specific soybean traits of water absorption (WA), cooked color (CC) and seed coat deficiency (SCD). The identified SNPs for these traits would facilitate the breeding and selection of high-quality natto lines that meet the expectations of both manufacturers and consumers.

2 Materials and Methods

2.1 Materials

Accessions were selected by filtering US-GRIN germplasm for small-seeded soybean (<10g/100 seed) with a yellow seed coat and yellow hilum across maturity groups IV, V, and VI. A total of 200 accessions were planted in Blacksburg, VA, Fayetteville, AR, and Portageville, MO in 2021 and 2022. They were planted in four row plots in a random complete block design (RCBD).

The center two rows were harvested and used for phenotypic data collection. Due to field conditions and weather, ultimately 146 accessions were successfully harvested and phenotyped.

2.2 Phenotypic Data Collection

One sample of each replication from each location was evaluated for the traits of WA, CC, SCD, protein and oil content. To determine WA, 20g of seed were submerged in 100 ml of deionized (DI) water for 14 hrs. After soaking, the seeds were drained, patted dry, and their final weight was recorded. WA was then calculated as percentage using the initial weight and final weight. Protein and oil content of the seed was measured using pre-established calibrations on the Perten DA7250 NIR Analyzer. The beans were soaked for 14 hrs. in DI water and then drained and patted dry. The seeds were then steamed in a autoclave for 20 minutes. Immediately after the autoclave cycle, the color of 10 beans of each sample was measured using a spectrophotometer (model, etc.). The average cooked color (CC) of the 10 samples was calculated and recorded in units of au. To determine SCD, 100 high-quality seeds were selected after removing any discolored, disfigured, damaged, or diseased seeds. The seeds were then soaked in 50 ml of a 1% bleach solution for 10 minutes. Seeds displaying a cracked, blistered or removed seed coat were counted and recorded as deficient [101].

2.3 Genotypic Data

The SNP marker data of 146 accession lines were obtained from the SoySNP50K SNPs data repository (Song *et al.*, 2015). A total of 42,291 initial SNPS were acquired and then filtered by missing genotypes and low minor allele frequency (MAF <0.05), resulting 26,927 SNPs used in final analysis.

2.4 Genome-Wide Association Analysis and Candidate Gene Identification

All genomic analyses were performed in RStudio. Associations between the phenotypic and genotypic data sets were analyzed using the rrBLUP package. A modified Šidák correction ($\alpha_{sid} = 1 - (1 - \alpha)^{1/m}$) was utilized to identify significant associations. The effective number of markers (M_{eff}) was calculated as 462 using the poolr package in R with the Li and Ji method (Li and Ji, 2005). M_{eff} was used in place of m , and thus, the adjusted significance threshold was $-\log_{10}(P) > 3.97$ at $\alpha = 5\%$ and a suggestive threshold at $-\log_{10}(P) > 3.27$ at $\alpha = 25\%$, respectively. Frequency histograms of each trait were generated using the ggplot2 package. Manhattan plots and QQ plots were generated using the qqman package.

Gene identification was performed using gene models from the Glyma.Wm82.a2.v1 (Williams 82) dataset, accessed through Soybase.org. Genes that flanked the significant SNPs within 10kb on either side were reported as candidate genes. TAIR homolog, PANTHER and GO databases were used to report gene descriptions when applicable.

3 Results

3.1 Phenotype Data

Trait evaluation showed that WA, CC, and protein had normal distribution across the 146 accessions over two-years (Figure 1). The mean range of WA across the two years and three locations was 53.35% - 163.93% (Table 1). While CC had a normal distribution, the mean range was only 63.26 au – 67.08 au. Protein displayed the expected mean range in soybean, of 39.66% - 46.76%. In contrast, SCD displayed a left skewed distribution with most accessions having low SCD values, showing a more durable seed coat (Figure 1). Of the 146 accessions, seven displayed

0% SCD and 35 displayed less than 10%. The mean SCD ranged from 0% - 37% (Table 1). Oil content had a non-symmetrical distribution, with a cluster around 10% and a larger cluster around 17% (Figure 1). However, the mean oil concentration of the two years ranged from 13.22 % - 18.96 %. From the 146 asseccions, five lines that performed well in three or more traits were identified for potential natto breeding lines (Table 2).

3.2 Significant SNPs and Candidate Genes for Natto Quality Traits (WA, CC, and SCD)

Four significant SNPs were associated with WA (Figure 2), linked to four candidate genes. Three SNPs are located on Chrom 7 (Glyma07g14330, Glyma07g14010, and an unknown) and one on Chrom 12 (Glyma12g31350 and Glyma12g31360). The candidate gene models on Chrom 7 are involved in mitochondrial transcription termination and fatty acid hydroxylase. The two candidate gene models on Chrom 20 are involved in tetratricopeptide and leucine rich repeat terminals.

For CC, two significant SNPs (Figure 2) were identified, linked to two candidate genes. Of the SNPs, one is located on Chrom 2 (candidate gene unknown) and the other on Chrom 18 (Glyma18g52120 and Glyma18g52130). The two candidate gene models associated on Chrom 18 are related to catalytic activity and amino acid binding.

Six significant SNPs were associated with the trait SCD (Figure 2), linked to three candidate genes. The SNPs were located on five different chromosomes, with one SNP on Chrom 3 (Glyma03g04960), one on Chrom 4 (Glyma04g40750), one on Chrom 7 (candidate gene unknown), two on Chrom 12 (Glyma12g16486 and candidate gene unknown), and one on Chrom 18 (candidate gene unknown). The three known candidate gene models are involved in seed storage proteins, nucleic acid binding and disease resistance proteins.

3.3 Significant SNPs and Candidate Genes for Seed Composition Traits (Oil and Protein)

A total of 25 significant SNPs were identified across the five evaluated traits (Figure 3), associated with 22 known candidate genes and 7 unknown genes. For protein content, nine significant SNPs were associated with eleven candidate genes. Of these SNPs, one was located on Chrom 2 (Glyma.02G059000), one on Chrom 3 (Glyma.03G009600 and Glyma.03G009700), two on Chrom 8 (Glyma.08G196200 and Glyma.08G196400), one on Chrom 10 (Glyma.10G058701), one on Chrom 13 (Glyma.13G072000), one on Chrom 14 (Glyma.14G028600), one on Chrom 18 (Glyma.18G061600 and Glyma.18G061700) and one on Chrom 20 (Glyma.20G176800). These candidate gene models are involved in various metabolic processes such as biosynthetic oxidation-reduction and catabolic processes, as well as DNA transcription and binding activities.

Regarding seed oil, four significant SNPs (Figure 3) were associated with two known and two unknown candidate gene models. One SNP was identified on Chrom 2 (candidate gene unknown), one on Chrom 19 (Glyma.19G219700) and two on Chrom 20 (Glyma.20G055200 and one unknown). The known candidate gene models are involved in the carbohydrate metabolic process and zinc ion binding.

4 Discussion

4.1 Phenotype

The accessions in the study displayed a normal distribution for WA, CC, and protein content. SCD was unbalanced, with a larger sample set displaying a low percentage of deficient seed coats, indicating strong seed coat integrity. Five accessions were identified as potential natto breeding lines (Table 2). PI506736 performed well in all traits but oil. It had a high-water absorption

(132.73%) a light CC (63.76 au) a low SCD (5.2%) and low protein (39.66%). However, it has low oil content at 16.08% [103]. PI471931 had the highest oil content (18.97%) and performed well in WA (138.77%), and CC (64.94 au). The SCD (16.17%) was higher than other accessions and the protein content (40.29%) is mid-range for soybeans. PI408340 had a high WA (44.11%), a light CC (64.68 au) and a low SCD (2.80%). However, its protein content was midrange (40.68%) and oil was low (16.50%).

4.2 Significant SNPs and Candidate Genes for Natto Quality Traits (WA, CC, and SCD)

For the trait of WA, essential to the cooking process of natto, four significant SNPs and four candidate genes were identified within 10kbp of the SNPs. Three SNPs are located on chromosome 7 with two candidate genes and one on chromosome 12 with two candidate genes. Notably, SNPs associated with WA have not been previously studied or identified. A trend was observed with each of the candidate genes being related to the plants' stress response. The first SNP (ss715596075) was associated with a candidate gene Glyma07g14330 involved in the function of mitochondrial transcription termination factor proteins. This family of proteins regulates organellar gene expression and plays a key role in stress responses [104]. The second SNP (ss715596035) identified on chromosome 7 also had a candidate gene (Glyma07g14010) involved in stress responses, as a part of the fatty acid hydroxylase superfamily. This superfamily plays a crucial role in the biosynthesis of suberin, a fatty acid that creates a waxy waterproof coating on plant tissues, specifically roots and seeds [105], providing a natural defense against waterlogging. The final SNP (ss715612503) was associated with two candidate genes (Glyma12g31350 and Glyma12g31360).. These genes are involved in the function of encoding tetratricopeptide repeats and leucine rich repeats, respectively. Tetratricopeptide repeat proteins are essential for cytokinesis and plasma membrane repair, while leucine rich repeats are vital for

plant development, growth, and stress responses. The involvement of candidate genes with stress-related functions is reasonable when considering that the prolonged submergence and waterlogging of a seed would essentially leave them non-viable. Additionally, 29 suggested SNPs were identified, with candidate gene functions in multiple representations of leucine-rich repeat receptors and heat-shock proteins.

The trait of CC, where a lighter is preferred by natto manufacturers and consumers, was associated with only two significant SNPs. The SNP (ss715581851) on chromosome 2 resulted in no candidate genes, however, two candidate genes were within 10kbp of the SNP (ss715632355) on chromosome 18. The most notable candidate gene, Glyma18g52130, is involved in catalytic activity, a process that accelerates biochemical reactions, which can be initiated by increased temperatures during cooking process. In this study, CC had limited variation ranging from 63.26 au – 67.08 au (Table 1), which may have influenced the identification of only a few significant SNPs. Additionally, 19 suggestive SNPs identified, with candidate gene functions involved in drought response proteins and drug transmembrane transporter activity.

For the trait of low SCD, which is essential to natto beans maintaining their shape after soaking, cooking and fermentation, six significant SNPs were identified on five chromosomes (3, 4, 7, 12, and 18). A previous study of SCD identified SNPs on chromosome 20, none of which were reported in this study [101]. However, only three of the six significant SNPs were associated with a candidate gene located within 10kbp distance. SNP ss715586760 was associated with the candidate gene Glyma03g04960, which is involved in the function of seed storage proteins. These proteins serve as the main nutrient source for germinating seedlings. The storage proteins previously identified in the seed coat are chitinase, which are essential for plant defense mechanisms, specifically in response to pathogens [106], [107]. Of the other two genes of note,

Glyma04g40750, is involved in nucleic acid binding. These proteins have been previously associated seed coat color as well as seed development and responses to environmental stimuli [108]. The final candidate gene Glyma12g16486 encodes disease resistance proteins. Because seed coat's role is to protect the seed until germination conditions are suitable, the three candidate genes identified for SCD are all related to either the appearance of the seed coat or the performance of it under adverse conditions. Additionally, 19 suggested SNPs were identified with the candidate gene functions centered around DNA binding, ADP binding, GTP binding, etc..

4.3 Significant SNPs and Candidate Genes for Seed Composition Traits (Protein and Oil)

The nine significant SNPs and eleven candidate genes associated with protein content of the seed had not been previously reported. Previous studies have also identified significant SNPs for protein on chromosomes 2, 3, 8, 13, and 20 [109], but SNPs on chromosomes 14 and 18 identified in this study were not previously reported for protein content.

Each SNP, candidate gene, and gene function are listed in Table 3. The SNP on Chrom 2 (ss715583586) is associated with a candidate gene involved in the isoprenoid biosynthesis process. This process is important for antioxidation in plants, specifically with carotenoids in soybeans, which has been shown to be positively correlated with seed protein content [110], [111]. Two SNPs (ss715587009 and ss715599749) and their associated candidate genes (Glyma.03G009600 and Glyma.08G196200) are involved in the oxidation-reduction process. This process is responsible for the formation of sugars (carbohydrates) which are crucial for the protein folding, stabilization and trafficking [112]. Two separate SNPs (ss715617991 and ss715638161) on different chromosomes were associated with the candidate genes Glyma.14G028600 and

Glyma.20G176800, respectively, and are both involved in DNA binding transcription activity, which regulates gene expression. Because protein content can be negatively affected by environmental stressors such as heat stress, drought stress, etc. [113], the genotype by environment interaction may influence gene expression and, consequently, protein content. Another SNP (ss715631691) was associated with the candidate gene (Glyma.18G061700) and is reported to be involved in flavin adenine dinucleotide binding, which plays a crucial role in the function of enzymes related to soybean metabolism and stress responses [114].

Notably, two SNPs significant for protein content were linked to two candidate genes with functions that would be negatively associated with protein but positively associated with oil. The gene Glyma.03G009700 associated with SNP ss715587009 codes for lipid catabolic process, breaking down lipids into fatty acids. An increase in fatty acids would increase oil content. Additionally, gene Glyma.13G072000 (ss715616659) codes for acyltransferase activity, a function known to increase soybean oil content [115]. While these three genes favor increased oil production, they remain significant for protein content as oil and protein have a well-established negative correlation [116]. Therefore, the identification of these genes can be beneficial for low protein content as a desired trait for natto soybean breeding. Additionally, 24 suggested SNPs were identified with candidate gene repeatedly displaying DNA binding, nuclease activity, hydrolase activity, etc.

Regarding seed oil, four significant SNPs were identified but only two had a candidate gene located within 10kbp of the SNP (Table 1). The significant SNPs are located on chromosomes 2, 19, 20, and 20. Although SNPs associated with oil production have been identified previously on chromosome 2 and 20, none of the SNPs identified in this study have been previously reported ([109], [117]). SNP ss715635641 was associated with candidate gene Glyma.19G219700 () which

is involved in the carbohydrate metabolic processes (Das, Rushton and Rohila, 2017). This process is crucial for many functions such as energy production, seedling growth, as well as drought and heat responses. Oil content, just as previously mentioned in protein, is influenced by the genotype by environment interactions. Increased heat and drought stress would likely reduce oil production, suggesting the significance of this SNP and its candidate gene (Li *et al.*, 2024). The second SNP, SNP ss715636739, and its associated candidate gene Glyma.20G055200, () is involved in zinc ion binding. Zinc finger proteins have been shown to enhance oil production in transgenic soybean plants through the increased activation of lipid-biosynthesis related genes [120]. Additionally, 19 suggested SNPs identified with candidate genes functions repeatedly reported as translationally controlled tumor proteins, leucine-rich repeat receptors, and acid-amine acid ligase activity.

5 Conclusions

Significant SNPs were identified for each of all five traits in the study and five accessions identified for use in natto breeding. The traits of WA, CC, and SCD are novel but essential for natto production. Regarding WA candidate genes with functions related to stress responses, membrane repairs, and waxy waterproof coating of plant tissues, all of which would be important to protect the seed under extended water submergence. The candidate gene encoding cooked color was responsible for accelerating biochemical reactions under increased temperature, such as during cooking. SCD candidate genes were also related to functions of stress and disease defense, which is critical for maintaining seed coat integrity. While protein and oil have been extensively studied, novel SNPs were identified in this study-some of which aligned with previously reported chromosome regions. The candidate gene functions for protein and oil were related to processes critical for production and responses to environmental stressors which can greatly impact both

protein and oil content. The SNPs identified in this study will facilitate the of natto soybean varieties for the international markets and beyond.

Tables

Table 1. The average range of phenotypic data across the two years and three locations.

Trait	Range	Mean	SD¹	CV²
WA	53.35% - 163.93 %	139.74	15.16	10.85
CC	63.26 au - 67.08 au	65.35	2.24	3.43
SCD	0% - 37%	16.61	16.12	97.00
Protein	39.66% - 46.76%	43.69	1.99	4.55
Oil	13.22% - 18.96%	16.12	3.06	18.99

¹SD: standard deviation

²CV: coefficient of variation

Table 2: The accessions identified as potential natto breeding lines based on their performance in four or more traits.

Line	WA¹ (%)	CC² (au)	SCD³ (%)	Protein (%)	Oil (%)
PI506736	132.73	63.76	5.20	39.66	16.08
PI594653	139.55	65.67	9.20	40.10	15.37
PI471931	138.77	64.94	16.17	40.29	18.97
PI408340	144.11	64.68	2.80	40.68	16.50
PI594568B	132.80	65.14	10.20	40.89	17.15

¹ WA: Water absorption

² Cooked Color

³ SDC: Seed coat deficiency

Table 3. A list of each significant SNP, candidate gene and gene function for each of the five traits. The significance threshold was $-\log_{10}(P) > 3.97$. Listed in this table is the SNP, chromosome (chr), position (pos), the allele from the WM82 alignment, the alternative allele associated with the SNP. Candidate genes were listed if located within 10kbp of the SNP.

Trait	chr	pos	SNP	WM82 Allele	Alternate Allele	Gene	Gene Function
Protein	2	5303576	ss715583586	T	C	Glyma.02G059000	isoprenoid biosynthetic process
	3	959652	ss715587009	A	G	Glyma.03G009600	obsolete oxidation-reduction process
						Glyma.03G009700	lipid catabolic process
	8	15811173	ss715599749	A	G	Glyma.08G196200	obsolete oxidation-reduction process
	8	15835838	ss715599751	C	T	Glyma.08G196400	Tudor/PWWP/MBT superfamily protein
	10	5424024	ss715608261	T	C	Glyma.10G058701	heparan-alpha-glucosaminide N-acetyltransferase protein
	13	17304314	ss715616659	A	G	Glyma.13G072000	acyltransferase activity
	14	2077690	ss715617991	T	C	Glyma.14G028600	DNA binding transcription factor activity
	18	5597675	ss715631691	A	G	Glyma.18G061600	transmembrane protein, putative
						Glyma.18G061700	flavin adenine dinucleotide binding
20	41389061	ss715638161	T	C	Glyma.20G176800	DNA binding transcription factor activity	
Oil	2	39122328	ss715582360	G	A	na	
	19	47194890	ss715635641	C	T	Glyma.19G219700	carbohydrate metabolic process
	20	12922198	ss715636739	A	G	Glyma.20G055200	zinc ion binding
	20	16712510	ss715639021	T	C	na	
WA	7	11434129	ss715596103	A	C	na	
	7	11584261	ss715596075	T	C	Glyma07g14330	Mitochondrial transcription termination protein
	7	11932808	ss715596035	A	G	Glyma07g14010	Fatty acid hydroxylase superfamily
	12	34910570	ss715612503	G	A	Glyma12g31350	Tetratricopeptide repeat
						Glyma12g31360	Leucine rich repeat N-terminal domain
SCD	3	5054960	ss715586760	C	A	Glyma03g04960	seed storage proteins
	4	49828162	ss715588770	A	G	Glyma04g40750	nucleic acid binding
	7	29123440	ss715597044	A	G	na	
	12	15521347	ss715611569	T	G	na	
	12	15895016	ss715611585	C	T	Glyma12g16486	disease resistance protein
	18	17662865	ss715629369	C	T	na	
Color	2	28871127	ss715581851	T	C	na	
	18	56545059	ss715632355	T	G	Glyma18g52120	amino acid binding
Glyma18g52130						catalytic activity	

Figures

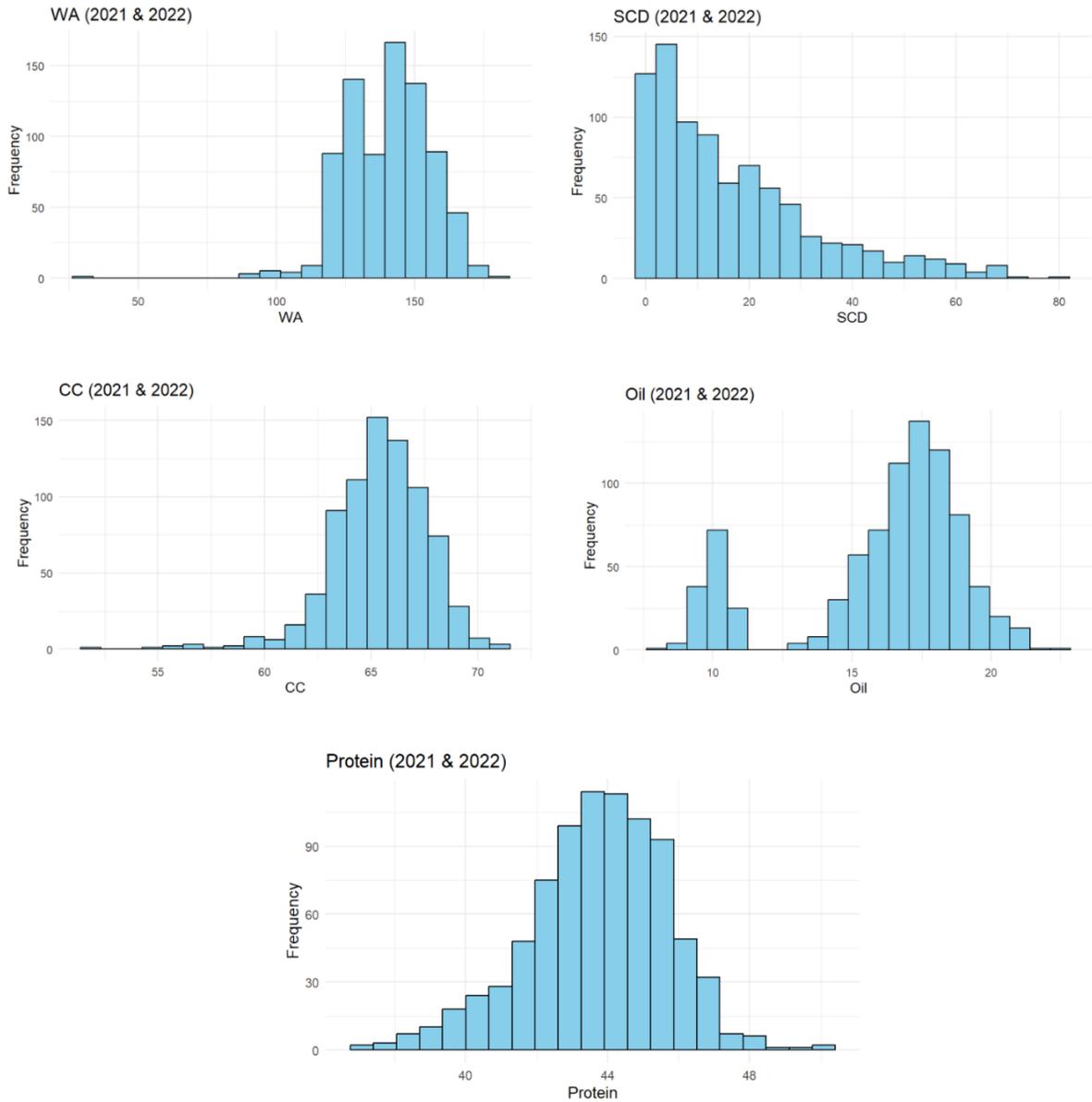


Figure 1. The frequency distribution of each trait from 2021 and 2022. The traits WA, SCD, CC, Oil and Protein were measured in percentage. Color was measured in absorbance units (AU).

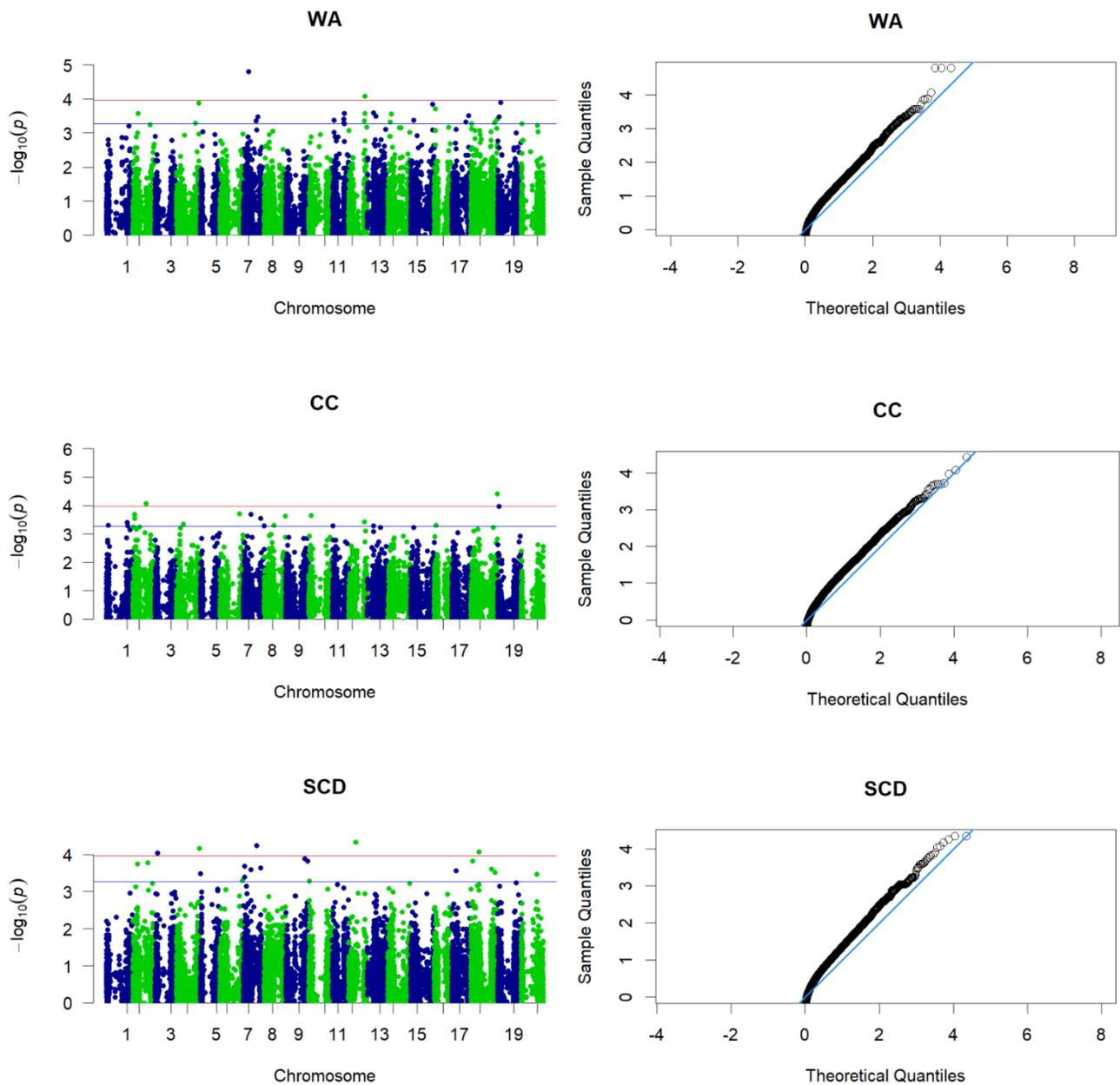


Figure 2. Manhattan plots showing the significant and suggestive SNPs for the three natto quality traits and the associated QQplots. A. The red line represents the significance threshold ($-\log_{10}(P) > 3.97$) and the blue line represents the suggestive threshold ($-\log_{10}(P) > 3.27$).

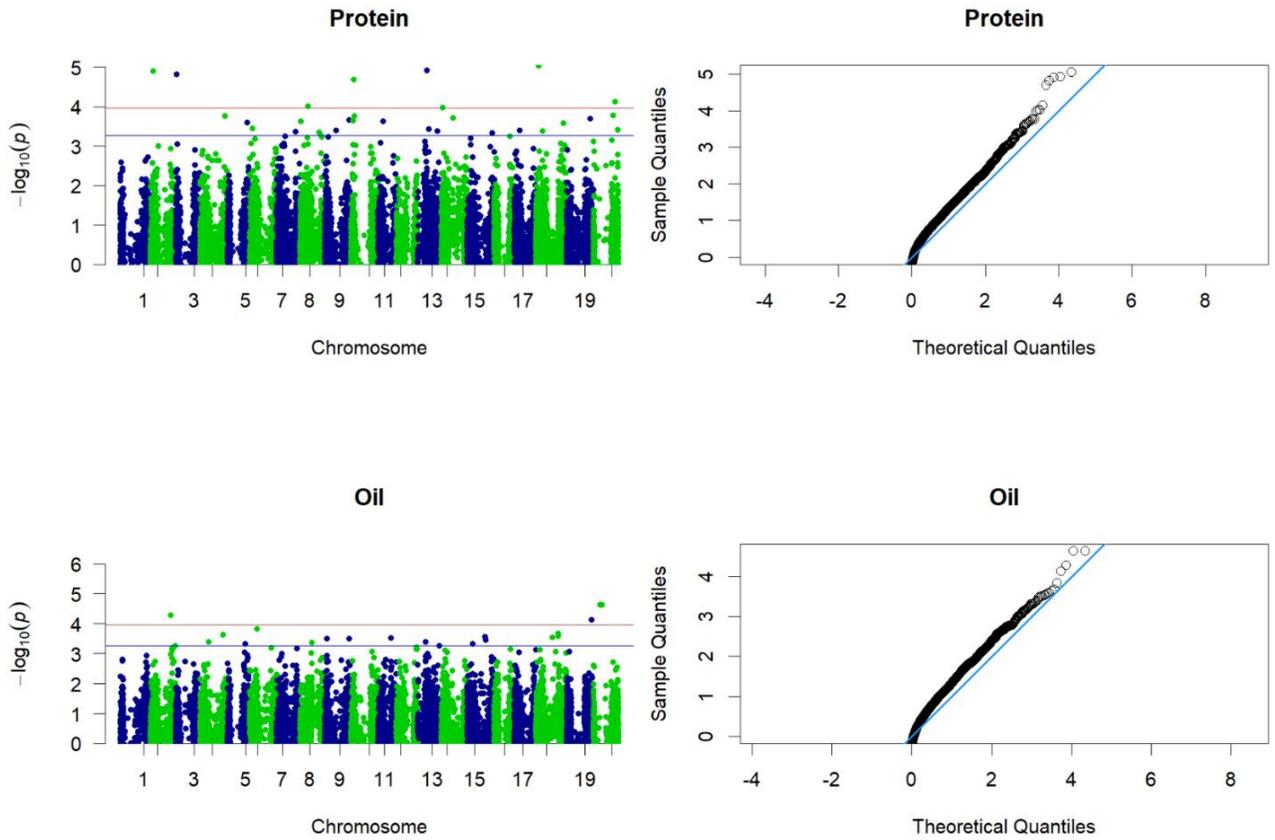


Figure 3. Manhattan plots showing the significant and suggestive SNPs for the seed composition traits (Oil and Protein) and the associated QQplots. The red line represents the significance threshold ($-\log_{10}(P) > 3.97$) and the blue line represents the suggestive threshold ($-\log_{10}(P) > 3.27$).

References

- [1] K. V. Flannery, “The Origins of Agriculture,” *Annual Review of Anthropology*, vol. 2, pp. 271–310, 1973.
- [2] J. Lee, J. H. Chin, S. N. Ahn, and H.-J. Koh, “Brief History and Perspectives on Plant Breeding,” in *Current Technologies in Plant Molecular Breeding: A Guide Book of Plant Molecular Breeding for Researchers*, H.-J. Koh, S.-Y. Kwon, and M. Thomson, Eds., Dordrecht: Springer Netherlands, 2015, pp. 1–14. doi: 10.1007/978-94-017-9996-6_1.
- [3] N. M. Laird and C. Lange, “Principles of Inheritance: Mendel’s Laws and Genetic Models,” in *The Fundamentals of Modern Statistical Genetics*, N. M. Laird and C. Lange, Eds., New York, NY: Springer, 2011, pp. 15–30. doi: 10.1007/978-1-4419-7338-2_2.
- [4] W. E. Castle, “Mendel’s Law of Heredity,” *Science*, vol. 18, no. 456, pp. 396–406, Sep. 1903, doi: 10.1126/science.18.456.396.
- [5] F. V. Monaghan and A. F. Corcos, “Tschermak: a non-discoverer of Mendelism II. A critique,” *Journal of Heredity*, vol. 78, no. 3, pp. 208–210, May 1987, doi: 10.1093/oxfordjournals.jhered.a110361.
- [6] K.-M. Kim, “On the Reception of Johannsen’s Pure Line Theory: Toward a Sociology of Scientific Validity,” *Soc Stud Sci*, vol. 21, no. 4, pp. 649–679, Nov. 1991, doi: 10.1177/030631291021004002.
- [7] D. Berry, “The plant breeding industry after pure line theory: Lessons from the National Institute of Agricultural Botany,” *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, vol. 46, pp. 25–37, Jun. 2014, doi: 10.1016/j.shpsc.2014.02.006.

- [8] D. Bar-Zvi, O. Lupo, A. A. Levy, and N. Barkai, “Hybrid vigor: The best of both parents, or a genomic clash?,” *Current Opinion in Systems Biology*, vol. 6, pp. 22–27, Dec. 2017, doi: 10.1016/j.coisb.2017.08.004.
- [9] R. E. Evenson and D. Gollin, “Assessing the Impact of the Green Revolution, 1960 to 2000,” *Science*, vol. 300, no. 5620, pp. 758–762, May 2003, doi: 10.1126/science.1078710.
- [10] P. Pinstруп-Andersen and P. B. R. Hazell, “The impact of the green revolution and prospects for the future,” *Food Reviews International*, Jan. 1985, doi: 10.1080/87559128509540765.
- [11] G. D. Herder, G. V. Isterdael, T. Beeckman, and I. D. Smet, “The roots of a new green revolution,” *Trends in Plant Science*, vol. 15, no. 11, pp. 600–607, Nov. 2010, doi: 10.1016/j.tplants.2010.08.009.
- [12] P. M. Visscher, M. A. Brown, M. I. McCarthy, and J. Yang, “Five Years of GWAS Discovery,” *The American Journal of Human Genetics*, vol. 90, no. 1, pp. 7–24, Jan. 2012, doi: 10.1016/j.ajhg.2011.11.029.
- [13] A. Korte and A. Farlow, “The advantages and limitations of trait analysis with GWAS: a review,” *Plant Methods*, vol. 9, no. 1, p. 29, Jul. 2013, doi: 10.1186/1746-4811-9-29.
- [14] J.-M. Ribaut and D. Hoisington, “Marker-assisted selection: new tools and strategies,” *Trends in Plant Science*, vol. 3, no. 6, pp. 236–239, Jun. 1998, doi: 10.1016/S1360-1385(98)01240-0.
- [15] S. M. Zargar *et al.*, “Recent advances in molecular marker techniques: Insight into QTL mapping, GWAS and genomic selection in plants,” *J. Crop Sci. Biotechnol.*, vol. 18, no. 5, pp. 293–308, Dec. 2015, doi: 10.1007/s12892-015-0037-5.

- [16] B. Vincent and P. Dardenne, “Application of NIR in Agriculture,” in *Near-Infrared Spectroscopy: Theory, Spectral Analysis, Instrumentation, and Applications*, Y. Ozaki, C. Huck, S. Tsuchikawa, and S. B. Engelsen, Eds., Singapore: Springer, 2021, pp. 331–345. doi: 10.1007/978-981-15-8648-4_14.
- [17] G. Qian and Z. Y. Wang, “Near-Infrared Organic Compounds and Emerging Applications,” *Chemistry – An Asian Journal*, vol. 5, no. 5, pp. 1006–1029, 2010, doi: 10.1002/asia.200900596.
- [18] M. L. Rosso, C. Shang, E. Correa, and B. Zhang, “An Efficient HPLC Approach to Quantify Kunitz Trypsin Inhibitor in Soybean Seeds,” *Crop Science*, vol. 58, no. 4, pp. 1616–1623, 2018, doi: 10.2135/cropsci2018.01.0061.
- [19] T. M. P. Cattaneo and A. Stellari, “Review: NIR Spectroscopy as a Suitable Tool for the Investigation of the Horticultural Field,” *Agronomy*, vol. 9, no. 9, Art. no. 9, Sep. 2019, doi: 10.3390/agronomy9090503.
- [20] M. T. V. Gonçalves, G. Morota, P. M. de A. Costa, P. M. P. Vidigal, M. H. P. Barbosa, and L. A. Peternelli, “Near-infrared spectroscopy outperforms genomics for predicting sugarcane feedstock quality traits,” *PLOS ONE*, vol. 16, no. 3, p. e0236853, Mar. 2021, doi: 10.1371/journal.pone.0236853.
- [21] X. Liu, S. Wu, J. Xu, C. Sui, and J. Wei, “Application of CRISPR/Cas9 in plant biology,” *Acta Pharmaceutica Sinica B*, vol. 7, no. 3, pp. 292–302, May 2017, doi: 10.1016/j.apsb.2017.01.002.
- [22] S. Gohil, A. Kumari, A. Prakash, N. Shah, S. Bhutani, and M. Singh, “CRISPER-Based Industrial Crop Improvements,” in *Industrial Crop Plants*, N. Kumar, Ed., Singapore: Springer Nature, 2024, pp. 123–162. doi: 10.1007/978-981-97-1003-4_5.

- [23] S. Ansari *et al.*, “Sculpting the Harvest: Genomics and Genome Editing Applications for Enhanced Oil Crop Development,” in *Omics and Genome Editing: Revolution in Crop Improvement for Sustainable Agriculture*, K. Sharma, Ed., Cham: Springer Nature Switzerland, 2025, pp. 237–253. doi: 10.1007/978-3-031-81639-0_16.
- [24] L. Gao *et al.*, “CRISPR/CasRx-mediated resistance to *Soybean mosaic virus* in soybean,” *The Crop Journal*, vol. 12, no. 4, pp. 1093–1101, Aug. 2024, doi: 10.1016/j.cj.2024.07.007.
- [25] Z. Wang *et al.*, “Development of new mutant alleles and markers for KTI1 and KTI3 via CRISPR/Cas9-mediated mutagenesis to reduce trypsin inhibitor content and activity in soybean seeds,” *Front Plant Sci*, vol. 14, p. 1111680, 2023, doi: 10.3389/fpls.2023.1111680.
- [26] “Soybean domestication: the origin, genetic architecture and molecular bases - Sedivy - 2017 - New Phytologist - Wiley Online Library.” Accessed: Apr. 08, 2025. [Online]. Available: <https://nph.onlinelibrary.wiley.com/doi/full/10.1111/nph.14418>
- [27] S. Zhao, “The spread of agriculture in China,” *Chinese Geographical Science*, vol. 3, no. 3, pp. 194–202, Sep. 1993, doi: 10.1007/BF02664272.
- [28] S.-C. Jeong *et al.*, “Genetic diversity patterns and domestication origin of soybean,” *Theor Appl Genet*, vol. 132, no. 4, pp. 1179–1193, Apr. 2019, doi: 10.1007/s00122-018-3271-7.
- [29] T. Hymowitz, “1 - The History of the Soybean,” in *Soybeans*, L. A. Johnson, P. J. White, and R. Galloway, Eds., AOCS Press, 2008, pp. 1–31. doi: 10.1016/B978-1-893997-64-6.50004-4.
- [30] T. Hymowitz, “On the domestication of the soybean,” *Econ Bot*, vol. 24, no. 4, pp. 408–421, Oct. 1970, doi: 10.1007/BF02860745.

- [31] “Domestication footprints anchor genomic regions of agronomic importance in soybeans - Han - 2016 - New Phytologist - Wiley Online Library.” Accessed: Apr. 08, 2025. [Online]. Available: <https://nph.onlinelibrary.wiley.com/doi/full/10.1111/nph.13626>
- [32] “Oilseed market: global trends - IOPscience.” Accessed: Apr. 08, 2025. [Online]. Available: <https://iopscience.iop.org/article/10.1088/1755-1315/274/1/012030/meta>
- [33] C. Valdes, J. Gillespie, and E. Dohlman, “Soybean Production, Marketing Costs, and Export Competitiveness in Brazil and the United States,” 2023. doi: 10.22004/ag.econ.340506.
- [34] “Global Market Report | Soybean prices and sustainability”.
- [35] K. Vaiknoras and T. Hubbs, “Characteristics and trends of U.S. soybean production practices, costs, and returns since 2002,” USDA Economic Research Service, Washington, D.C., Jun. 2023. doi: 10.32747/2023.8023698.ers.
- [36] “Comparison of the quality of soybean meal and oil by soybean production origin - Liu - 2024 - Journal of the American Oil Chemists’ Society - Wiley Online Library.” Accessed: Apr. 08, 2025. [Online]. Available: https://aocs.onlinelibrary.wiley.com/doi/full/10.1002/aocs.12835?casa_token=S3E-LCcgGd0AAAAA%3AkzvEX6SJyePbuH5iwYUGyr3oO3LB3SgOMtqJ6B_OEqb7iXssSXMpiHCfM7qfv06HrE0yDCf2RXgG5ZE
- [37] B. H. Vagadia, S. K. Vanga, and V. Raghavan, “Inactivation methods of soybean trypsin inhibitor – A review,” *Trends in Food Science & Technology*, vol. 64, pp. 115–125, Jun. 2017, doi: 10.1016/j.tifs.2017.02.003.

- [38] I. E. Liener, "Implications of antinutritional components in soybean foods," *Critical Reviews in Food Science and Nutrition*, vol. 34, no. 1, pp. 31–67, Jan. 1994, doi: 10.1080/10408399409527649.
- [39] M. Friedman and D. L. Brandon, "Nutritional and health benefits of soy proteins," *J Agric Food Chem*, vol. 49, no. 3, pp. 1069–1086, Mar. 2001, doi: 10.1021/jf0009246.
- [40] Z. Shea *et al.*, "A Review of Bioactive Compound Effects from Primary Legume Protein Sources in Human and Animal Health," *Current Issues in Molecular Biology*, vol. 46, no. 5, Art. no. 5, May 2024, doi: 10.3390/cimb46050257.
- [41] J. Chen, K. Wedekind, and J. Escobar, "Trypsin Inhibitor and Urease Activity of Soybean Meal Products from Different Countries and Impact of Trypsin Inhibitor on Ileal Amino Acid Digestibility in Pig," *Journal of American Oil Chemists Society*, 2020, Accessed: Aug. 27, 2024. [Online]. Available: <https://aocs.onlinelibrary.wiley.com/doi/abs/10.1002/aocs.12394>
- [42] M. Hirota, M. Ohmuraya, and H. Baba, "The role of trypsin, trypsin inhibitor, and trypsin receptor in the onset and aggravation of pancreatitis," *J Gastroenterol*, vol. 41, no. 9, pp. 832–836, Sep. 2006, doi: 10.1007/s00535-006-1874-2.
- [43] M. Kunitz, "CRYSTALLINE SOYBEAN TRYPSIN INHIBITOR," *J Gen Physiol*, vol. 30, no. 4, pp. 291–310, Mar. 1947.
- [44] S. Onesti, P. Brick, and D. M. Blow, "Crystal structure of a Kunitz-type trypsin inhibitor from *Erythrina coffra* seeds," *Journal of Molecular Biology*, vol. 217, no. 1, pp. 153–176, Jan. 1991, doi: 10.1016/0022-2836(91)90618-G.

- [45] Y. Birk, “Purification and some properties of a highly active inhibitor of trypsin and α -chymotrypsin from soybeans,” *Biochimica et Biophysica Acta*, vol. 54, no. 2, pp. 378–381, Dec. 1961, doi: 10.1016/0006-3002(61)90387-0.
- [46] C. M. DiPietro and I. E. Liener, “Heat inactivation of the Kunitz and Bowman-Birk soybean protease inhibitors,” *J. Agric. Food Chem.*, vol. 37, no. 1, pp. 39–44, Jan. 1989, doi: 10.1021/jf00085a010.
- [47] H. Rafiee-Yarandi, M. Alikhani, G. R. Ghorbani, and A. Sadeghi-Sefidmazgi, “Effects of temperature, heating time and particle size on values of rumen undegradable protein of roasted soybean,” *South African Journal of Animal Science*, vol. 46, no. 2, pp. 170–179, 2016, doi: 10.4314/sajas.v46i2.8.
- [48] N. Ruiz, C. M. Parsons, H. H. Stein, C. N. Coon, J. E. van Eys, and R. D. Miles, “A historical look at the soybean and its use for animal feed.,” 2020.
- [49] Samiksha, D. Singh, A. K. Kesavan, and S. K. Sohal, “Exploration of anti-insect potential of trypsin inhibitor purified from seeds of *Sapindus mukorossi* against *Bactrocera cucurbitae*,” *Sci Rep*, vol. 9, no. 1, Art. no. 1, Nov. 2019, doi: 10.1038/s41598-019-53495-6.
- [50] “Soybeans: are they used for food, feed or fuel?,” Our World in Data. Accessed: Mar. 17, 2025. [Online]. Available: <https://ourworldindata.org/grapher/soybean-production-and-use>
- [51] K. Makwana, “Soybeans for Global Nutrition: A Numbers Story,” Sustainable Nutrition Initiative®. Accessed: Mar. 13, 2025. [Online]. Available: <https://sustainablenutritioninitiative.com/soybeans-for-global-nutrition-a-numbers-story/>

- [52] S.-I. Park, "Preparation of Natto(Unripe Chungkukjang) Using Small Soybeans and *Bacillus subtilis* KCCM 11315," *Culinary science and hospitality research*, vol. 12, no. 4, pp. 225–235, 2006.
- [53] C. Nagata *et al.*, "Dietary soy and natto intake and cardiovascular disease mortality in Japanese adults: the Takayama study1," *The American Journal of Clinical Nutrition*, vol. 105, no. 2, pp. 426–431, Feb. 2017, doi: 10.3945/ajcn.116.137281.
- [54] M. Afzaal *et al.*, "Nutritional Health Perspective of Natto: A Critical Review," *Biochemistry Research International*, vol. 2022, no. 1, p. 5863887, 2022, doi: 10.1155/2022/5863887.
- [55] M. Milner and K. Makise, "Natto and Its Active Ingredient Nattokinase: A Potent and Safe Thrombolytic Agent," *Alternative and Complementary Therapies*, vol. 8, no. 3, pp. 157–164, Jun. 2002, doi: 10.1089/107628002760091001.
- [56] L. Yuan, C. Liangqi, T. Xiyu, and L. Jinyao, "Biotechnology, Bioengineering and Applications of *Bacillus Nattokinase*," *Biomolecules*, vol. 12, no. 7, Art. no. 7, Jul. 2022, doi: 10.3390/biom12070980.
- [57] "Natto Market Research Report: Market size, Industry outlook, Market Forecast, Demand Analysis, Market Share, Market Report 2024-2030." Accessed: Sep. 23, 2024. [Online]. Available: <https://www.industryarc.com/Report/17816/natto-market.html>
- [58] H. El-Shemy, *Soybean and Nutrition*. BoD – Books on Demand, 2011.
- [59] "Power BI Report." Accessed: Aug. 23, 2024. [Online]. Available: <https://app.powerbi.com/view?r=eyJrIjoiNTU0MDIzZmMtYmMzYi00NmM3LTk3ODgtZjBiZGQzOTgxY2MwIiwidCI6ImQyNTQyNGQwLTY1MGUtNDZmYi1iZGYzLTQzNGQ1N2Y3YmE3ZCIsImMiOjN9>

- [60] J. T. Yen, A. H. Jensen, and J. Simon, “Effect of dietary raw soybean and soybean trypsin inhibitor on trypsin and chymotrypsin activities in the pancreas and in small intestinal juice of growing swine,” *J Nutr*, vol. 107, no. 1, pp. 156–165, Jan. 1977, doi: 10.1093/jn/107.1.156.
- [61] H. Fekadu Gemedo, “Antinutritional Factors in Plant Foods: Potential Health Benefits and Adverse Effects,” *IJNFS*, vol. 3, no. 4, p. 284, 2014, doi: 10.11648/j.ijnfs.20140304.18.
- [62] D. M. Blow, J. Janin, and R. M. Sweet, “Mode of action of soybean trypsin inhibitor (Kunitz) as a model for specific protein–protein interactions,” *Nature*, vol. 249, no. 5452, Art. no. 5452, May 1974, doi: 10.1038/249054a0.
- [63] D. E. Bowman, “Fractions Derived from Soy Beans and Navy Beans Which Retard Tryptic Digestion of Casein,” *Proceedings of the Society for Experimental Biology and Medicine*, vol. 57, no. 1, pp. 139–140, Oct. 1944, doi: 10.3181/00379727-57-14731P.
- [64] M. Deshimaru, S. YOSHIMI, S. SHIOI, and S. TERADA, “Multigene Family for Bowman–Birk Type Proteinase Inhibitors of Wild Soja and Soybean: The Presence of Two BBI-A Genes and Pseudogenes,” *Bioscience, Biotechnology, and Biochemistry*, vol. 68, no. 6, pp. 1279–1286, Jan. 2004, doi: 10.1271/bbb.68.1279.
- [65] Y. Wang, X. Chen, and L. Qiu, “Novel alleles among soybean Bowman-Birk proteinase inhibitor gene families,” *Sci China C Life Sci*, vol. 51, no. 8, pp. 687–692, Aug. 2008, doi: 10.1007/s11427-008-0096-7.
- [66] J.-F. Hsieh and S.-T. Chen, “Comparative studies on the analysis of glycoproteins and lipopolysaccharides by the gel-based microchip and SDS-PAGE,” *Biomicrofluidics*, vol. 1, no. 1, p. 14102, Jan. 2007, doi: 10.1063/1.2399892.

- [67] A. M. Torbica, D. R. Živančev, Z. T. Nikolić, V. B. Đorđević, and B. G. Nikolovski, “Advantages of the Lab-on-a-Chip Method in the Determination of the Kunitz Trypsin Inhibitor in Soybean Varieties,” *J. Agric. Food Chem.*, vol. 58, no. 13, pp. 7980–7985, Jul. 2010, doi: 10.1021/jf100830m.
- [68] T. Zhou, S. Han, Z. Li, and P. He, “Purification and Quantification of Kunitz Trypsin Inhibitor in Soybean Using Two-Dimensional Liquid Chromatography,” *Food Anal. Methods*, vol. 10, no. 10, pp. 3350–3360, Oct. 2017, doi: 10.1007/s12161-017-0902-6.
- [69] “The Best HPLC Systems: A Buyer’s Review of Price and Features.” Accessed: Apr. 04, 2025. [Online]. Available: <https://www.labx.com/resources/the-best-hplc-systems-a-buyers-review-of-price-and-features/4819>
- [70] “The Best NIR Systems: A Buyer’s Guide to Price and Features.” Accessed: Apr. 04, 2025. [Online]. Available: <https://www.labx.com/resources/the-best-nir-systems-a-buyers-guide-to-price-and-features/4960>
- [71] D. D. Le Pevelen and G. E. Tranter, “NIR Spectroscopy - an overview | ScienceDirect Topics.” Accessed: Aug. 24, 2024. [Online]. Available: <https://www.sciencedirect.com/topics/chemistry/nir-spectroscopy>
- [72] M. Blanco and I. Villarroya, “NIR spectroscopy: a rapid-response analytical tool,” *TrAC Trends in Analytical Chemistry*, vol. 21, no. 4, pp. 240–250, Apr. 2002, doi: 10.1016/S0165-9936(02)00404-1.
- [73] K. H. Norris, “History of NIR.” Accessed: Aug. 24, 2024. [Online]. Available: <https://journals.sagepub.com/doi/abs/10.1255/jnirs.941>

- [74] J. U. Porep, D. R. Kammerer, and R. Carle, “On-line application of near infrared (NIR) spectroscopy in food production,” *Trends in Food Science & Technology*, vol. 46, no. 2, Part A, pp. 211–230, Dec. 2015, doi: 10.1016/j.tifs.2015.10.002.
- [75] B. G. Osborne, T. Fearn, and P. T. Hindle, *Practical NIR spectroscopy with applications in food and beverage analysis*, 2nd ed. in Longman food technology. Harlow, Essex, England, New York: Longman Scientific & Technical ; Wiley, 1993.
- [76] A. International, *Approved methods of analysis, 11th Ed. Method 39-21.01. Near-infrared reflectance method for protein and oil determination in soybeans.*, 11th ed. AACC International, 2010. [Online]. Available: <https://doi.org/10.1094/aaccintmethod-39-21.01>
- [77] “Home – SoyBase.” Accessed: Apr. 04, 2025. [Online]. Available: <https://www.soybase.org/>
- [78] N. Lord, C. Shang, L. Rosso, and B. Zhang, “Development of near-infrared reflectance spectroscopy calibration for sugar content in ground soybean seed using Perten DA7250 analyzer,” *Crop Science*, vol. 61, no. 2, pp. 966–975, 2021, doi: 10.1002/csc2.20358.
- [79] S. K. Setarehdan, J. J. Soraghan, D. Littlejohn, and D. A. Sadler, “Maintenance of a calibration model for near infrared spectrometry by a combined principal component analysis–partial least squares approach,” *Analytica Chimica Acta*, vol. 452, no. 1, pp. 35–45, Jan. 2002, doi: 10.1016/S0003-2670(01)01446-5.
- [80] M. R. Smith, R. D. Jee, A. C. Moffat, D. R. Rees, and N. W. Broad, “Optimisation of partial least squares regression calibration models in near-infrared spectroscopy: a novel algorithm for wavelength selection,” *Analyst*, vol. 128, no. 11, pp. 1312–1319, Nov. 2003, doi: 10.1039/B309233J.

- [81] T. O. Hodson, “Root-mean-square error (RMSE) or mean absolute error (MAE): when to use them or not,” *Geoscientific Model Development*, vol. 15, no. 14, pp. 5481–5487, Jul. 2022, doi: 10.5194/gmd-15-5481-2022.
- [82] M. B. Pesic, B. V. Vucelic-Radovic, M. B. Barac, S. P. Stanojevic, and V. A. Nedovic, “Influence of Different Genotypes on Trypsin Inhibitor Levels and Activity in Soybeans,” *Sensors*, vol. 7, no. 1, Art. no. 1, Jan. 2007, doi: 10.3390/s7010067.
- [83] J. N. Losso, “The Biochemical and Functional Food Properties of the Bowman-Birk Inhibitor,” *Critical Reviews in Food Science and Nutrition*, vol. 48, no. 1, pp. 94–118, Jan. 2008, doi: 10.1080/10408390601177589.
- [84] A. Gitlin-Domagalska, A. Maciejewska, and D. Dębowski, “Bowman-Birk Inhibitors: Insights into Family of Multifunctional Proteins and Peptides with Potential Therapeutical Applications,” *Pharmaceuticals (Basel)*, vol. 13, no. 12, p. 421, Nov. 2020, doi: 10.3390/ph13120421.
- [85] A. P. Herwade, S. S. Kasar, N. R. Rane, S. Ahmed, J. S. Maras, and P. K. Pawar, “Characterization of a Bowman–Birk type trypsin inhibitor purified from seeds of *Solanum surattense*,” *Sci Rep*, vol. 11, no. 1, p. 8648, Apr. 2021, doi: 10.1038/s41598-021-87980-8.
- [86] Y. Zhu, L. Chen, X. Chen, J. Chen, and H. Zhang, “Near-infrared spectroscopy identification method of cashmere and wool fibers based on an optimized wavelength selection algorithm,” *Heliyon*, vol. 10, no. 14, p. e34537, Jul. 2024, doi: 10.1016/j.heliyon.2024.e34537.
- [87] D. Hoffmann, D. Brugger, W. Windisch, and S. Thurner, “Calibration Model for a Near Infrared Spectroscopy (NIRS) System to Control Feed Quality of Soy Cake Based on Feed

- Value Assessments In-Vitro,” *Chemical Engineering Transactions*, vol. 58, Jul. 2017, doi: 10.3303/CET1758064.
- [88] M. Otsuka, Y. Fukui, and Y. Ozaki, “Comparative evaluation of bioactivity of crystalline trypsin for drying by Fourier-transformed infrared spectroscopy,” *Colloids and Surfaces B: Biointerfaces*, vol. 69, no. 2, pp. 194–200, Mar. 2009, doi: 10.1016/j.colsurfb.2008.11.016.
- [89] K. J. Wedekind, J. Chen, F. Yan, J. Escobar, and M. Vazquez-Anon, “Efficacy of a mono-component protease is affected by trypsin inhibitor concentration in soybean meal,” *Animal Feed Science and Technology*, vol. 265, p. 114502, Jul. 2020, doi: 10.1016/j.anifeedsci.2020.114502.
- [90] R. L. Bernard and D. A. Lindahl, “Registration of Williams Soybean (Reg. No. 94),” *Crop Science*, vol. 12, no. 5, p. crops1972.0011183X001200050067x, 1972, doi: 10.2135/crops1972.0011183X001200050067x.
- [91] R. D. Riggs and D. P. Schmitt, “Complete Characterization of the Race Scheme for *Heterodera glycines*,” *J Nematol*, vol. 20, no. 3, pp. 392–395, Jul. 1988.
- [92] K. Ghimire *et al.*, “Inoculation Method Impacts Symptom Development Associated with *Diaporthe aspalathi*, *D. caulivora*, and *D. longicolla* on Soybean (*Glycine max*),” *Plant Disease*, vol. 103, no. 4, pp. 677–684, Apr. 2019, doi: 10.1094/PDIS-06-18-1078-RE.
- [93] “Soybean Meal - an overview | ScienceDirect Topics.” Accessed: Apr. 01, 2025. [Online]. Available: <https://www.sciencedirect.com/topics/agricultural-and-biological-sciences/soybean-meal>
- [94] S. Mehmood *et al.*, “Crystal structure of Kunitz-type trypsin inhibitor: Entomotoxic effect of native and encapsulated protein targeting gut trypsin of *Tribolium castaneum* Herbst,”

- Computational and Structural Biotechnology Journal*, vol. 23, pp. 3132–3142, Dec. 2024, doi: 10.1016/j.csbj.2024.07.023.
- [95] Y. Arjoune *et al.*, “Soybean cyst nematode detection and management: a review,” *Plant Methods*, vol. 18, no. 1, p. 110, Sep. 2022, doi: 10.1186/s13007-022-00933-8.
- [96] B. Zhang, D.-F. Wang, H. Wu, L. Zhang, and Y. Xu, “Inhibition of endogenous α -amylase and protease of *Aspergillus flavus* by trypsin inhibitor from cultivated and wild-type soybean,” *Ann Microbiol*, vol. 60, no. 3, Art. no. 3, Sep. 2010, doi: 10.1007/s13213-010-0056-x.
- [97] C. Zhang *et al.*, “A fungal effector and a rice NLR protein have antagonistic effects on a Bowman–Birk trypsin inhibitor,” *Plant Biotechnology Journal*, vol. 18, no. 11, pp. 2354–2363, 2020, doi: 10.1111/pbi.13400.
- [98] “Microorganisms | Free Full-Text | Fermented Soy Products and Their Potential Health Benefits: A Review.” Accessed: Aug. 15, 2024. [Online]. Available: <https://www.mdpi.com/2076-2607/10/8/1606>
- [99] Y. Yoshikawa, P. Chen, B. Zhang, A. Scaboo, and M. Orazaly, “Evaluation of seed chemical quality traits and sensory properties of natto soybean,” *Food Chemistry*, vol. 153, pp. 186–192, Jun. 2014, doi: 10.1016/j.foodchem.2013.12.027.
- [100] Y. Yoshikawa *et al.*, “Evaluation of Natto Soybean for Agronomic and Seed Quality Traits,” *Journal of Crop Improvement*, vol. 29, no. 1, pp. 40–52, Jan. 2015, doi: 10.1080/15427528.2014.960056.
- [101] Q. Zhu *et al.*, “Identification and validation of major QTLs associated with low seed coat deficiency of natto soybean seeds (*Glycine max* L.),” *Theor Appl Genet*, vol. 133, no. 11, pp. 3165–3176, Nov. 2020, doi: 10.1007/s00122-020-03662-5.

- [102] Q. Song *et al.*, “Fingerprinting Soybean Germplasm and Its Utility in Genomic Research”, Accessed: Mar. 18, 2025. [Online]. Available: <https://dx.doi.org/10.1534/g3.115.019000>
- [103] G. Patil *et al.*, “Dissecting genomic hotspots underlying seed protein, oil, and sucrose content in an interspecific mapping population of soybean using high-density linkage mapping,” *Plant Biotechnology Journal*, vol. 16, no. 11, pp. 1939–1953, 2018, doi: 10.1111/pbi.12929.
- [104] V. Quesada, “The roles of mitochondrial transcription termination factors (MTERFs) in plants,” *Physiol Plant*, vol. 157, no. 3, pp. 389–399, Jul. 2016, doi: 10.1111/ppl.12416.
- [105] T. L. A. Tully, P. Kaushik, J. O’Connor, and M. A. Bernards, “Fatty acid ω -hydroxylases of soybean: CYP86A gene expression and aliphatic suberin deposition,” *Botany*, vol. 98, no. 6, pp. 317–326, Jun. 2020, doi: 10.1139/cjb-2019-0198.
- [106] N. C. Silva *et al.*, “Soybean seed coat chitinase as a defense protein against the stored product pest *Callosobruchus maculatus*,” *Pest Manag Sci*, vol. 74, no. 6, pp. 1449–1456, Jun. 2018, doi: 10.1002/ps.4832.
- [107] P. Sharma, L. Malhotra, and R. K. Dhamija, “Comprehensive amino acid composition analysis of seed storage proteins of cereals and legumes: identification and understanding of intrinsically disordered and allergenic peptides,” *J Biomol Struct Dyn*, vol. 43, no. 7, pp. 3715–3727, Apr. 2025, doi: 10.1080/07391102.2023.2300126.
- [108] J. J. Todd and L. O. Vodkin, “Pigmented Soybean (*Glycine max*) Seed Coats Accumulate Proanthocyanidins during Development,” *Plant Physiology*, vol. 102, no. 2, pp. 663–670, 1993.
- [109] E.-Y. Hwang *et al.*, “A genome-wide association study of seed protein and oil content in soybean,” *BMC Genomics*, vol. 15, no. 1, p. 1, Jan. 2014, doi: 10.1186/1471-2164-15-1.

- [110] T. Kuzuyama and H. Seto, “Two distinct pathways for essential metabolic precursors for isoprenoid biosynthesis,” *Proc Jpn Acad Ser B Phys Biol Sci*, vol. 88, no. 3, pp. 41–52, 2012, doi: 10.2183/pjab.88.41.
- [111] M. A. Schmidt *et al.*, “Transgenic soya bean seeds accumulating β -carotene exhibit the collateral enhancements of oleate and protein content traits,” *Plant Biotechnol J*, vol. 13, no. 4, pp. 590–600, May 2015, doi: 10.1111/pbi.12286.
- [112] M. A. Mensink, H. W. Frijlink, K. van der Voort Maarschalk, and W. L. J. Hinrichs, “How sugars protect proteins in the solid state and during drying (review): Mechanisms of stabilization in relation to stress conditions,” *European Journal of Pharmaceutics and Biopharmaceutics*, vol. 114, pp. 288–295, May 2017, doi: 10.1016/j.ejpb.2017.01.024.
- [113] D. Shi, J. Hang, J. Neufeld, S. Zhao, and J. D. House, “Effects of genotype, environment and their interaction on protein and amino acid contents in soybeans,” *Plant Science*, vol. 337, p. 111891, Dec. 2023, doi: 10.1016/j.plantsci.2023.111891.
- [114] Y. Zhao *et al.*, “*GmGPDH12*, a mitochondrial FAD-GPDH from soybean, increases salt and osmotic stress resistance by modulating redox state and respiration,” *The Crop Journal*, vol. 9, no. 1, pp. 79–94, Feb. 2021, doi: 10.1016/j.cj.2020.05.008.
- [115] K. S. Flyckt *et al.*, “A Novel Soybean Diacylglycerol Acyltransferase 1b Variant with Three Amino Acid Substitutions Increases Seed Oil Content,” *Plant and Cell Physiology*, vol. 65, no. 6, pp. 872–884, Jun. 2024, doi: 10.1093/pcp/pcad148.
- [116] S. Kambhampati *et al.*, “On the Inverse Correlation of Protein and Oil: Examining the Effects of Altered Central Carbon Metabolism on Seed Composition Using Soybean Fast Neutron Mutants,” *Metabolites*, vol. 10, no. 1, p. 18, Dec. 2019, doi: 10.3390/metabo10010018.

- [117] M. Sung *et al.*, “Identification of SNP markers associated with soybean fatty acids contents by genome-wide association analyses,” *Mol Breeding*, vol. 41, no. 4, p. 27, Mar. 2021, doi: 10.1007/s11032-021-01216-1.
- [118] A. Das, P. J. Rushton, and J. S. Rohila, “Metabolomic Profiling of Soybeans (*Glycine max* L.) Reveals the Importance of Sugar and Nitrogen Metabolism under Drought and Heat Stress,” *Plants (Basel)*, vol. 6, no. 2, p. 21, May 2017, doi: 10.3390/plants6020021.
- [119] H. Li *et al.*, “Soybean Oil and Protein: Biosynthesis, Regulation and Strategies for Genetic Improvement,” *Plant, Cell & Environment*, vol. n/a, no. n/a, doi: 10.1111/pce.15272.
- [120] Q.-T. Li *et al.*, “Selection for a Zinc-Finger Protein Contributes to Seed Oil Increase during Soybean Domestication,” *Plant Physiology*, vol. 173, no. 4, pp. 2208–2224, Apr. 2017, doi: 10.1104/pp.16.01610.